IDS

Introduction to Data Science

Robert Gould

Suyen Machado

Terri Anna Johnson

James Molyneux

Sponsors & Supporters

This curriculum was created under the auspices of the National Science Foundation, Mathematics and Science Partnership grant, "MOBILIZE: Mobilizing for Innovative Computer Science Teaching and Learning." Lead Principal Investigator: Robert Gould (UCLA, Statistics).

Contributing Authors

LAUSD: Monica Casillas, Heidi Estevez, and Carole Sailer

UCLA: Amelia McNamara and Linda Zanontian

Acknowledgments and Special Thanks

Co-Principal Investigators: Deborah Estrin (UCLA, CENS), Joanna Goode (University of Oregon), Mark Hansen (UCLA, Statistics), Jane Margolis (UCLA, Center X), Thomas Philip (UCLA, Center X), Jody Priselac (UCLA, GSEIS), Derrick Chau (LAUSD), Gerardo Loera (LAUSD) and Todd Ullah (LAUSD); Mobilize Project Director: LeeAnn Trusela

LAUSD IDS Pilot Teachers

Robert Montgomery
Carole Sailer
Joy Lee
Monica Casillas

Roberta Ross
Velia Valle
Jose Guzman
Pamela Amaya

Arlene Pascua
Chris Marangopoulos

This material is based upon work supported by the National Science Foundation under Grant Number 0962919.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

For additional information related to IDS visit: www.idsucla.org

Mobilize, an innovative partnership between the University of California, Los Angeles (UCLA) and the Los Angeles Unified School District (LAUSD), was funded in 2010 by the National Science Foundation to develop barrier-breaking curriculum in science, mathematics, and computer science to teach students to think creatively, constructively, and critically about the role of data in science and in everyday life. The Mobilize curricula center around Participatory Sensing campaigns, through which students use their mobile devices to collect and share data about their community and their lives, and analyze these data to gain a greater understanding about their world. Mobilize broke barriers by teaching students to apply concepts and practices from computer science and statistics to learning science and mathematics, and it was uniquely dynamic in that each Mobilize class collects its own data, and each class has the opportunity to make unique discoveries. Across all Mobilize curricula, mobile devices are used not as gimmicks to capture students' attention, but as legitimate tools that bring scientific enquiry into their everyday lives. Since 2011, LAUSD high school mathematics, science, and computer science teachers have attended the summer institutes designed by the Mobilize grant to learn to use the participatory sensing (PS) methods, tools, and materials to deepen their knowledge of computer science (CS) concepts and to support student CS, math, and science learning.

First implemented in 2014 under the auspices of the Mobilize grant, Introduction to Data Science (IDS) began as a pilot program with 10 LAUSD mathematics teachers, and by the 5th printing of the curriculum in 2018 has expanded to 30+ schools in seven Southern California public school districts, serving over 4,000 students and counting. In addition to addressing the Common Core State Standards (CCSS) for High School Statistics and Probability IDS leads students to:

- understand how data are used by professionals to address real-world problems;
- understand that data are used in all facets of modern life;
- understand how data support science to identify and tackle real-world problems in our communities;
- analyze statistical graphics to identify patterns in data and to connect these patterns back to the real world;
- understand that by treating photos, words, numbers, and sounds as data, we can gain insight into the real world;
- learn to analyze data, including: posing questions that can be answered by considering relations among variables in a data set, using collected data to generate hypotheses for future data collection, critically evaluating shortcomings and strengths in the data and the data collection process, and informally evaluating hypotheses using data at hand.

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Introduction to Data Science: Overview & Philosophy

Course Overview

Goals

Introduction to Data Science (IDS) is designed to introduce students to the exciting opportunities available at the intersection of data analysis, computing, and mathematics through hands-on activities. Data are everywhere, and this curriculum will help prepare students to live in a world of data. The curriculum focuses on practical applications of data analysis to give students concrete and applicable skills. Instead of using small, tailored, curated data sets as in a traditional statistics curriculum, this curriculum engages students with a wider world of data that fall into the "Big Data" paradigm and are relevant to students' lives. In contrast to the traditional formula-based approach, in IDS, statistical inference is taught algorithmically, using modern randomization and simulation techniques. Students will learn to find and communicate meaning in data, and to think critically about arguments based on data.

This curriculum was developed in partnership with the Los Angeles Unified School District for a culturally, linguistically, and socially diverse group of students. Upon first publication of the IDS curriculum in 2015, the district-wide student ethnicities included .3% American Indian, 3.7% Asian, .4% Pacific Islander, 2.3% Filipino, 73.0% Latino, 10.9% African American, 8.8% White, and .6% other/multiple responses. Over 38% of students were English-language learners – most of whom spoke Spanish as their primary language – and 74% of students qualified for free or reduced lunches.

Standards

The standards used for the IDS curriculum are based on the High School Probability and Statistics Mathematics Common Core State Standards **(CCSS-M)** and include the Standards for Mathematical Practice **(SMP)**. Specific standards are delineated in the scope and sequence section. The Computer Science Teachers Association **(CSTA)** K-12 Computer Science Standards were also consulted and incorporated. Applied Computational Thinking Standards **(ACT)** delineate the application of Data Science concepts using technology.

Hardware

An ideal laboratory environment has a 1:1 computer to student ratio. The computers can be either Apple, PC, or Chromebook, depending upon availability. Internet access is required for the use of RStudio on an external server. The IDS instructor must have access to a computer and a projector for daily use.

Software

Each computer (tablets are not recommended) in the classroom should have a modern, updated web browser installed (such as Firefox or Google Chrome). This will allow students to access RStudio via the RStudio Cloud platform, and to perform searches and make use of a variety of websites and internet tools. RStudio is accessible at https://posit.cloud/ or through the IDS home page at https://portal.idsucla.org. The IDS team will provide the remainder of the software used in the IDS curriculum, also available at https://portal.idsucla.org.

This software includes the IDS UCLA app, which is deployed for Android and iOS (Apple) smartphones and tablets, as well as through a web browser on a desktop or laptop computer via the IDS home page. The app allows students to collect the Participatory Sensing data that is a motivational foundation for the course. In addition to the app, students will use the IDS software to access and manipulate their Participatory Sensing data, and to author their own campaigns.

All computer-based assignments are intended to be completed in class to avoid the assumption that students have access to computers at home. However, if a student misses a lab assignment, they will need to make it up on their own time. All the software required for the curriculum is available via the Internet, so students can complete the assignment on any Internet-enabled computer (e.g., at the school or public library).

Prerequisites

It is recommended that students successfully complete a first-year Algebra course prior to taking IDS. With this background, the curriculum provides a rigorous but accessible introduction to data science and statistics. No previous statistics or computer science courses are required to take this course.

The Instructional Philosophy of Introduction to Data Science

IDS uses a project-based learning approach to instruction. Finkle and Torp (1955) define Project-Based Learning (PBL) as a curriculum development and instructional system that simultaneously develops both problem-solving strategies and disciplinary knowledge bases and skills by placing students in the active role of problem solvers confronted with an ill-structured problem that mirrors real-world problems. PBL, therefore, is a model for teaching and learning that focuses on the main concepts and principles of a discipline, involves students in problem-solving investigations and other meaningful tasks, allows students to construct their own knowledge through inquiry, and culminates in a project.

Because IDS is a mathematical science, the BSCS 5-E Instructional Model provides a planned sequence of instruction that places students at the center of their learning experiences. This model encourages students to explore, create their own meaning of concepts, and relate their understanding to other concepts. The units in IDS contain lessons that, together, fit the 5-E Instructional Model:

Stage of Inquiry in an Inquiry-Based Science Program	Possible Student Behavior	Possible Teacher Strategy
Engage	Asks questions such as, Why did this happen? What do I already know about this? What can I find out about this? How can I solve this problem? Shows interest in the topic.	Creates interest. Generates curiosity. Raises questions and problems. Elicits responses that uncover student knowledge about the concept/topic.
Explore	Thinks creatively within the limits of the activity. Tests predictions and hypotheses. Forms new predictions and hypotheses. Tries alternatives to solve a problem and discusses them with others. Records observations and ideas. Suspends judgment. Tests ideas.	Encourages students to work together without direct instruction from the teacher. Observes and listens to students as they interact. Asks probing questions to redirect students' investigations when necessary. Provides time for students to puzzle through problems. Acts as a consultant for students.
Explain	Explains their thinking, ideas, and possible solutions or answers to other students. Listens critically to other students' explanations. Questions other students' explanations. Listens to and tries to comprehend explanations offered by the teacher. Refers to previous activities. Uses recorded data in explanations.	Encourages students to explain concepts and definitions in their own words. Asks for justification (evidence) and clarification from students. Formally provides definitions, explanations, and new vocabulary. Uses students' previous experiences as the basis for explaining concepts.
Elaborate	Applies scientific concepts, labels, definitions, explanations, and skills in new, but similar situations. Uses previous information to ask questions, propose solutions, make decisions, and design experiments. Draws reasonable conclusions from evidence. Records observations and explanations.	Expects students to use vocabulary, definitions, and explanations provided previously in new context. Encourages students to apply the concepts and skills in new situations. Reminds students of alternative explanations. Refers students to alternative explanations.
Evaluate	Checks for understanding among peers. Answers open-ended questions by using observations, evidence, and previously accepted explanations. Demonstrates an understanding or knowledge of the concept or skill. Evaluates his or her own progress and knowledge. Asks related questions that would encourage future investigations.	Refers students to existing data and evidence and asks, What do you know? Why do you think...? Observes students as they apply new concepts and skills. Assesses students' knowledge and/or skills. Looks for evidence that students have changed their thinking. Allows students to assess their learning and group process skills. Asks open-ended questions such as, Why do you think...? What evidence do you have? What do you know about the problem? How would you answer the question?

IDS is designed to develop students' computational and statistical thinking skills. Computationally, students will learn to write code to enhance analyses of data, to break large problems into smaller pieces, and to understand and employ algorithms to solve problems. Statistical thinking skills include developing a data "habit of mind" in which one learns to seek data to answer questions or support (or undermine) claims; thinking critically about the ability of particular data to support claims; learning to interpret analyses of data; and learning to communicate findings.

IDS employs Participatory Sensing to give students control of the data collection process, and to enable them to collect data about things that are important to them. The curriculum is organized around a series of Participatory Sensing "campaigns" in which students engage in all stages of the statistical process, which we call the Data Cycle: asking questions, examining and collecting data, analyzing data, interpreting data and, if necessary, beginning again. As students progress, they engage in the Data Cycle in a deeper way. Initially, analysis and interpretation is purely descriptive. Later, randomization-based algorithms and simulations are used to develop notions of inference and to make students more critical of the data collection process. By engaging in the Data Cycle repeatedly in different contexts - some of which include the students' own designs - students will learn to think like data scientists.

Student Team Collaboration

Many of the activities in the IDS curriculum are based on students collaborating with each other. Activities may call on pairs or teams of students. **It is imperative that teams and team roles be established as close to the beginning of the course as possible**. Expectations about teamwork should be introduced as soon as teams are formed. The ideal team comprises four students. The Teacher Resources section provides a list of instructional strategies and a description of team roles to use for effective student team collaboration. If student teams are unfamiliar with these instructional strategies, it is important for the instructor to take the time to model each strategy.

Classroom Discussions

Because this is an inquiry-based curriculum, classroom discussion will be especially important. It is important to set classroom discussion norms from the beginning of the course. All students should be encouraged to contribute to the classroom discussion, and the learning environment should be as non-judgmental and as open as possible. Instead of one right answer, most questions in this class have many right answers. In fact, even yes/no questions could have two right answers, both with valid supporting evidence. Teachers should create an environment to help students hold each other accountable so that all voices are heard, meaning that if there are a few students who tend to share a lot, invite them to encourage their peers so other voices can be heard. If there are students who tend to avoid contributing to the class discussion, encourage them to share so that their voices are heard.

Assignments & Homework

As much as possible, IDS work will take place in the classroom. Lessons are designed for a 50-60 minute class period. Classes on block schedule will need to complete two lessons; however, it is up to the teacher to decide where to stop in each lesson. There will be open-ended assignments that are sent home. Assignments that require the computer will be completed in class, to avoid the assumption that students have access to computers at home. The exception to this is if a student misses lab time, in which case they will need to find a time to complete the assignment outside of class. As discussed in the software section above, they can use an Internet-enabled computer to do their make-up work.

IDS assignments will not be drill-based. Instead, they will follow the inquiry-based instructional model. Again, most questions will not have one right answer. Instead, students will learn to support their claims with evidence and to participate in data-based discussions. Newspaper or other periodical or digital articles are available via links in the lessons. If desired, articles may be downloaded and printed.

On average, students will complete a lab assignment in RStudio approximately once per week. It will be at the discretion of the teacher whether or not to collect lab assignments. Calculators should be available every day for students to use.

Every day, students will be expected to bring their Data Science (DS) journal, a notebook where they record their notes, work on small assignments, and sketch plots. Teachers may choose to check DS journals and other assignments in the curriculum for credit.

End of Unit Projects, oral presentations, and Practicums are designed as application exercises. Scoring guides are provided as an aid for student performance expectations. It will be up to the teacher to score or attach a grade to these assignments.

Overview of Instructional Topics

The purpose of IDS is to introduce students to dynamic data analysis. The four major components of this curriculum are based on the conceptual categories called upon by the Common Core State Standards High School - Statistics and Probability:

I. **Interpreting Categorical and Quantitative Data**

II. **Making Inferences and Justifying Conclusions**

III. **Conditional Probability and the Rules of Probability**

IV. **Using Probability to Make Decisions**

IDS will emphasize the use of statistics and computation as tools for creative work, and as a means of telling stories with data. Seen in this way, its content will also prepare students to "read" and think critically about existing data stories. Ultimately, this course will be about how we discern good stories from bad through a practice that involves compiling evidence from one or more sources, and which often requires hands-on examination of one or more data sets.

IDS will develop the tools, techniques, and principles for reasoning about the world with data. It will present a process that is iterative and authentically inquiry-based, comparing multiple "views" of one or more data sets. Inevitably, these views are the result of some kind of computation, producing numerical summaries or graphical displays. Their interpretation relies on a special kind of computation known as simulation to describe the uncertainty in each view. This kind of reasoning is exploratory and investigatory, sometimes framed as hypothesis evaluation, and sometimes as hypothesis generation.

Interpreting Categorical and Quantitative Data

A handful of data interpretations are standard. Some, including summaries of shape, center, and spread of one or more variables in a data set - as well as graphical displays like histograms and scatterplots - are standard in the sense that they provide interpretable information in a number of research contexts. They are portable from one set of data to the next, and the rules for their use are simple. And yet, our interpretation of data is rarely “standard.” Data have no natural look - even a spreadsheet or a table of numbers embeds within it a certain representational strategy. We construct multiple views of data in an attempt to uncover stories about the world.

In addition to numerical data, this course will consider time, location, text, and image as data types, and will examine views that uncover patterns or stories. Throughout the course, simulation will be used to calibrate our interpretation of a view, or of a numerical or graphical summary, so that we understand what “story-less” data (i.e., pure noise, no association) look like.

In addition to summaries and simple graphics, students will engage in a modeling practice aligned with the CCSS mathematical practices in order to learn how statistical analyses can explain and describe real-world phenomena. Students will practice fitting and evaluating standard mathematical and statistical models, such as the least-squares regression line. Modeling comes into play when students are asked to design and implement probabilistic simulations in order to test and compare hypothetical chance processes to real-world data.

Making Inferences and Justifying Conclusions

Data are becoming increasingly plentiful, supported by a host of new "publication" techniques or services. Post-Web 2.0, data are interoperable, flowing out of one service and into another, helping us easily build a detailed data version of many phenomena in the world. Reasoning with data, then, starts with the sources and the mechanics of this flow. Which sources do we trust? How do data from different organizations compare? What stories have been told previously with these data, and by whom?

This course answers these questions, in part, by using the tools and techniques already mentioned. The ability to read and critique published stories and visualizations are additions to these tools and techniques. Finally, as an act of comparison, students should also be able to formulate questions, identify existing data sets, and evaluate how the new stories stack up against the old. To support this cycle of inquiry, students will examine the basic publication mechanisms for data and develop a set of questions to ask of any data source - computation meets critical thinking. In some cases, data will exhibit special structures that can be used to aid in inference. The simulation techniques for calibrating different views of a data set take on new life when some form of random process was followed to generate the data. Polls, for example, rely on random samples of the population, and clinical trials randomly assign patients to treatment and control groups. A simulation strategy that repeats these random mechanisms can be used to assess uncertainty in the data, assigning a margin of error to poll results, or identifying new drugs that have a "significant" effect on some health outcome.

In many cases, data will not possess this kind of special origin story. A census, for example, is meant to be a complete enumeration of a population, and we can reason in a very direct way from the data. In other cases, no formal principle was applied, perhaps being a sample "of convenience." The techniques for telling stories from these kinds of data will also rely on a mix of simulation and subsetting or filtering.

Finally, this course will introduce Participatory Sensing as a technique for collecting data. The idea of a data collection campaign will be introduced as a means of formalizing a question to be addressed with data. Campaigns will be informed by research and data analysis, and will build on, augment, or challenge existing sources. The "culture" behind the existing sources and the summaries or views they promote will be part of the classroom discussions.

It is worth noting that everything described so far depends on computation, using a piece of statistical software on a computer. Students will be taught simple programming tools for accessing data, creating views or fitting models, and then assessing their importance via simulation. Computation becomes a medium through which students learn about data. The more expressive the language, the more elaborate the stories we can tell.

Probability

Since simulation is our main tool for reasoning with data, interpreting the output of simulations requires understanding some basic rules of probability. First and foremost, this course will discuss the ways in which a computer can generate random phenomena (e.g., How does a computer toss a coin?). Simple probability calculations will be used to describe what we expect to see from random phenomena, then students will compare their results to simulations. The point is to both rehearse these basic calculations and to make a formal tie between simulation and theory in simple cases.

In that vein, this course will motivate the relationship between frequency and probability. Students will essentially be simulating independent trials and creating summaries of those simulations. In turn, they should understand that the frequency with which an event occurs in a series of independent simulations tends to the probability for that event as the number of simulations gets large (the Law of Large Numbers, a topic that is often taught in introductory statistics courses).

From here, students will simulate a variety of random processes to aid in formal statistical inference when some random mechanism was applied as part of the data design. In short, probability becomes a ruler of sorts for assessing the importance of any story we might tell. In this approach to probability, a combination of direct mathematical calculation and computer simulations will be used in order to give students a deep sense of the underlying statistical concepts.

Topic Outline
This outline describes only the scope of the course; the sequence is described in each unit.

I. Interpreting Data

A. Types of data

B. Numerical and graphical summaries

1. Measures of center and spread, boxplots
2. Bar plots
3. Histograms
4. Scatterplots
5. Graphical summaries of multivariate data

C. Simulation and visual inference

1. Side-by-side bar plots and association
2. Scatterplots

D. Models

1. Linear models
2. k-means
3. Smoothing
4. Learning and tree-based models

II. Making Inferences and Justifying Conclusions

A. Aggregating data

1. Identification of sources
2. Mechanics of Web 2.0
3. Comparison of sources

B. Data with special structures

1. Random sampling
2. Random assignment and A/B testing
3. Simulation-based inference

C. Participatory Sensing

1. Designing a campaign
2. Participation as a data collection strategy

III. Probability

A. Computers and randomness

1. Web services
2. Pseudo-random numbers (optional)

B. Frequency and probability

C. Probability calculations

IV. Algebra in RStudio

1. Vectors
2. Algorithms
3. Functions
4. Evaluating and fitting models to data
5. Graphical representations of multivariate data
6. Numerical summaries of distributions and interpreting in context

Scope and Sequence

Unit 1

This unit will introduce the idea of “data,” fundamental to the rest of the course. While most people think of data simply as a spreadsheet or a table of numbers, almost anything can be considered data, including images, text, GPS coordinates, and much more. Our world has become increasingly data-centric, and we are constantly generating data, whether we know it or not. From posts on Facebook, to shopping records created when you swipe your credit card, to driving over sensors embedded in highway on-ramps, we leave behind a stream of data wherever we go. These data are used to generate stories about our world, whether it is for political forecasting, marketing, scientific research, or even Netflix recommendations. Traditional statistics courses consist of understanding data from only a small subset of data generation processes, namely those collected through random sampling or random assignment in scientific experiments. This unit exposes students to a wider world of data and will help students see how to make sense of these ubiquitous data types.

This unit will motivate the idea that data and data products (charts, graphs, statistics) can be analyzed and evaluated just like other arguments, such as those used by journalists. We want to know how the evidence was collected, what the perspective or bias of the creator might be and look behind the scenes to the process used to create the product. Even the way data are represented embeds within it decisions on the part of the data creator.

Using the techniques of descriptive statistics, students will begin learning how to construct multiple views of data in an attempt to uncover new insights about the world. This will require the introduction of the computational tool R through the interface of RStudio. Standard graphical displays like histograms and scatterplots will be introduced in RStudio, as well as measures of center and spread.

Focus Statistics CCSS-M

S-ID 1. Represent data with plots on the real number line (dotplots, histograms, and boxplots).

S-ID 2: Use statistics appropriate to the shape of the data distribution to compare center (median, mean) of two or more different data sets (measures of spread will be studied in Unit 2).

S-ID 3: Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

S-ID 5. Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

S-ID 6. Represent data on two quantitative variables on a scatterplot and describe how the variables are related.

S-IC 6. Evaluate reports based on data.*
*This standard is woven throughout the course. It is a recurring standard for every unit.

Focus Standards for Mathematical Practices

SMP-3. Construct viable arguments and critique the reasoning of others.
SMP-5. Use appropriate tools strategically.

Upon completion of Unit 1, students will be able to:

- Give examples of where they leave data traces.
- Understand that rows and columns are a form of data structure.
- Explain why the relationship between the variables might exist, or, if there is no relationship, why that might be so.
- Construct and interpret a frequency table.

- Critically read reports from media sources to evaluate their claims.
- Read plots (identify the name of the plot, interpret the axes, look for trends, identify confounding factors).
- Calculate conditional and marginal probabilities using frequency tables.
- Provide a real-world explanation for why the conditional or independent probabilities make sense, using critical thinking skills and background knowledge.
- Communicate their evaluations in written or verbal form using different types of media.
- Load data into RStudio.
- Create basic plots in RStudio.
- Create frequency tables in RStudio.

Unit 2

This unit deepens the informal reasoning skills developed in Unit 1 by enriching students' technical vocabulary and developing more precise analytical tools. Most importantly, this unit introduces the formal concept of probability as a tool for understanding that sometimes patterns observed in data are not "real." Traditional courses attempt to develop this understanding through the development of abstract mathematical probability concepts, but IDS creates enduring understanding by teaching students to design and implement simulations using pseudo-random number generators. This activity also develops computational thinking by teaching students about some basic programming structures. Then, the use of models will come to the foreground. Students will be introduced to linear models - the most common form of modeling in introductory statistics classes - which will serve as the foundation to learn more complex modeling techniques that use the computer technology available to them later in the course, including smoothing techniques and tree-based models.

Focus Statistics CCSS-M

S-ID 2: Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.

S-ID 3: Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

S-ID 4. Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Understand that there are data sets for which such a procedure is not appropriate. Use calculators, spreadsheets, and tables to estimate areas under the normal curve.

S-IC 2. Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation.

S-IC 6. Evaluate reports based on data.*
*This standard is woven throughout the course. It is a recurring standard for every unit.

S-CP 2. Understand that two events A and B are independent if the probability of A and B occurring together is the product of their probabilities, and use this characterization to determine if they are independent.

S-CP 9. (+) Use permutations to perform [informal] inference.
*This standard will be addressed in the context of data science.

Focus SMPs

SMP-4. Model with mathematics.
SMP-5. Use appropriate tools strategically.

Upon completion of Unit 2, students will be able to:

- Create a boxplot by calculating the five-number summary, upper and lower fences, and determining outliers.
- Explain what “standard deviation” means in context.
- Explain why the measures of central tendency and spread may or may not be accurate descriptions of the data from which they came.
- Use permutations of data to solve problems.
- Read/interpret a normal curve/distribution.
- Explain where the normal distribution came from.
- Describe situations where the normal distribution may model the phenomena, and others where it may not.
- Simulate normal distribution.
- Simulate from a model.
- Compare real data to simulation.
- Determine if model and data appear consistent.
- Merge data by columns/rows, and verify that merging is successful.
- Learn for() loops and apply() functions in RStudio.
- Create functions.

Unit 3

Unit 3 focuses on data collection methods, including traditional methods of designed experiments and observational studies and surveys. It introduces students to sampling error and bias, which cause problems in analysis made from survey data. Participatory Sensing is presented as another method of data collection, and students learn to design Participatory Sensing campaigns that will allow them to address particular statistical questions. Participatory Sensing is a unique data collection method because it uses sensors. Furthermore, this method emphasizes the involvement of citizens and community groups in the process of sensing and documenting where they live, work, and play. Triggers play an important role in the Participatory Sensing data collection process. The response to the triggers may or may not be the same each time. Data takes on a variety of forms online and requires a different style of representation. Students enhance computing skills by learning about modern data structures, and by learning to "scrape" data stored in XML format.

Focus Statistics CCSS-M

S-IC 1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

S-IC 3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6. Evaluate reports based on data.*
*This standard is woven throughout the course. It is a recurring standard for every unit.

Focus SMPs

SMP-1. Make sense of problems and persevere in solving them.
SMP-4. Model with mathematics.
SMP-8. Look for and express regularity in repeated reasoning.

Upon completion of Unit 3, students will be able to:

- Provide a loose definition of “statistics” in their own words.
- Compare and contrast population vs. sample.

- Compare and contrast parameter vs. statistic.
- Explain the difference between special data structures, particularly as they relate to inference.
- Exploit special data structures for re-randomization analysis.
- Explain situations where one measure of central tendency or spread may be more appropriate than others.
- Read/interpret boxplots (In-depth look into samples size and their relationship to the population parameters).
- Identify reports that use special data structures (census, survey, observational study, and randomized experiment).
- Do data scraping.
- Use HTML and XML formats.
- Use RStudio to re-randomize data.
- Compute measures of central tendency and spread in RStudio.

Unit 4

This unit will develop modeling skills, beginning with learning to fit and interpret least squares regression lines and learning to use regression to make predictions. Students will learn to evaluate the success of these predictions and so compare models for their predictive accuracy. Modern algorithmic approaches to regression are presented, and students will strengthen algorithmic thinking skills by understanding how and why these algorithms help data scientists make accurate predictions from data. Students engage in a complete modeling experience in which they apply the skills and concepts learned in the previous units. The modeling experience is designed to make students' thinking visible and audible by encouraging them to be metacognitive about the process of inventing and testing a model, ask questions as they go through the process, and recognize the iterative nature of modeling.

Focus Statistics Standards

S-IC 2. Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation.

S-ID 6. Represent data on two quantitative variables on a scatter plot and describe how the variables are related.

a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. *Use given functions or choose a function suggested by the context. Emphasize linear models*.
b. Informally assess the fit of a function by plotting and analyzing residuals.
c. Fit a linear function for a scatter plot that suggests a linear association.

S-ID 7. Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

S-ID 8. Compute (using technology) and interpret the correlation coefficient of a linear fit.

S-IC 6. Evaluate reports based on data.*
*This standard is woven throughout the course. It is a recurring standard for every unit.

Focus SMPs

SMP-2. Reason abstractly and quantitatively.
SMP-4. Model with mathematics.
SMP-7. Look for and make use of structure.

Upon completion of Unit 4, students will be able to:

- Describe how well the linear model fits the data (or does not).

- Provide a real-world explanation of why the model may or may not fit, using critical thinking skills and background knowledge.
- Interpret the slope and intercept on a plot.
- Compute the correlation coefficient using RStudio.
- Interpret linear models in reports, including the correlation coefficient.
- Determine if a trend is “real” or if it could have arisen from randomness.
- Use critical thinking skills to explain why a trend may or may not make sense.
- Fit a regression line.
- Extract the slope, intercept, correlation coefficient, coefficient of determination, and residuals using RStudio.
- Use RStudio to predict y given an x value.
- Explore what happens to the line and the response variable if we multiply (divide) or add (subtract) a constant from the predictor.
- Design and execute their own Participatory Sensing Campaigns.
- Use RStudio to compute permutations and combinations.
- Create Classification and Regression Tree (CART) models.
- Understand non-linear models.

Introduction to Data Science

Unit 1

Introduction to Data Science
Daily Overview: Unit 1

Theme	Day	Lessons and Labs	Campaign	Topics	Page
Data Are All Around (7 days)	1	Lesson 1: Data Trails		Defining data, consumer privacy	26
	2	Lesson 2: Stick Figures		Organizing & collecting data	28
	3	Lesson 3: Data Structures		Organizing data, rows & columns, variables	30
	4	Lesson 4: The Data Cycle		Data cycle, statistical questions	33
	5	Lesson 5: So Many Questions		Statistical questions, variability	38
	6^	Lesson 6: What Do I Eat?	Food Habits	Collecting data, statistical questions	42
	7	Lesson 7: Setting the Stage	Food Habits – data	Participatory sensing	45
Visualizing Data (14 days)	8	Lesson 8: Tangible Plots	Food Habits – data	Dotplots, minimum/maximum, frequency	52
	9	Lesson 9: What Is Typical?	Food Habits – data	Typical value, center	56
	10	Lesson 10: Making Histograms	Food Habits – data	Histograms, bin widths	58
	11	Lesson 11: What Shape Are You In?	Food Habits - data	Shape, center, spread	61
	12	Lesson 12: Exploring Food Habits	Food Habits – data	Single & multi-variable plots	63
	13	Lesson 13: RStudio Basics	Food Habits – data	Intro to RStudio	65
	14	Lab 1A: Data, Code & RStudio	Food Habits – data	RStudio basics	68
	15+	Lab 1B: Get the Picture?	Food Habits – data	Variable types, bar graphs, histograms	71
	16	Lab 1C: Export, Upload, Import	Food Habits – data	Importing data	74
	17	Lesson 14: Variables, Variables, Variables		Multi-variable plots	79
	18	Lab 1D: Zooming Through Data		Subsetting	84
	19	Lab 1E: What's the Relationship?		Multi-variable plots	88
	20	Practicum: The Data Cycle & My Food Habits	Food Habits	Data cycle, variability	91
	21	Practicum Presentations	Food Habits	Data cycle, variability	-
Would You Look at the Time? (9 Days)	22^	Lesson 15: Americans' Time on Task	Time Use – data	Evaluating claims	95
	23	Lab 1F: A Diamond In the Rough	Time Use - data	Cleaning names, categories, and strings	100
	24	Lesson 16: Categorical Associations	Time Use - data	Joint relative frequencies in 2-way tables	105
	25	Lesson 17: Interpreting Two-Way Tables	Time Use - data	Marginal & conditional relative frequencies	107
	26+	Lab 1G: What's the FREQ?	Time Use – data	2-way tables, tally	112
	27	Practicum: Teen Depression	Time Use	Statistical questions, interpreting plots	115
	28	Practicum Presentations		Statistical questions, interpreting plots	-
	29-30	Lab 1H: Our Time		Data cycle, synthesis	117
Unit 1 Project (5 Days)	31-35	End of Unit Project and Oral Presentation: Analyzing Data to Evaluate Claims		Data cycle	118

^=Data collection window begins.
+=Data collection window ends.

IDS Unit 1: Essential Concepts

Lesson 1: Data Trails

Data are a collection of recorded observations. Data are gathered by people and by sensors. Patterns in data can reveal previously unknown patterns in our world. Data play a large, and sometimes invisible, role in our lives.

Lesson 2: Stick Figures

Data consist of records of particular characteristics of people or objects. Data can be organized in many different ways, and some ways make it easier than others for achieving particular purposes.

Lesson 3: Data Structures

Variables record values that vary. By organizing data into rectangular format, we can easily see the characteristics of observations by reading across a row, or we can see the variability in a variable by reading down the column. Computers can easily process data when it is in rectangular format.

Lesson 4: The Data Cycle

A statistical investigation consists of cycling through the four stages of the Data Cycle. The term statistical questions encompasses the variety of questions asked during the statistical problem-solving process which support statistical thinking and reasoning. Statistical investigative questions are perhaps the most important because they are challenging to learn and are the types of questions that determine whether an analysis is productive or not. Statistical investigative questions are questions that address variability and are productive in that they motivate data collection, analysis, and interpretation. The Data Collection phase might consist of collecting data through Participatory Sensing or some other means, or it might consist of examining previously collected data to determine the quality of the data for answering the statistical investigative questions. Data Analysis is almost always done on the computer and consists of creating relevant graphics and numerical summaries of the data. Data Interpretation is involved with using the analysis to answer the statistical investigative questions.

Lesson 5: So Many Questions

Statistical investigative questions typically begin with a vague general question, then develop into a precise question. The process of developing or creating a good investigative question is iterative and requires time and effort to get right. In her 2021 paper, What Makes a Good Statistical Question, Dr. Pip Arnold identified the following as features of a good investigative question:

(1) The variable(s) of interest is/are clear
(2) The group or population we are interested in is clear
(3) The question can be answered with the data
(4) The question asks about the whole group, not an individual or portion of the group
(5) The intention is clear (e.g., summary, comparison, association, time series)
(6) The question is one that is worth investigating, is interesting, and has a purpose

Lesson 6: What Do I Eat? [The Data Cycle: Consider Data]

After raising statistical questions, we examine and record data to see if the questions are appropriate.

Lesson 7: Setting the Stage [The Data Cycle: Collect Data]

In Participatory Sensing, we humans behave as if we are robot sensors, collecting data whenever a "trigger" event occurs. Our ability to learn about the patterns in our life through these data depends on our being reliable data collectors.

Lesson 8: Tangible Plots [The Data Cycle: Analyze Data]

Distributions organize data for us by telling us (a) which values of a variable were observed, and (b) how many times the values were observed (their frequency).

Lesson 9: What Is Typical?

The “center” of a distribution is a deliberately vague term, but it is one way to answer the subjective question "what is a typical value?" The center could be the perceived balancing point or the value that approximately cuts the area of the distribution in half.

Lesson 10: Making Histograms

Histograms can be created through the use of an algorithm. The distributions displayed in a histogram can be classified using the technical terms for the shapes of distributions. Learning to describe routine tasks through an algorithm is an important component of computational thinking.

Lesson 11: What Shape Are You In?

Identifying the shape of a histogram is part of the **interpret** step of the Data Cycle.

Lesson 12: Exploring Food Habits

Once Participatory Sensing data has been collected, the Dashboard and PlotApp perform the analysis step of the Data Cycle, though humans need to tell the computer which plots to examine.

Lesson 13: RStudio Basics

The computer has a syntax, and it can only understand if you speak its language.

Lesson 14: Variables, Variables, Variables

To examine whether two (or more) variables are related, we can plot their distributions on the same graph.

Lesson 15: Americans’ Time on Task

Learning to examine other analyses is an important part of statistical thinking.

Lesson 16: Categorical Associations

A two-way table is a summary of the association/relationship between two categorical variables. Joint relative frequencies answer questions of the form "what proportion of the people/objects had *this* value on the first variable and *this* value on the second?"

Lesson 17: Interpreting Two-Way Tables

Marginal (relative) frequencies tell us about the distribution of a single variable. Conditional relative frequencies tell us about the distribution of one variable when "subsetting" the other.

Data Are All Around

Instructional Days: 7

Enduring Understandings

Data play an important role in our everyday lives. Organizing it can provide evidence about real-life events and people. The data collected by answering survey questions produce variability. Distributions, graphs, and plots are useful tools for organizing data to understand variability. Statistical questions address people, processes, and/or events that contain variability. Situations with variability can sometimes be simplified with some basic statistics.

Engagement

The Target Story will introduce students to the idea that data are ubiquitous. The advent of computers has transformed the way data are collected, used, and analyzed. Video can be found at: https://www.youtube.com/watch?v=XvSA-6BJkx4&feature=youtu.be

Note: Pre-loading the video on your computer prior to the beginning of class is highly recommended to avoid any technical difficulties.

Learning Objective

Statistical/Mathematical:

S-ID: Summarize, represent, and interpret data on a single count or measurement variable.

S-ID 1: Represent data with plots on the real number line (dotplots, histograms, bar plots, and boxplots.

S-ID 2: Use statistics appropriate to the shape of the data distribution to compare center (median, mean) of two or more different data sets. (Measures of spread will be studied in unit 2.)

S-ID 6: Represent data on two quantitative variables on a scatterplot, and describe how the variables are related.

Focus Standards for Mathematical Practice for All of Unit 1:

SMP-3: Construct viable arguments and critique the reasoning of others.

SMP-5: Use appropriate tools strategically.

Data Science:

Experience data handling using ubiquitous data and organize data using rectangular or spreadsheet format as data storage structures.

Everyday activities can be observed and recorded as data. Become aware of the difference between plots used for categorical and numerical variables. Interpret and understand graphs of distributions for numerical and categorical variables.

Applied Computational Thinking using RStudio:

- Work effectively in teams.
- Explain how data, information, and knowledge are represented for computational use.
- Collect, upload, and share personal data via a Participatory Sensing campaign.
- Learn about different representations of distributions using software.
- Utilize software to begin to analyze plots of data collected via Participatory Sensing.

Real-World Connections:

Students begin to develop an awareness that data are all around us. Information can be collected and organized. Computers are powerful tools that make organizing, storing, retrieving, and analyzing data accessible to use in problem solving and decision making. Students will begin to see the relevance of data collection to their own lives. They will begin to understand that data on its own is just collected; but once interpreted, it can lead to discoveries or understandings.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it is used, and how to work with it.

2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

3. Students will use complex sentences to write informative short reports that use data science concepts and skills.

Data File or Data Collection Method

Data Collection Method:

1. Students will keep a Data Diary for 24 hours to track their daily data output.
2. Students will gather data from the cards in the Stick Figures file.
3. As a class, students will determine how to organize the Stick Figures data.
4. Students will collect data using paper and pencil on the *Food Habits Data Collection* activity sheet.
5. **Food Habits Participatory Sensing Campaign**: Students will collect data about their snacking habits.

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 1: Data Trails

Objective:

Students will understand what are data, how they are collected, and possible effects of sharing data.

Materials:

1. Video: *The Target Story* found at: https://www.youtube.com/watch?v=XvSA-6BJkx4&feature=youtu.be
2. Data Science (DS) journal (quad-ruled composition book or similar); MUST be available for every lesson
3. *Data Diary* handout (LMR_1.1_Data Diary)
4. Video: *Terms and Conditions* found at: https://www.youtube.com/watch?v=ZcjtEKNP05c

Vocabulary:

data, observations, data trails, privacy

Essential Concepts: Data are a collection of recorded observations. Data are gathered by people and by sensors. Patterns in data can reveal previously unknown patterns in our world. Data play a large, and sometimes invisible, role in our lives.

Lesson:

Before You Begin!

Before implementing the IDS curriculum, ensure that:

a) Students have been placed in teams and each student understands his or her role in the team.
b) Each student knows who his/her partner is within each team.
c) Expectations regarding collaborative teamwork are discussed and understood (see Team Roles in Teacher Resources).

1. Introduce the lesson by showing *The Target Story* video: https://www.youtube.com/watch?v=XvSA-6BJkx4&feature=youtu.be

2. In pairs, ask students to discuss the following question using the *TPS* strategy (see Instructional Strategies in Teacher Resources):

    a. How do you think Target knew about the daughter? In other words, how did Target know the daughter was pregnant before her father did? ***Target used the information gathered from the daughter's Red Card and compared it to information about other shoppers. Typically, women who bought those particular products were pregnant.***

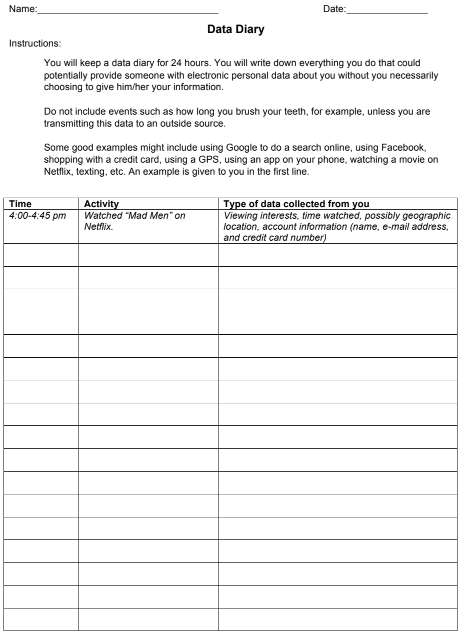
3. After students have had time to share their responses, engage in a whole class discussion regarding:

    a. What are **data**? ***Data are information, or observations, that have been gathered and recorded.***
    b. Where do data come from? ***Data can come from a variety of places. Some examples might include: cell phones, computers, school records, surveys, etc.***
    c. Give an example of data. ***Answers will vary. One example might be information about a person – including their age, height, weight, eye color, etc.***
    d. Give an example of something that is not data (e.g., something that was never written down). ***Answers will vary. One example might be just watching an event happen. If it wasn't recorded in some way, it cannot be counted as data.***

4. Explain to the students that we create "**data trails**" as we go through life. A data trail is the data collected about us as individuals that could be used to see the patterns in our personal lives.

Inform the students that they will learn about their own data trails by keeping a data diary and logging entries over the next 24 hours. It is likely that students do not realize how often they leave a data trail or what information is being collected about them on a regular basis.

5. Distribute the *Data Diary* handout (LMR_1.1) and be sure to go over the instructions, along with the first example to give the students an idea of how to proceed.



Name:______________________________ Date:______________

Data Diary

Instructions:

You will keep a data diary for 24 hours. You will write down everything you do that could potentially provide someone with electronic personal data about you without you necessarily choosing to give him/her your information.

Do not include events such as how long you brush your teeth, for example, unless you are transmitting this data to an outside source.

Some good examples might include using Google to do a search online, using Facebook, shopping with a credit card, using a GPS, using an app on your phone, watching a movie on Netflix, texting, etc. An example is given to you in the first line.

Time	**Activity**	**Type of data collected from you**
4:00-4:45 pm	*Watched "Mad Men" on Netflix.*	*Viewing interests, time watched, possibly geographic location, account information (name, e-mail address, and credit card number)*

LMR_1.1

6. Inform the students that you will collect the handouts during the next class in order to assess their understanding of data.

7. To get students thinking about what happens to their data, show the *Terms and Conditions* video: https://www.youtube.com/watch?v=ZcjtEKNP05c

8. Engage the students in a whole class discussion about the video, particularly noting:

   a. What terms in the **privacy** statements were concerning or worrisome?
   b. Do you read the agreements when you download phone apps?

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were. Be prepared to facilitate a good discussion and to ask probing questions in order for students to elaborate on their thinking so that vague responses such as "we learned about data" can be avoided.

Homework

Students will complete the *Data Diary* handout. When grading the homework, be aware of whether the data really could be collected, and whether the students' ideas about how the data might be used are reasonable. For instance, students will often imagine that there is a "spy" watching them; this is not what we are after. We are after actual instances in which sensors or electronic surveillance records their actions or records information about them. For example, "someone saw me going into the store" is not valid data for this exercise, but "a camera recorded me entering the store" is valid data.

Lesson 2: Stick Figures

Objective:

Students will learn how to observe, record, and organize data.

Materials:

1. *Stick Figures* cutouts (LMR_1.2_Stick Figures)
   Advanced preparation required (see step 3 below)
2. Poster paper
3. Markers
4. Sticky notes

Vocabulary:

collect, record, organize, representations, variables

Essential Concepts: Data consist of records of particular characteristics of people or objects. Data can be organized in many different ways, and some ways make it easier than others for achieving particular purposes.

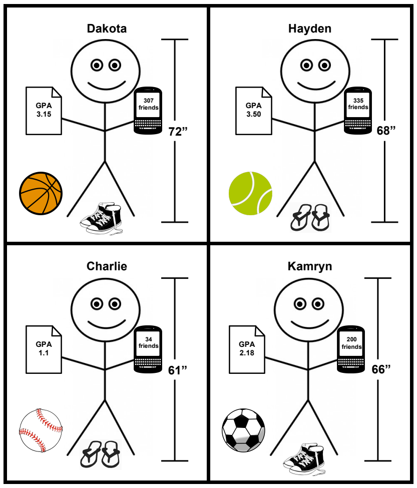
Lesson:

1. Engage students in a *Think-Pair-Share* (see Instructional Strategies) of the *Data Diary* handout that the students completed for homework. Ask them to think about the following questions as they reflect on their data collection homework:
   - How many observations did you make?
   - Where do you leave the most data trails?
   - What could someone learn about you if that person had all of this data?
2. Explain to students that they are going to act as researchers and **collect** data on a strange group of people who appear to be completely two-dimensional. Their goal is to **record** as much information as possible about these people, and to then **organize** the information in any way they choose.
3. Distribute one full set of 8 cards from the *Stick Figures* file (LMR_1.2) to each student team.

   Advanced preparation required:

   Print the *Stick Figures* file (LMR_1.2). The handout can then be cut into the 8 cards. You will need enough sets of the cards for each student team to share one full set. For example, if there are 5 student teams in a class, then 5 copies of the file will need to be printed so that each team gets all 8 cards.

Stick Figures



LMR_1.2

4. Every student from the team will select one of the cards from the team’s pile of 8, and should record all possible information in their DS journal. Once each student has completed this, the team should come together to share individual findings.
5. Distribute one piece of poster paper and a set of markers to each team. The students will then begin to organize the data from all 8 cards into a visual that they think represents the data. It is important that no guidance is given during this portion of the lesson. Students should be free to come up with their own schema for organizing the data.
6. Display all the posters around the room and allow students to participate in a Gallery Walk (see Instructional Strategies in Teacher Resources) to view other teams’ **representations** of the Stick Figure data. For each poster, the teams should write either a comment or a question on a sticky note and add it to the poster to provide feedback for the original team.
7. Afterwards, engage the students in a discussion with the following questions:
   a. Describe some similarities among the team posters. Were the data organized in similar ways? ***Answers will vary by class.***
   b. Describe some differences among the team posters. How were the data organized differently across teams? ***Answers will vary by class.***
   c. What information was available about the stick figures on each card? ***The person’s name, height, GPA, shoe style, sport, and number of friends on social media.***
   d. Which representations made it easy to see what (or who) the objects were that were observed? Which representations made it easy to see whether different stick figures had different characteristics? ***Answers will vary by class.***
   e. Which representation makes it easiest to see which stick figure is tallest? ***Answers will vary by class.***
   f. If you were handed a blank stick figure and knew only the person's name, could you fill in the rest of the information? ***No. You would not know a person’s height, GPA, shoe preference, etc. just by knowing their name.***
8. Explain to the students that the general categories of information, such as a person’s height, are called **variables**. Variables are simply characteristics of an object or person. As statisticians, we use variable names to organize data into a simplified form so that a computer can read them. This will be discussed further in Lesson 3.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 3: Data Structures

Objective:

Students will learn that data can be represented in rectangular format.

Materials:

1. DS journals (must be available during every lesson)
2. *Stick Figures* cutouts (see Lesson 2)

Vocabulary:

variables, numerical variables, categorical variables, rows, columns, rectangular or spreadsheet format, variability

Essential Concepts: Variables record values that vary. By organizing data into rectangular format, we can easily see the characteristics of observations by reading across a row, or we can see the variability in a variable by reading down the column. Computers can easily process data when it is rectangular format.

Lesson:

1. Remind students that they briefly learned what **variables** are during the previous lesson. Have students create their own definitions of the term “variables” and share their responses with their teams. Select a few students in the class to share out their definitions and discuss what could be modified (if anything) to create a more complete definition.
2. Using the *Stick Figure* information from Lesson 2, allow the class to come up with a set of variable names that describe the different categories of information. Note that it is best when variable names are short (one to three words). The variable names for the *Stick Figures* data could possibly be:
   a. Name
   b. Height
   c. GPA
   d. Shoe or Shoe Type
   e. Sport
   f. Friends or Number of Friends
3. Next, have a class discussion about how the values from “Shoe” are different than the values from “Height.”
   a. The values from “Shoe” are either “sneakers” or “sandals”.

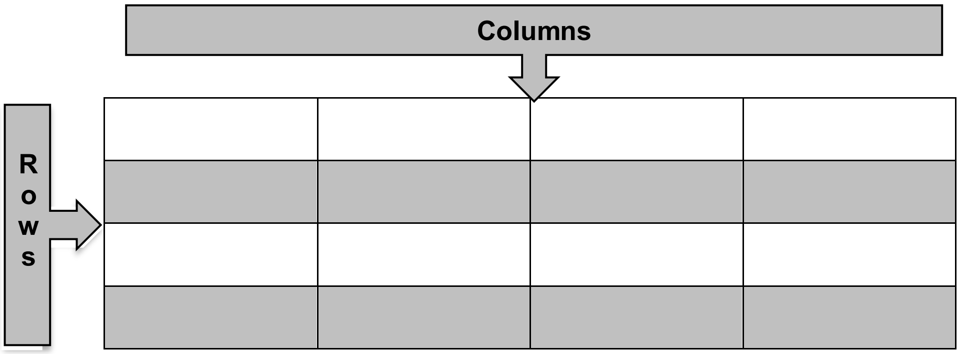
      Note: Other terms for these shoes are acceptable – e.g., tennis shoes, flip flops, closed-toe, open-toe, etc.
   b. The values from “Height” are 72, 68, 61, 66, 65, 61, 67, and 64.
4. Students should notice that the “Shoe” variable consists of categories or groupings, and the “Height” variable consists of numbers. Therefore, we can classify variables into two types: **categorical variables** and **numerical variables**. Typically, categorical variables represent values that have words, while numerical variables represent values that have numbers. **Note**: Categorical variables can sometimes be coded as numbers (e.g., “Gender” could have values 0 and 1, where 0=Male and 1=Female).
5. As a class, determine which variables from the *Stick Figures* data are numerical, and which variables are categorical. The students should create two lists in their DS journals similar to the ones below (the correct classifications are in grey):

Numerical	**Categorical**
1. ***Height***	1. ***Name***
2. ***GPA***	2. ***Shoe***
3. ***Friends***	3. ***Sport***

6. Explain that although we can understand many different representations of data (as evidenced by the posters from Lesson 2), computers are not as capable. Instead, we need to organize data in a structured way so that a computer can read and interpret them.

7. One way to organize the data is to create a **data table** that consists of **rows** and **columns**. We can define this type of organization as **rectangular format**, or **spreadsheet format**.

8. Display a generic table on the board (see example below) and explain that the columns are the vertical portions of the table, while the rows are the horizontal portions. Another way to think of it is that columns go from top to bottom, and rows go from left to right.



9. Ask students:
   a. What should each row represent? ***Each row should represent one observation, or one stick figure person in this case.***
   b. What should each column represent? ***Each column should represent one variable. As you go down a column, all the values represent the same characteristic (e.g., Height).***

10. On the board, draw the following table and have the students copy it into their DS journals (be sure to use variable names agreed upon by the class):

Name	Height	GPA	Shoe	Sport	Friends

11. In teams, students should complete the data table using all 8 of the *Stick Figures* cards. Each row of the table should represent one person on a card.

12. Engage the class in a discussion with the following questions:
   a. Do any of the people in the data have the same value for a given variable? In other words, does a value appear more than once in a column? Give two examples. ***Answers will vary. One example could be that Dakota, Kamryn, Emerson, and London all wear sneakers. Another example could be that Charlie and Jessie are both 61 inches tall.***

b. Do any of the people in the data have different values for a given variable? ***Absolutely. There are many instances of this in the data table.***

13. Discuss the term **variability**. As in question (b) above, the values for each variable vary depending on which person we are observing. This shows that the data has variability, and the first step in any investigation is to notice variability. We can see the relationship between the terms **variable** and **variability**. The word “variable” indicates that values vary.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 4: The Data Cycle

Objective:

Students will learn about the stages of the Data Cycle.

Materials:

1. *The Data Cycle* file (LMR_1.3_Data Cycle)
2. Computer, projector, or board and markers/chalk
3. Printed description of each stage of the Data Cycle (refer to step 3 in the lesson)
4. *The Data Cycle Spinners* handout (LMR_1.4_Data Cycle Spinners)
5. RStudio: https://portal.idsucla.org
6. Article headline: *People Who Order Coffee Black Are More Likely To Be Psychopaths found at:* https://www.huffpost.com/entry/black-coffee-psychopath_n_561baf08e4b0dbb8000f150f
7. *Dude Map* found at: https://qz.com/316906/the-dude-map-how-american-men-refer-to-their-bros/
8. *Bros & Dudes Graphics* handout (LMR_1.5_Bros & Dudes Graphics)
9. Sticky notes
10. Poster paper

Vocabulary:

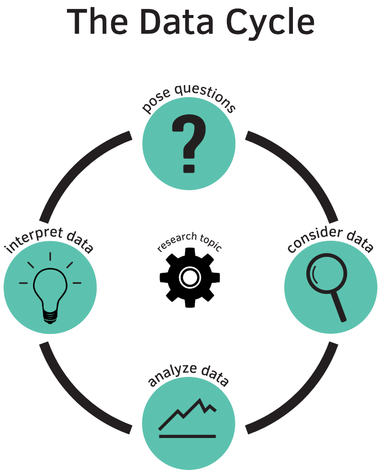
data cycle, statistical questions, investigative questions, data collection, data analysis, data interpretation

Essential Concepts: A statistical investigation consists of cycling through the four stages of the Data Cycle. The term statistical questions encompasses the variety of questions asked during the statistical problem-solving process which support statistical thinking and reasoning. Statistical investigative questions are perhaps the most important because they are challenging to learn and are the types of questions that determine whether an analysis is productive or not. Statistical investigative questions are questions that address variability and are productive in that they motivate data collection, analysis, and interpretation. The Data Collection phase might consist of collecting data through Participatory Sensing or some other means, or it might consist of examining previously collected data to determine the quality of the data for answering the statistical investigative questions. Data Analysis is almost always done on the computer and consists of creating relevant graphics and numerical summaries of the data. Data Interpretation is involved with using the analysis to answer the statistical investigative questions.

Lesson:

1. During the past few lessons, we have discussed what data are, how to collect and organize them, and how their values can vary. But what do we do with all this data? How can we navigate it and turn it into something useful to us?

2. Inform students that they will be learning about the **Data Cycle** today. The Data Cycle is a guide we can use when learning to think about data. We usually start with posing statistical investigative questions. Display the graphic from *The Data Cycle* file (LMR_1.3):



3. Display the Data Cycle on the board or on a projector, and give a brief explanation of the 4 components (listed below).

   Note: we will explore each component of the Data Cycle more explicitly throughout the course.

   a. **Pose Statistical Investigative Questions**: Statistical Investigative questions are questions that address variability and can be answered with data.

   b. **Consider Data**: This is the process of observing and recording data, or of examining previously collected data to make sure it meets the needs of the investigation.

   c. **Analyze Data**: During analysis, tables, graphs, and summaries of the data are produced to help us find patterns and relationships.

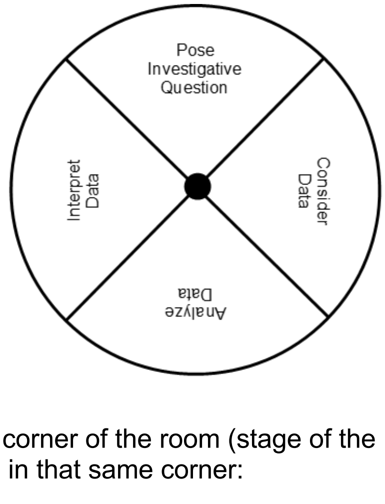
   d. **Interpret Data**: The statistical investigative questions are answered by referring to the tables, graphs, and summaries made in the Data Analysis phase.

4. Almost all statistical investigations begin with statistical investigative questions. There are times when the questions may be given to us, so we might start at the data collection step, but this should ideally be our starting point.

5. As an example, explain that you might ask a person “How old are you?” Although this is a question, it is NOT a statistical investigative question because we are only asking one person so there is no variability in the data. The question “How old are you?” is a survey question that you might ask if you were trying to answer the investigative question “How old are the students in my school?” We would need to collect data to answer the question and we would expect student’s ages to vary.

6. To help students get a firm understanding of the Data Cycle and how each component is connected, they will participate in a *Four Corners* strategy (see Instructional Strategies in Teacher Resources). Write down the name of each stage in the Data Cycle on a sheet of paper and include the description of that particular stage (see step 3 for descriptions). Then tape each sheet on a different corner of your room.

7. Explain to the students that you are going to display different artifacts from statistical investigations on the projector. For each artifact, they will move to the corner of the room they feel that artifact represents (posing a statistical investigative question, consider data, analyze data, interpret data). If you have limited space in your classroom or for students that cannot physically participate, you may consider printing LMR_1.4_Data Cycle Spinners. Students can participate by pointing to the spinner.



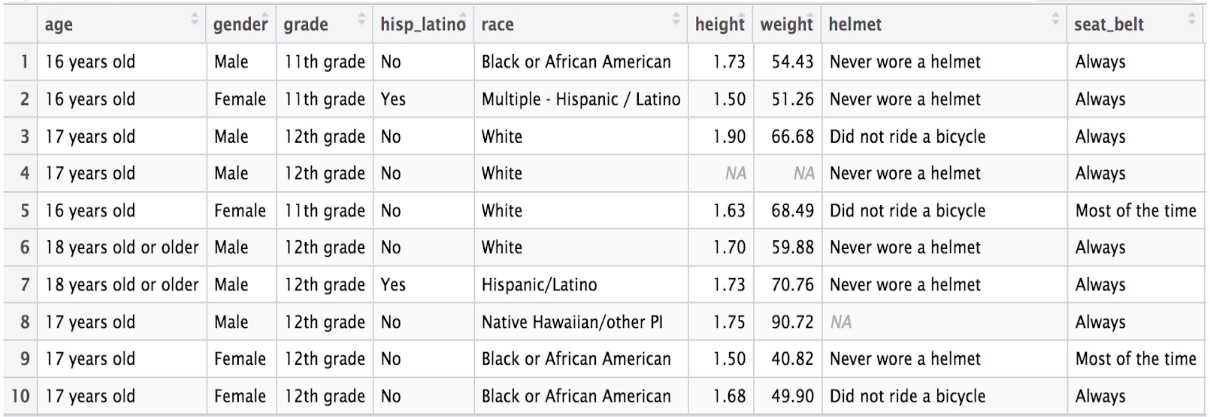
8. Once students have chosen a corner of the room (stage of the Data Cycle) they will discuss the following with their classmates in that same corner:

   a. What part of the data cycle does the artifact represent (posing a statistical investigative question, consider data, analyze data, interpret data)? Why do we think that?

   b. What questions or wonderings do we have about the artifact?

9. Allow each group time to discuss the questions and have one member from each team (corner) share the answers to the questions. This activity is not about having a correct answer. It is about having students begin to think critically about statistical artifacts that they are constantly consuming. Data are encountered through visualizations, reports from scientific studies, journalists' articles and websites. This activity is meant to begin to develop students' statistical habits of mind.

10. Artifact 1: Spreadsheet of the CDC data.



	age	gender	grade	hisp_latino	race	height	weight	helmet	seat_belt
1	16 years old	Male	11th grade	No	Black or African American	1.73	54.43	Never wore a helmet	Always
2	16 years old	Female	11th grade	Yes	Multiple - Hispanic / Latino	1.50	51.26	Never wore a helmet	Always
3	17 years old	Male	12th grade	No	White	1.90	66.68	Did not ride a bicycle	Always
4	17 years old	Male	12th grade	No	White	NA	NA	Never wore a helmet	Always
5	16 years old	Female	11th grade	No	White	1.63	68.49	Did not ride a bicycle	Most of the time
6	18 years old or older	Male	12th grade	No	White	1.70	59.88	Never wore a helmet	Always
7	18 years old or older	Male	12th grade	Yes	Hispanic/Latino	1.73	70.76	Never wore a helmet	Always
8	17 years old	Male	12th grade	No	Native Hawaiian/other PI	1.75	90.72	NA	Always
9	17 years old	Female	12th grade	No	Black or African American	1.50	40.82	Never wore a helmet	Most of the time
10	17 years old	Female	12th grade	No	Black or African American	1.68	49.90	Did not ride a bicycle	Always

Note: You can display this spreadsheet below using RStudio by running the following commands: `data(cdc)` `View(cdc)`

   a. What part of the data cycle does the artifact represent (posing a statistical investigative question, consider data, analyze data, interpret data)? Why do we think that? ***Answers will vary.***

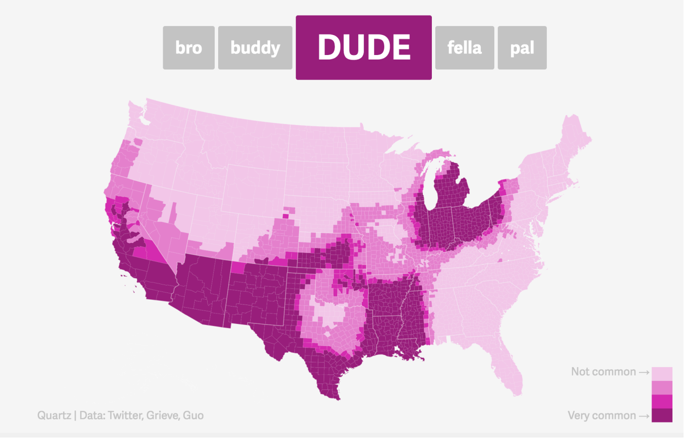
   b. What questions or wonderings do we have about the artifact? ***Students should begin developing statistical habits of mind. They should be interrogating the data by asking questions such as: Who is this data about? What was the purpose of collecting the data? What was the survey question asked to collect the data?***

11. Artifact 2: Headline from Huffington Post *People Who Order Coffee Black Are More Likely To Be Psychopaths* found at: https://www.huffpost.com/entry/black-coffee-psychopath_n_561baf08e4b0dbb8000f150f

   a. What part of the data cycle does the artifact represent (posing a statistical investigative question, consider data, analyze data, interpret data)? Why do we think that? ***Answers will vary.***

   b. What questions or wonderings do we have about the artifact? ***Students should be interrogating this headline with questions like: What type of study was this? Who funded the study? What was the purpose of the study? How was the variable measured?***

12. Artifact 3: The Dude map found at: https://qz.com/316906/the-dude-map-how-american-men-refer-to-their-bros/



Quartz | Data: Twitter, Grieve, Guo

a. What part of the data cycle does the artifact represent (posing a statistical investigative question, consider data, analyze data, interpret data)? Why do we think that? ***Answers will vary.***

b. What questions or wonderings do we have about the artifact? ***Students should be asking questions like: What was the purpose of this study? What variables were measured and how were they measured?***

13. Inform students that the *Dude Map* was created for the *Quartz* website by Nikhil Sonnad as a data visualization. He collected the data via Twitter. The graphic shows how common the terms: bro, buddy, dude, fella, and pal are when referring to friends throughout the United States.

14. Ask students to return to their seats, take out their DS journal and make a sketch of the Data Cycle making sure to include the names of the four stages (Pose statistical investigative question, consider data, analyze data, interpret data).

15. Ask students to write *Dude Map* under the analyze data of their data cycle and the information about where the data came from (see #13) in the consider data part of their Data Cycle sketch.

16. Have each team discuss a possible investigative question that could be answered using the *Dude Map* graphic. Have the reporter/ recorder write the question on a sticky note and the resource manager bring it up to the board.

17. Lead a class discussion around the investigative questions the student teams created, and as a class, choose one to write down as an example. ***Example: Where in the United States is the term dude more common to use when referring to a friend?***

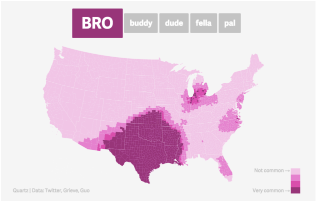
18. Allow the teams to work together to answer the investigative question. Ask the reporter/ recorder to share their team's interpretation. Have students write down the answer that resonated the most with them under the interpret part of the data cycle.

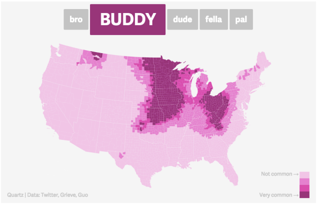
19. Assign ONE of the pages from the *Bros & Dudes Graphics* handout (LMR_1.5) to each team. There are 10 different versions of word pairings (10 combinations of 2 words chosen from the 5 options), so multiple teams will have the same graphic if there are more than 10 teams in a class.

Team Members: ______________________ ______________________ Date: ______________

______________________ ______________________

The dude map: How Americans refer to their bros
(http://qz.com/316906/the-dude-map-how-american-men-refer-to-their-bros/)
Created by Nikhil Sonnad, December 23, 2014, for *Quartz* website.
The data originally came from accessing Twitter feeds.





What statistical question(s) would you ask given the above graphics? Give 2 examples.

(1) ____

____

(2) ____

____

Version (a)

LMR_1.5

20. The goal of this activity is for each team to complete a full statistical investigation with the *Bros & Dudes Graphics* assigned to them. Tell the teams that they need to create a Data Cycle poster using their assigned graphic for the analyze stage. The cycle should be clearly labeled and have appropriate responses for each of the 4 components.

    For example, a team given the "Bro" and "Buddy" graphics might come up with the following questions: Which region of the US is most likely to use the term "Bro" when referring to a friend? Do the coastal areas prefer different terms than the Midwest? Is there a difference between the northern states versus southern states?

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students reformulate any investigative questions generated by their team about the *Bros & Dudes Graphics* that could not be answered so that they can be answered.

***Lesson 5: So Many Questions* [The Data Cycle: Pose Questions]**

Objective:
Students will learn the features of a good statistical investigative question.

Materials:

1. Post-its

Vocabulary:
(statistical) investigative questions

Essential Concepts: Statistical investigative questions typically begin with a vague general question, then develop into a precise question. The process of developing or creating a good investigative question is iterative and requires time and effort to get right. In her 2021 paper, What Makes a Good Statistical Question, Dr. Pip Arnold identified the following as features of a good investigative question:
(1) The variable(s) of interest is/are clear
(2) The group or population we are interested in is clear
(3) The question can be answered with the data
(4) The question asks about the whole group, not an individual or portion of the group
(5) The intention is clear (e.g., summary, comparison, association, time series)
(6) The question is one that is worth investigating, is interesting, and has a purpose

Lesson:

1. Entrance Slip (see Instructional Strategies in Teacher Resources): Each student should submit a ticket that displays the 4 components of the Data Cycle.
2. Inform students that they will learn about what makes a good investigative question. Ask them recall the definition of an investigative question:

   Investigative questions are questions that address variability and can be answered with data. They are questions we ask of the data. A good way to determine this is to ask: *Do we need to see the data to answer the question?*

3. Remind the students of the two questions from the previous lesson, noting that one of the questions was an investigative question, and the other was not:
   a. How old am I?
   b. How old are the students in my school?
4. In pairs, ask students to analyze each question using the definition of an investigative question and come to an agreement about which one is an investigative question.
5. Using Agree/Disagree (see Instructional Strategies in Teacher Resources), ask a pair of students for their results. Discuss why the first question **IS NOT** an investigative question (***there is only one possible value so there is no variability in the data***) and why the second question **IS** an investigative question (***not all students are the same age. The ages vary, so there is variability in the data***).
6. Ask students to think of the data they collected about the stick figures (name, GPA, friends, sport, height, shoe). Inform them that the researchers used the following survey questions to collect the data:
   a. What is your name?
   b. What is your GPA?
   c. How many friends do you have?

d. What sport do you play?

e. How tall are you in inches?

f. What type of shoe do you mostly wear?

Survey questions are another example of a type of statistical question, but with a different purpose to investigative questions. Survey questions are questions we ask to get the data.

7. Tell students that it is important to know exactly what survey questions were asked to collect the data before asking investigative questions. For example, we saw an image of a ball next to each stick figure but we don't know if that represents a sport they like to watch, their favorite sport, or a team they are on.
8. In teams, ask students to create investigative questions that could be answered using the data collected about the stick figures. Introduce the sentence stem "I wonder…" to help students get started. Have the Recorder/Reporter record the questions on post-its.
9. Ask the teams to identify which variable(s) each question is investigating by having them circle the variable name(s) within their investigative questions.
10. Have the Task Manager organize their group's investigative questions on the board, placing investigative questions that incorporate only one variable on the left-hand side of the board and investigative questions that incorporate two or more variables on the right-hand side.

    Note: This sorting activity will help students begin to distinguish between different types of investigative questions.

    Summary investigative questions are questions about a single variable.

    Comparison investigative questions compare a numerical variable across groups.

    Association investigative questions look for a relationship between paired numerical or paired categorical variables.

11. As a class, begin the process of transforming some of the summary investigative questions so that they have all of the features of a good investigative question. Here is an example to get you started:

    Initial investigative question: ***I wonder who has the most friends?***

Feature	Explanation
The variable(s) of interest is/are clear	Yes. The variable of interest is the number of friends.
The group or population we are interested in is clear	No. Teacher should ask: "Who did the researchers want to learn about?" These stick figures.
The question can be answered with the data	Yes. The researchers collected data on the number of friends.
The question asks about the whole group, not an individual or portion of the group	No. This question is about an individual stick figure. Teacher should ask: "How can we reword the question to include the whole group?" How many friends do ... have?
The intention is clear (e.g., summary, comparison, association)	It is clear that this is a summary investigative question (single variable), specifically the number of friends.
The question is one that is worth investigating, is interesting, and has a purpose	For students, this might be something interesting.

Reworded investigative questions after going through the criteria: ***I wonder how many friends this group of stick figures have?***

12. As a class, apply the same process to a few of the comparison and association questions. Initial investigative question: ***I wonder if someone who plays a specific sport has more friends?***

Feature	Explanation
The variable(s) of interest is/are clear	Yes. This seems to be a comparison investigative question comparing the number of friends within the sport played.
The group or population we are interested in is clear	No. Teacher should ask: “Who did the researchers want to learn about?” These stick figures.
The question can be answered with the data	Yes. The researchers collected data on the number of friends and the sport the stick figures played.
The question asks about the whole group, not an individual or portion of the group	No. The word someone gives the impression that we are interested in one observation.
The intention is clear (e.g., summary, comparison, association)	The intent is somewhat clear. This seems to be a comparison investigative question between the sport the stick figures played and the number of friends each stick figure had. Teacher should ask: “What is the variable that is being compared? Which groups within the sport variable are you comparing (all groups, specific groups)?”.
The question is one that is worth investigating, is interesting, and has a purpose	For students, this might be something interesting.

Reworded investigative question after going through the criteria:

I wonder if there is a difference in the typical number of friends the stick figures have based on the sport they play?

I wonder if these stick figures who play soccer tend to have more friends that these stick figures who play tennis?

13. Using the criteria of a good statistical investigative question, student teams will go back and modify their statistical investigative questions. Facilitators will ensure the team goes through the criteria for each investigative question. Task Managers will encourage everyone to contribute. Resource Managers will ensure all materials are easily accessible for recording and reporting. Recorders in each team will capture team members’ responses while the teacher circulates the room to check for understanding.

14. Ask the Reporters of selected teams to share their revised statistical investigative question(s). Students in the audience will listen to the presentations and provide feedback about each team’s statistical investigative question(s). Be sure to discuss disagreements before moving on to different questions.

15. Inform students that in the next lesson, they will begin using the Data Cycle to learn about their food habits. To prepare for this, students should begin collecting the “Nutrition Facts” labels from foods/snacks they typically eat.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Ask students to bring at least 2 cutouts of the “Nutrition Facts” labels of the snacks they typically eat (e.g., chips, yogurt, blended drinks, etc.).

Note: An alternative to collecting “Nutrition Facts” labels is to print them from an online source and bring the printouts to class.

***Lesson 6: What Do I Eat?* [The Data Cycle: Consider Data]**

Objective:

Students will collect data using paper and pencil to understand the challenges of organizing, storing, and sharing data. They will learn that there must be an agreement about the variables that need to be recorded in order to attain consistency.

Materials:

1. Video: Jamie Oliver's *Food Revolution* found at: https://youtu.be/I0vYwqkoktM

   Note: If the video is unavailable, search for “"Jamie Oliver's Food Revolution What's In a Sundae". The video should be 5-6 minutes in length.

2. Nutrition Facts labels or pictures (collected previously by students)
   Note: If needed, use *Nutrition Facts Cutouts* handout (LMR_1.7_Nutrition Facts Cutouts)

3. *Food Habits Data Collection* handout (LMR_1.8_Food Habits Data Collection)

Vocabulary:
data set(s)

Essential Concepts: After raising statistical questions, we examine and record data to see if the questions are appropriate.

Lesson:

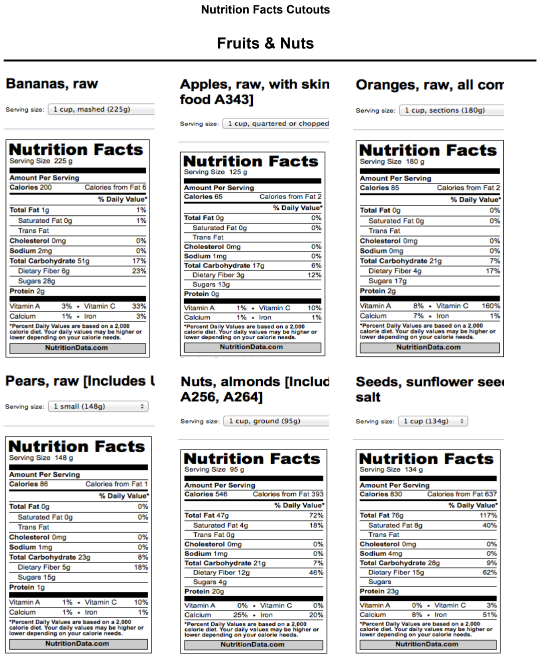
1. Inform students that today's lesson will focus on the Data Collection component of the Data Cycle.

2. To motivate this, the students will watch a short video of an episode of Jamie Oliver's show titled *Food Revolution* found at: https://youtu.be/I0vYwqkoktM. This video was recorded at a Los Angeles high school.

   a. As the students watch the video, they should use their DS journals to write down their comments and/or reactions to what they see and hear and be ready to share out.
   b. After sharing out some of their responses, have the students respond to the following question in their DS journals: **Why should I care about what I eat?**
   c. Student teams will share their reactions and responses by engaging in a Silent Discussion (see Instructional Strategies in Teacher Resources).

3. Have students recall the *Stick Figures* activity from Lesson 2. During that activity, they collected data about other people. But today, they are going to be collecting data about themselves and the foods they eat.

4. Students should have Nutrition Facts labels available from food/snacks they consumed at home between the previous lesson and today. Note: If some students forgot to bring any, then you can pass out some of the *Nutrition Facts Cutouts* (LMR_1.7) for them to use instead.



Nutrition Facts Cutouts

Fruits & Nuts

Bananas, raw

Serving size: 1 cup, mashed (225g)

Nutrition Facts
Serving Size 225 g

Amount Per Serving	
Calories 200	Calories from Fat 6
	% Daily Value*
Total Fat 1g	1%
Saturated Fat 0g	1%
Trans Fat	
Cholesterol 0mg	0%
Sodium 2mg	0%
Total Carbohydrate 51g	17%
Dietary Fiber 6g	23%
Sugars 28g	
Protein 2g	
Vitamin A 3% • Vitamin C	33%
Calcium 1% • Iron	3%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

NutritionData.com

Apples, raw, with skin food A343]

Serving size: 1 cup, quartered or chopped

Nutrition Facts
Serving Size 125 g

Amount Per Serving	
Calories 65	Calories from Fat 2
	% Daily Value*
Total Fat 0g	0%
Saturated Fat 0g	0%
Trans Fat	
Cholesterol 0mg	0%
Sodium 1mg	0%
Total Carbohydrate 17g	6%
Dietary Fiber 3g	12%
Sugars 13g	
Protein 0g	
Vitamin A 1% • Vitamin C	10%
Calcium 1% • Iron	1%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

NutritionData.com

Oranges, raw, all com

Serving size: 1 cup, sections (180g)

Nutrition Facts
Serving Size 180 g

Amount Per Serving	
Calories 85	Calories from Fat 2
	% Daily Value*
Total Fat 0g	0%
Saturated Fat 0g	0%
Trans Fat	
Cholesterol 0mg	0%
Sodium 0mg	0%
Total Carbohydrate 21g	7%
Dietary Fiber 4g	17%
Sugars 17g	
Protein 2g	
Vitamin A 8% • Vitamin C	160%
Calcium 7% • Iron	1%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

NutritionData.com

Pears, raw [Includes U

Serving size: 1 small (148g)

Nutrition Facts
Serving Size 148 g

Amount Per Serving	
Calories 86	Calories from Fat 1
	% Daily Value*
Total Fat 0g	0%
Saturated Fat 0g	0%
Trans Fat	
Cholesterol 0mg	0%
Sodium 1mg	0%
Total Carbohydrate 23g	8%
Dietary Fiber 5g	18%
Sugars 15g	
Protein 1g	
Vitamin A 1% • Vitamin C	10%
Calcium 1% • Iron	1%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

NutritionData.com

Nuts, almonds [Includ A256, A264]

Serving size: 1 cup, ground (95g)

Nutrition Facts
Serving Size 95 g

Amount Per Serving	
Calories 546	Calories from Fat 393
	% Daily Value*
Total Fat 47g	72%
Saturated Fat 4g	18%
Trans Fat 0g	
Cholesterol 0mg	0%
Sodium 1mg	0%
Total Carbohydrate 21g	7%
Dietary Fiber 12g	46%
Sugars 4g	
Protein 20g	
Vitamin A 0% • Vitamin C	0%
Calcium 25% • Iron	20%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

NutritionData.com

Seeds, sunflower see salt

Serving size: 1 cup (134g)

Nutrition Facts
Serving Size 134 g

Amount Per Serving	
Calories 830	Calories from Fat 637
	% Daily Value*
Total Fat 76g	117%
Saturated Fat 8g	40%
Trans Fat	
Cholesterol 0mg	0%
Sodium 4mg	0%
Total Carbohydrate 28g	9%
Dietary Fiber 15g	62%
Sugars	
Protein 23g	
Vitamin A 0% • Vitamin C	3%
Calcium 8% • Iron	51%

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

NutritionData.com

LMR_1.7

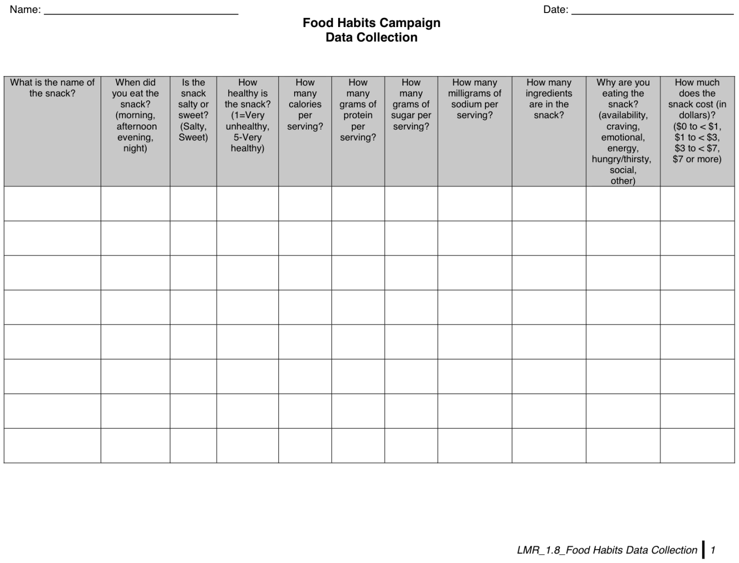
5. For 3-5 minutes, allow students to collect any data they can from the label and record it in their DS journals. This should be done individually.

6. Once they have collected their facts, ask students to compare and contrast their data with their team members. They need to respond to:

    a. How are their **data sets** similar?
    b. How are their **data sets** different?

7. Gather the students as a whole group and ask them to share out the similarities and differences they discussed. Be sure to draw responses that show that while some facts collected were the same, there were others that were collected by some students and not by others. Also point to differences in the variables collected and the data structure used.

    a. Ask students to engage in the following individual Quickwrite (see Instructional Strategies in Teacher Resources): How can the data you just gathered be quickly displayed and easily read?
    b. Distribute the *Food Habits Data Collection* handout (LMR_1.8). Ask students to record 8 observations. They can use their own 2 labels for the first observations, and then use some of their team members' labels to complete the table.



Name: ____________________________

Date: ________________________

Food Habits Campaign
Data Collection

What is the name of the snack?	When did you eat the snack? (morning, afternoon evening, night)	Is the snack salty or sweet? (Salty, Sweet)	How healthy is the snack? (1=Very unhealthy, 5-Very healthy)	How many calories per serving?	How many grams of protein per serving?	How many grams of sugar per serving?	How many milligrams of sodium per serving?	How many ingredients are in the snack?	Why are you eating the snack? (availability, craving, emotional, energy, hungry/thirsty, social, other)	How much does the snack cost (in dollars)? ($0 to < $1, $1 to < $3, $3 to < $7, $7 or more)

LMR_1.8_Food Habits Data Collection | 1

LMR_1.8

8. Once they are finished, in pairs, ask students to give a one-word identifier to each variable. For example: “What’s the name of your snack?” = Name
9. Share the one-word variable identifiers with the class by conducting a quick team Whip Around (see Instructional Strategies in Teacher Resources).
10. For homework, students will begin to formulate statistical questions based on their *Food Habits* data.
11. Inform students that they are permitted to bring mobile devices to the next class.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Ask students to examine the data in their *Food Habits Data Collection* handout (LMR_1.8) and to generate two simple and two complex statistical questions that they think can be answered by the data they collected. A simple statistical question involves one variable, whereas a complex statistical question involves two or more variables.

Students may bring their mobile devices to the next class for data collection purposes.

***Lesson 7: Setting the Stage* [The Data Cycle: Collect Data]**

Objective:

Students will begin to collect and record data to learn more about their own eating habits, as well as those of their classmates. They will learn about data that is collected by a Participatory Sensing campaign, and also about privacy issues and photo ethics when collecting and sharing data.

Materials:

1. Students' own mobile devices (smartphone or tablet compatible with iOS or Android)
2. Access to App Store or Google Play Store in student devices to download IDS UCLA app
3. Login information (username and password) for each student—**generated and ready for distribution prior to the lesson**
4. *Food Habits Campaign* guidelines (LMR_U1_Campaign_Food Habits)

Vocabulary:

Participatory Sensing, campaign, surveys, images, GPS, ethics, photo ethics

Essential Concepts: In Participatory Sensing, we humans behave as if we are robot sensors, collecting data whenever a "trigger" event occurs. Our ability to learn about the patterns in our life through these data depends on our being reliable data collectors.

Lesson:

1. Become familiar with the *Food Habits Campaign* guidelines (shown at the end of this lesson), especially the big questions found under "The Issue," to help guide students during the campaign (see Campaign Guidelines in Teacher Resources).

2. Distribute the usernames and passwords to student team leaders (make sure safeguards are in place so that only the owner of the username and password is able to see this information). Ask team leaders to distribute their team's information when you are ready to download the IDS UCLA app.

3. Ask students to think about the Nutritional Facts labels from which they collected data in the previous lesson, and answer the following in their DS journals:

   a. What questions would you want answered about eating habits?
   b. What can you do to find out about your own eating habits?

4. Ask students to refer back to their reactions and comments from Jamie Oliver's video and have them ponder the question: **What are we really eating?**

5. Over the next 9 days, they will engage in a **Participatory Sensing campaign** in which they will act as human sensors to collect data about themselves. The data collected will be used to analyze their classmates' and their own snacking habits.

6. For this unit, they will collect data about every snack they eat.

   __**Note:**__ Students should **NOT** collect data for full meals like breakfast, lunch, or dinner. Data should only be collected for anything eaten in between meals, like fruit, chips, cookies, nuts, sodas, etc.

7. Ask students why it makes sense to study snacks specifically. Brainstorm some questions that could be answered using the snack-only data that would be hard to answer if the data included meals as well.

8. Inform students that they will be taking part in a specific data collection method know as **Participatory Sensing** via a mobile application. This application can gather data via **surveys**, **images,** and **GPS** tracking.

9. Make it clear to students that the reason they are collecting the data is to learn more about themselves and their classmates, NOT to provide data for an external data collection team. Students occasionally have the misconception that when they use the Participatory Sensing app, they are providing data to external researchers, such as UCLA.

10. Inform students that they are now going to engage in their own first Participatory Sensing data collection experience, in which they will collect their own data using a smart device. Depending on the device, there are 3 different options available:

    a. **Android**. A native Android application called "IDS UCLA" is available from the Google Play Store.
    b. **iOS (Apple devices**) The mobile application called "IDS UCLA" is available from the iOS App Store.
    c. **No mobile device - browser-based version.** For students that do not have a mobile device (or an unsupported device, such as a Windows phone or Blackberry), a browser-based version to perform data collection is available at https://portal.idsucla.org
    Click on the **Survey Taking icon** on the page.

11. Once students have downloaded the app or have found the website, ask team leaders to distribute the login information. Students will need to keep this information in a safe place for the entire duration of this course. **Emphasize the importance of keeping their username and password confidential.** When students receive their login information, they can log in to the app. If students have trouble with their logins, the teacher has the ability to reset a student's password.

12. Once logged into the app or the browser-based version, students will see the **Campaigns** in which they will participate. They will then select the campaign by tapping the name of the campaign. If no campaigns are visible, ask them to tap the refresh option, located on the top right-hand side of the screen.

13. Using one of their nutrition facts cutouts or pictures, ask students to complete their first survey by going through the questions in the app.

14. After every student has had the opportunity to complete at least one survey, ask students the meaning of the word **ethics**. For this course, they will need to understand **photo ethics**. They may NOT take pictures of any person's identifying features such as faces, hair, hands, tattoos, etc. For this campaign, they may only take pictures of their snacks and/or the nutrition facts labels.

    Note to teacher: Inspect students' data collection photos throughout the data collection period and before each data collection, monitoring to ensure that no inappropriate images are shared. If you believe a photo is inappropriate, please delete the data entry immediately.

15. Setting reminders: The IDS UCLA app has a reminder feature to help students in their data collection journey. Show students that they can set reminders directly on the app by tapping the menu button on the top left-hand side of the screen and selecting **Reminders** from the menu.

16. Data collection norms: Ask students how many snacks they think they eat a day. From this, come up with an approximate number of surveys they think each student should complete during the data collection period (days 7 through 15).

17. Inform students that you will be monitoring their data collection to make sure that everyone is submitting surveys regularly.

18. Go over the previous day's homework. Ask the facilitator from each table group conduct a round robin during which each team member shares one simple statistical question and one complex statistical question. The recorder/reporter will select and share out one of the team's simple statistical questions and one complex statistical question with the class.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

For the next 9 days, students will collect nutritional facts data using the *Food Habits* Participatory Sensing campaign on their smart devices or via web browser.

Campaign Guidelines – Food Habits

1. The Issue:

Although we might take it existence for granted today, the Nutrition Facts label was not always required to be on food packages. It wasn't until 1990 that the Nutrition Labeling Education Act mandated food companies to provide information on food label to help consumers make wiser choices about what they eat. This raises some interesting questions:

1) Does knowing nutritional information about my snacks help me change my habits?
2) What is my snacking pattern?
3) How good am I at rating the healthiness of my snack?
4) Do I tend to eat healthy? How do I compare to my class? How does my class compare to the rest of the country?

2. Objectives:

Upon completing this campaign, students will have the enduring understanding that interpreting graphs provides useful information about the real world as viewed through the data represented in the graphs. We can explore the relationship between two variables, and if there is a relationship, it is driven by the change in the independent variable, x, which causes a change in the dependent variable, y.

3. Survey Questions: (students will enter data about the snacks they consume):

Consider Data: Before students submit a survey for their first snack, a class consensus of the meaning of the variables must be reached so that proper analysis and interpretations can be made. Two examples are listed below:

when - If students have different definitions of "evening", it might make it hard to compare snacking patterns across students. As a class, come to a consensus about what time intervals are considered morning, afternoon, evening and night.

cost - If a student has a bowl of cereal as a snack, are they going to include the cost of the entire box or are they going to calculate the unit cost for one serving? This needs to be a class decision.

Prompt	Variable	Data Type
What's the name of your snack?	name	text
When did you eat the snack?	when	categorical morning afternoon evening night
Is your snack salty or sweet?	salty_sweet	categorical Salty Sweet
How healthy is the snack? (1 = Very unhealthy, 5 = Very healthy)	healthy_level	numerical 1 2 3 4 5

How many calories per serving?	calories	numerical
How many grams of protein per serving?	protein	numerical
How many grams of sugar per serving?	sugar	numerical
How many milligrams of sodium per serving?	sodium	numerical
How many ingredients are in the snack?	ingredients	numerical
Why are you eating the snack?	why	categorical availability craving emotional energy hungry/thirsty social other
How much does the snack cost (in dollars)?	cost	categorical $0 to < $1 $1 to < $3 $3 to < $7 $7 or more
Take a picture (optional).	snack_image	photo
AUTOMATIC	location	lat, long
AUTOMATIC	time	time
AUTOMATIC	date	date
AUTOMATIC	user	user id

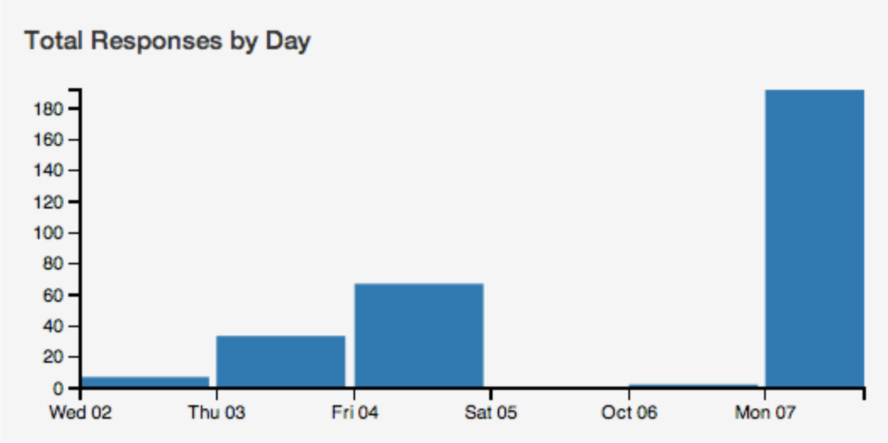
When should you take the survey? If possible, take the survey every time you eat a snack or at the end of the day. Reminders can be set to ensure survey completion.

How long should the campaign last? About nine days. Ideally, two of these days will include a weekend.

4. **Motivation:**

   As a class, come to an agreement about how many surveys each student should submit. All students should submit roughly the same number of surveys, and each student should submit at least four surveys. After the first day, use the campaign monitoring tool to see who has collected data. After two to three days, direct students' attention to the Total Responses by Day plot and comment on any patterns. For example, if they see a plot like the one below, ask "What story does this tell us about our data collection?"

Story: They collected a lot of data together in class. Data collection increased every day from Wednesday to Friday. There was little to no data collection over the weekend. Data collection peaked on Monday - there were over 180 responses!



Discuss data collection issues. What makes it hard? Does this affect the quality of data? What sort of snacks are you less likely to enter?

5. **Technical Analysis:**

   Students will use the Dashboard and Plot App as well as RStudio.

6. **Guiding Questions:**

   a. At what time of day do we eat the healthiest snacks?
   b. When did you snack? How does this compare to the rest of the class?
   c. Typically, how healthy were your snacks? How does this compare to the class as a whole?
   d. How good are we at identifying healthy and unhealthy snacks?

7. **Report:**

   Students will complete a practicum in which they answer a statistical question based on the Food Habits data collected.

Visualizing Data

Instructional Days: 14

Enduring Understandings

Data collection methods affect what we can know about the real world. Visual representations help tell stories with data. Distributions of numerical and categorical variables help describe variability in the data. Technology and computers allow us to visualize complex relationships in data.

Engagement

Students will view the video called *The Value of Data Visualization* to help them understand the importance of graphical representations of data. Discussion questions will allow students to begin to think about how they would want see a data set visualized. The video can be found at: https://www.youtube.com/watch?v=xekEXM0Vonc.

Learning Objectives

Statistical/Mathematical:

S-ID 1: Represent data with plots on the real number line (dotplots, histograms, bar plots, and boxplots).

S-ID 3: Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

S-ID 6: Represent data on two quantitative variables on a scatterplot and describe how the variables are related.

Data Science:

Create visualizations with data. Learn the difference between plots used for categorical and numerical variables. Interpret and understand graphs of distributions for numerical and categorical variables.

Applied Computational Thinking Using RStudio:

- Learn to download, load, upload, and work with data using RStudio syntax and structure.
- Create appropriate graphical displays of data.
- Differentiate between observations and variables.
- Learn to use objects, functions, and assignments.

Real-World Connections:

Students will continue to understand that data on its own is just collected; but once interpreted, it can lead to discoveries or understandings.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it is used and how to work with it.

2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

3. Students will use complex sentences to write informative short reports that use data science concepts and skills.

Data File or Data Collection Method

Data Files:

1. Students' *Food Habits Campaign* Data
2. CDC Data File

Legend for Activity Icons

🎥	🗨	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 8: Tangible Plots **[The Data Cycle: Analyze Data]**

Objective:

Students will learn how distributions help us organize and visualize data values, and that the shapes of the distributions give us information about the variability of the data.

Materials:

1. Computer and projector for Campaign Monitoring
2. Video: *Value of Data Visualization* found at:
   https://www.youtube.com/watch?v=xekEXM0Vonc
3. Nutrition facts labels or pictures (collected previously by students)
4. *Food Habits Data Collection* handout (from Lesson 6, LMR_1.8)
5. 3 pieces of tape per student
6. Poster paper
7. Dot stickers or sticky notes
8. *Tangible Plot* handout (LMR_1.9_Tangible Plot)

Vocabulary:

x-axis, y-axis, visualization, range, minimum, maximum, frequency, distribution, typical, symmetric, range, data points, dotplot

Essential Concepts: Distributions organize data for us by telling us (a) which values of a variable were observed, and (b) how many times the values were observed (their frequency).

Lesson:

1. **Food Habits Campaign Data Collection Monitoring:**
   a. Display the IDS Campaign Monitoring Tool, found at https://portal.idsucla.org
      Click on **Campaign Monitor** and sign in.
   b. Inform students that you will be monitoring their data collection. This is a good opportunity for teachers to remind students that if their data are not shared, they cannot be used in analysis.
      i. See *User List* and sort by *Total*. Ask: Who has collected the most data so far?
      ii. Click on the pie chart. Ask: How many active users are there? How many inactive users are there?
      iii. See *Total Responses*. How many responses have been submitted?
      iv. Using TPS, ask students to think about what can they do to increase their data collection.
2. Inform students that today they will be learning how to visualize their data.
3. Show the *Value of Data Visualization* video at https://www.youtube.com/watch?v=xekEXM0Vonc, which describes the importance of graphical representations of data. As they watch the video, students should respond to the following in their DS journals:
   a. What is data visualization?
   b. List one example of how visualization can be used to increase data comprehension.
4. Have a whole class discussion regarding the video's last statement: "Your message is only as good as your ability to share it." Ask students:
   a. What does this statement mean?
   b. What makes a good message in terms of data and visualizations?
5. Have students take out their nutrition facts labels or pictures, and also their *Food Habits Data Collection* handout (from lesson 6).

6. On poster paper, make the first quadrant of a coordinate plane. Leave the labels for the **x-axis** and **y-axis** blank for now (see step 10).
7. Distribute 3 pieces of tape to each student. Make sure they fold each piece of tape to make two sticky sides. Have each student tape one sticky side to the back of each label and ask them to have the labels ready to tape onto the poster paper.
8. As a class, ask students to select 2 numerical variables and 1 categorical variable from the *Food Habits Data Collection* handout whose data they would like to see in a **visualization**, which is a picture of the data. For example, students may vote to see a visualization of the following numerical variables: calories per serving, protein per serving; categorical variable: salty_sweet
9. Once students select the variables, inform them that they will be creating a plot with the nutrition facts labels for each of the variables they selected.
10. Create a bargraph of the categorical variable chosen by the students. Begin by showing students how to clearly label the x-axis with the categories. For instance, if salty_sweet is the variable, ask students to identify the categories for that variable. Then mark the y-axis with the label **frequency**, which simply means the number of times an outcome occurs. Do not put tick-marks on the y-axis. The frequency will be measured by the number of labels plotted.
11. Have students come up and place their nutrition fact label above the category that describes their snack. Have students stack their nutrition label so that is easy to calculate the frequency. Once all the labels have been placed, create bars with the appropriate height (frequency) for each category. Make sure to leave spaces between the bars, and that bars are the same width.
12. Ask students to respond to the following questions in their DS journals:
    a. How many **data points** does this distribution have? Why?
    b. What information is this visualization telling us about [insert categorical variable name] in the snacks we consume?
13. Use another piece of poster paper to create a distribution for the first numerical variable chosen by the students.
14. Create a dotplot of the first numerical variable chosen by the students. Begin by showing students how to clearly label the x-axis. For instance, if calories per serving is the variable, ask students for the range of values for calories per serving and determine the **minimum** and **maximum** values for the data set. Clearly label the x-axis with adequate intervals and the variable's name. Then mark the y-axis with the label **frequency**, which simply means the number of times a value occurs. Do not put tick-marks on the y-axis. The frequency will be measured by the number of labels plotted.
15. For each value in the data set, put a nutrition facts label above that value on the x-axis. When a value occurs more than once, stack the nutrition facts labels. For example, if there are three nutrition facts labels with 120 calories per serving in the data set, there will be three nutrition facts labels above the 120 mark on the x-axis.
16. Once all the labels have been placed, ask students to observe the **distribution** of the data in the dotplot. Ask students to respond to the following questions in their DS journals:
    a. What are the minimum and the maximum values of the data set? ***Answers will vary by class.***
    b. **The range** is the largest value minus the smallest value. It is one way of measuring the variability of a variable. What is the range, and why do you think this measures the variability? ***Answers will vary by class. The range measures variability because if the values of the variable are really different, the range will be a big number (because the max and min will be far apart); but if there is little variability, the range will be small. For example, if all of the values were the same, we would have no variability and the range would be 0 because the max and min would be the same number.***
    c. How many **data points** does this distribution have? Why? ***Answers will vary by class.***

d. What is the amount of [insert variable name] that appears most often in a snack? Why? ***Answers will vary by class.***

e. What do you think the phrase *distribution of the data set* means? ***It shows us how values are distributed. We learn where there are many values, where there are only a few values, and where there are no values at all.***

f. What information is this distribution telling us about the [insert variable name] in the snacks we consume? ***Answers will vary. We see how the value of [variable name] varies. For example, we can see whether all foods have the same number of calories, or if they differ.***

g. A distribution tells us two things: the values of the variable and the frequency of the values. "Frequency" is just another way of saying "the count." Why is this dotplot a picture of the distribution of [variable name]? ***Because the location of the labels on the x-axis tells us the values we saw, and the number of labels at that value tells us the frequency for that value.***

17. Review the students' responses in a class discussion. Ask students to put a check mark next to the answers that were validated, and to revise the answers that need to be corrected.

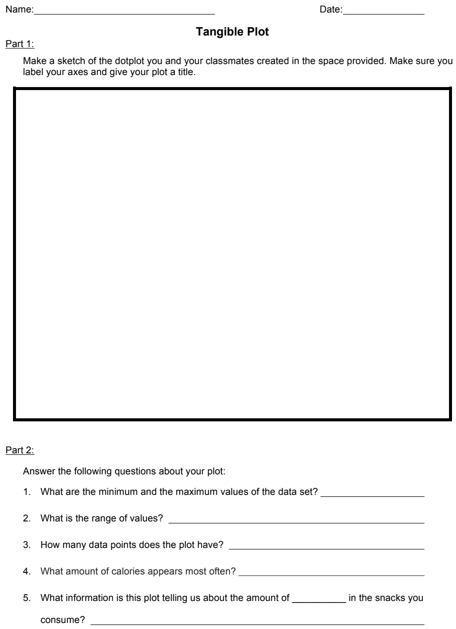
18. Use another piece of poster paper to create a distribution for the second numerical variable chosen by the students. Repeat steps 14-16 with this variable.

19. On the first visualization for the numerical variable, show students how to convert the nutrition facts labels into something more readable. Draw another horizontal line on the plot above the nutrition facts labels. Explain that we can represent each label with an item such as a dot sticker or a sticky note.

20. Then, start with the minimum x-value on the plot and place the new sticker above the second horizontal axis. Do this for each nutrition facts label in the plot. Once all values have been represented, ask the students how this new plot IS or IS NOT better than the original. Explain that we can call this type of plot a **dotplot** since we're using dots to represent each observation.

21. Distribute the *Tangible Plot* handout (LMR_1.9). Each student should pick one of the 2 numerical variables plotted on the poster paper. Then, they should complete part 1 of the *Tangible Plot* handout before leaving class. They will complete part 2 of the handout for homework.

22. Ask students to think about the statistical questions they came up with. Can the visualizations they created in class help answer their statistical question? If yes, answer the question; if not, what visualization would be appropriate?



Name:______________________________ Date:______________

Tangible Plot

Part 1:

Make a sketch of the dotplot you and your classmates created in the space provided. Make sure you label your axes and give your plot a title.

Part 2:

Answer the following questions about your plot:

1. What are the minimum and the maximum values of the data set? __________________
2. What is the range of values? ____
3. How many data points does the plot have? _________________________________
4. What amount of calories appears most often? ______________________________
5. What information is this plot telling us about the amount of __________ in the snacks you consume? ____

LMR_1.9

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

☑

Students will complete part 2 of the *Tangible Plot* handout (LMR_1.9) and bring it to the next class for assessment.

Students should continue to collect nutritional facts data using the *Food Habits* Participatory Sensing campaign on their smart devices or via web browser.

Lesson 9: What is Typical?

Objective:

Students will learn about the typical value when looking at a distribution by finding the “center” and determining any point clusters.

Materials:

1. Nutrition facts dotplot (from lesson 8)
2. Poster paper
3. Markers, dot stickers, or sticky notes

Vocabulary:

typical, center, shape, spread

Essential Concepts: The “center” of a distribution is a deliberately vague term, but it is one way to answer the subjective question "what is a typical value?" The center could be the perceived balancing point or the value that approximately cuts the area of the distribution in half.

Lesson:

1. **Food Habits Campaign Data Collection Monitoring:**
   a. Display the IDS Campaign Monitoring Tool, found at https://portal.idsucla.org/ Click on **Campaign Monitor** and sign in.
   b. Inform students that you will be monitoring their data collection again today.
      i. See *User List* and sort by *Total*. Ask: Who has collected the most data so far?
      ii. Click on the pie chart. Ask: How many active users are there? How many inactive users are there?
      iii. See *Total Responses*. How many responses have been submitted?
      iv. Using TPS, ask students to think about what they can do to increase their data collection.
2. Inform students that today they will be learning about a distribution’s **typical** value.
3. Ask the class to brainstorm characteristics of the "typical" student. Does the typical 12th grader differ from the typical 9th grader? How so? ***They may say that everyone is different, and that there's no typical student. Keep pressing them to identify characteristics that are typical. The idea is to get them to recognize that there is variability, and yet we might still have an opinion about what is typical. For instance, not all students walk to school, but this might still be the typical experience.***
4. Give students 3 minutes to write down synonyms to the word “typical**”** in their DS journals. After time is up, have the students share their responses and keep a record on the board. ***Some possible synonyms might be: normal, average, usual, standard, representative, regular, ordinary, natural, etc.***
5. Once students share their synonyms, ask students to think about which terms apply best to categorical variables and which terms apply best to numerical variables. Ask volunteers to share out their thoughts and give a brief explanation of why they categorized the term as either applying best to categorical variables or numerical variables. Create a T-chart on the board to keep track of their categories.
6. Next, display the dotplot created by the class with their nutrition facts labels during the previous class (from lesson 8). Ask: what value might we consider to be the typical value of this distribution? ***Answers will vary by class. Common answers will be to identify the mode (the value with the most labels) or the value in the center. A common wrong answer will be to confuse the frequency with the value. For example, they will say the most typical number of calories was "3" because, perhaps, 100 calories occurred 3 times, and that was more often than any other value. Students may also identify "clumps” of data: "it's somewhere***

between 110 and 120." That's ok but probe them as to why they chose that chunk and not another. The point is to get them to see that chunk as being in the middle or center of the distribution.

   a. Hopefully, at least one student will choose a value close to the center of the distribution. If not, point to a value near the extreme and ask them if they think this is typical. Then move closer to the center until they agree on which values are typical.
   b. It is ok to be vague in the definition of typical for today's lesson. The discussion needs to be very teacher-driven. Some possible points of discussion might be:
      i. Clustering/clumps of data.
      ii. Most of the observations are between ________ and _________.
      iii. Overall range of the data.

7. Ask students to reconsider the typical number of sugar grams. What is the typical amount of sugar (in grams) in our snacks? ***For example, students may come up with the same answer for different reasons: "The typical amount of sugar grams is 10." The reasons may include the data points are half below and half above; it's the mode; it has plurality***. Then, tie it back to the synonyms they provided. Ask: Which synonym are you using?
8. In pairs, ask students to discuss the question:
   a. Which synonyms are associated with "center"? Is this concept of **center** useful for numerical or categorical variables? **Center is useful for numerical variables. The center of the distribution often corresponds to our notion of 'typical value.' For example, the typical height of the students in our class might be centered around 5'5.**
9. Inform students that the value at the center of the distribution often matches up with our everyday notion of the typical value of a distribution. The middle observation is not always the typical value. Similarly, the middle person would not always be the center value.
10. Defining the center of a distribution depends on many things, such as the placement of points in the distribution (known as the **shape**) and how dense the distribution is at certain values (known as the **spread**).
11. Ask the students to write down the number of hours of sleep they got last night. They will be creating a dotplot of this data, so ask them:
    a. What do you think the typical value will be?
    b. What do you think the lowest value will be?
    c. What do you think the highest value will be?
    d. What do you think the shape of the distribution will look like?
12. One-by-one, have them come up to the board (or poster paper) and put a dot above the correct value on the dotplot. After each student has placed a dot on the board, have a discussion about the distribution. Is the typical value similar to what they originally thought? The shape? The variability? Why or why not?
13. Next, have the students write down the number of hours of sleep they hope to get this Saturday. How do they think this plot will differ from the first plot? Focus discussion on the shape, center, and spread of the distributions. Repeat steps 8-9 and discuss how this plot is similar and/or different than the first plot.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students should continue to collect nutritional facts data using the *Food Habits* Participatory Sensing campaign on their smart devices or via web browser.

Lesson 10: Making Histograms

Objective:

Students will understand that a histogram represents observations grouped into bins, and that bars are drawn to show how many observations (or what proportion of the observations) lie in each bin, rather than representing individual observations, as in a dotplot.

Materials:

1. Peanut butter
2. Jelly
3. Loaf of sliced bread
4. Butter knife
5. Plate
6. *Sleep* dotplots (from lesson 9)
7. Poster paper
8. Markers

Vocabulary:

algorithm, histogram, bin(s), bin widths, output, input, left-hand rule, right-hand rule

Essential Concepts: Histograms can be created through the use of an algorithm. The distributions displayed in a histogram can be classified using the technical terms for the shapes of distributions. Learning to describe routine tasks through an algorithm is an important component of computational thinking.

Lesson:

1. Inform students that they will be telling us how to make a sandwich today. Giving clear, concise instructions to others is an important skill for students to learn. In this activity, students will practice using descriptive vocabulary, communicating ideas to others, recognizing steps in a process, and recognizing the importance of the use of clear language.
2. Prepare for this task by gathering the necessary materials for making a peanut butter and jelly sandwich and arranging them in a way that makes them easy to use. You may want to wear an apron and have a trash bag smock —this can get messy but that's most of the fun!

   Note: **Be aware of peanut allergies!** If any of your students are allergic to peanut butter, DO NOT ALLOW STUDENTS TO HANDLE THE PEANUT BUTTER! Peanut allergies can be very serious and children can have reactions without even eating it. So be aware and be careful!
3. Ask your students if they have ever followed a recipe before.
   a. What kinds of things have they made?
   b. Does anyone know how to make a peanut butter and jelly sandwich?
   c. Would they teach you how?
   d. Would they give you all the steps to make a sandwich?
4. Show your students the materials you have for making a sandwich. Have students take out paper and pencils and ask each student (or pair of students) to write down their instructions for making a peanut butter and jelly sandwich. We can also call these instructions an **algorithm**.
5. Explain to students that precise instructions for any process are like a formula to follow in order to get the same results each time. Also, an algorithm is how we communicate with the computer. The teacher will function as the computer. Your job is to give him/her rules so that he/she can carry out and successfully make a PB&J sandwich.
6. Every algorithm needs input and produces output. The output here will be a PB&J sandwich. What is the input? ***Steps, or actions to follow***.

7. Tell students that when they are done you will select someone to share their instructions and you will make a sandwich following the instructions.

8. Select a student to read their instructions, and do EXACTLY what it says. For example, if it says “put the peanut butter on the bread,” you can literally put the jar of peanut butter on the bag of bread. There was no instruction to open the bread or the jar of peanut butter, no instruction to use the knife in any way, etc. Listen for other examples of unclear instructions and think of how you might act them out. If students are not clear about where to spread the peanut butter, put it on the crust. The more literal you are by doing exactly what the instructions say, the funnier the activity will be and the more likely you are to get your point across about the importance of clear instructions.

9. After your first sandwich, ask you students if they think their instructions were clear or not. What are some things they might have done differently?

10. Select another student to read his/her instructions. They will be sure to use clarifications of the instructions you acted upon before - this is a good thing!

11. After your finish the sandwich, ask you students if they think clear instructions are important. Why?

12. Let students know that they will now develop an algorithm for building a histogram to represent the sleep dotplots they created in the previous lesson.

13. Explain that a **histogram**, rather than showing the frequency for each value, shows the frequency (or percent, but we will focus on frequency) of all the values that fall in a certain range, called a **bin**. For example, we might choose bins that go from 0-5, 5-10, 10-15, 15-20, 20-25. ***Bin widths will vary by class.***

14. Model how to create a histogram using the data from the dotplot “hours of sleep last night”. On a blank chart, create the x-axis with bin widths 0-3, 3-6, 6-9, etc. and place marks on the plot at those intervals and ask students: “What are the frequencies in each bin?”

    Notice that multiples of three appear in more than one bin. Let’s take the value of 6 hours as an example. Should those observations be included in the second bin (3-6) or the third bin (6-9)?

    - If students include the values of 6 hours in the second bin then they are using the **left-hand rule**.

    - If students include the values of 6 hours in the third bin then they are using the **right-hand rule**.

15. Once the frequencies have been determined, draw the bars with corresponding heights. Do not include spaces between the bars as time is a continuous variable.

16. Next, student teams will create an algorithm that gives directions for how to construct a histogram for the data from the dotplot for “hours of sleep they hope to get on Saturday.” Remember, an algorithm is a set of rules that can always be applied. Similar to the way they wrote a process for making a PB&J sandwich, students will write a process for creating a histogram. Tell students to continue thinking of the process to transform the data in the dotplot to create a histogram. The algorithm will produce an **output**, which will be a histogram. What's the **input**? ***Data, or maybe the dotplot.***

17. Inform the students that you will provide a piece of input: how wide the bin will be. For instance, it might be 5 hours, it might be 1 hour, or it might be 10 hours (or half an hour!). Whatever it is, their algorithm should work for any input value.

18. Let students work for a bit. They should write out Step 1, Step 2, etc. Then choose a group and ask them to get you started. Give them a bin width of 4.

19. Teachers should sketch the histogram on the board or chart paper as students read their algorithms. Again, teachers should take things very literally. For example, if they do not tell you exactly where the bins should start, start one way off to the left. If they are vague and say "divide the number line into groups of 10,” then make them arbitrary sizes. If they have to be the same size, ask them how to do that. Points to consider:

   a. Where do we start drawing the bins? Always at the location of the smallest dotplot? Always at the greatest? A little to the left?
   b. What do we do with points that fall exactly on a boundary? Do they go to the bin on the left or on the right? Does it matter? ***No***.
   c. Can we do it differently every time? ***No. We need to be consistent. This is called either the left-hand rule or the right-hand rule, depending on which is chosen.***

20. After following 2 or 3 algorithms, ask students if they feel their algorithm is precise enough. Allow students time to revise their algorithms.

21. Have a class discussion about the similarities and differences between the original dotplot and a histogram. Ask:
    a. What have we gained from the histogram? ***We now can see the shape of the distribution as a whole.***
    b. What have we lost? ***We lost each individual observation by grouping them into bins.***

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students should continue to collect nutritional facts data using the *Food Habits* Participatory Sensing campaign on their smart devices or via web browser.

Lesson 11: What Shape Are You In?

Objective:

Students will learn to classify distributions in terms of shape, and can suggest theories for why a distribution might be one shape or another.

Materials:

1. *Sorting Histograms* handout (LMR_1.10_Sorting Histograms) - one copy per group of 4 students. (This activity comes from the AIMS project, University of Minnesota, J. Garfield.)
   Advanced preparation required (see step 1 below)

Vocabulary:

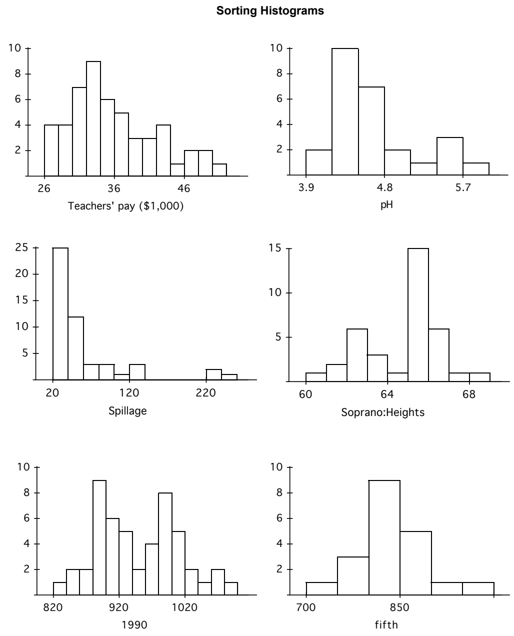
symmetric, left-skewed, right-skewed, unimodal, bimodal

Essential Concepts: Identifying the shape of a histogram is part of the **interpret** step of the Data Cycle.

Lesson:

1. Distribute the cutouts from the *Sorting Histograms* handout (LMR_1.10). Give each student team all of the 24 histograms (can be paper-clipped together or put in small zippered bags).

   Advanced preparation required: Print the *Sorting Histograms* file (LMR_1.10). Cut each histogram so that it is on its own piece of paper. Create enough sets for each team to have all 24 histograms. They can be paper clipped together, or put in small zippered bags.



LMR_1.10

2. Inform students that the type of data being measured is indicated on the horizontal axis, and the vertical axis represents how many observations are in each bar.
3. The students will then sort their stack of plots into different piles according to their shapes. Histograms that have similar shapes should be sorted into the same stack.

4. Once the student teams have agreed upon the histogram shape groupings, they should discuss and write down answers to the following in their DS journals:

    a. Describe what’s similar about the plots in each group. ***Answers will vary, but should be grouped by the overall shape of the distribution. For example, plots with a higher density of bars on the right side of the plot should all be in the same group.***
    b. Pick one graph in each group that is the best example of that group. ***Answers will vary.***
    c. Give the group a name that you think describes the general shape. ***Answers will vary.***
    d. If there are graphs that do not fit into any group, try to determine why it was impossible to place them. What is different or confusing about them? ***Answers will vary.***

5. After each team has had time to discuss and write down their observations, have a class discussion about the histogram groupings. Do the students agree about the general shapes?

6. In statistics, we use specific terminology when discussing the shapes of distributions, such as **symmetric**, **right-skewed**, **left-skewed**, **unimoda**l, **bimodal**, etc. Did any of the teams use these terms? If not, introduce each one and ask which of the 24 histograms could be classified as that shape.

7. Next, introduce the following scenarios and ask students to determine what a corresponding histogram might look like. They should use statistical terms to describe their answer.

    a. The grades on an easy test. ***Left-skewed, unimodal***
    b. The grades on a difficult test. ***Right-skewed, unimodal***
    c. The number of times IDS students study during the first week of class. ***Answers will vary.***
    d. The age of cars on a used car lot. ***Right-skewed, probably unimodal***
    e. The amount of time spent by students on a difficult test (max time allowed is 50 mins). ***Left-skewed, but may also just be one bar with all observations at 50 mins, unimodal***
    f. The heights of students in your high school band. ***Symmetric, bimodal***
    g. The salaries of all persons employed at the Los Angeles Unified School District. ***Right-skewed, potentially bimodal (teachers vs. LAUSD administrators)***

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students should continue to collect nutritional facts data using the *Food Habits* Participatory Sensing campaign on their smart devices or via web browser.
`

Lesson 12: Exploring Food Habits

Objective:
Students will experience the full Data Cycle, and for the first time will do so with data they have collected. They will use the Dashboard and PlotApp, tools that are easy to learn. This first time, the teacher will “navigate and steer" so that students can focus on asking questions and interpreting the plots.

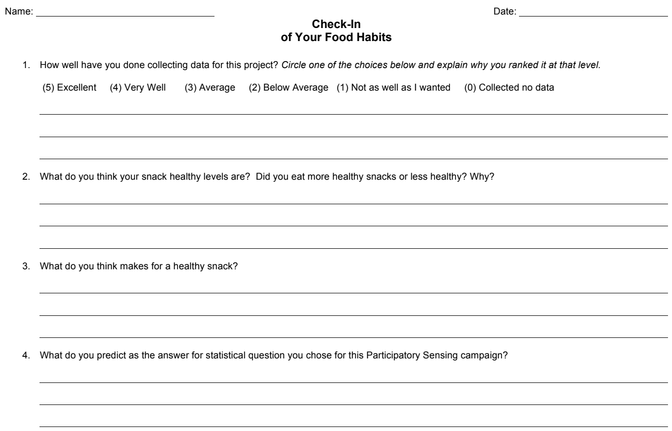
Materials:

1. Computers
2. Projector
3. *Food Habits Check-In* handout (LMR_1.11_Food Habits Check-In)
4. *Exploring Our Food Habits* handout (LMR_1.12_Exploring Our Food Habits)

Essential Concepts: Once Participatory Sensing data has been collected, the Dashboard and PlotApp perform the analysis step of the Data Cycle, though humans need to tell the computer which plots we wish to examine.

Lesson:

1. Ask students to reflect about their experience so far with the Food Habits Participatory Sensing campaign by completing the *Food Habits Check-In* handout (LMR_1.11).



Name: ____________________ Date: ____________________

Check-In
of Your Food Habits

1. How well have you done collecting data for this project? *Circle one of the choices below and explain why you ranked it at that level.*

   (5) Excellent (4) Very Well (3) Average (2) Below Average (1) Not as well as I wanted (0) Collected no data

   ____
   ____
   ____

2. What do you think your snack healthy levels are? Did you eat more healthy snacks or less healthy? Why?

   ____
   ____
   ____

3. What do you think makes for a healthy snack?

   ____
   ____
   ____

4. What do you predict as the answer for statistical question you chose for this Participatory Sensing campaign?

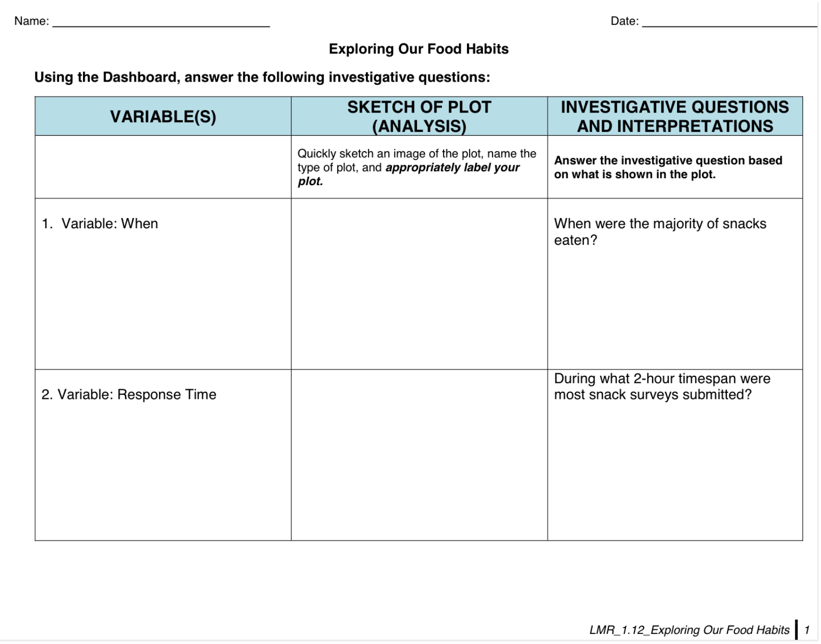
   ____
   ____
   ____

LMR_1.11

2. Demonstrate how to access the **IDS Homepage** found at https://portal.idsucla.org/
3. Explain to students that all of the IDS web tools can be accessed through this page.
4. For this lesson, students will need to observe the teacher using the **Campaign Manager**, the **Dashboard**, and the **PlotApp**.
5. Click on the Campaign Manager. Explain that selecting any of the web tools on the IDS page without logging in first will redirect them to the login prompt.
6. Demonstrate how to log in to access the IDS software suite. Inform students that they will use the same login information they have used to collect data on their mobile devices or the browser-based version.

   Inform students that the Campaign Manager is the place where they will access, download, and export their campaign data in subsequent lessons. It also provides shortcuts to the Dashboard and PlotApp. The Campaign Manager allows them to view and learn about the campaigns in which they are participating, and to edit the campaigns they will be creating later in this course. Show them the drop-down menu on the right-hand side, and explain that they will only be concerned with the **Responses** tab for this lesson. Explain that the Responses tab allows them to view, delete, or share their data, and to view shared data contributed by other users of the campaign.

7. Distribute the *Exploring Our Food Habits* handout (LMR_1.12).



Name: ____________________________ Date: ________________________

Exploring Our Food Habits

Using the Dashboard, answer the following investigative questions:

VARIABLE(S)	SKETCH OF PLOT (ANALYSIS)	INVESTIGATIVE QUESTIONS AND INTERPRETATIONS
	Quickly sketch an image of the plot, name the type of plot, and ***appropriately label your plot.***	**Answer the investigative question based on what is shown in the plot.**
1. Variable: When		When were the majority of snacks eaten?
2. Variable: Response Time		During what 2-hour timespan were most snack surveys submitted?

LMR_1.12_Exploring Our Food Habits | 1

LMR_1.12

8. Inform students that the teacher will be navigating through the IDS website as the students follow along in the *Exploring Our Food Habits* handout (LMR_1.12).

9. Once completed, students should turn in the handout for assessment.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students should continue to collect nutritional facts data using the *Food Habits* Participatory Sensing campaign on their smart devices or via web browser.

Lesson 13: RStudio Basics

Objective:

Students will learn RStudio’s interface, as well as a few basic commands to discover the structure behind a data set.

Materials:

1. Computer
2. Projector
3. RStudio: https://portal.idsucla.org/

Vocabulary:

pane, preview, console, plot, environment

RStudio Commands:

`data( ), View( ), names( ), help( ), dim( ), tally( ), load_labs( )`

Essential Concepts: The computer has a syntax, and it can only understand if you speak its language.

Lesson:

1. Inform students that the Dashboard and PlotApp are data visualization tools that are coded in R, the statistical programming software that academics and professional statisticians use. The Introduction to Data Science course will utilize RStudio, which also runs on R. They will learn the programming language of RStudio for data analysis.
2. Demonstrate how to access RStudio by projecting the URL: https://portal.idsucla.org/ on a screen. Then, click on the RStudio icon on the page.
3. Inform students that they will log into RStudio using the “Log In with Google” option. Note that it is not the same as their IDS App & IDS Homepage login.
4. Once logged in, show each **pane**, or rectangular area, of the RStudio interface:
    a. **preview** (spreadsheet) - where they will be able to see the variables and observations (index); rows and columns of data
    b. **console** - where they will be entering their code
    c. **plot** - where their plots/graphs/visualizations will be generated
    d. **environment** - where they will see values and objects
5. Inform students that they will be looking at a data set from The Centers for Disease Control and Prevention (CDC), a government agency that collects data about teenagers on a variety of topics.
6. Demonstrate how to load and view the CDC data file to the workspace by typing the following command in the console:

    ```
    >data(cdc)
    >View(cdc)
    ```

7. Examine the environment pane. Ask a student to describe how the data are displayed. ***The data are displayed in rows and columns.***
8. Demonstrate how to list the variables found in the CDC data set. Students may take notes and write down commands in their DS journals:
    a. `>names(cdc)`

   b. Ask: What do you notice? What is one variable of this data set? How many variables are there? How does this output compare to the information in the preview pane? ***This command lists the names of each variable in the data set.***

9. Demonstrate how to obtain more detailed information about the data set by typing the following command in the console

   a.
   ```
   >help(cdc)
   ```

   b. Ask: What unit of measurement is height reported in? ***Height was reported in meters.***

10. Demonstrate how to find the number of rows and columns in the data set.

    a.
    ```
    >dim(cdc)
    ```

    b. Ask: Which number do you think represents the rows? Which one represents the columns? How does this output compare to the information in the preview and environment panes? How many observations are there in the data set? How many variables does this data set contain? ***There are 15,624 rows, or 15,624 observations; and there are 33 columns, or 33 variables. This information is also visible in the preview pane.***

11. Next, show students how to access the number of observations of a specific variable.

    a.
    ```
    >tally(~seat_belt, data = cdc)
    ```

    b. Ask: What do you notice? Describe the output. ***Notice that six categories are displayed. Each category shows the number of observations contained in it. E.g,. "Never" has 326 observations, meaning 326 teens reported never wearing their seat belt as a passenger in a motor vehicle. <NA> = Not Available, represents teens that did not provide information about their seat belt habits.***

12. Change the variable to *height*.

    a.
    ```
    >tally(~height, data = cdc)
    ```

    b. Ask: What do you notice? Describe the output. ***The levels are missing. It happened because the variable*** *height* ***contains numbers, not categories.***

13. Let's take a closer look at the variables *seat_belt* and *height*. Maximize the console. Ask teams to discuss the following question:

    What is the difference between the data from the variables *seat_belt* and *height*? ***The data from the*** *seat_belt* ***variable is categorical, which means it consists of groupings. The data from the variable*** *height* ***is numerical, which means it consists of numbers***.

14. Summarize: In data science, the variable *seat_belt* is what we call a **categorical variable**, and the variable *height* is what we call a **numerical variable**.

15. Let's look at the other variables in this data set. In pairs, categorize each variable as categorical or numerical:

    a. eat_fruit ***(categorical)***
    b. weight ***(numerical)***
    c. grade ***(categorical)***
    d. gender ***(categorical)***

16. Inform students that they will be learning RStudio code to work with data. They will be completing RStudio labs throughout the course.

17. Demonstrate how to load the menu of labs by typing the following code:

    ```
    >load_labs( )
    ```

18. The load labs command displays a list of available labs and a selection prompt. To select Lab 1A, type number 1 after the selection prompt.
19. Next, direct students' attention to the plot pane. Show them the location of Lab 1A's presentation.
20. Click on the arrows at the bottom right-hand side of the presentation to view each slide. Pause on a slide titled "R's most important syntax." There are 3 boxes, each containing a line of code.
21. Explain that every time they see a grey box with a line of code, they are to type the code in the console. The output will appear either on the console itself or on the plot pane.
22. Type in one of the lines of code. In this particular case, the output will be a plot. Show students the location of the plot and demonstrate how to toggle between the plots and presentation tabs.
23. Inform students that they will be completing the first lab, 1A, the next day.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next 3 Days

Students should continue to collect nutritional facts data using the *Food Habits* Participatory Sensing campaign on their smart devices or via web browser.

Lab 1A: Data, Code & RStudio

Lab 1B: Get the Picture?

Lab 1C: Export, Upload, Import

Complete Labs 1A, 1B and 1C prior to Lesson 14

Lab 1A - Data, Code & RStudio

Directions: Follow along with the slides and answer the questions in **bold (red bold in lab)** font in your journal.

Welcome to the labs!

- Throughout the year, you'll be putting your data science skills to work by completing the labs.
- You'll learn how to program in the `R` programming language.
  - The programming language used by actual data scientists.
- Your code will be written in RStudio which is an easy to use interface for coding using `R`.

So let's get started!

- The data for our first few labs comes from the Centers for Disease Control (CDC)
  - The CDC is a federal institution that studies public health.
- Type these two commands into your console:

```
data(cdc)

View(cdc)
```

- **Describe the data that appeared after running `View(cdc)`:**
  - ***Who* is the information about?**
  - **What sorts of information about them was collected?**
- To find out more information about the `cdc` data, type the command below into your console.
  - To get back to the slides find and click on the `Viewer` tab

```
?cdc
```

Data: Variables & Observations

- Data can be broken up into two parts.
  1. *Observations*
  2. *Variables*
  - *Observations* are the *who* or *what* we are collecting data from/ about.
  - *Variables* are the measurements or characteristics about our *observations*.
- If need be, re-type the command you used to `View` your data. Then answer the following:
  - **Based on the data, describe a few characteristics about the first observation.**
  - **What does the first column tell us about our observations?**
- In order to describe the first observation, notice that you had to look at the first row of the spreadsheet. Each row, in this case, describes a person.
- The columns of the spreadsheet represent variables.

Uncovering our Data's Structure

- Now that we've looked at our data, let's look at how RStudio is organized.
- RStudio's main window is composed of four *panes*
- Find the pane that has a *tab* titled *Environment* and click on the *tab*.
  - This pane contains a list of everything that's currently available for R to use.
  - Notice that R knows we have our `cdc` data loaded.
- **How many students are in our `cdc` data set?**

- **How many variables were measured for each student?**

Type the following commands into the console

```
dim(cdc)

nrow(cdc)

ncol(cdc)

names(cdc)
```

- **Which of these functions tell us the number of observations in our data?**
- **Which of these functions tell us the number of variables?**

First Steps

- Typing commands into the console is your first step into the larger world of *programming* or *coding* (terms which are often used interchangeably).
- Coding is all about learning how to send instructions to your computer.
  - The way we *speak* to the computer, using a coding language, is *syntax*.
- R is one of many coding languages. Each coding language is slightly different, and these differences are reflected in the syntax.
- *Capitalization*, *spelling* and *punctuation* are REALLY important.

Syntax matters

- **Run the following commands and write down what happens after each. Which does R understand?**

```
Names(cdc)

NAMES(cdc)

names(cdc)

names(CDC)
```

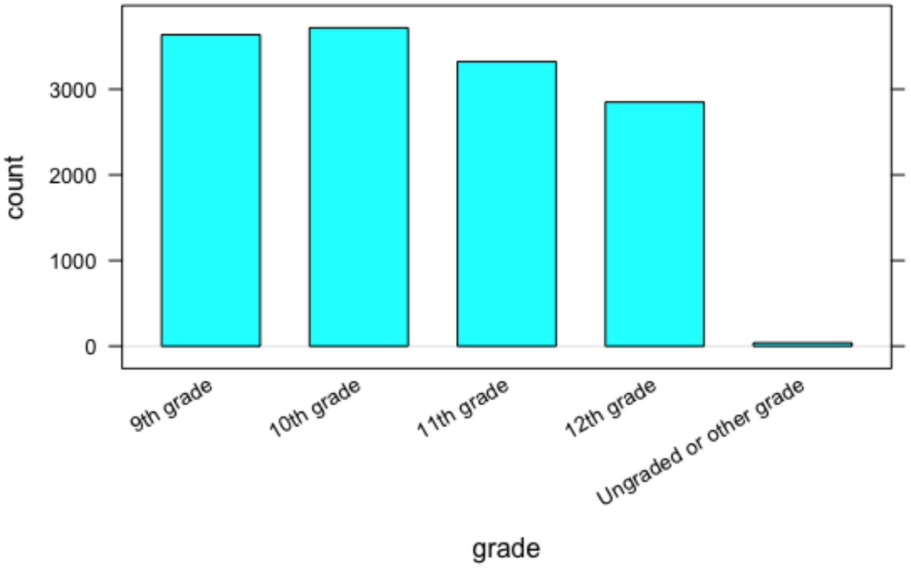
R's most important syntax

- Most of the commands you will be using follow the syntax below:

function (y~x, data = ____)

- To create graphs or plots you need to provide R with the following:
  - The name of the R function, often the plot’s name, that tells the computer how to create your graph.
  - The variable(s) containing the information we want the function to use.
  - The data set containing the variables.
- Notice that when we analyze a single variable the value for y is left blank.

```
bargraph(~grade, data = cdc)
```



- Later on, we'll see we can use this syntax to do more than create graphs.

Syntax in action

function (y~x, data = ____)

- Search through the different panes. Find and then click on the *Plots* tab.
    - To get back to the slides, find and then click on the *Viewer* tab.
- **Which one of these plots would be useful for answering the question: *Is it unusual for students in the CDC dataset to be taller than 1.8 meters?***
- Run the three commands below then answer the question that follows.

```
histogram(~height, data = cdc)

bargraph(~drive_text, data = cdc)

xyplot(weight~height, data = cdc)
```

- **Do you think it's unusual for students in the data to be taller than 1.8 meters? Why or why not?**
- Hint: Use the arrow keys on the `Plots` tab to toggle between the plots.

On your own:

- After completing the lab, answer the following questions:
    - **What is *public health* and do we collect data about it?**
    - **How do you think our data was collected? Does it include every high school aged student in the US?**
    - **How might the CDC use this data? Who else could benefit from using this data?**
    - **Write the code to visualize the distribution of weights of the students in the CDC data with a histogram. What is the *typical* weight?**
    - **Write the code to create a bargraph to visualize the distribution of how often students ate fruit. About how many students did not eat fruit over the previous 7 days?**

Lab 1B - Get the picture?

Directions: Follow along with the slides and answer the questions in **bold (red bold in lab)** font in your journal.

Where'd we leave off ...

- In the previous lab, we started to get acquainted with the layout of RStudio and some of the commands.
- In this lab, we'll learn about different *types* of variables.
  - Such as those that are measured by numbers and others that have values that are categories.
- We'll also look at ways to visualize these different types of data using *plots* (A word data scientists use interchangeably with the word *graph*).
- Find the *History* tab in RStudio and click on it. Figure out how to use the information to reload the `cdc` data.

Variable Types

- Numerical variables have values that are measured in units.
- Categorical Variables have values that describe or categorize our observations.
- `View` your `cdc` data and find the columns for `height` and `gender` (Use the *History* pane again if you need help to `View` your data).
  - **Is `height` a numerical or categorical variable? Why?**
  - **Is `gender` a numerical or categorical variable? Why?**
  - **List either the different categories or what you think the measured units are for `height` and `gender`.**

Which is which?

- Run the code you used in the previous lab to display the `names` of your `cdc` data's variables (Use the code displayed in the *History* pane to resubmit previously typed commands). Use the code's output to help you complete the following:
  - **Write down 3 variables that you think are *categorical* variables and why.**
  - **Write down 3 variables that you think are *numerical* variables and why.**

Data Structures

- One way to get a good summary of your data is to look at the data's *structure*.
  - One way to view this info would be to click on the little blue arrow next to `cdc` in the *Environment* pane.
  - Another way would be to run the following in the console:

```
str(cdc)
```

- Look at the `structure` of your `cdc` data and answer:
- **List all the types of info the `str()` function outputs**
- **Were you able to correctly guess which variables were categorical and numeric? Which ones did you mis-label?**

Visualizing data

- Visualizing data is a really helpful way to learn about our variables.

  - Making a picture of the distribution of a variable is a good way to begin visualizing data.
  - Remember: A distribution gives us the values of the variable and tells us how many of these values we have in our data set.
- Choose one numeric and one categorical variable from the data and create both a `bargraph` and a `histogram` for each variable.
  - **Which function, either `bargraph` or `histogram` is better at visualizing categorical variables? Which is better at visualizing numerical variables?**

We have options

- **Make a graph that shows the distribution of people's `weight`.**
- **Describe the distribution of `weight`. Make sure to describe the shape, center and spread of the distribution.**
- Options can be added to plotting functions to change their appearance. The code below includes the `nint` option which controls the number of *intervals* in a numerical plot.
  - Type the command below on your console and then answer the questions that follow.

```
histogram(~weight, data = cdc, nint = 3)
```

- **How did including the option `nint = 3` change the `histogram`?**
- **Does setting `nint = 3` impact how you would describe the shape, center and spread?**
- **Try other values for `nint`. What value produced the best graph? Why?**

How often do people text & drive?

- Make a graph that shows how often people in our data texted while driving.
  - **What does the y-axis represent?**
  - **What does the x-axis tell us?**
  - **Would you say that *most* people *never* texted while driving? What does the word *most* mean?**
  - **Approximately what percent of the people texted while driving for 20 or more days? (Hint: There's 13677 students in our data.**

Does texting and driving differ by gender?

- Fill in the blanks with the correct variables to create a side-by-side bargraph:
  bargraph (~ ____ , data = ____ , groups = ____)

- **Write a sentence explaining how boys and girls differ when it comes to texting while driving.**
- **Would you say that most girls never text and drive? Would you say that most boys never text and drive?**
- **How did including the `groups` argument in your code change the graph?**

Do males/females have similar heights?

- To answer this, what we'd like to do is visualize the distributions of heights, separately, for males and females.
  - This way, we can easily compare them.
- Use the `groups` argument to create a `histogram` for the `height` of males and females.

  - **Can you use this graphic to answer the question at the top of the slide? Why or why not?**
  - **Is grouping numeric values, such as heights, as helpful as grouping categorical variables, such as texting & driving?**

Do males/females have similar heights?

- `groups` uses color to differentiate between groups.
  - **Why does this work for bargraphs but not for histograms?**
- Fill in the blanks with the correct variables to create a split histogram (The " | " symbol is usually between the delete and enter keys on a keyboard) to answer the questions below:

histogram (~ ____ | ____ , data = ____)

- **Do you think males & females have similar heights? Use the plot you create to justify your answer.**
- **Just like we did for the histogram, is it possible to create a *split* bargraph? Try to create a bargraph of `drive_text` that's split by gender to find out.**

On your own:

- In this lab, we looked at boy's and girl's texting & driving habits:
- **What other factors do you think might affect how often people text and drive?**
  - **Choose one variable from the `cdc` data, make a graph, and use the graph to describe how `drive_text` use differs with this variable.**

Lab 1C - Export, Upload, Import

Directions: Follow along with the slides and answer the questions in **bold (red bold in lab)** font in your journal.

Whose data? Our data.

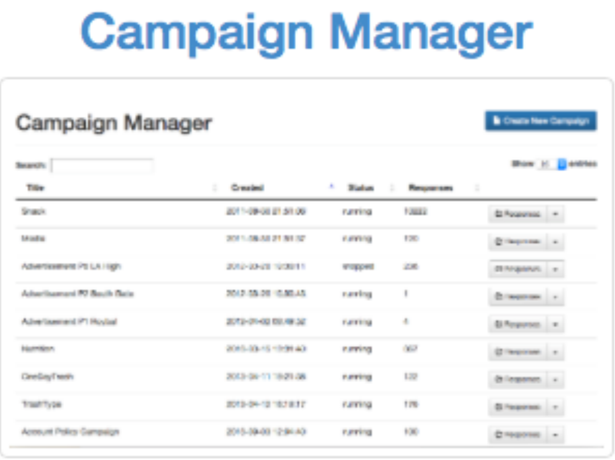
- Throughout the previous labs, we've been using data that was already loaded in RStudio.
  - But what if we want to analyze our own data?
- This lab is all about learning how to load our own participatory sensing data into RStudio

Export, upload, import`

- Before we can perform any analysis, we have to load data into R.
- When we want to get our participatory sensing data into RStudio, we:
- Export the data from your class’ campaign page.
- Upload data to *RStudio* server
- Import the data into R's working memory

Exporting

- To begin, go to the *IDS* Tools page.
  - Click on the Campaign Manager
  - Fill in your username and password and click "Sign in."

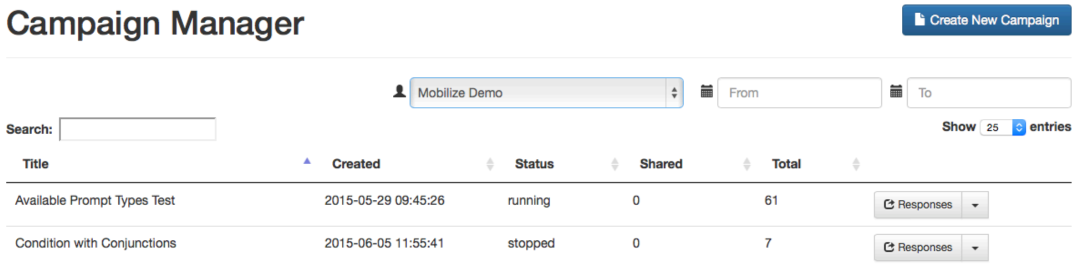


Title	Created	Status	Responses	
Snack	[illegible]	running	[illegible]	Responses
Media	[illegible]	running	[illegible]	Responses
Advertisement P[illegible] LA High	[illegible]	stopped	[illegible]	Responses
Advertisement P[illegible] South Gate	[illegible]	running	1	Responses
Advertisement P1 Roybal	[illegible]	running	4	Responses
Nutrition	[illegible]	running	[illegible]	Responses
OneDayTeeth	[illegible]	running	[illegible]	Responses
TrashType	[illegible]	running	[illegible]	Responses
Account Policy Campaign	[illegible]	running	[illegible]	Responses

Manage and create campaigns

If you forget your username or password, ask your teacher to remind you.

Campaign Manager



Campaign Manager

Create New Campaign

Mobilize Demo | From | To

Search:

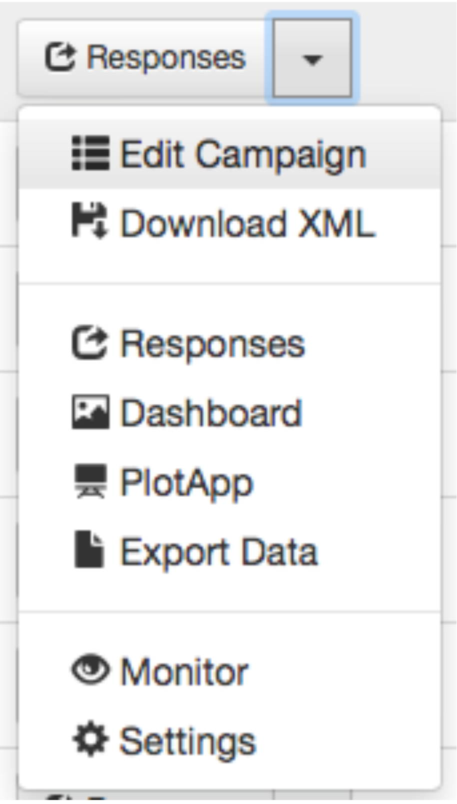
Show 25 entries

Title	Created	Status	Shared	Total	
Available Prompt Types Test	2015-05-29 09:45:26	running	0	61	Responses
Condition with Conjunctions	2015-06-05 11:55:41	stopped	0	7	Responses

- After logging in, your screen should look similar to this.
- Click on the dropdown arrow for the campaign you are interested in downloading.
  - At this point in the course, it will most likely be the Food Habits campaign

Dropdown Arrow

- The options for the dropdown menu will look like this.



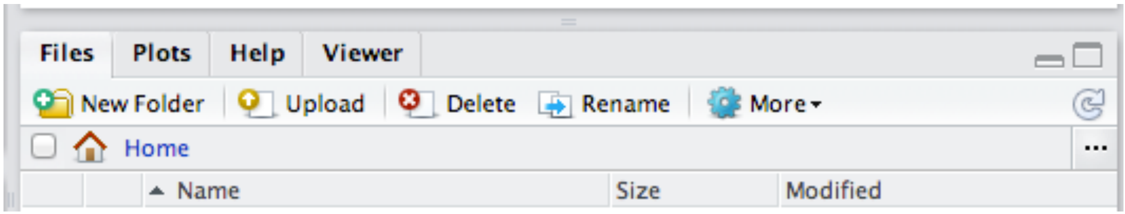
- Look for the option labeled Export Data. Click it.
  - Remember where you save your file!

Exporting

- When you clicked the Export link a *.csv* file was saved on your computer.
- Now that the file is on your computer, we need to upload it into RStudio.

Uploading

- Look at the four different *panes* in RStudio.
  - Find the *pane* with a Files tab.
  - Click it!

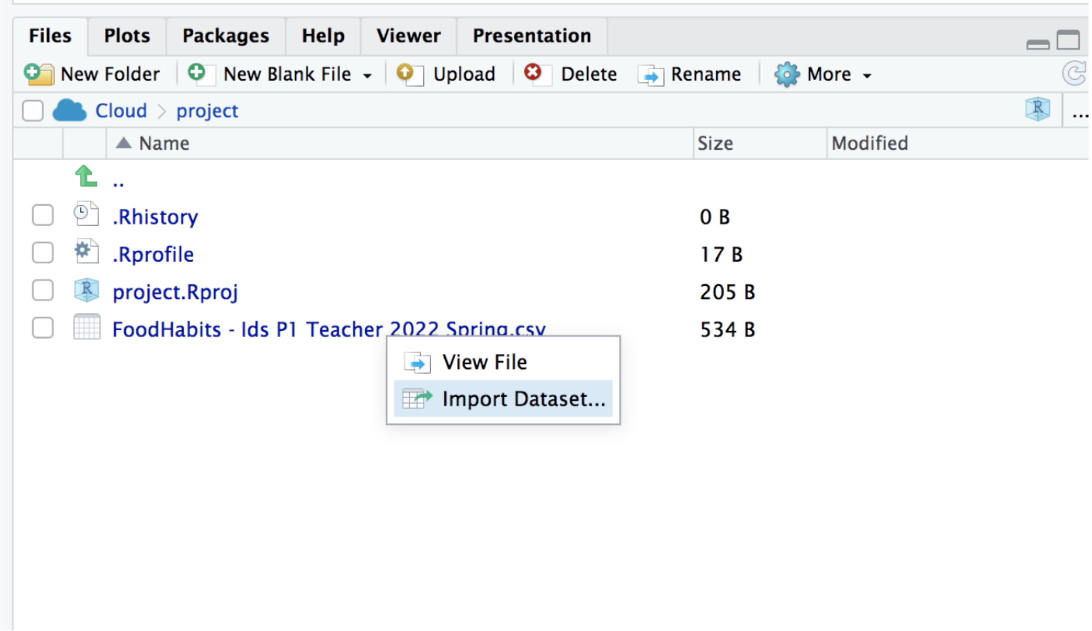


- Click the button on the *Files* pane that says “Upload”.
  - Click on “Choose File” and find the SurveyResponses.csv file you saved to your computer.
  - Hit the *OK* button.
- Voila!
  - If you look in the Files pane, you should be able to find your data!

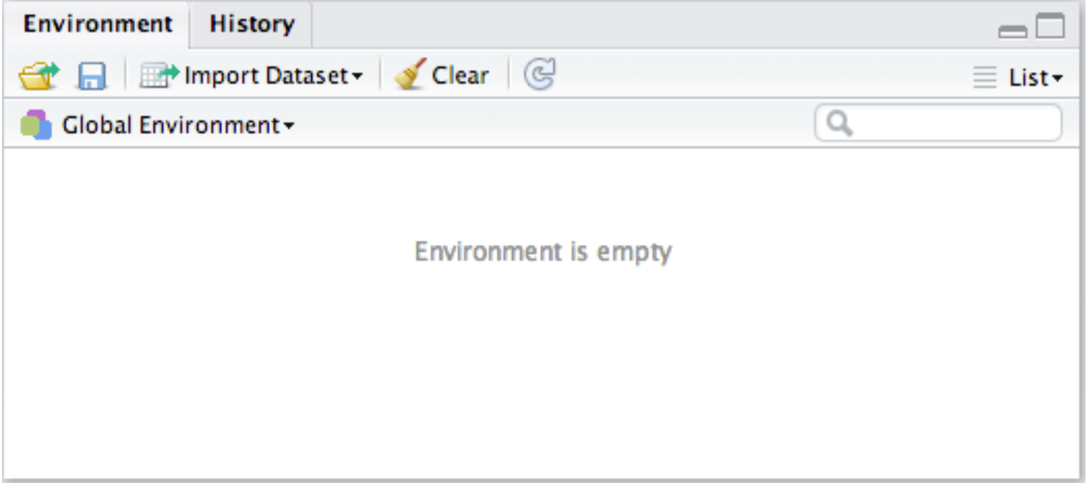
Upload vs. Import

- By Uploading your data into RStudio you've really only given yourself access to it.
  - Don't believe me? Look at the Environment pane ... where's your data?
- To actually use the data we need to Import it into your computer's memory.
- To compute more quickly and efficiently, R will only keep a few data sets stored in its memory at a time.
  - By importing data, you are telling R that this is a data set that is important to store it in its memory so you can use it.

Importing



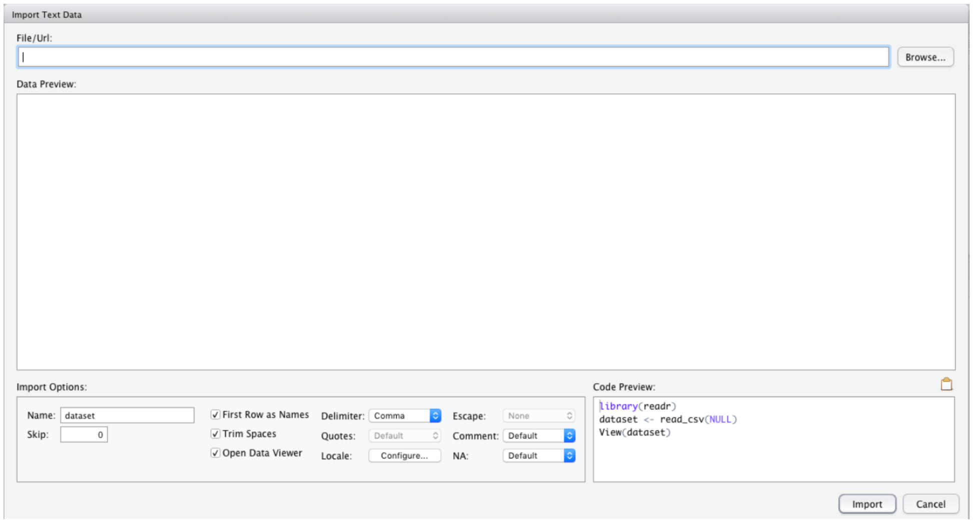
Name	Size	Modified
..		
.Rhistory	0 B	
.Rprofile	17 B	
project.Rproj	205 B	
FoodHabits - Ids P1 Teacher 2022 Spring.csv	534 B	



Environment is empty

- On the Files pane, find the data you want to import.
- Click on the name of the file and choose the option “Import Data set...”

Data Preview



File/Url:

Data Preview:

Import Options:

Name: dataset
Skip: 0
☑ First Row as Names
☑ Trim Spaces
☑ Open Data Viewer
Delimiter: Comma
Quotes: Default
Locale: Configure...
Escape: None
Comment: Default
NA: Default

Code Preview:

```
library(readr)
dataset <- read_csv(NULL)
View(dataset)
```

- You can give your data a name using the Name: field in the lower left corner.

What's in a name?

- The name you give your data is what you will use when you write code to analyze your data.
  - Good names are short and descriptive.
  - For your food habits campaign, some good names to use would be "foodhabits" or even just "food".
- When you're ready, click the *Import* button.

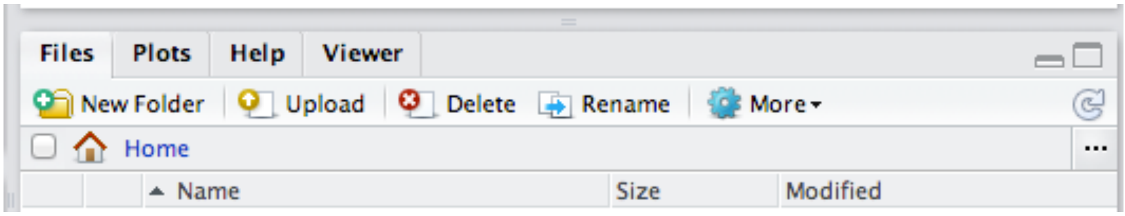
read.csv()

- After you click Import you might notice something appeared in your console.

```
data.file <- read_csv("~/SurveyResponse.csv")
View(data.file)
```

- This is the actual code `RStudio` uses to read your data when you clicked the Import button.
  - So instead of using the `RStudio` buttons, we can actually Import by writing code similar to what was output into the console!
  - This will come in handy later in the course.

A word on staying organized...



- The Files tab has a few other features to help keep you organized.
  - *SurveyResponse* probably isn't the best name for your data. Click Rename to give it a clearer name.
  - Often, it’s helpful to give your data file the same name as when you import your data.
  - So in this case, we could name our data file *foodhabits.csv*

Export, upload, import

- After you *export*, u*pload*, i*mport* your data you're ready to analyze.
- **`View` your data, select a variable and try to make an appropriate plot for that variable.**
  - If you're having issues, make sure you're spelling the name of your data correctly.

Lesson 14: Variables, Variables, Variables

Objective:

Students will learn how to read and interpret multiple variable plots: bivariate scatterplots, multiple variable scatterplots, stacked bar plots, and side-by-side bar plots. They will summarize their learning about multiple variable plots using a four-fold graphic organizer.

Materials:

1. *Scatterplot of Heights & Weights* (LMR_1.13)
2. *Scatterplot of Heights & Weights, Split by Gender* (LMR_1.14)
3. *Side-by-Side Bar Chart* (LMR_1.15)
4. *Faceted Histogram of Height by Gender (LMR_1.16)*
5. *Summarizing Multi-Variable Plots* graphic organizer (LMR_1.17)

Vocabulary:

scatterplot, grouping, side-by-side bar plot

Essential Concepts: To examine whether two (or more) variables are related, we can plot their distributions on the same graph.

Lesson:

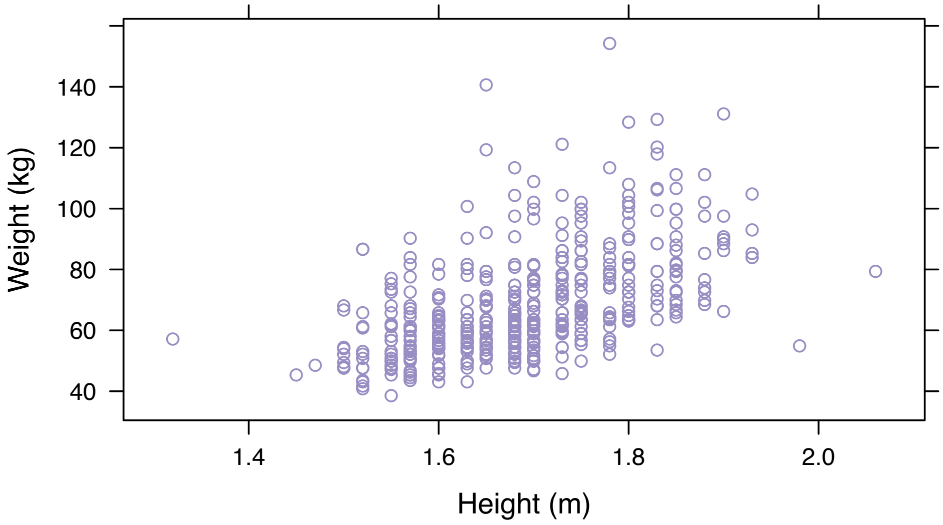
1. Begin by informing students that there will learn how to make visual displays using more than one variable, and by asking them to ponder the following questions:

    a. What do you think is the relation between people’s heights and weights?
    b. Are taller people heavier? Always? Or is this just a tendency?

    What do you think is the relation between people's heights and weights? Are taller people heavier? Always? Or is this just a tendency? Let's look at some data.

2. Display the following plot to the class (LMR_1.13) so they can see some actual data:

Scatterplot of Heights vs. Weights for a Sample of Teens Ages 13–18



LMR_1.13

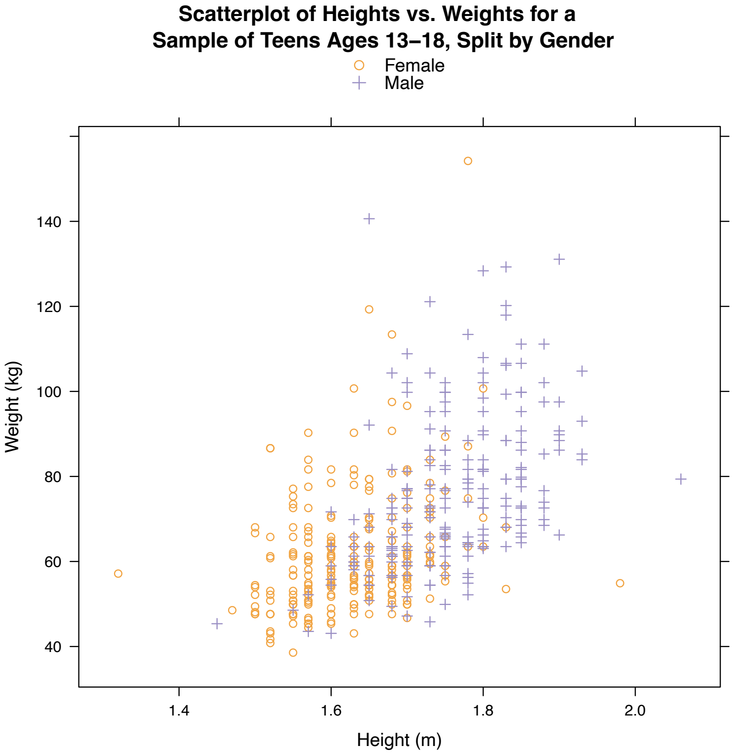
3. Ask students to individually answer the following questions about the plot on the handout (LMR_1.13):
    a. What kind of plot is this and how will you remember its features? ***Scatterplot.***
    b. How many variables are displayed in this plot? Name the variable(s) and identify the type of variable(s). ***Two variables. Weight in kilograms and height in meters. Numerical variables.***
    c. What do the axes show? ***The x-axis shows the height of teens in meters, and the y-axis shows the weight of teens in kilograms.***
    d. Do taller people weigh more? ***Not necessarily, but there is a tendency for this to be true.***

4. Discuss this plot with the class by eliciting students' responses to the questions. Students actively listen to the discussion by confirming, correcting, or adding to their own responses.

5. Close the discussion by asking students: What questions might you have about this plot? What additional information would be helpful?

6. Now, suppose we could see which of these dots represented girls and which represented boys. Where do you think most of the girls' dots would be relative to the boys?

7. Display the following plot to the class (LMR_1.14):



LMR_1.14

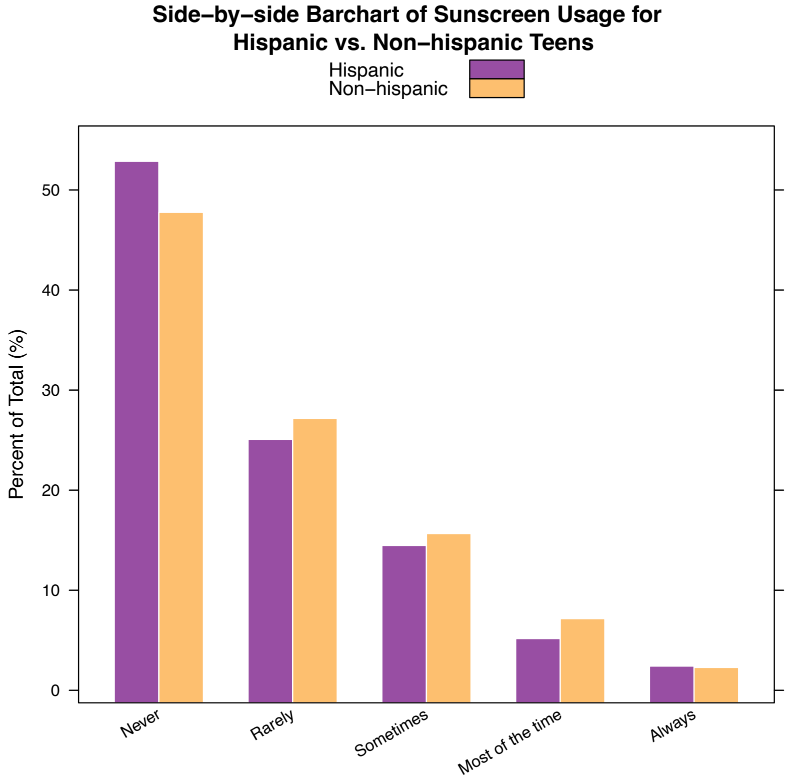
8. Ask students to individually answer the following questions about the plot on the handout (LMR_1.14):
    a. What kind of plot is this and how will you remember its features? ***Scatterplot.***
    b. How many variables are displayed in this plot? Name the variable(s) and identify the type of variable(s). ***Three variables. Weight in kilograms and height in meters are numerical variables. Gender is categorical.***
    c. What do the axes show? ***The x-axis shows the height of teens in meters, and the y-axis shows the weight of teens in kilograms.***

d. What questions can we ask that this graph might answer? ***Who is taller, boys or girls? Who weighs more? Is the association between height and weight the same for boys as it is for girls?***

9. Discuss this plot with the class by eliciting students' responses to the questions. Students actively listen to the discussion by confirming, correcting, or adding to their own responses. Follow-up discussion: when the data are split into categories, it is called **grouping.**

10. Close the discussion by asking students: What questions might you have about this plot? What additional information would be helpful?

11. Display the following plot to the class (LMR_1.15):



LMR_1.15

12. Ask students to individually answer the following questions on the handout (LMR_1.15):

a. What kind of plot is this and how will you remember its features? ***Side-by-side bar chart.***

b. How many variables are displayed in this plot? Name the variable(s). ***Two variables: whether or not someone is Hispanic, and how often they wear sunscreen.***

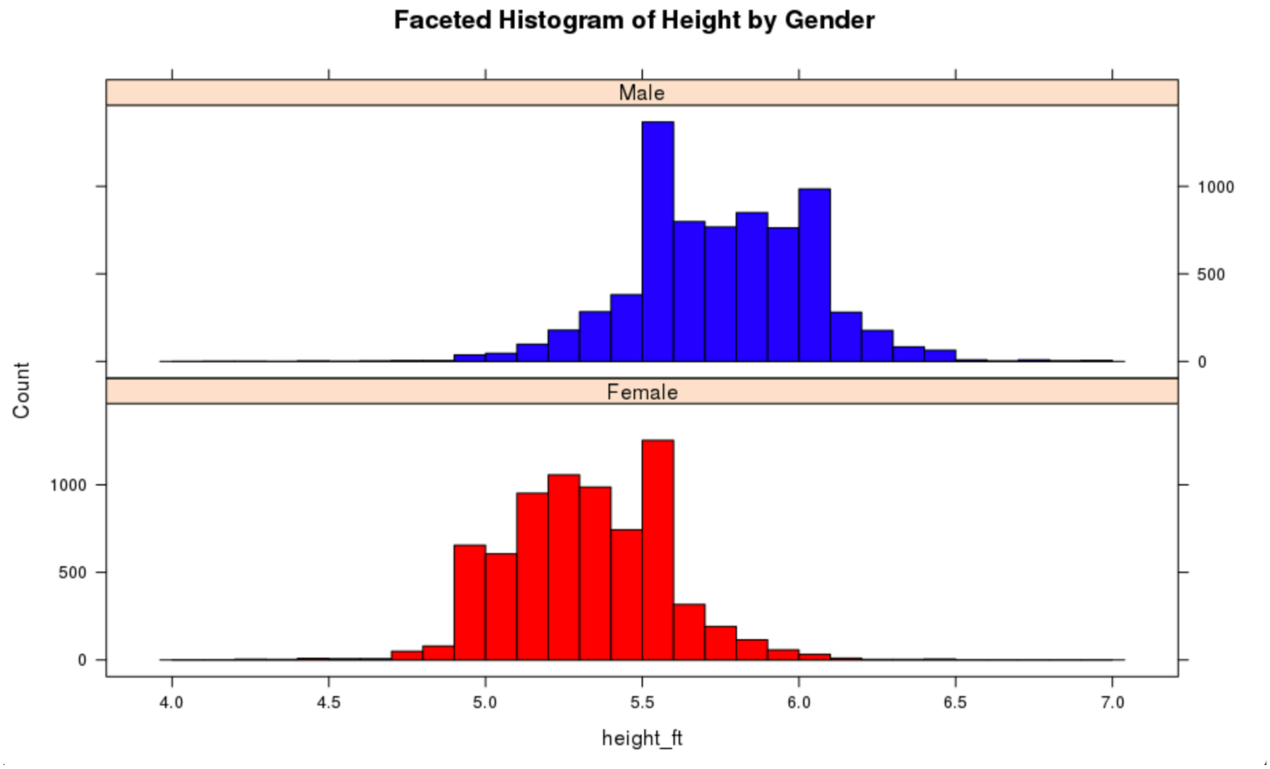
c. What are the x-axis and y-axis telling us? ***The x-axis shows how often a student wears sunscreen, and the y-axis shows the percentage of the total that fall into that category (broken into two bars, one for Hispanic and one for non-Hispanic).***

d. What statistical questions can you answer with this graph? ***Do Hispanics and non-Hispanics have different approaches to sunscreen? What percent of Hispanics always/never wear sunscreen? How does that compare to non-Hispanics?***

13. Discuss this plot with the class by eliciting students' responses to the questions. Students actively listen to the discussion by confirming, correcting, or adding to their own responses.

14. Close the discussion by asking students: What questions might you have about this plot? What additional information would be helpful?

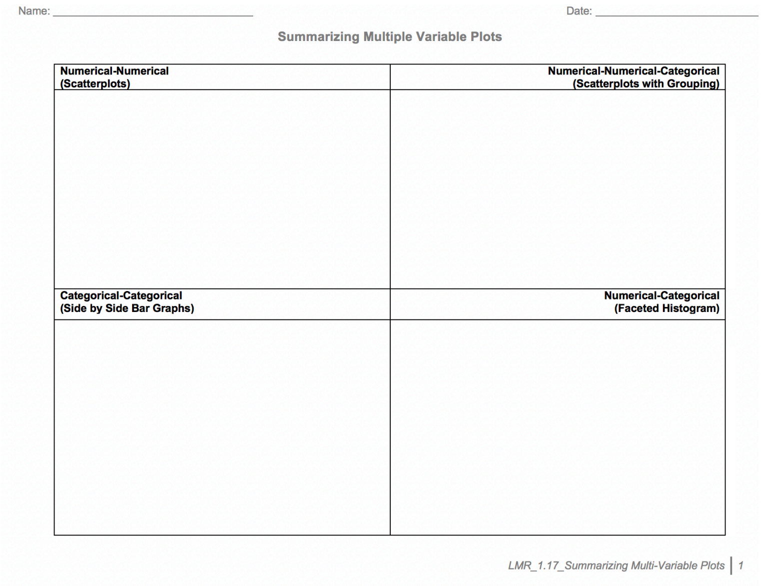
15. Display the following plot to the class (LMR_1.16)



16. Ask students to individually answer the following questions on the handout (LMR_1.16):
    a. What kind of plot is this and how will you remember its features? ***Split or faceted histogram.***
    b. How many variables are displayed in this plot? Name the variable(s). ***Two variables: height and gender.***
    c. What are the x-axis and y-axis telling us? ***The x-axis shows height in feet, and the y-axis shows the total that fall into a certain range of heights (broken into two histograms, one for males and one for females).***
    d. What statistical questions can you answer with this graph? ***Do males and females differ in height? What is the typical female height? What is the typical male height?***

17. Discuss this plot with the class by eliciting students' responses to the questions. Students actively listen to the discussion by confirming, correcting, or adding to their own responses.

18. Using the notes and sketches in their DS journals, students will summarize their learning of how to read and interpret basic multiple variable plots by completing the *Multiple Variable Plots* four-fold graphic organizer (LMR_1.17):



Name: ______________________________ Date: ________________________

Summarizing Multiple Variable Plots

Numerical-Numerical **(Scatterplots)**	**Numerical-Numerical-Categorical** **(Scatterplots with Grouping)**
Categorical-Categorical **(Side by Side Bar Graphs)**	**Numerical-Categorical** **(Faceted Histogram)**

LMR_1.17_Summarizing Multi-Variable Plots | *1*

LMR_1.17

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next 2 Days

LAB 1D: Zooming through Data

LAB 1E: What's the Relationship?

Complete Lab 1D and 1 E prior to the Practicum.

Lab 1D - Zooming Through Data

Directions: Follow along with the slides and answer the questions in **bold (red bold in lab)** font in your journal.

Data with Clarity

- Previously, we've looked at graphs of entire variables (By looking at all of their values).
  - Doing this is helpful to get a *big picture* idea of our data.
- In this lab, we'll learn how to *zoom in* on our data by learning how to subset.
  - We'll also learn a few ways to manipulate the plots we've been making to make them easier to use for analyses.
- Import the data from your class' *Food Habits* campaign and name it `food`.

Another plotting function

- A `dotPlot` is another plot that can be used to analyze a numerical variable.
  - Dotplots are better suited for smaller data sets. If data sets are too large, the dot become too small to see.
  - Similarly, distributions with a large spread might impact the readability of the plot.
- **Use the `dotPlot()` function to create a `dotPlot` of the amount of `sugar` in our food data.**
  - The code to create a `dotPlot` is exactly like you'd use to make a `histogram`.
  - Make sure to use a capital *P* in `dotPlot`.

More Options

- While a `dotPlot` should conserve the exact value of each data point, sometimes it behaves like a histogram in that it lumps values together.
- **Create a more accurate `dotPlot` by using the `nint` option.**
  - Set nint equal to max sugar – min sugar +1
    - On your `food` data spreadsheet, click on the sugar header to sort in ascending order (to obtain minimum)
    - Click on the sugar header again to sort in descending order (to obtain maximum)
  - Use your history pane to see how we included the option `nint` with the `histogram` function
- Pro-tip: If the `dotPlot` comes out looking wonky, try changing the value of the *character expansion option*, `cex`.
  - The default value is `1`. Try a few values between `0` and `1` and a few more values larger than `1`.

Splitting data sets

- In lab 1B, we learned that we can *facet* (or split) our data based on a categorical variable.
- **Split the `dotPlot` displaying the distribution of grams of sugar in two, by faceting on our observations' `salty_sweet` variable.**
  - **Describe how R decides which observations go into the left or right plot.**
  - **What does each *dot* in the plot represent?**

Altering the layout

- It would be much easier to compare the sugar levels of salty and sweet snacks if the dotPlots were stacked on top of one another.
- We can change the **layout** of our separated plots by including the `layout` option in our `dotPlot` function.
    - Add the following option to the code you used to create the `dotPlot` split by `salty_sweet`

```
layout = c(1,2)
```

- *Hint*: Use a similar syntax used with the `nint` option to add the `layout` option to the `dotPlot` function.

Subsetting

- Subsetting is a term we use to describe the process of looking at only the data that conforms to some set of rules:
    - Geologists may subset earthquake data by looking at only large earthquakes.
    - Stock market traders may subset their trading data by looking only at the previous day's trades.
- There's *many* ways to subset data using RStudio, we'll focus on learning the most common methods.

The filter function

- Creating two plots, one for salty and one for sweet is useful for comparing salty and sweet but what if we want to examine only one group by itself?
- Start by creating a subset of the data:
    - Fill in the blanks below with the data and variable names needed to filter the `Salty` snacks from our `food` data:

```
food_salty <- filter(____ , ____ == "Salty")
```

- **View `food_salty` and write down the number of observations in it. Then use the subset data to make a `dotPlot` of the `sodium` in our `Salty` snacks.**

So what's really going on?

- Coding in R is really just about supplying directions in a way that R understands.
    - We'll start by focusing on everything to the right of the "<-" symbol

```
food_salty <- filter(____ , ____ == "Salty")
```

- `filter()` tells R that we're going to look at only the values in our data that follow a *rule*.
- The first blank should be the data we're going to filter down into a smaller set (Based on our rule).
- `salty_sweet == "Salty"` is the rule to follow.

3 parts of defining rules

- We can decompose our rule, `salty_sweet == "Salty"`, into 3 parts:
    (1) `salty_sweet`, is the particular *variable* we want to use to select our subset.
    (2) `"Salty"`, is the *value* of the variable that we want to select. We only want to see data with the value `Salty` for the variable `salty_sweet`.

(3) `==` describes how we want to relate our variable (`salty_sweet`) to our value (`"Salty"`). In this case, we want values of `salty_sweet` that are *exactly equal* to `"Salty"`.

- Notice: *Values* (that are also words) have quotation marks around them. *Variables* do not.

More on ==

- We can use the `head()` function to help us see what's happening when we write `salty_sweet == "Salty"`.
  - `head()` returns the values of the first 6 observations.
  - The `tail()` function returns the last 6 observations.
- Run the following code and answer the question below:

```
head(~salty_sweet == "Salty", data = food)
```

- **What do the values `TRUE` and `FALSE` tell us about how our *rule* applies to the first six snacks in our data? Which of the first six observations were `Salty`?**

Saving values

- To use our subset data we need to save it first.
  - When we *save* something in R what we are really doing is giving a value, or set of values, a specific name for us to use later.
- The arrow `<-` is called the "assignment" operator. It assigns names (on the left) to values (on the right)
  - We now focus on everything to the left of, and including, the "<-" symbol

```
food_salty <- filter(____ , ____ == "Salty")
```

Saving our subset

```
food_salty <- filter(____ , ____ == "Salty")
```

- This code then:
  - takes our subset data, (everything to the right of "<-") ...
  - and assigns the subset data, by using the arrow "<-" ...
  - the name `food_salty`.
- We can now use `food_salty` to do anything we could do with the regular `food` data ...
  - but only including those snacks who reported being `Salty`.

Including more filters

- We often want to filter our data based on multiple rules.
  - For instance, we might want to filter our `food` data based on the food being salty AND having less than 200 calories.
- We can include multiple filters to our subsets by separating each rule with a comma like so:

```
my_sub <- filter(food , salty_sweet == "Salty", calories < 200)
```

- `View` the `my_sub` data we filtered in the above line of code and verify that it only includes salty snacks that have less than 200 calories.

Put it all together

- **Use an appropriate `dotPlot` to answer each of the following questions:**
  - **About how much sugar does the typical sweet snack have?**
  - **How does the typical amount of sugar compare when `healthy_level < 3` and when `healthy_level > 3`?**
- Because you are now working with subsets of data, it is important to be able to label our plots and make this distinction.
  - We can use the `main` option to add a title to our plots
    - Add the following option to the code you used to create the `dotPlot` of the `sugar` in `Sweet` snacks.

```
main = "Distribution of sugar in sweet snacks"
```

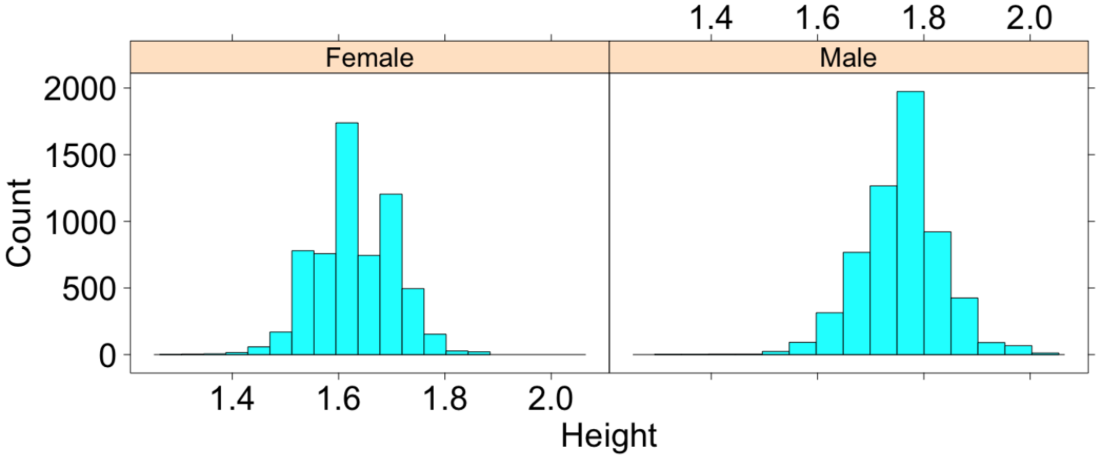
Lab 1E - What's the Relationship?

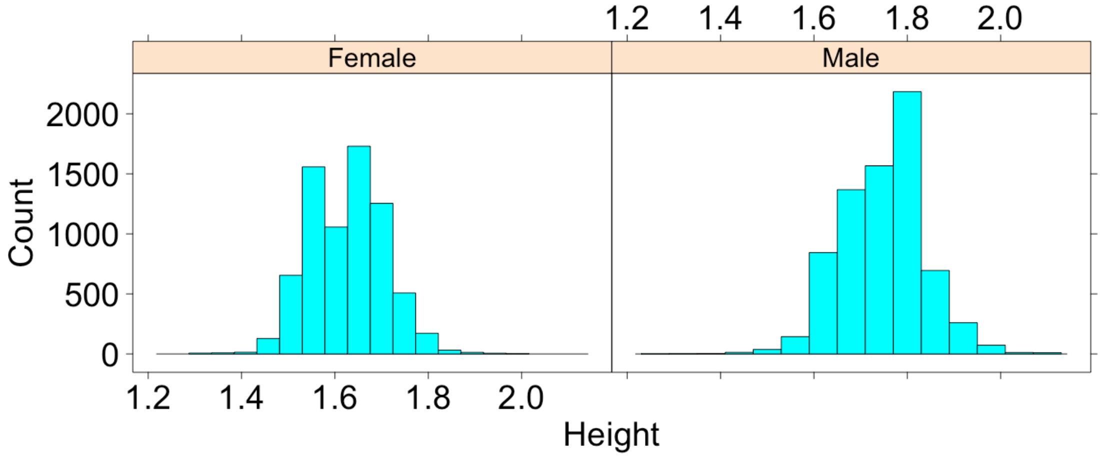
Directions: Follow along with the slides and answer the questions in **bold (red bold in lab)** font in your journal.

Finding patterns in data.

- To discover (*really*) interesting observations or relationships in data, we need to find them!
  - Which is difficult if we only look at the raw data.
- The best tool for finding patterns is often ... your own eyes.
  - Plots are an excellent way to help your eye search for patterns.
- In this lab, we'll learn how to include more variables in our plots to make them more informative.
- Import the data from your class' *Food Habits* campaign and name it `food`.

Where's the variables?



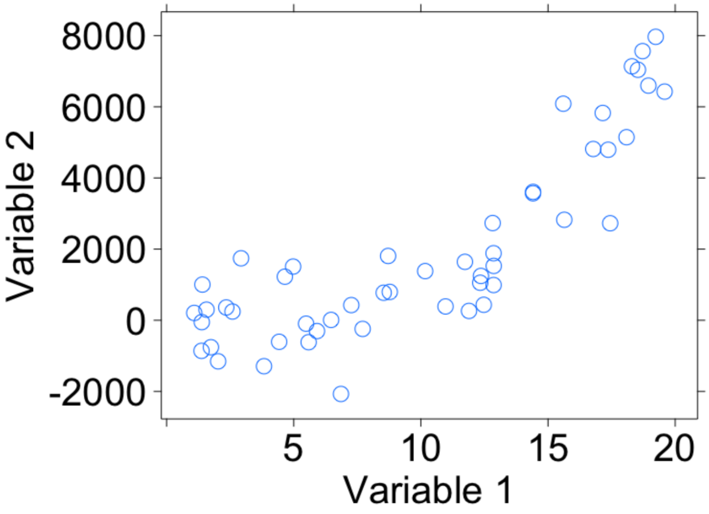


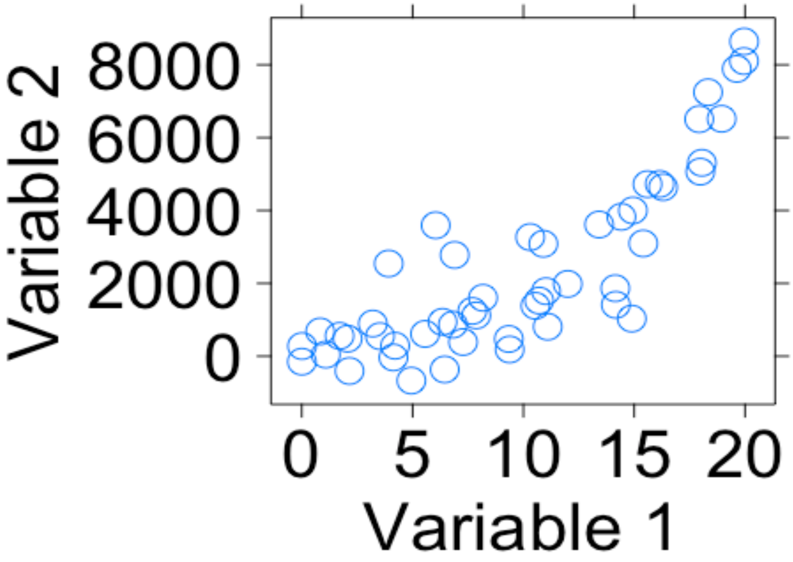
- **How many variables were used to create this plot? Which variables were used and how were they used?**

Multiple variable plots

- The previous graph is an example of a *multiple variable plot*, which means that more than a single variable was used. In this case:
  - Variable 1: *height*
  - Variable 2: *gender*
- Multiple variable plots are tools for finding *relationships* between data.
- Let's take our `food` data and make some new multiple variable plots you haven't created before!

Scatterplots





Creating scatterplots

- Scatterplots are useful for viewing how one *numerical* variable relates to another *numerical* variable.
- Fill in the blanks to create a scatterplot with `sodium` on the y-axis and `sugar` on the x-axis.

```
xyplot(____ ~ ____, data = food)
```

Scatterplots in action

- Use a scatterplot to answer the following questions:
  - **Do snacks that have more `protein` also have more `calories`? Why do you think that?**
  - **What happens if you swap the `protein` and `calories` variables in your code? Does the relationship between the variables change?**
  - **Does the relationship between `protein` and `calories` change when the snack is either `Salty` or `Sweet`? Write down the code you used to answer this question.**

4-variable scatterplots

- When we make scatterplots, we can include:
  - 1 numerical variable on the x-axis
  - 1 numerical variable on the y-axis
  - Use 1 categorical variable to facet our scatterplot
  - Change the color of the points based on another categorical variable
- To change the color of our points, we can include the `groups` argument much like we did for bargraphs (use the *search* feature in the *History* pane if you need help).
- **Create a scatterplot that uses these 4 variables: `sodium`, `sugar`, `cost`, `salty_sweet`.**

Multiple facets

- It can sometimes be helpful to facet on more than 1 variable.
  - Splitting the data using 2 facets can give us additional insights that might otherwise be hidden.

Create a `dotPlot` or `histogram` of the `calories` variable, but facet the data using:

```
healthy_level + salty_sweet
```

How does the `healthy_level` of a `Salty` or `Sweet` snack impact the number of `calories` in the snack?

- Although we are treating `healthy_level` as a categorical variable, `R` recognizes it as a numerical variable.
  - Use the `str` command to confirm
  - Notice that the faceted `histograms` or `dotPlots` do not have labels but rather tick-marks
  - You will have the opportunity to convert the `healthy_level` variable into a factor later on
- Faceting your data on a numerical variable is NOT recommended
  - Numerical variables often have so many different values that they overwhelm the plot and make it hard to read

On your own

- Answer the following questions by creating an appropriate graph or graphs.
  - **Do healthier snacks have more or less `ingredients` than less healthy snacks?**
  - **What other variables seem to be related to the number of `ingredients` of a snack? Describe their relationships.**

Practicum: The Data Cycle & My Food Habits

Objective:

☑ Students will apply what they have learned by engaging in the Data Cycle using the data they collected from the *Food Habits* campaign. Students will present their findings to the class.

Materials:

1. *The Data Cycle Practicum* (LMR_U1_Practicum_Data Cycle)
2. Poster paper
3. Markers

Practicum
The Data Cycle & My Food Habits

Instructions:

With a partner, you will engage in the Data Cycle to address the Research Topic:

What do our snacking habits reveal about us?

Task:

1. Create a Data Cycle poster.
2. The poster should illustrate how the Data Cycle is used to address the Research Topic.
3. Use RStudio to create at least one statistical graphic. The graphic MUST be included on the poster.
4. You and your partner will present your findings with appropriate evidence from the data.

Awards:

Your teacher will select the top posters in the following categories:

- Best Statistical Question
- Most Interesting Statistical Graphic
- Best Illustration of the Data Cycle

Scoring Guide

Below you will find some parameters to assist you in scoring. They are meant only as a guide.

4-point response:

- The poster correctly illustrates how the Data Cycle is used to address the big question.
- A histogram, bar chart, scatterplot, or other graphical representation was correctly created.
- An answer and a justification for the answer to the statistical question are presented.
- A justification includes mention of statistics concepts learned thus far. For example, "The variables are…" **AND** it includes acknowledgment of variability. For example, "There are between ____ and ___."

3-point response:

- The poster correctly illustrates how the Data Cycle is used to address the big question.
- A histogram, bar chart, scatterplot, or other graphical representation was correctly created.
- An answer and a justification for the answer to the statistical question are presented.
- A justification includes mention of statistics concepts learned thus far. For example, "The variables are…" **OR** it includes acknowledgment of variability. For example, "There are between ____ and ___."

2-point response:

- The poster partially illustrates how the Data Cycle is used to address the big question.
- A histogram, bar chart, scatterplot, or other graphical representation was created.
- An answer and a justification for the answer to the statistical question are presented.

1-point response:

- The poster incorrectly illustrates how the Data Cycle addresses the big question.
- An answer to the statistical question is presented but a justification is missing.
- A histogram, bar chart, scatterplot, or other graphical representation was correctly created.

0-point response:

- The Data Cycle is missing **OR** does not show how it addresses the big question.
- A histogram, a dot plot, or other graphical representation was incorrectly created **OR is** missing.
- No answer **AND/OR** no justification for the answer to the statistical question is presented.

Would You Look at the Time!

Instructional Days: 9

Enduring Understandings

Data are useful for evaluating claims and reports. Summaries of categorical and numerical data show important features and patterns in the data. Data summaries provide evidence to make claims.

Engagement

The Bureau of Labor Statistics (BLS) collects data about daily time-use of Americans. Students will explore an interactive graphic titled *How Men and Women Spend Their Days* created by Nathan Yau, that uses data from the American Time Use Survey, to spark their curiosity about how they spend their own time. The graphic can be found at https://flowingdata.com/2021/09/21/how-men-and-women-spend-their-days/.

Learning Objectives

Statistical/Mathematical:

S-ID 5: Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand that data are collected and stored in particular formats. Before data can be analyzed, it must be cleaned so it can be read.

Applied Computational Thinking Using RStudio:

- Create tabular displays of categorical data and summaries of numerical data.
- Create two-way frequency (and relative frequency) tables.
- Use RStudio to calculate joint, marginal, and conditional relative frequencies.
- Subset data frames and create new categorical variables from numerical variables.
- Clean and polish data to make it readable.

Real-World Connections:

Make claims that are based on data and begin to evaluate reports that make claims based on data.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it is used, and how to work with it.

2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

3. Students will use complex sentences to write informative short reports that use data science concepts and skills.

4. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data Collection Method:

Time-Use Participatory Campaign: Students will monitor the amount of time they devote to activities such as sleeping, studying, eating, and partaking in media.

Data Files:

1. Students' *Time-Use* campaign data
2. American Time-Use Survey (ATUS) data

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 15: Americans' Time on Task

Objective:
Introduction to *Time Use Campaign*. Students will explore a multimedia graphic that incorporates data from the American Time Use Survey to spark their interest about how they spend their time. They will begin to learn how to evaluate reports that make claims based on data by reading The Washington Post article Teens Are Spending More Time Consuming Social Media, on Mobile Devices.

Materials:

1. Computers
2. Data Collection Devices
3. Interactive multimedia graphic titled How Men and Women Spend Their Days found at: https://flowingdata.com/2021/09/21/how-men-and-women-spend-their-days/
4. Article: *The Washington Post*'s *Teens Are Spending More Time Consuming Social Media, on Mobile Devices* found at: https://www.washingtonpost.com/postlive/teens-are-spending-more-time-consuming-media-on-mobile-devices/2013/03/12/309bb242-8689-11e2-98a3-b3db6b9ac586_story.html
5. K-L-W Graphic Organizer (LMR_TR_K-L-W Chart)

Vocabulary:
evaluate, claim

Essential Concepts: Learning to examine other analyses is an important part of statistical thinking.

Lesson:

1. Become familiar with the *Time-Use Campaign Guidelines* (shown at the end of this lesson), particularly the big questions, to help guide students during the campaign (see Campaign Guidelines in Teacher Resources).

2. In pairs, ask students to make predictions based on the big questions in the *Time-Use Campaign Guidelines*.

3. Next, inform students that *The Bureau of Labor Statistics* (BLS) collects data about Americans' daily time use and that they will be exploring time use through an interactive graphic.

4. Ask students to go to the multimedia graphic at the following URL: https://flowingdata.com/2021/09/21/how-men-and-women-spend-their-days/

5. Students will spend 10 minutes exploring the interactive graphic. Their task is to answer the following questions (display questions to students):

   a. What variables are represented in this graphic? ***The variables represented are activities that Americans spend their time doing. These include sleeping, eating, traveling, socializing, etc.***

   b. Explain what the graphic is telling you. ***The graphic shows how much time Americans over the age of 15 are spending doing these activities. This information is broken down by different categories of Americans (e.g., gender, ethnicity) and the percentage of Americans doing particular activity at a particular time (e.g., 5% of Americans are working at 6:00 am). The average time spent on a particular activity is also shown (e.g., average time spent at work for all Americans is 3 hours and 25 minutes).***

   c. Where did the data come from? ***The data come from thousands of Americans over the age of 15 who took a survey recalling every minute of a day in 2008.***

   d. What are some interesting findings? Be prepared to share. ***Answers will vary.***

6. Ask students to share their findings in pairs. Each pair will agree on and select one finding to share with the class. In a *Whip Around*, ask each pair to share their finding.

7. Inform students that they will continue to investigate Americans' daily time use. Using the KLW graphic organizer, read out loud the title of *The Washington Post* article: *Teens Are Spending More Time Consuming Social Media, On Mobile Devices*. Ask them to write what they know about the topic in the Know column.

   Note to Teacher: If this is the first time using KLW, please take time to provide an overview of the graphic organizer.

8. Next, ask students to read the article individually: https://www.washingtonpost.com/postlive/teens-are-spending-more-time-consuming-media-on-mobile-devices/2013/03/12/309bb242-8689-11e2-98a3-b3db6b9ac586_story.html

9. As they read, students may complete the Learn column of the KLW graphic organizer.

10. Ask students to complete the Want to Learn column when they finish reading the article.

11. When reading a newspaper, magazine, or blog that includes statistical analysis, it is important to **evaluate,** or think carefully, about **claims** that these articles state as fact.

12. Ask students to work in teams to evaluate the article based on the questions below:

    a. Who was observed and what were the variables observed? ***A group of 8 to 18-year-olds were observed, and the variables observed had to do with consuming media - watching TV, listening to music, surfing the Web, playing video games, and time spent on mobile devices.***

    b. What statistical questions were they trying to answer***? Possible statistical question: How much time per day does today's typical 8 to 18-year-old spend consuming media?***

    c. Who collected the data? ***There were 3 sources cited. The Kaiser Family Foundation collected data in a 2010 study, the Pew Internet and American Life Project collected data in a 2011 study, and the Bureau of Labor Statistics collected data in 2011 with the American Time Use Survey.***

    d. How was the data collected? ***Two were studies whose data collection method is not stated, and one was a survey.***

    e. What claim(s) did the article make? ***Main claim: "Today's teens spend more than 7.5 hours a day consuming media."***

    f. What are some statistics that the article used to make the claim(s)? ***Examples include: Teens use their cellphones to send an average of 60 texts a day. On average, high school students spent less than one hour per weekday on sports, exercise, and recreation.***

13. Select a whole group share-out/discussion strategy from the Instructional Strategies Teacher Resource to discuss the answers to the evaluation questions.

14. Inform students that they will engage in the Time-Use Participatory Sensing campaign and will begin to collect data about their own time use. Follow the *Time-Use Guidelines*.

    Reminder: Once logged into the app or the browser-based version, students may go to **Campaigns** to see the campaigns in which they are participating. They can then add the campaign by tapping the name of the campaign. If no campaigns are visible, ask them to click the refresh option.

15. Emphasize that this data will be tracked throughout the day via a log of some sort – it might be helpful to split the log into three intervals where students pause and think about what they did before school, after school and in the evening. Once the log is complete and accounts for all 1,440 minutes of their day, students should then submit the survey corresponding to that day. They will keep a log for at least 5 days (of which 2 days include a weekend) but no more than 10 days.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Campaign Guidelines – Time Use

1. **The Issue:**

   There have been many reports lately about people spending a large amount of time interacting with technology and the Internet. This raises some questions about time use:

   1) How do I spend my time?
   2) Is there a difference between how females and males spend their time?
   3) Do we spend too much time doing homework?
   4) How is my time use similar or different to other Americans?

2. **Objectives:**

   Upon completing this campaign, students will have compared themselves to the U.S. population to find how they are similar to and/or different from other people in terms of time use. They will use single and multivariable plots, summary statistics, and frequency tables to find similarities and differences between groups of students, and between students and other residents of the United States.

3. **Survey Questions: (**Students will enter data for the activities in which they participate.)

   Consider Data: The categories below are similar to the categories found in the American Time Use Survey (ATUS), which provides nationally representative estimates of how Americans spend their time. Having similar variables allows students to compare the way they spend their time to the official ATUS dataset. Before students begin collecting data, it is important to discuss different activities in their day and how they might be classified. A class consensus of the meaning of the variables must be reached so that proper analysis and interpretations can be made.

   Note: Students cannot double dip their time. For example, if they read during class, then those minutes spent reading do not count towards “read” but instead toward “school”.

   Below are the definitions of some of the variables in the ATUS documentation.

   socialize - This category includes face-to-face social communication and hosting or attending social functions.

   consumer purchases - Time spent purchasing or renting consumer goods, regardless of the mode or place of purchase or rental (in person, online, via telephone, at home, or in a store) is classified into this category. Subcategories in this section include those for time spent purchasing gasoline, time spent purchasing groceries, time spent purchasing other food items, and time spent on all other shopping activities.

   Note: The ATUS variable “leisure” combines many activities in which people might participate, such as watching television, reading, relaxing or thinking, playing on a computer, board, or card games, using a computer or the Internet for personal interest, playing or listening to music, and other activities, such as attending arts, cultural, and entertainment events.

   We have opted to list specific leisure activities that high school students might be more likely to engage in and made them separate variables.

   Students will respond to the following questions:

Prompt	Variable	Data Type
For which day are you collecting data?	day	ordinal category (integers 1-10)
What activities did you participate in?	activities	n/a
a. How many MINUTES did you sleep?	sleep	number
b. How many MINUTES did you spend eating/drinking?	meals	number
c. How many MINUTES did you spend in classes at school?	school	number
d. How many MINUTES did you spend doing homework?	homework	number
e. How many MINUTES did you spend working at a job?	work	number
f. How many MINUTES did you spend grooming yourself?	grooming	number
g. How many MINUTES did you spend traveling/commuting?	travel	number
h. How many MINUTES did you spend doing household chores?	chores	number
i. How many MINUTES did you spend watching television (includes streaming)?	television	number
j. How many MINUTES did you spend playing video games?	videogames	number
k. How many MINUTES did you spend participating in sports/exercise/physical activity?	sports	number
l. How many MINUTES did you spend reading (not for class)?	read	number
m. How many MINUTES did you spend communicating (includes texting, emails, video and voice calls)?	communicate	number
n. How many MINUTES did you spend socializing (outside of class, in person)?	socialize	number
o. How many MINUTES did you spend on a spiritual activity?	spiritual	number
p. How many MINUTES did you spend purchasing items online or in a store?	purchases	number
q. How many MINUTES did you spend on hobbies/volunteering/leisure/extra-curricular activities (excluding sports and physical activity)?	extra	number
r. How many MINUTES did you spend on social media?	social_media	number
AUTOMATIC	location	lat, long
AUTOMATIC	time	time
AUTOMATIC	date	date

When? It is recommended that students keep a log of their time and submit one survey at the end of each day, accounting for every minute of each day of the campaign. It might be helpful to split the log into three intervals where students pause and think about what they did before school, after school and in the evening. Once the log is complete and accounts for all 1,440 minutes of their day, students should then submit the survey corresponding to that day.

How Long? At least five days (maximum of ten days). Ideally, two of these days would include a weekend.

4. **Motivation:**

   Use the https://flowingdata.com/2021/09/21/how-men-and-women-spend-their-days/ Interactive Time Use graphic to explore how Americans spend their time.

   After the first day, monitor the data collection and ensure that each student has submitted a survey for Day 1.

   Discuss data collection issues. What makes it hard? Does this affect the quality of data?

5. **Technical Analysis:**

   RStudio and American time use graphics

   Single/Multivariable plots: histograms, bar graphs, scatterplots, etc.

   Numerical summaries: mean, median, MAD, standard deviation.

   Frequency tables: One and two-way tables.

6. **Guiding Questions:**

   1) On average, how long do students think they spend on homework?
   2) Do males or females take longer to groom themselves?
   3) Are there groups of students who spend their time similarly to one another?

7. **Report:**

   Students will complete a practicum in which they answer a statistical question based on the time-use data collected.

Homework & Next Day

For the next 5 days, students will collect data using the Time Use campaign on their smart devices or via web browser.

LAB 1F: A Diamond in the Rough

and

Data Collection Monitoring

1. **Data Collection Monitoring:** Display the IDS Campaign Monitoring Tool, found at https://portal.idsucla.org/ Click on **Campaign Monitor** and sign in.
   a. See *User List* and sort it by *Total*. Ask: Who has collected the most data so far?
   b. Click on the pie chart. Ask: How many active users are there? How many inactive users are there?
   c. See *Total Responses*. How many responses have been submitted?
   d. Using TPS, ask students to think about what they can do to increase their data collection.

2. Inform students that you will conduct another data collection check with the whole class in a couple of days, and that they will understand the private vs. shared data after they have completed the campaign collection.

Complete Lab 1F prior to Lesson 16

Lab 1F - A Diamond in the Rough

Directions: Follow along with the slides and answer the questions in **bold (red bold in lab)** font in your journal.

Messy data? Get used to it

- Since lab 1, the data we've been using has been pretty *clean*.
- Why do we call it *clean*?
  - Variables were named so we could understand what they were about.
  - There didn't seem to be any *typos* in the values.
  - Numerical variables were considered numbers.
  - Categorical variables were composed of categories.
- Unfortunately, more often than not, data is *messy* until YOU clean it.
- In this lab, we'll learn a few essentials for cleaning *dirty* data.

Messy data?

- What do we mean by messy data?
- Variables might have *non-descriptive names*
  - *Var01*, *V2*, *a*, ...
- Categorical variables might have *misspelled categories*
  - *"blue"*, *"Blue"*, *"blu"*, ...
- Numerical variables might have been *input incorrectly*. For example, if we're talk about people's height in inches:
  - *64.7*, *6.86*, *676*, ...
- Numerical variables might be *incorrectly coded* as categorical variables (Or vice-versa)
  - "64.7", "68.6", "67.6"

The American Time Use Survey

- To show you what *dirty* data looks like, we'll check out the *American Time Use Survey*, or *ATU* survey.
- What is ATU survey?
  - It's a survey conducted by the US government (Specifically the Bureau of Labor Statistics).
  - They survey thousands of people to find out exactly what activities they do throughout a single day.
  - These thousands of people combined together give an idea about how much time the typical person living in the US spends doing various activities.

Load and go:

- Type the following commands into your console:

```
data(atu_dirty)

View(atu_dirty)
```

- **Just by viewing the data, what parts of our ATU data do you think need cleaning?**

Description of ATU Variables

- The description of the actual variables:
  - `caseid`: Anonymous ID of survey taker.
  - `V1`: The age of the respondent.
  - `V2`: The gender of the respondent.
  - `V3`: Whether the person is employed full-time or part-time.
  - `V4`: Whether the person has a physical difficulty.
  - `V5`: How long the person sleeps, in minutes.
  - `V6`: How long the survey taker spent on homework, in minutes.
  - `V7`: How long the respondent spent socializing, in minutes.

New name, same old data

- To fix the variable names, we need to *assign* a new set of names in place of the old ones.
  - Below is an example of the `rename` function:

```
atu_cleaner <- rename(atu_dirty, age = V1,
                      gender = V2)
```

- **Use the example code and the variable information on the previous slide to rename the rest of the variables in `atu_dirty`.**
  - Names should be short, contain no spaces and describe what the variable is related to. So use abbreviations to your heart's content.

Next up: Strings

- In programming, a *string* is sort of like a *word*.
  - It's a value made up of *characters* (i.e. letters)
- The following are example of strings. Notice that each **string** has quotes before and after.

```
"string"

"A1B2c3"

"Hot Cocoa"

"0015"
```

Numbers are words? (Sometimes)

- In some cases, R will treat values that look like *numbers* as if they were *strings*.
- Sometimes we do this on purpose.
  - For example, we can code `Yes/No` variables as `"1"`/`"0"`.
- Sometimes we don't mean for this to happen.
  - The *number of siblings* a person has should not be a string.
- Look at the `structure` of your data and the variable descriptions from a few slides back:
  - **Write down the variables that should be *numeric* but are improperly coded as *strings* or *characters*.**

Changing strings into numbers

- To fix this problem, we need to tell R to think of our "*numeric*" variables as numeric variables.
- We can do this with the `as.numeric` function.
  - An example using this function is below:

```
as.numeric("3.14")

## [1] 3.14
```

- Notice: We started with a string, `"3.14"`, but `as.numeric` was able to turn it back into a number.

Mutating in action

- Look at the variables you thought should be *numeric* and select one. Then fill in the blanks below to see how we can correctly code it as a number:

```
atu_cleaner <- mutate(atu_cleaner,
        age = as.numeric(age),
        ___ = as.numeric(___))
```

- **Once you have this code working, use a similar line of code to correctly code the other *numeric* variables as numbers.**

Deciphering Categorical Variables

- We mentioned earlier that we sometimes code categorical variables as numbers.
  - For example, our gender variable uses `"01"` and `"02"` for `"Male"` and `"Female"`, respectively.
- It's often much easier to analyze and interpret when we use more descriptive categories, such as `"Male"` and `"Female"`.

Factors and Levels

- R has a special name for *categorical* variables, called *factors*.
- R also has a special name for the different *categories* of a *categorical* variable.
  - The individual categories are called *levels*.
- To see the levels of `gender` and their counts type:

```
tally(~gender, data = atu_cleaner)
```

- **Use similar code as we used above to write down the levels for the three factors in our data.**

A level by any other name...

- If we know that '`01`' means '`Male`' and '`02`' means '`Female`' then we can use the following code to recode the *levels* of *gender*.
- Type the following command into your console:

```
atu_cleaner <- mutate(atu_cleaner, gender =
        recode(gender,
                "01"="Male",
                "02" = "Female"))
```

- This code is definitely a bit of a mouthful. Let's break it down.

Allow me to explain

```
        atu_cleaner <- mutate(atu_cleaner, gender =
                recode(gender, "01"="Male",
                  "02" = "Female"))
```

- This code is saying:
  - Replace my current version of `atu_cleaner`…
  - with a mutated one where ...
  - the `gender` variable's levels ...
  - have been recoded..."
  - where `"01"` will now be `"Male"`…
  - and `"02"` will now be `"Female"`.

Finish it off!

- **Recode the categorical variable about whether the person surveyed had a physical challenge or not. The coding is currently:**
  - `"01"`: Person surveyed *did not* have a physical challenge.
  - `"02"`: Person surveyed *did* have a physical challenge.
- **Write a script that:**
  1. Loads the `atu_dirty` data set
  2. Cleans the data as we have in this lab
  3. Saves a copy of the cleaned data (see next slide).

The final lines

- The last few lines of your script are extremely important because they will save all your work.
- Be sure to `View` your data and check its `structure` to make sure it looks clean and tidy before saving.

Run the code below:

- `atu_clean <- atu_cleaner`This code will create a new data frame in your `Environment` called `atu_clean` which is a final copy of `atu_cleaner`
  - If `atu_clean` is swept from your `Environment` all of the changes you made will NOT be saved
  - You would need to re-run the script to clean the data again
- To permanently save your changes you need to save the file as an R data file or `.Rda`

Run the code below:

```
        save(atu_clean, file = "atu_clean.Rda")
```

- Look in your `Files` pane for the `atu_clean.Rda` file
  - This is a permanent copy of your clean atu data
  - To load the data onto your `Environment` click on the file
  - A pop-up window confirming the upload will appear

Flex your skills

- Now that you have learned some cleaning data basics, it’s time to revisit the `food` data.

Run the code below:

```
        histogram(~calories | healthy_level, data = food)
```

- **Use the `as.factor()` function to convert `healthy_level` into a categorical variable and re-run the `histogram` function.**

Notice that the healthy_level categories are now numbers as opposed to tick-marks. This is an improvement but an even better solution would be to recode the categories.

- **Recode the healthy_level categories and re-run the histogram function.**
  - "1" = "Very Unhealthy"
  - "2" = "Unhealthy"
  - "3" = "Neutral"
  - "4" = "Healthy"
  - "5" = "Very Healthy"
- If your food data is cleared from your Environment, the changes that you made to the healthy_level variable will not be saved.
- To save your changes permanently save your food file as an R data file.

Lesson 16: Categorical Associations

Objective:
Students will learn to construct, interpret, and calculate the joint relative frequencies of two-way frequency tables.

Vocabulary:
two-way frequency table, joint relative frequency

Essential Concepts: A two-way table is a summary of the association/relationship between two categorical variables. Joint relative frequencies answer questions of the form "what proportion of the people/objects had *this* value on the first variable and *this* value on the second."

Lesson:

1. Launch the lesson by displaying the following scenario:

   Rosa has a theory that cat owners are also musical. To find out, she decided to collect data that would help her understand the relationship between cat ownership and instrument playing among the students in her art class. She conducted a survey and found that out of the 35 students in her art class, 16 owned a cat and out of those that owned a cat, 7 played an instrument. She also discovered that 9 owned a cat, but did not play an instrument. There were also 9 students who neither owned a cat nor played an instrument.

2. Inform students that Rosa asked two questions that provided the data for her two-way frequency table. What could those two questions be?

   Possible Answer:* *Question 1—Do you play an instrument?
   Question 2—Do you own a cat?

3. What variables did Rosa collect? What were the values of those variables?

4. In pairs, write out on paper what the original data must have looked like.

   Answer: Variable 1: Owns Cat. Variable 2: Plays Instrument

Owns Cat	***Plays Instrument***	
Yes	***Yes***	***(There are 7 of these)***
Yes	***No***	***(9 of these)***
No	***Yes***	***(10 of these)***
No	***No***	***(9 of these)***

5. Inform students that today they will be looking at associations in categorical variables.

   Note: If necessary, review the difference between questions for categorical and numerical variables.

6. Explain that their task is to summarize Rosa's findings in one table that shows totals. Remind students to use their knowledge of data structures from Lesson 2, especially organizing in rows and columns.

7. Allow time for teams to wrestle with how to organize their data in one table. As teams work, walk around monitoring their data tables.

8. Select a few data tables to display and share with the entire class.

9. Explain the *Anonymous Author* strategy to students (see Instructional Strategies in Teacher Resources).

10. Display the data tables and ask students to engage in the *Anonymous Author* strategy. You may want to start the discussion by asking about the total number of students Rosa surveyed.

11. Make sure the last data table you display correctly shows a **two-way frequency table**. A two-way frequency table displays the data that pertains to two categories from one group. One category is represented in rows and the other is represented in columns. In this exercise, the group is a class of art students.

Cat Ownership and Instruments

	Plays an instrument	**Does not play instrument**	**Total**
Owns a cat	7	9	16
No cats	10	9	19
Total	17	18	35

12. Based on the Cat Ownership and Instruments table, ask student teams to generate questions that can be asked and answered by the data.
13. In a *Whip Around*, ask student teams to share one of their questions.
14. Explain that a two-way frequency table can show relative frequencies. A **relative frequency** is how often something occurs in relation to the total number of occurrences, and is expressed as a proportion or percentage of a total. For example, what is the relative frequency of those who own a cat and play an instrument? ***Answer: 7/35 or 0.2 or 20%***.

    Note: Review how to write a proportion and how to express a proportion as a percent.
15. Ask students to calculate the relative frequencies for the entire table. They may check their calculations with a partner.

Cat Ownership and Instruments
Relative Frequencies

	Plays an instrument	**Does not play instrument**	**Total**
Owns a cat	7/35 = 0.20	9/35 ≈ 0.26	16/35 ≈ 0.46
No cats	10/35 ≈ 0.29	9/35 ≈ 0.26	19/35 ≈ 0.54
Total	17/35 ≈ 0.49	18/35 ≈ 0.51	35/35 = 1.00

16. In teams, students will generate 2 questions about 2 categorical variables. Allow teams 3-5 minutes to generate their questions. Students should not choose two random categorical variables. Rather, they should choose two categorical variables that they predict might be associated.
17. The team will create a two-way table that corresponds to their categorical variables.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Rosa posed this statistical question:

What proportion of students did not play an instrument and did not own a cat?

Use what you know about two-way tables to answer her question.

Lesson 17: Interpreting Two-Way Tables

Objective:
Students will calculate conditional, marginal, and joint frequencies and explain what they mean in the context of the data.

Materials:

1. Poster paper
2. Markers
3. *Analyzing Categorical Variables* (LMR_1.18)
   Advanced preparation required (see step 19 below)
4. *Interpreting Categorical Variables* (LMR_1.19)

Vocabulary:
marginal frequency, joint frequency, conditional relative frequency

Essential Concepts: Marginal (relative) frequencies tell us about the distribution of a single variable. Conditional relative frequencies tell us about the distribution of one variable when "subsetting" the other.

Lesson:

1. **Time Use Campaign Data Collection Monitoring:**
   a. Display the IDS Campaign Monitoring Tool, found at https://portal.idsucla.org/
      Click on **Campaign Monitor** and sign in.
   b. Inform students that you will be monitoring their data collection again today.
      i. See *User List* and sort it by *Total*. Ask: Who has collected the most data so far?
      ii. Click on the pie chart. Ask: How many active users are there? How many inactive users are there?
      iii. See *Total Responses*. How many responses have been submitted?
      iv. Using TPS, ask students to think about what they can do to increase their data collection.
2. Remind students that this is the last day to collect data.
3. Ask student teams to take out the 2 questions and the two-way table that they created in the previous day's lesson.
4. Before teams ask the class their questions, ask them to strategize about how they will collect and record their data, because they can only ask the 2 questions.
5. Students in the class will respond to each question by raising their hands.
6. In a *Whip Around*, have each team ask their 2 questions. Pause briefly between teams so that the asking team has time to collect and record their data.
7. Students will use their frequency tables before the end of the lesson.
8. Recall that in the previous lesson, students learned to calculate relative frequencies. Now it's time to look at other ways of understanding a two-way frequency table.
9. Display the *Cat Ownership and Instruments* table:

Cat Ownership and Instruments

	Plays an instrument	Does not play instrument	Total
Owns a cat	7	9	16
No cats	10	9	19
Total	17	18	35

10. Suppose that we want to know the following information (display questions):

    a. How many students own a cat? ***16***
    b. What is the proportion of students who own a cat? ***16/35 ≈ 0.46***
    c. What is the proportion of students who do not play an instrument? ***18/35 ≈ 0.51***

11. In teams, discuss where on the table you would find this information and how you would calculate it. The specific answers are not important; but what is important is to know how to obtain the information***. Possible answer: You would find the proportion of students who do not play an instrument by dividing the number in the "Total" row that is in the "Does not play instrument" column by the total number of students (35).***

12. After a few minutes, ask a team to volunteer a response. Mark up the margins on the table to show that the cells with the initial total counts are called **marginal frequencies**. Note: 10 b and c are marginal relative frequencies.

13. Now suppose that we want to know the following information (display questions):

    a. How many students own a cat and play an instrument? ***7***
    b. What is the proportion of students that own a cat and play and instrument? ***7/35=0.2***
    c. What is the proportion of students who do not own a cat and play an instrument? ***10/35*** ≈ ***0.286***

14. In teams, discuss where on the table you would find this information and how you would calculate it. The specific answers are not important; but what is important is to know how to obtain the information***. Possible answer: You would find the answers in the cells that make up the body of the table. The value for each proportion is the frequency for each cell over the total number of observations***.

15. After a few minutes, ask a team to volunteer a response. Mark up the cells in the body of the table to show that the cells with the initial counts are called **joint frequencies**. Note: 13 b and c are joint relative frequencies.

16. Finally, suppose that we wanted to answer the question: Do a greater proportion of students in Rosa's art class who do not own cats prefer to play an instrument than those who do own cats?

17. In teams, discuss where on the table you would find this information and how you would calculate it. The specific answers are not important; but what is important is to know how to obtain the information.

18. After a few minutes, ask a team to volunteer a response. Encourage students to agree or disagree with the explanations provided. Lead students to see that the total for the "No cats" row is important because we are only concerned with that subset of the group. Mark up "No cats" row on the table to show that we have conditioned, or are bound, by this variable. Compare the values that show the **conditional relative frequency** for the row. More non-cat owners slightly prefer to play an instrument (display table below).

Cat Ownership and Instruments

Conditional Relative Frequencies by Row

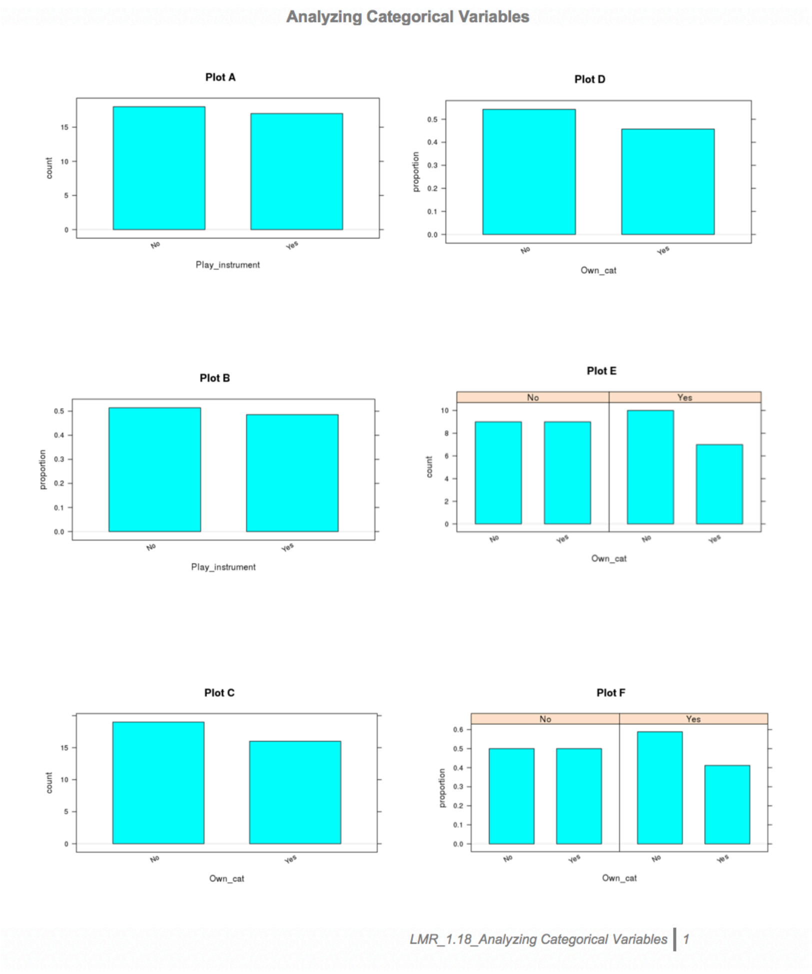
	Plays an instrument	**Does not play instrument**	**Total**
Owns a cat	7/16 ≈ 0.44	9/16 ≈ 0.56	16/16 ≈ 1.00
No cats	10/19 ≈ 0.53	9/19 ≈ 0.47	19/19 ≈ 1.00
Total	17/35 ≈ 0.49	18/35 ≈ 0.51	35

Note: This is a conditional relative frequency by row. We can also calculate conditional relative frequencies by column if we were interested in knowing the difference in cat preference for those who play instruments versus those who don't.

19. Distribute one full set of cards from the *Analyzing Categorical Variables* file (LMR_1.18) to each student team.

    Advanced preparation required:

    Print the *Analyzing Categorical Variables* file (LMR_1.18). The handouts can then be cut into a total of 20 cards (12 visuals, 8 numerical summaries). You will need enough sets of the cards for each student team to share one full set. For example, if there are 5 student teams in a class, then 5 copies of the file will need to be printed so that each team gets all 20 cards.

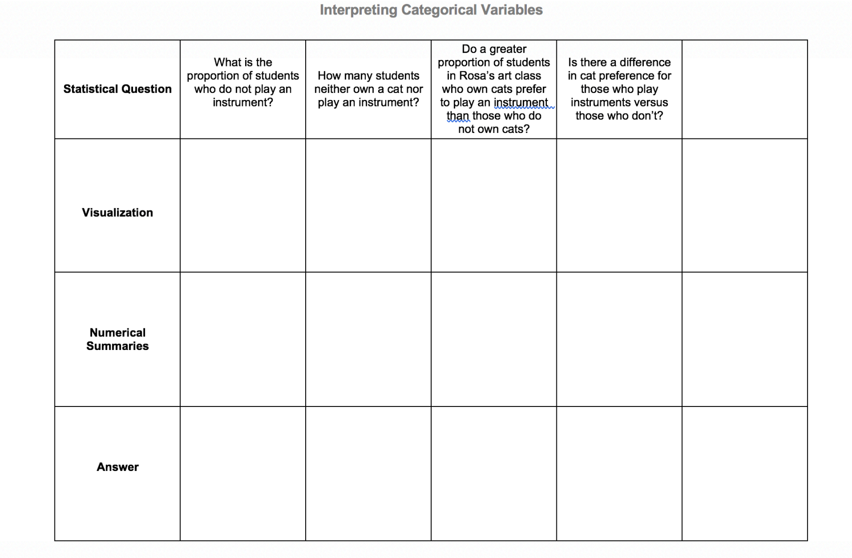


Analyzing Categorical Variables

LMR_1.18_Analyzing Categorical Variables | *1*

LMR_1.18

20. Distribute LMR_1.19_*Interpreting Categorical Variables* to each student team.



Interpreting Categorical Variables

Statistical Question	What is the proportion of students who do not play an instrument?	How many students neither own a cat nor play an instrument?	Do a greater proportion of students in Rosa's art class who own cats prefer to play an instrument than those who do not own cats?	Is there a difference in cat preference for those who play instruments versus those who don't?	
Visualization					
Numerical Summaries					
Answer					

LMR_1.19

21. Each student team will work together and decide which visualization(s) and numerical summaries can be used to answer each statistical question. They will then answer each statistical question, citing a numerical summary as evidence.

    Note: Student teams may tape or glue visuals and numerical summaries onto LMR_1.19, or they can simply write the plot letter and table number in the appropriate box. The blank column is for student teams to write a statistical question than can be answered with a visual and a numerical summary that was not used.

22. After student teams have been allotted ample time to complete LMR_1.19, lead a class discussion to go over the answers. It is extremely important to have students justify their answers by referring to their visuals and tables. For example, the statistical question "How many students neither own a cat or play an instrument?" can be answered with Plot E, Plot G, Plot K, Plot I, and with Tables 1 and 7.

23. Ask students to refer back to the two-way frequency tables they created earlier. Have each team create one poster that shows their two-way frequency table. Then, ask each team to ask 4 questions about the data in their table that must be answered by a:

    a. marginal frequency
    b. marginal relative frequency
    c. joint relative frequency
    d. conditional relative frequency (either by row or column)

24. If time permits, pair teams up and ask them to present their findings to each other.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

Using the data below, generate 2 questions: one must be answered with a marginal relative frequency and the other must be answered by a conditional relative frequency.

Gender and the Color Red

Which emotion do you most relate with the color red?

	Love	Anger	Fear	Total
Male	7	11	5	23
Female	12	15	10	37
Total	19	26	15	60

Lab 1G: What's the FREQ?

Complete Lab 1G prior to the Practicum.

Lab 1G - What's the FREQ?

Directions: Follow along with the slides and answer the questions in **bold (red bold in lab)** font in your journal.

Clean it up!

- In Lab 1F, we saw how we could *clean* data to make it easier to use and analyze.
  - You cleaned a small set of variables from the American Time Use (ATU) survey.
  - The process of cleaning and then analyzing data is *very* common in Data Science.
- In this lab, we'll learn how we can create frequency tables to detect relationships between categorical variables.
  - For the sake of consistency, rather than using the data you cleaned, you will use the pre-loaded ATU data.
  - Use the `data()` function to load the `atu_clean` data file to use in this lab.

How do we summarize categorical variables?

- When we're dealing with categorical variables, we can't just calculate an **average** to describe a *typical* value.
  - (Honestly, what's the average of categories *orange*, *apple* and *banana*, for instance?)
- When trying to describe categorical variables with numbers, we calculate **frequency tables**

Frequency tables?

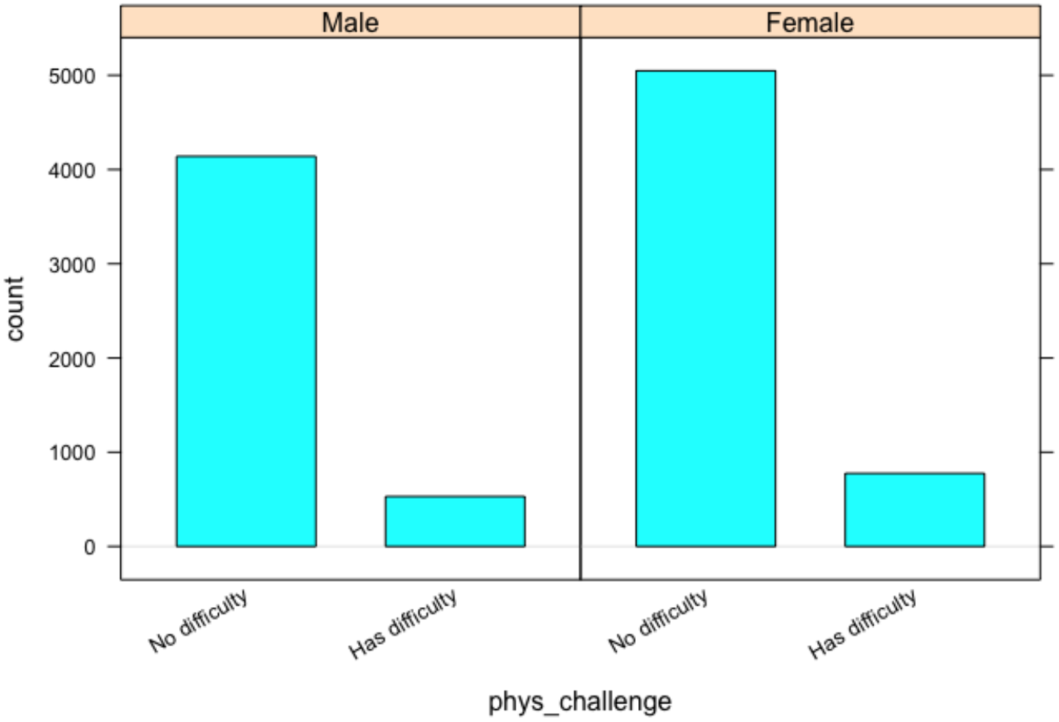
- When it comes to categories, about all you can do is *count* or *tally* how often each category comes up in the data.
- Fill in the blanks below to answer the following: **How many more *females* than *males* are there in our ATU data??**

```
tally(~ ____, data = ____)
```

2-way Frequency Tables

- Counting the categories of a single variable is nice, but often times we want to make comparisons.
- For example, what if we wanted to answer the question:
  - **Does one `gender` seem to have a higher occurrence of physical challenges than the other? If so, which one and explain your reasoning?**
- We could use the following plot to try and answer the question:

```
bargraph(~phys_challenge | gender, data = atu_clean)
```



- The split bargraph helps us get an idea of the answer to the question, but we need to provide precise values.

Use a line of code, that's similar to how we facet plots, to obtain a tally of the number of people with physical challenges and their genders.

Interpreting 2-way frequency tables

- Recall that there were 1153 more women than men in our data set.
  - If there are more women, then we might expect women to have more physical challenges (compared to men).
- Instead of using *counts* we use *percentages*.
- Include: `format = "percent"` as option to the code you used to make your 2-way frequency table. Then answer this question again:
  - **Does one `gender` seem to have a higher occurrence of physical challenges than the other? If so, which one and explain your reasoning?**
  - **Did your answer change from before? Why?**
- It's often helpful to display totals in our 2-way frequency tables.
  - To include them, include `margins = TRUE` as an option in the tally function.

Conditional Relative Frequencies

- There is a difference between `phys_challenge | gender` and `gender | phys_challenge`

```
tally(~phys_challenge | gender, data = atu_clean, margin = TRUE)
##                      gender
## phys_challenge      Male   Female
##  No difficulty      4140    5048
##  Has difficulty      530     775
##  Total              4670    5823
tally(~gender | phys_challenge, data = atu_clean, margin = TRUE)
##              phys_challenge
## gender       No difficulty   Has difficulty
##  Male                4140              530
##  Female              5048              775
##  Total               9188             1305
```

- At first glance, the two-way frequency tables might look similar (epecially when the `margin` option is excluded). Notice, however, that the totals are different.
- The totals are telling us that `R` calculates conditional frequencies by column!
- What does this mean?
  - In the first two-way frequency table the groups being compared are `Male` and `Female` on the distribution of physical challenges.
  - In the second two-way frequency table the groups being compared are the people with `No difficulty` and those that `Has difficulty` on the distribution of gender.

Add the option `format = “percent”` to the first tally function. How were the percents calculated? Interpret what they mean.

On your own

- **Describe what happens if you create a 2-way frequency table with a numerical variable and a categorical variable.**
- **How are the types of statistical questions that 2-way frequency tables can answer different than 1-way frequency tables?**
- **Which gender has a higher rate of *part time employment*?**

Practicum: Teen Depression

Objective:

☑ Using the CDC data set, students will apply their learning of statistical concepts to determine possible factors that might be associated with depression in teens. They will create graphical representations to analyze and interpret the data. Students will present their findings to their teams the following day.

Materials:

1. *Teen Depression Practicum* (LMR_U1_Practicum_Depression)
2. *Depression Fact Sheet* (LMR_U1_Practicum_Depression_Fact Sheet)
3. Poster paper
4. Markers

Practicum
Teen Depression

Background:

The Centers for Disease Control and Prevention (CDC) collect data about teenagers on a variety of topics. One of these topics is depression. According to the fact sheet published by the National Institute for Mental Health, depression is a real problem among teens.

Instructions:

With a partner, you will read the depression fact sheet and then use the CDC data to address the Research Topic.

Research Topic:

What factors are associated with depression in teens?

Task:

1. Create a poster that addresses the Research Topic.
2. Generate a statistical question that might address the Research Topic.
3. Use RStudio to create at least one statistical graphic. The graphic MUST be included on the poster.
4. You and your partner will present your findings with appropriate evidence from the data.

Awards:

Your teacher will select the top posters in the following categories:

- Best Statistical Question
- Most Interesting Statistical Graphic

Scoring Guide:

4-point response:

- The poster identifies possible factors that are in the data set that might be associated with depression in teens.
- A graphical representation that shows an association was created.
- An answer and a justification for the answer to the statistical question are presented.
- A justification includes mention of statistics concepts learned thus far. For example, “The variables are…” ; **AND** it includes acknowledgment of variability. For example, “There are between ____ and ___.”

3-point response:

- The poster identifies possible factors in the data set that might be associated with depression in teens.
- A graphical representation that shows an association was created.
- An answer and a justification for the answer to the statistical question are presented.
- A justification includes mention of statistics concepts learned thus far. For example, “The variables are…” ; **OR** it includes acknowledgment of variability. For example, “There are between ____ and ___.”

2-point response:

- The poster identifies possible factors that might be associated with depression in teens.
- A graphical representation that shows an association was created.
- An answer and a justification for the answer to the statistical question are presented.

1-point response:

- The poster identifies possible factors that might be associated with depression in teens.
- An answer to the statistical question is presented but a justification is missing.
- A histogram, bar chart, scatterplot, or other graphical representation was correctly created.

0-point response:

- The poster does not identify possible factors that might be associated with depression in teens
- A histogram, a dot plot or other graphical representation was incorrectly created **OR** is missing.
- No answer **AND/OR** no justification for the answer to the statistical question is presented.

Next Day

Lab 1H: Our time.

Complete Lab 1H prior to the End of Unit Project.

Lab 1H - Our time.

Directions: Follow along with the slides and answer the questions in **bold (red bold in lab)** font in your journal.

We've come a long way

- The labs until now have covered a huge range of topics:
  - We've learned how to make plots for different types of variables.
  - We know how to subset our data to get a more refined view of our data.
  - We've covered cleaning data and making two-way frequency tables.
- In this lab, we're going to combine all of these ideas and topics together to find out how we spend our time.

First steps first.

- *Export*, *Upload*, *Import* the data from your class' *Time Use* campaign.
- The data, as-is, is very messy and hard to interpret/analyze.
  - Fill in the blank with the name of your imported data to format it:

```
timeuse <- timeuse_format( _____ )
```

- This function formats/cleans the data so that each row represents a typical day for each student in the class
- Hint: Search your `History` tab for the code to save your formatted timeuse data as an R data file (.Rda)

timeuse_format specifics

- In case you're wondering, the timeuse_format function:
  - Takes each student's daily data and adds up all of the time spent doing each activity for each day.
  - The time spent on each activity for each day is then averaged together to create a *typical day* in the life of each student.

Exploring your data

- Start by getting familiar with your `timeuse` data:
  - **How many observations and variables are there?**
  - **What are the names of the variables?**
  - **Which row represents YOUR *typical day*?**

How do we spend our time?

- We would like to investigate the *research question*: "How did our class spend our time?"
  - To do this, we'll perform a statistical investigation.
- **State and answer two statistical questions based on our *research question*.**
  - **Also, state one way in which your personal data is *typical* and one way that it *differs* from the rest of the class.**
- **Justify your answers by using appropriate statistical graphics and summary tables.**
  - **If you subset your data, explain why and how it benefited your analysis.**

End of Unit Project and Oral Presentation: Analyzing Data to Evaluate Claims

Objective:

☑ Students will apply their learning of the first unit in the curriculum by completing an End of Unit Project.

Materials:

1. *IDS Unit 1—End of Unit Project* (LMR_U1_End of Unit Project)

End of Unit 1 Project and Oral Presentation: Analyzing Data to Evaluate Claims

Congratulations! You are on your way to becoming a Data Scientist. You have now learned some basic statistics concepts - along with RStudio skills - to help you analyze and interpret data. It is time to apply what you have learned so far.

You will apply what you have learned by engaging in the following:

1. Use an article from the list provided below, or find an article, report, blog post, etc., in a magazine, newspaper, or other media related to the topic of nutrition or time use that makes a claim. Use an article we have not used in class.
    a. *How Americans Eat Today:* http://www.cbsnews.com/news/how-americans-eat-today/
    b. *Why do we still eat this way?* https://www.washingtonpost.com/news/to-your-health/wp/2014/08/04/why-do-we-still-eat-this-way
    c. *Americans Snack Differently Than Other Nations*: http://www.usatoday.com/story/money/business/2014/09/29/snacking-consumer-eating-habits-nielsen/16263375/?siteID=je6NUbpObpQ-3jFHwYITZ99FE23ytK_q9g
    d. *The Surprising Amount of Time Kids Spend Looking at a Screen* *http://www.theatlantic.com/education/archive/2015/01/the-surprising-amount-of-time-kids-spend-looking-at-screens/384737/*
    e. *Youths Spend 7+ Hours/Day Consuming Media:* http://www.cbsnews.com/news/youths-spend-7-plus-hours-day-consuming-media/
2. Analyze the article or report based on the following questions:
    a. What claim(s) did the article make?
    b. What statistical questions were they trying to answer?
    c. Does the article cite data? If so:
        i. Who was observed and what were the variables observed?
        ii. Who collected the data?
        iii. How was the data collected?
        iv. What are some statistics that the article used to make the claim(s)?
    d. If there was no data, how did the article justify its claim?
3. Determine whether the class's Food Habits or Time Use campaign data supports, refutes, or is inconclusive of the claim(s) made in the article.
4. Use RStudio to do your analysis using either the Food Habits or Time Use campaign data and create graphics/plots that support your reasoning.
5. Generate other statistical questions that you would like to investigate further after you reach your conclusion.
6. Write a summary of your analysis that is no more than 4 pages long. Include graphics/plots/tables that provide evidence to support your reasoning. Be sure to include everything in items 1-5.
7. Prepare a 2-minute presentation of your report. Make sure you refer to your graphics/plots/tables during your presentation.

IDS

Introduction to Data Science

Unit 2

Introduction to Data Science
Daily Overview: Unit 2

Theme	Day	Lessons and Labs	Campaign	Topics	Page
What Is Your True Color? (10 days)	1	Lesson 1: What Is Your True Color?	Personality Color - data	Subsets, relative frequency	125
	2	Lesson 2: What Does *Mean* Mean?	Personality Color	Measures of center – mean	128
	3	Lesson 3: Median In the Middle	Personality Color	Measures of center – median	132
	4	Lesson 4: How Far Is It from Typical?	Personality Color	Measures of spread – MAD	136
	5	Lab 2A: All About Distributions	Personality Color	Measures of center & spread – mean, median, MAD	140
	6	Lesson 5: Human Boxplots		Boxplots, IQR	142
	7	Lesson 6: Face Off		Comparing distributions	145
	8	Lesson 7: Plot Match		Comparing distributions	147
	9	Lab 2B: Oh, the Summaries…	Personality Color	Boxplots, IQR, numerical summaries, custom functions	150
	10	Practicum: The Summaries	Food Habits or Time Use	Statistical questions, comparing distributions	153
How Likely Is It? (7 days)	11	Lesson 8: How Likely Is It?		Probability, simulations	157
	12	Lesson 9: Bias Detective		Simulations to detect bias	161
	13	Lesson 10: Marbles, Marbles		Probability, with replacement	165
	14	Lab 2C: Which Song Plays Next?		Probability of simple events, do loops, set.seed()	167
	15	Lesson 11: This AND/OR That		Compound probabilities	170
	16	Lab 2D: Queue It Up!		Probability with & without replacement, sample()	174
	17	Practicum: Win, Win, Win		Probability estimation through repeated simulations	177
Are You Stressing or Chilling? (8 Days)	18^	Lesson 12: Don't Take My Stress Away	Stress/Chill – data	Introduction to campaign	180
	19	Lesson 13: The Horror Movie Shuffle	Stress/Chill – data	Chance differences – cat var	184
	20	Lab 2E: The Horror Movie Shuffle	Stress/Chill – data	Inference for categorical variable, do loops, shuffle()	188
	21	Lesson 14: The Titanic Shuffle	Stress/Chill – data	Chance differences – num var	191
	22	Lab 2F: The Titanic Shuffle	Stress/Chill – data	Inference for numerical variable, do loops, shuffle()	195
	23+	Lesson 15: Tangible Data Merging	Stress/Chill – data	Merging data sets	197
	24	Lab 2G: Getting It Together	Stress/Chill & Personality Color	Merging data sets, stacking vs. joining	199
	25	Practicum: What Stresses Us?	Stress/Chill & Personality Color	Answering statistical questions of merged data	201
What's Normal? (5 Days)	26	Lesson 16: What Is Normal?		Introduction to normal curve	204
	27	Lesson 17: Normal Measure of Spread		Measures of spread - SD	208
	28	Lesson 18: What's Your Z-Score?		z-scores, shuffling	211
	29	Lab 2H: Eyeballing Normal		Normal curves overlaid on distributions & simulated data	216
	30	Lab 2I: R's Normal Distribution Alphabet		Normal probability, rnorm(), pnorm(), quantiles, qnorm()	218
Unit 2 Project (5 Days)	31-35	End of Unit Project and Oral Presentations: Asking and Answering Statistical Questions of Our Own Data	Stress/Chill, Personality Color, Food Habits, or Time Use	Synthesis of above	220

^=Data collection window begins.
+=Data collection window ends.

IDS Unit 2: Essential Concepts

Lesson 1: What Is Your True Color?

Students will understand that the 'typical' value is a value that can represent the entire group, even though we know that not all members of the group share the same value.

Lesson 2: What Does Mean Mean?

The center of a distribution is the 'typical' value. One way of measuring the center is with the mean, which finds the balancing point of the distribution. The mean gives us the typical value, but does not tell the whole story. We need a way to measure the variability to understand how observations might differ from the typical value.

Lesson 3: Median In the Middle

Another measure of center is the median, which can also be used to represent the typical value of a distribution. The median is preferred for skewed distributions or when there are outliers, because it better matches what we think of as 'typical.'

Lesson 4: How Far Is It from Typical?

MAD measures the variability in a sample of data - the larger the value, the greater the variability. More precisely, the MAD is the typical distance of observations from the mean. There are other measures of spread as well, notably the standard deviation and the interquartile range (IQR).

Lesson 5: Human Boxplots

A common statistical question is “How does this group compare to that group?” This is a hard question to answer when the groups have lots of variability. One approach is to compare the centers, spreads, and shapes of the distributions. Boxplots are a useful way of comparing distributions from different groups when all of the distributions are unimodal (one hump).

Lesson 6: Face Off

Writing (and saying) precise comparisons between groups in which variability is present based on the (a) center, (b) spread, (c) shape, and (d) unusual outcomes help to make statements in context of the data. Actual comparison statements should use terms such as "less than," "about the same as," etc.

Lesson 7: Plot Match

Boxplots are an alternative visualization of histograms or dot plots. They capture most, but not all, of the features we can see in a dotplot or histogram.

Lesson 8: How Likely Is It?

Probability is an area about which we humans have poor intuition. Probability measures a long-run proportion: 50% chance means the event happens 50% of the time *if you repeated it forever*. When we don't repeat forever, we see variability.

Lesson 9: Bias Detective

In the short-term, actual outcomes of chance experiments vary from what is 'ideal.' An ideal die has equally likely outcomes. But that does not mean we will see exactly the same number of one-dots, two-dots, etc.

Lesson 10: Marbles, Marbles…

There are two ways of sampling data that model real-life sampling situations: with and without replacement. Larger samples tend to be closer to the "true" probability.

Lesson 11: This AND/OR That

What does "A or B" mean versus "A and B" mean? These are compound events and two-way tables can be used to calculate probabilities for them.

Lesson 12: Don’t Take My Stress Away!

Generating statistical questions is the first step in a Participatory Sensing campaign. Research and observations help create applicable campaign questions.

Lesson 13: The Horror Movie Shuffle

We can "shuffle" data based on categorical variables. The statistic we use is the difference in proportions. The distribution we form by shuffling represents what happens if chance were the only factor at play. If the actual observed difference in proportions is near the center of this shuffling distribution, then we would conclude that chance is a good explanation for the difference. But if it is extreme (in the tails or off the charts), then we should conclude that chance is NOT to blame. Sometimes, the apparent difference between groups is caused by chance.

Lesson 14: The Titanic Shuffle

We can also "shuffle" data based on numerical variables. The statistic we use is the difference in means. The distribution we form by this form of shuffling still represents what happens if chance were the only factor at play. When differences are small, we suspect that they might be due to chance. When differences are big, we suspect they might be 'real.'

Lesson 15: Tangible Data Merging

We can enhance the context of a statistical problem by merging related data sets together. To merge data, each data set must have a "unique identifier" that tells us how to match up the lines of the data.

Lesson 16: What Is Normal?

The Normal curve, also called the Gaussian distribution and the "bell curve," is a model that describes many real-life distributions and is usually called the Normal Model.

Lesson 17: A Normal Measure of Spread

The standard deviation is another measure of spread. This is commonly used by statisticians because of its role in common models and distributions, such as the Normal Model.

Lesson 18: Shuffling with Normal

Z-scores allow us a way to measure how extreme a value is, regardless of the units of measurement. Usually, z-scores will range between -3 and +3, and so values that are at or more extreme than -3 or +3 standard deviations are considered extremely rare.

What is Your True Color?

Instructional Days: 10

Enduring Understandings

Statistics enable us to make sense of large amounts of data. Numerical summaries capture important elements of a distribution. Measures of center, also known as measures of central tendency, show the tendency of quantitative data to gather around a central value. Measures of spread, also known as measures of variability, show how much the quantitative data is spread out. Measurements of the propensity for the data to cluster on a central location and the range of variability within the data can provide insightful indicators about the data.

Engagement

Students will complete the *True Colors Personality Test* to discover the qualities and characteristics of their personality styles. Students will use the results from the personality color test to learn about subsetting data and finding measures of center and spread. The data from their personality test will be collected in a survey using the IDS UCLA App or via web browser at https://portal.idsucla.org/

Learning Objective

Statistical/Mathematical:

S-ID 2: Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.

S-ID 3: Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

S-IC 6: Evaluate reports based on data.

Focus Standards for Mathematical Practice for All of Unit 2:

SMP-4: Model with mathematics.

SMP-5: Use appropriate tools strategically.

Data Science:

Understand the information that numerical summaries provide about the data. Understand that a boxplot is a graphical representation of a numerical summary.

Applied Computational Thinking Using RStudio:

- Calculate numerical summaries (mean, median, Sum of Absolute Deviations (SAD), and Mean of Absolute Deviations (MAD)).
- Create graphical representations to compare two or more data sets, including boxplots.

Real-World Connections:

We must be able to synthesize vast amounts of data into coherent, comprehensible measures. Today's media is continuously publishing articles that include statistical references. Critical consumerism requires that we understand the information provided in summaries of data.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.

2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

Data File or Data Collection Method

Data Collection Method:

1. **True Colors Personality Test**: Students will complete the *Personality Color* survey that will collect their data about their personality styles.

Data Files:

1. Students' *Personality Color* survey data `(colors)`

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 1: What is Your True Color?

Objective:

Students will collect data that might tell them about their personality type, and will understand how to subset their data.

Materials:

1. *True Colors Personality Test* (LMR_2.1_True Colors Personality Test)
2. Posted signs for each Personality Color: Blue, Gold, Green, and Orange
   Advanced preparation required (see step 5 below)
3. Poster paper
4. Markers
5. Data collection devices

Vocabulary:

subsets

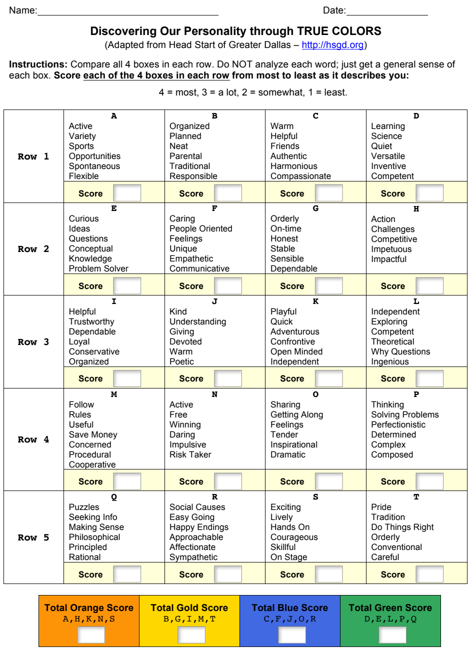
Essential Concepts: Students will understand that the 'typical' value is a value that can represent the entire group, even though we know that not all members of the group share the same value.

Lesson:

1. Ask students to consider the following questions (they do not need to record any responses):

   a. How well do you know yourself?
   b. How well do you know your classmates?

2. There are things students know and don't know about themselves. The *True Colors Personality Test* (LMR_2.1) claims to identify personality types (Later, students can gather more evidence to test these claims). Students will use these data to explore fundamental statistical concepts.



Name:______________________________ Date:______________

Discovering Our Personality through TRUE COLORS

(Adapted from Head Start of Greater Dallas – http://hsgd.org)

Instructions: Compare all 4 boxes in each row. Do NOT analyze each word; just get a general sense of each box. **Score each of the 4 boxes in each row from most to least as it describes you:**

4 = most, 3 = a lot, 2 = somewhat, 1 = least.

Row 1	**A** Active Variety Sports Opportunities Spontaneous Flexible	**B** Organized Planned Neat Parental Traditional Responsible	**C** Warm Helpful Friends Authentic Harmonious Compassionate	**D** Learning Science Quiet Versatile Inventive Competent
	Score □	**Score** □	**Score** □	**Score** □
Row 2	**E** Curious Ideas Questions Conceptual Knowledge Problem Solver	**F** Caring People Oriented Feelings Unique Empathetic Communicative	**G** Orderly On-time Honest Stable Sensible Dependable	**H** Action Challenges Competitive Impetuous Impactful
	Score □	**Score** □	**Score** □	**Score** □
Row 3	**I** Helpful Trustworthy Dependable Loyal Conservative Organized	**J** Kind Understanding Giving Devoted Warm Poetic	**K** Playful Quick Adventurous Confrontive Open Minded Independent	**L** Independent Exploring Competent Theoretical Why Questions Ingenious
	Score □	**Score** □	**Score** □	**Score** □
Row 4	**M** Follow Rules Useful Save Money Concerned Procedural Cooperative	**N** Active Free Winning Daring Impulsive Risk Taker	**O** Sharing Getting Along Feelings Tender Inspirational Dramatic	**P** Thinking Solving Problems Perfectionistic Determined Complex Composed
	Score □	**Score** □	**Score** □	**Score** □
Row 5	**Q** Puzzles Seeking Info Making Sense Philosophical Principled Rational	**R** Social Causes Easy Going Happy Endings Approachable Affectionate Sympathetic	**S** Exciting Lively Hands On Courageous Skillful On Stage	**T** Pride Tradition Do Things Right Orderly Conventional Careful
	Score □	**Score** □	**Score** □	**Score** □

Total Orange Score	**Total Gold Score**	**Total Blue Score**	**Total Green Score**
A,H,K,N,S	B,G,I,M,T	C,F,J,O,R	D,E,L,P,Q
□	□	□	□

LMR_2.1

3. Distribute the first 2 pages of the *True Colors Personality Test* (LMR_2.1). DO NOT include the final page, which contains the descriptions of each color. Instruct students on how to complete the test, and allow time for them to complete it (see page 2 in handout).

   Note: When adding scores for each color at the bottom of the test, make sure that students have **NOT** added straight down each column.

4. Students should have a score for each of the 4 colors. Ask students to record each color and its respective score in their IDS journal. Inform them that the color with the *highest* score describes their personality. We can refer to this as their predominant color. They should record their predominant personality color in their IDS journal as well. Tell students that you will show them what each color means at the end of the lesson.
5. Post a sign for each personality color on different walls of the classroom. For example, Blue on the north wall, Gold on the east wall, etc.
6. Ask students to gather by the wall corresponding to their predominant personality color. The students should record answers to the following questions in their IDS journals.
   a. How many students are in your color group?
   b. How many students are in each of the other color groups?
   c. What is the predominant personality color in your class?
7. Then ask students to determine some common characteristics of the people in their group. Questions to help steer the discussions are included below. Each team should come up with a consensus to describe their team and will share their descriptions with the whole class. The goal is to get the students to think about "what is typical?" within their groups.
   a. What are your likes and dislikes?
   b. What things do you have in common?
   c. What are your favorite activities?
   d. What's your favorite color?
   e. Do you prefer mornings or nights?
   f. What's your favorite type of music?
8. As the groups are presenting, record some dominant characteristics on the board for each color. The students will be able to compare their perceived traits with the actual descriptions from the activity at the end of the class.
9. Next ask students in each color group to gather into two **subsets**: introvert and extrovert. Inform them that subsetting is another way to organize collected data. Create a two-way frequency table like the one below on the board to record the results.

		Color				
		Orange	Gold	Blue	Green	**Total**
Introvert/ Extrovert	Introvert					
	Extrovert					
	Total					

10. Distribute poster paper and markers to each team.
11. Inform the students that they will be creating visuals for this data by comparing subsets. The Orange and Gold groups should create visualizations that subset the color variable by introverts and extroverts, and the Blue and Green groups should create visualizations that subset introverts and extroverts by color.
12. If students are confused or stuck, have them recall the topic of two-way tables and relative frequencies from Unit 1 (Lessons 16 & 17). The Orange and Gold groups will be looking at the columns and comparing introverts/extroverts, while the Blue and Green groups will be looking at the rows and comparing colors.

13. Once all groups have completed their visuals, the Orange and Gold teams should choose one of their 2 posters to display to the class. The Blue and Green groups should do the same and select one of their visuals.

14. Display both visuals on the board and discuss their similarities and differences. Ask students to analyze and interpret the visualizations by discussing the following questions for each of the visualizations:

    a. What type of plot is this and how many variables are present? ***Answers will vary by class.***
    b. What information about this subset can I gather from this visualization? ***Answers will vary by class.***
    c. What do I see the most/least of? ***Answers will vary by class.***
    d. What is the typical personality color for this subset? Or, what is the typical group (introverts/extroverts) for this subset? ***Answers will vary by class.***

15. Ask students to summarize their impressions of the class's personality color data by writing this summary in their IDS journals.

16. Distribute the description of each personality color to students (page 3 of LMR_2.1). Remind them that the highest score is considered their predominant color and the second highest score is considered their secondary color. If there is a tie for their predominant or secondary colors, ask students to choose the color that describes their personality better.

17. Compare the given descriptions on the handout to the characteristics listed on the board for each group during step 7. Do the descriptions match what the students originally thought? How accurate are the descriptions? If time allows, ask a couple of students to share their comparisons.

18. Students will now record their data by completing the *Personality Color* campaign on the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

If not finished in class, students should complete the *Personality Color* survey either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org/

Lesson 2: What Does Mean Mean?

Objective:

Students will learn that values that gather around the center of a distribution show the typical value. This value is also referred to as the mean, or average.

Materials:

1. *Pennies on a Ruler* handout (LMR_2.2_Pennies on a Ruler)
2. Markers (1 for each table)
3. Rulers (1 for each table)
4. Pennies (6 for each table group)
5. Tape

Digital Option:

IDS Balancing Point app
Balancing Point handout (LMR_2.2b)

6. Exported, printed, and reproduced class's *Personality Color* survey data
   Advanced preparation required: The teacher must share students' data on the IDS Home page (https://portal.idsucla.org) before it can be exported and printed. Students will keep for use in subsequent lessons.
7. *Mr. Jones Mile Run Times* handout (LMR_2.3_Mr. Jones Run Times)

Vocabulary:

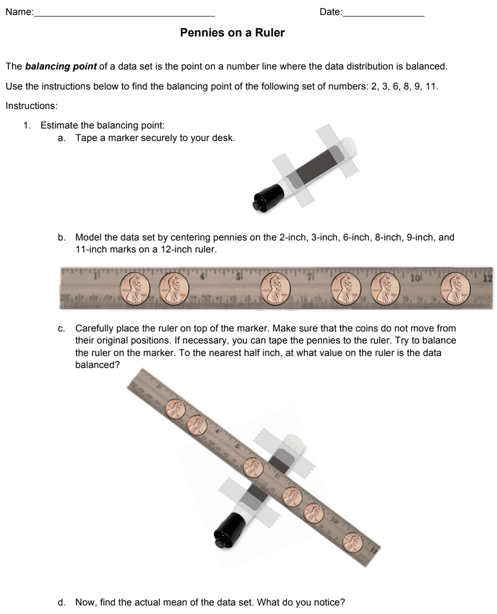
measures of central tendency (or center), typical, measures of variability (or spread), mean, average, balancing point

Essential Concepts: The center of a distribution is the 'typical' value. One way of measuring the center is with the mean, which finds the balancing point of the distribution. The mean gives us the typical value, but does not tell the whole story. We need a way to measure the variability to understand how observations might differ from the typical value.

Lesson:

1. In student pairs, ask students to discuss what they think the following terms mean:
   a. Measures of central tendency. ***A value that shows the tendency of quantitative data to gather around a central, or typical, value. Also known as measures of center. Students will learn about two such measures: the mean and the median.***
   b. Measures of variability. ***Values that show how much the quantitative data varies. Also known as measures of spread. Note: This is not taught during this lesson, but will be addressed as part of Lesson 4.***
2. Ask a pair to share what they think these two terms mean. Pairs who are listening must decide whether they agree or disagree with the pair that shared. Lead a discussion based on their statements of agreement or disagreement.
3. Communicate to the class that they will be learning more about these measures and what they tell us about data as we progress through this unit.
4. By a show of hands, ask students how many are familiar with finding the **mean**, or **average**.
5. Select a student to share his/her process for finding the mean. ***Possible answer: Add up all of the numbers. Then divide by how many numbers there are.***
6. Another way to find the mean is to find the **balancing point** of a distribution. They will learn about the balancing point via the activity in Steps 7 & 8.

7. Distribute the *Pennies on a Ruler* handout (LMR_2.2) along with a marker, ruler, tape, and 6 pennies to each table group. If you prefer to not print the document, you can project it on the board instead.



Name:______________________________ Date:______________

Pennies on a Ruler

The ***balancing point*** of a data set is the point on a number line where the data distribution is balanced.

Use the instructions below to find the balancing point of the following set of numbers: 2, 3, 6, 8, 9, 11.

Instructions:

1. Estimate the balancing point:
   a. Tape a marker securely to your desk.
   b. Model the data set by centering pennies on the 2-inch, 3-inch, 6-inch, 8-inch, 9-inch, and 11-inch marks on a 12-inch ruler.
   c. Carefully place the ruler on top of the marker. Make sure that the coins do not move from their original positions. If necessary, you can tape the pennies to the ruler. Try to balance the ruler on the marker. To the nearest half inch, at what value on the ruler is the data balanced?
   d. Now, find the actual mean of the data set. What do you notice?

LMR_2.2

8. Guide the students through the handout and have them share their findings throughout the activity. Be sure to emphasize the idea that the mean of a distribution can be identified by finding its balancing point.

9. Next, distribute the class's *Personality Color* survey data to the students.

10. Have student pairs find the variable ***Blue*** (whether or not that was their predominant color) in the class's printed data.

11. As a class, make a dot plot on the board to show the distribution of ***Blue*** values. Each student should come to the board and draw a dot to indicate where their value is in the distribution. Ask the students:
    a. What do you think the typical ***Blue*** score is? ***Answers will vary by class. They should be driven to an answer in the center of the distribution.***
    b. Are the data roughly symmetric? Where is the balancing point of this distribution? ***Answers will vary by class. Once a value is chosen, indicate the location on the dot plots.***

12. As a class, compute the mean ***Blue*** score for the entire class on the board and compare this value to the class's prediction of the balancing point. Students may not remember exactly how to compute the mean, so you can remind them of the general algorithm or refer them back to their responses from Step 5 above.

13. Show the students the formula for calculating the mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

14. Now that they have calculated the mean for the ***Blue*** score, ask them to identify each symbol in the formula with a step in their algorithm for finding the mean, and discuss the meaning of the

symbols in the formula as a class. x_i ***represents each individual data point and*** n ***represents the total number of observations.***

15. Indicate the location of the calculated mean on the dot plots by drawing a vertical line at the value on the x-axis. Ask student pairs to engage in a conversation about how close the mean value is to their predicted balancing point and why their prediction was made that way. Select a pair to share their discussion with the whole class.

16. Using the *Personality Color* survey data from Step 9, ask student pairs to compute the mean score for each of the other three personality colors.

17. Inform the students that, during the next lesson, they will learn about another method that can be used for measuring the center of a distribution.

18. Now, you can inform the class about an even easier method of calculating the mean – using RStudio! Explain that the command RStudio uses to calculate the mean incorporates the algorithm of summing up all the data and dividing by the total number of observations. Students will be able to use this command for quick calculations now.

    Note: If you have already *"Exported, Downloaded, Imported"* the class's *Personality Color* campaign data, you can simply use the exact command below to calculate the mean ***Blue*** score:

    ```
    > mean(~blue, data = colors)
    ```

    In general, the function can be denoted as follows:

    ```
    > mean(~variable, data = datafile)
    ```

    So, for our specific example, **`blue`** is the **`variable`** we want to find the mean value of, and **`colors`** is the **`datafile`**.

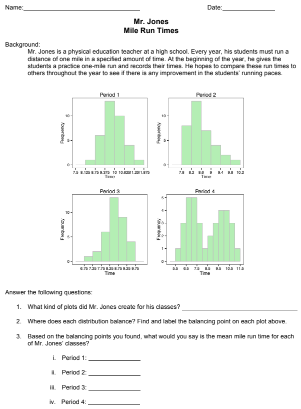
19. Have the students *Think-Pair-Share* to discuss how the mean value of a group of data could be used to easily describe complicated things. For example, instead of giving someone the entire class's ***Blue*** scores, we could just tell him/her the mean score and he/she would have a general idea about the class.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students should complete the *Mr. Jones Mile Rule Times* handout (LMR_2.3) for homework. They can practice finding the mean of distributions by determining a balancing point for the data. Answers to the handout are below. **Note:** The mean values in part (3) do NOT need to be exact.



Name:______________________________ Date:_____________

Mr. Jones
Mile Run Times

Background:

Mr. Jones is a physical education teacher at a high school. Every year, his students must run a distance of one mile in a specified amount of time. At the beginning of the year, he gives the students a practice one-mile run and records their times. He hopes to compare these run times to others throughout the year to see if there is any improvement in the students' running paces.

Answer the following questions:

1. What kind of plots did Mr. Jones create for his classes? ______________________________
2. Where does each distribution balance? Find and label the balancing point on each plot above.
3. Based on the balancing points you found, what would you say is the mean mile run time for each of Mr. Jones' classes?
    i. Period 1: _____________
    ii. Period 2: _____________
    iii. Period 3: _____________
    iv. Period 4: _____________

LMR_2.3

1. What kind of plots did Mr. Jones create for his classes? ***Histograms.***
2. Where does each distribution balance? Find and label the balancing point of each distribution. ***The balancing point for all of these distributions is at the mean.***
3. Based on the balancing points you found, what would you say the mean mile run time is for each class?

    i. Period 1: ***9.91***
    ii. Period 2: ***8.48***
    iii. Period 3: ***8.45***
    iv. Period 4: ***8.17***

Lesson 3: Median in the Middle

Objective:

Students will learn that the median is another way to measure the center, or typical-ness, of a distribution, and will understand how medians compare and contrast with the mean.

Materials:

1. Sticky notes (one per student)
   Advanced preparation required (see step 6 below)
2. Poster paper
3. Graphics from *Medians – Dotplots or Histograms?* (LMR_2.4_Medians – Dotplots or Histograms)
4. *Where is the Middle?* handout (LMR_2.5_Where is the Middle)
5. Exported, printed, and reproduced class's *Personality Color* survey data

Vocabulary:

median

Essential Concepts: Another measure of center is the median, which can also be used to represent the typical value of a distribution. The median is preferred for skewed distributions or when there are outliers because it better matches what we think of as 'typical.'

Lesson:

1. Remind students that, during the previous lesson, they learned about the mean as the balancing point of a distribution and as a measure of center. In statistics, there are a few values that can be considered as measures of center – the mean is one, and another is the **median**. The median is the middle value in a group of ordered observations.
2. As a simple example, write or display the following group of numbers on the board:

   8, 2, 6, 3, 7, 4, 9, 5, 5
3. Since there are 9 numbers in the list above, we should use the 5th number as the median because it is directly in the middle and there are 4 numbers above it, and 4 numbers below it.
4. However, students should realize that they cannot simply pick the middle number of the list as it is currently written (this would give a median value of 7). Instead, they must first arrange the numbers in numerical order (from lowest to highest).

   2, 3, 4, 5, 5, 6, 7, 8, 9
5. Now they can identify that the true median value of this list of numbers is 5.
6. Next, randomly distribute one sticky note to each student.

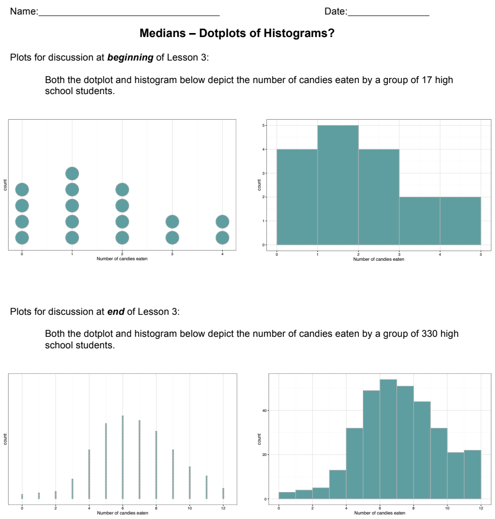
   Advanced preparation required: There should be one card for every student in the class. All of the cards, except one, need to have the value 0 written on them. One card should have the value 1,000,000 written on it.
7. Place poster paper on the board and have the students create a dot plot by placing their sticky notes at the corresponding values on the axis. Then, ask and record answers to the following questions:
   a. What is the typical value of these data? ***0 – all sticky notes but one have a value of 0.***
   b. Using the formula we learned in class, calculate the mean, or average, value of this distribution. ***Answer will vary by class/class size. Example: for a class with 28 students enrolled, there would be 27 values of 0 and 1 value of 1,000,000. Therefore, the mean value would be (0*27 + 1,000,000)/28 ≈ 35,714.3.***
   c. Does the mean you calculated match your understanding of "typical?" Why is the mean not capturing our notion of "typical?" ***The 1,000,000 value is heavily skewing the***

calculation of the mean. It is pulling the mean to a higher value than what we consider to be typical for these data.

8. Since we introduced the idea of the median as a measure of center at the beginning of class, have the students find the median value of the data on their sticky notes. If time permits, have them place the sticky notes in a line across the board in order (from least to greatest) and have them find the middle number. The median value will be 0.

9. Ask students why there is such a large difference between the mean and median values even though they are both measures of center? Is there a specific reason why the mean is larger than the median for this particular set of data? ***In this case, there was an outlier value that skewed the distribution and forced the balancing point to move to the right.***

10. Display the first 2 plots in the *Medians – Dotplots or Histograms?* file (LMR_2.4). They are labeled as plots for discussion for the *beginning* of class. Both the dotplot and histogram depict the number of candies eaten by a group of 17 high school students.



Name:______________________________ Date:______________

Medians – Dotplots of Histograms?

Plots for discussion at ***beginning*** of Lesson 3:

Both the dotplot and histogram below depict the number of candies eaten by a group of 17 high school students.

Plots for discussion at ***end*** of Lesson 3:

Both the dotplot and histogram below depict the number of candies eaten by a group of 330 high school students.

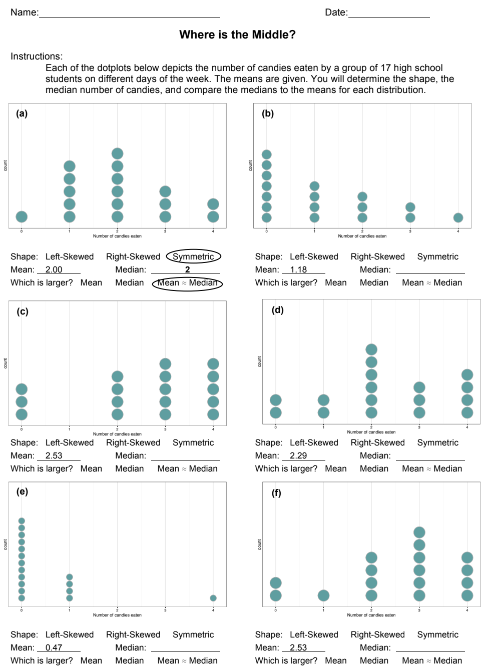
LMR_2.4

11. For the first 2 plots, ask students:

    a. Which plot makes it easier to find the median number of candies eaten – the dot plot or the histogram? Why? ***The dot plot is easier because we can simply find the middle dot and record the value. It is harder on the histogram, because we would have to add up amount in each bar to find the middle person.***
    b. What is the median value? ***The median number of candies eaten is 1 candy.***

12. Inform the students that they will practice finding medians of distributions using the *Where is the Middle?* handout (LMR_2.5). They will be determining medians when distributions have different shapes (e.g., symmetric, left-skewed, right-skewed).

13. Distribute the *Where is the Middle?* handout (LMR_2.5). Students should complete the handout individually first, then compare answers with their team members. Once each team has agreed upon their answers, discuss the handout as a class.



Name:______________________________ Date:______________

Where is the Middle?

Instructions:

Each of the dotplots below depicts the number of candies eaten by a group of 17 high school students on different days of the week. The means are given. You will determine the shape, the median number of candies, and compare the medians to the means for each distribution.

(a)

Shape: Left-Skewed Right-Skewed [✓ Symmetric]
Mean: 2.00 Median: 2
Which is larger? Mean Median [✓ Mean ≈ Median]

(b)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 1.18 Median: ____________
Which is larger? Mean Median Mean ≈ Median

(c)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 2.53 Median: ____________
Which is larger? Mean Median Mean ≈ Median

(d)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 2.29 Median: ____________
Which is larger? Mean Median Mean ≈ Median

(e)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 0.47 Median: ____________
Which is larger? Mean Median Mean ≈ Median

(f)

Shape: Left-Skewed Right-Skewed Symmetric
Mean: 2.53 Median: ____________
Which is larger? Mean Median Mean ≈ Median

LMR_2.5

14. Ask the following questions to elicit a team discussion about the relationship between means and medians:

    a. What did you notice about the relationship between the mean and median values for the symmetric distributions? ***The mean and median values in the symmetric distributions - plots (a) and (d) - are fairly similar. For plot (a), the mean and median are exactly equal. For plot (d), the mean is actually larger than the median, but not by much (2.29 > 2).***
    b. What did you notice about the relationship between the mean and median values for the left-skewed distributions? ***The mean value was smaller than the median value in both of the left-skewed distributions - plots (c) and (f). Both plots had the same values for the mean (2.53) and the median (3.00) - clearly, the mean is much smaller than the median (2.53 < 3).***
    c. What did you notice about the relationship between the mean and median values for the right-skewed distributions? ***The mean value was larger than the median value in both of the right-skewed distributions - plots (b) and (e). For plot (b), the mean was only slightly higher than the median (1.18 > 1). For plot (e), the mean was a decent amount higher than the median (0.47 > 0).***

15. Steer the discussion towards the relationship between the shape of a distribution and its corresponding mean and median values.

    a. Is there a pattern that emerges between the mean and median values for differently shaped distributions? ***Yes! It seems that symmetric distributions will produce similar mean and median values, left-skewed distributions will produce smaller means and higher medians, and right-skewed distributions will produce higher means and smaller medians.***
    b. For each of the plots in the *Where is the Middle?* handout (LMR_2.5), which value better matches your idea of "typical" for that specific distribution? ***For plot (a), both the mean and median agree and appear to be the balancing point of the distribution – both match what we think is typical. For plot (b), the median seems to be more typical, but the values are very close. For plot (c), the median appears to be a more typical value. For plot (d), both the mean and median appear to be capturing our idea of***

typical. For plot (e), the median is a better match to typical. For plot (f), the median is also a better match.

16. Steer the discussion so that students recognize that the better measures of center for skewed distributions are typically medians, and the better measures for center for symmetric distributions are typically means.
17. Display the last 2 plots in the *Medians – Dotplots or Histograms* file (LMR_2.4). They are labeled as plots for discussion for the *end* of class. Both the dot plot and histogram depict the number of candies eaten by a group of 330 high school students.
18. For the last 2 plots, ask students:
    a. Which plot makes it easier to find the median number of candies eaten – the dot plot or the histogram? Why? ***The histogram is easier because we can estimate based on the distribution's shape. There are too many dots in the dot plot to find the exact middle person.***
    b. What is the median value? ***The median number of candies eaten is 7 candies.***

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students should calculate the median values for each of their personality color scores. They should compare the median values to the mean values (calculated in Lesson 2) and make a decision about the possible shape of the distribution if we were to create a dot plot of the scores.

Lesson 4: How Far is it from Typical?

Objective:

Students will understand that the mean of the absolute deviations (MAD) is a way to assess the degree of variation in the data from the mean and adjusts for differences in the number of points in the data set (*n*). The MAD measures the total distance between all the data values from the mean and divides it by the number of observations in the data set.

Materials:

1. Masking tape (or painter's tape) – approximately 4-5 feet long – one for each student team
2. *How Far Apart?* handout (LMR_2.6_How Far Apart) – will be used again in Lesson 17
3. Exported, printed, and reproduced class's *Personality Color* survey data

Vocabulary:

measures of variability (spread), deviation, mean of absolute deviations (MAD)

Essential Concepts: MAD measures the variability in a sample of data - the larger the value, the greater the variability. More precisely, the MAD is the typical distance of observations from the mean. There are other measures of spread as well, notably the standard deviation and the interquartile range (IQR).

Lesson:

1. Remind students that they learned about 2 different measures of center during the previous 2 lessons: the mean and the median. Have the students recall when it is appropriate to use each value based on the shape of the distribution.

    a. Mean – use with symmetric distributions.
    b. Median – use with skewed distributions or when there are outliers.

2. Inform the students that, during today's lesson, they will learn about **measures of variability** – also known as measures of **spread**. These values show us how much the quantitative data varies from the center of a distribution. Similar to measures of center, we will use two different measures of spread: (1) the mean of absolute deviations (MAD), and (2) the interquartile range (IQR).

    Note: IQR will be discussed in detail during Lesson 5.

3. Introduce the term **deviation**. Using *Think, Pair, Share*, ask students what they think this word means and how it could relate to variability. ***A deviation is the act of departing from an established course or accepted standard. Common synonyms include departure, detour, difference, digression, divergence, fluctuation, inconsistency, modification, shift, etc.***

4. On the classroom floor next to each student team, place a 4-5 foot long piece of masking tape (or painter's tape). Then, propose the following scenario:

    > Your team has been invited to guest star at the circus! You have been asked to perform as part of the tightrope act – a routine that requires tremendous focus and balance to walk across a tightly pulled rope that is suspended high in the air. In order to practice your balancing skills, the circus has provided your team with a line of tape that will represent the tightrope.

5. Have the students consider the piece of tape (aka the rope) to be the "typical" path they must take to finish the circus act. Since they do not want to fall from the suspended tightrope while performing at the actual circus, they will need to practice walking directly on the middle of the line at all times. If they *deviate* from the line, they will no longer be walking the "typical" path, and will likely fall.

6. Each team should select one student to be their starting performer.

7. In teams of 4, one student is the performer, two are measuring the distance of the deviation (one on each side of the tape), and one is the recorder.

8. Place a ruler perpendicular to the “rope” and measure the distance, in centimeters, from the path to the center of the back of their heel as the student walks and attempts to balance across the “rope.”

9. The performer will walk the tightrope by looking straight up to the sky – first they look to place a foot on the line, then walk naturally while looking up to the sky, and repeating one step at a time for 4 steps, measuring after each step. Any time the performer missteps, this is considered a variation from the typical value. *You can have students take turns so everyone gets a chance to balance, walk, and to measure, depending on time in your class.*

10. Now that the students have an idea about what it means to deviate from something they consider “typical,” they can start looking at distributions to see how data points vary from their typical value.

11. Inform students that they were observing deviations from typical while calculating actual differences between the rope and the performer’s steps. When data are quantified with numbers, we can then calculate how far away each value is from the center.

12. One such calculation that is popular among data scientists is the mean of absolute deviations (MAD). Ask students to consider the components of the MAD in math terms, and brainstorm what the MAD value might represent.

    mean – an average

    absolute – in mathematics, we talk about absolute value, the positive difference between 2 numerical values

    deviation – as discussed earlier in the lesson, deviation represents how much things vary

13. Using the 3 components in Step 12, explain that the MAD measures the absolute distance of each data point from the mean, and then finds the average of all those distances.

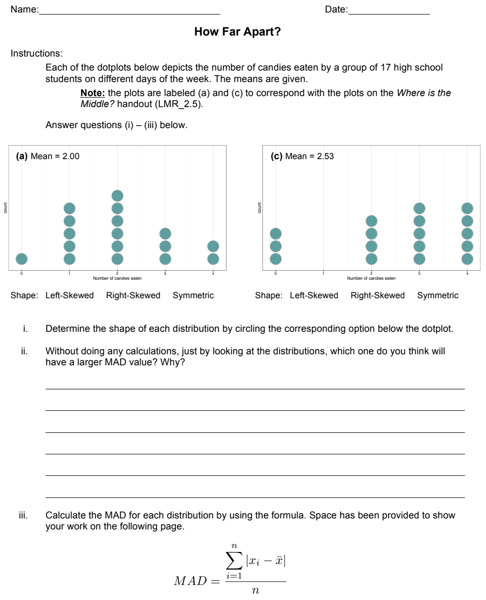
14. Display the formula for the MAD distribution for the whole class to see.

$$MAD = \frac{\displaystyle\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

15. Discuss what each symbol in the formula means and how we use it to perform the calculation. x_i ***represents each individual data point,*** $\bar{x}$ ***represents the mean value, and*** n ***represents the total number of observations. The*** $\sum$ ***symbol represents the summation – this tells us to add up all the absolute distances from each point to the mean.***

16. To practice using this formula with actual data, students will calculate and compare the MAD values for 2 distributions.

17. Distribute the *How Far Apart?* handout (LMR_2.6), which contains 2 of the dot plots - plots (a) and (c) from the *Where is the Middle?* handout (LMR_2.5) used in Lesson 3. As before, the dot plots depict the number of candies eaten by a group of 17 high school students on different days of the week. The means are also given.



Name:________________________________ Date:_______________

How Far Apart?

Instructions:

Each of the dotplots below depicts the number of candies eaten by a group of 17 high school students on different days of the week. The means are given.

<u>Note:</u> the plots are labeled (a) and (c) to correspond with the plots on the *Where is the Middle?* handout (LMR_2.5).

Answer questions (i) – (iii) below.

(a) Mean = 2.00

Shape: Left-Skewed Right-Skewed Symmetric

(c) Mean = 2.53

Shape: Left-Skewed Right-Skewed Symmetric

i. Determine the shape of each distribution by circling the corresponding option below the dotplot.

ii. Without doing any calculations, just by looking at the distributions, which one do you think will have a larger MAD value? Why?

____

____

____

____

____

____

iii. Calculate the MAD for each distribution by using the formula. Space has been provided to show your work on the following page.

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

LMR_2.6

The calculations for each plot are shown below for the teacher's reference.

<u>MAD for plot (a)</u>

$$\begin{aligned}
MAD &= \frac{1\,|0-2| + 5\,|1-2| + 6\,|2-2| + 3\,|3-2| + 2\,|4-2|}{17} \\
&= \frac{1(2) + 5(1) + 6(0) + 3(1) + 2(2)}{17} \\
&= \frac{2 + 5 + 0 + 3 + 4}{17} \\
&= \frac{14}{17} \\
&\approx 0.8235
\end{aligned}$$

<u>MAD for plot (c)</u>

$$MAD = \frac{3|0-2.53| + 0|1-2.53| + 4|2-2.53| + 5|3-2.53| + 5|4-2.53|}{17}$$

$$= \frac{3(2.53) + 0(1.53) + 4(0.53) + 5(0.47) + 5(1.47)}{17}$$

$$= \frac{7.59 + 0 + 2.12 + 2.35 + 7.35}{17}$$

$$= \frac{19.41}{17}$$

$$\approx 1.1418$$

18. Students may work in pairs to complete the handout. After all student pairs have come to an agreement on their answers, pose the following questions to the class as a whole:

    a. Which MAD value did you think would be larger based only on the look/shape of the distributions? Why? ***Since plot (c) is skewed to the left, it probably has a larger MAD because more points will be further away from the mean than in plot (a).***
    b. Which MAD value was actually larger when you calculated it? ***The MAD value for plot (c) was larger (1.1418 > 0.8253).***
    c. Did your prediction match the actual calculated values, or were you surprised by the results? ***Yes. The distribution with the wider spread (more variability) had the larger MAD value.***

19. To continue exploring with the class's Personality Color survey data, student teams should calculate the MAD value for their ***Blue*** scores. Does the MAD value seem reasonable based on the dot plot they created during Lesson 2?

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

☑ Students should calculate the MAD values for each of the other 3 personality color scores and compare the values of the 4 color scores.

LAB 2A: All About Distributions

Complete Lab 2A prior to Lesson 5.

Lab 2A - All About Distributions

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

In the beginning...

- Most of the labs thus far have covered how to visualize, summarize, and manipulate data.
  - We used visualizations to explore how your class spends their time.
  - We also learned how to clean data to prepare it for analyzing.
- Starting with this lab, we'll learn to use R to answer statistical questions that can be answered by calculating the mean, median and MAD.

How to talk about data

- When we make plots of our data, we usually want to know:
- Where is the *bulk* of the data?
- Where is the data more *sparse*, or *thin*?
- What values are *typical*?
- How much does the data *vary*?
- To answer these questions, we want to look at the *distribution* of our data.
  - We describe *distributions* by talking about where the *center* of the data are, how *spread* out the data are, and what sort of *shape* the data has.

Let's begin!

- *Export*, *upload* and *import* your class' *Personality Color* data.
  - Name your data `colors` when you load it.
- Before analyzing a new data set, it's often helpful to get familiar with it. So:
  - **Write down the names of the 4 variables that contain the point-totals, or *scores*, for each personality color.**
  - **Write down the names of the variables that tell us an observation's *introvert/extrovert designation* and whether they participated in playing *sports*.**
  - **How many variables are in the data set?**
  - **How many observations are in the data set?**

Estimating centers

- Create a `dotPlot` of the scores for your *predominant color*.
  - Pro-tip: If the `dotPlot` comes out looking wonky, include the `nint` and `cex` options.
- Based on your `dotPlot`:
  - **Which values came up the most frequently? About how many people in your class had a score similar to yours?**
  - **What, would you say, was a *typical* score for a person in your class for your predominant color? How does your own score for this color compare?**

Means and medians

- *Means* and *medians* are usually good ways to describe the *typical* value of our data.
- Fill in the blank to calculate the `mean` value of your predominant color score:

```
    mean(~____, data = colors)
```

- **Use a similar line of code to calculate the `median` value of *your* predominant color.**
  - **Are the `mean` and `median` roughly the same? If not, use the `dotPlot` you made in the last slide to describe why.**

Estimating Spread

- Now that we know how to describe our data's *typical* value we might also like to describe how closely the rest of the data are to this *typical* value.
  - We often refer to this as the **variability** of the data.
  - Variability is seen in a `histogram` or `dotPlot` as the horizontal *spread*.
- Re-create a `dotPlot` of the scores for your predominant color and then run the code below filling in the blank with the name of your predominant color.

```
add_line(vline = mean(~____, data = colors))
```

- **Look at the spread of the scores from the mean score then complete the sentence below:**

Data points in my plot will usually fall within ____ units of the center.

Mean Absolute Deviation

- The **mean absolute deviation** finds how far away, on average, the data are from the mean.
  - We often write *mean absolute deviation* as *MAD*.
- Calculate the MAD of your *predominant color* by filling in the blanks:

```
MAD(~_____, data = colors)
```

- **How close was your estimate of the spread for your predominant color (from the previous slide) to the actual value?**

Comparing introverts/extroverts

- Do introverts and extroverts differ in their typical scores for your predominant color?
  - Answer this investigative question using a dotPlot and numerical summaries.
- Make a `dotPlot` of your predominant color again; but this time, facet the plot by the introvert/extrovert variable. Include the `layout` option to stack the plots as well as the `nint` and `cex` options.
- **Describe the shape of the distribution of scores for the extroverts. Do the same for the introverts.**
- Using similar syntax to how you facet plots, calculate either the `mean` or `median` to describe the center of your predominant color for introverts and extroverts.
- Do introverts and extroverts differ in their typical scores for your predominant color?
- Based on the MAD, which group (introverts or extroverts) has more variability for your predominant color's scores?

On your own

- Do introverts and extroverts in your class differ in their color scores?
  - **Perform an analysis that produces *numerical summaries* and *graphs*.**
  - **Then, write a few sentences that address this statistical question and considers the *shape*, *center* and *spread* of the distributions of the graphs you create.**

Lesson 5: Human Boxplots

Objective:

Students will learn how and when to use boxplots to compare groups of data. They will learn how to compute and interpret another measure of spread: the IQR.

Materials:

1. Poster paper, 3-4 feet long
   Advanced preparation required (see Step 9 below)
2. Tape
3. Poster paper
4. Markers
5. *Ages of Oscar Winners* handout (LMR_2.7_Oscar Ages)

Vocabulary:

boxplot, quartiles, first quartile (Q_1), third quartile (Q_3), quantiles, minimum, maximum, five-number summary, range, interquartile range (IQR)

Essential Concepts: A common statistical question is "How does this group compare to that group?" This is a hard question to answer when the groups have lots of variability. One approach is to compare the centers, spreads, and shapes of the distributions. Boxplots are a useful way of comparing distributions from different groups when all of the distributions are unimodal (one hump).

Lesson:

1. Remind students that we have been using the following numerical and graphical summaries to look at data:
   a. Measures of center – mean, median
   b. Measures of spread – MAD
   c. Graphing – dotPlots, histograms
2. Explain that all of these tools help us describe data to someone who may not actually be viewing it. Today, we will explore another way to summarize and describe data to others with the use of another type of statistical plot that involves breaking data up into distinct pieces: a **boxplot**.
3. For the next activity, students will need to carry their IDS journals and a pen with them.
4. Instruct students to stand up and move their chairs away from the longest wall in the classroom. Ask them to line up against the wall (in no particular order).

   Note: If there isn't enough room for everyone to line up together inside the classroom, you may do this activity outside along a building wall.
5. Say, "I want to know which person represents the typical height of students in our class. Can I tell by looking at the line as it currently stands? How would I be able to tell?" Students should discuss with a partner.
6. Ask students to share their discussions. Call on students to contribute to what has been shared if needed. Guide students to see that organizing the data (in other words, themselves) can give you a visual for their heights. Then tell them to line up in height order from shortest to tallest along the wall.
7. Once students are arranged (and this may take a little time—allow students to develop their own algorithm for finding the ordering), ask them how they might be able to describe their distribution of heights. ***Possible answers include: mean, median, MAD.***
8. Ask them to split themselves into two groups, one half that is taller and one half that is shorter, and have them decide which student represents the class's median height.

9. Have the median student stand next to the wall directly in front of the poster paper.

   Advanced preparation required: Before class begins, tape a piece of poster paper, approximately 3-4 feet long, vertically to a wall in the classroom. The students will be creating a plot using lines drawn at certain students' heights.

10. Draw a horizontal line on the poster paper to mark the location of the median by having the actual student stand in front of the poster paper so you can mark his/her exact height. Be sure to label this point as the median and include the student's actual height, in inches.

11. Next, ask the two halves to split again, so there are now four groups of students.

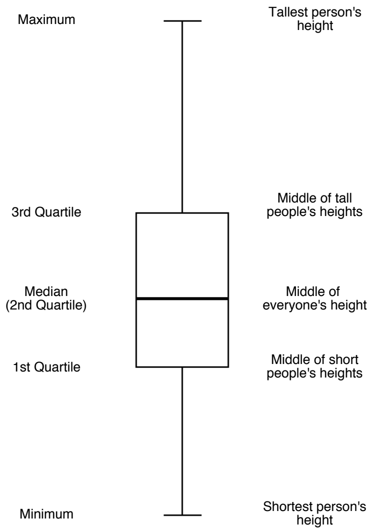
12. The breaks between each group are called **quartiles** because they break the data into four groups (*quartile* comes from the Latin word *quartus*, which is also the root of the Spanish word *cuatro*). The lower break represents the **first quartile** (because 25% of the class is shorter than this student's height), and the upper break represents the **third quartile** (because 75% of the class is shorter than this height). Another term that can be used in place of percentiles is **quantiles** because they represent the *quantity* of data that is lower than that value.

13. Using the student who represents the first quartile, draw another horizontal line on the poster paper marking his/her height. The student should stand in the same spot as the student who represented the median so that the line for this student is drawn underneath the median line. Be sure to label this point as the first quartile (or $\mathbf{Q_1}$) and include the student's actual height, in inches.

14. Using the student who represents the third quartile, draw another horizontal line on the poster paper marking his/her height. The student should stand in the same spot as the student who represented the median so that the line for this student is drawn above the median line. Be sure to label this point as the third quartile (or $\mathbf{Q_3}$) and include the student's actual height, in inches.

15. Finally, ask the tallest and shortest student to stand in front of the poster paper and draw horizontal lines at their heights. The shortest person represents the **minimum** height of the students in the class, and the tallest person represents the **maximum** height. Be sure to label the points as the minimum and maximum, and include the students' actual heights, in inches.

16. When you finish, you should have five lines, which represent the **five-number summary**: minimum, first quartile, median, third quartile, and maximum. Draw a box using the first and third quartiles as the edges of the box. The median line will be contained within the box. Extend a line from the first quartile down to the minimum and extend a line from the third quartile up to the maximum. Your class's boxplot should look similar to the following:



17. Students should now be facing the newly created boxplot. Allow students time to sketch the boxplot in their IDS journals, with the appropriate labels.

18. Ask students:

    a. What is the difference between the largest and smallest heights? Is there a large difference between the tallest and shortest person? ***Students should calculate maximum – minimum.*** Inform students that this difference is known as the **range** of the data set.

    b. What is the difference between the quartiles Q_1 and Q_3? What percent of our class falls within these two values? ***Students should calculate Q_3 – Q_1. 50% of the class falls between these two height values.*** Inform students that this difference is known as the **interquartile range (or IQR)**.

19. Remind students that they learned about one measure of spread (the MAD) during the previous lesson, and tell them that we now have another measure of spread – the IQR. Pose the following questions to the students:

    a. What does it mean when the IQR is small? ***The middle 50% of heights are close to each other.***

    b. What does it mean when the IQR is large? ***The middle 50% of heights are more spread out.***

20. Finally, subset the class into introverts and extroverts. Ask each group of students (the introverts and extroverts) to create a boxplot of their group's heights on a piece of poster paper using the techniques they just learned as a class.

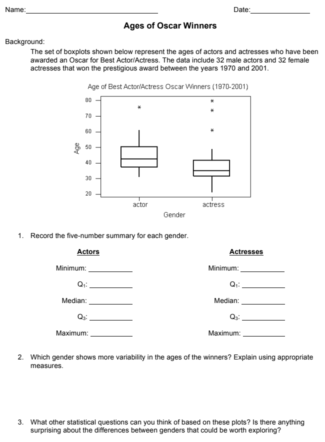
21. Ask each group to share their boxplot with the class. Lead a discussion about the similarities and differences between the plots, and be sure to include how they compare to the overall combined boxplot of heights they created earlier. In the discussion, have the students calculate the IQR for both plots and make a comparison by asking: What does the IQR tell us about each group? ***Answers will vary by class.***

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students should complete the *Ages of Oscar Winners* handout (LMR_2.7) for homework using their newly acquired knowledge of boxplots.



Name:________________________ Date:____________

Ages of Oscar Winners

Background:

The set of boxplots shown below represent the ages of actors and actresses who have been awarded an Oscar for Best Actor/Actress. The data include 32 male actors and 32 female actresses that won the prestigious award between the years 1970 and 2001.

1. Record the five-number summary for each gender.

Actors	**Actresses**
Minimum: ________	Minimum: ________
Q_1: ________	Q_1: ________
Median: ________	Median: ________
Q_3: ________	Q_3: ________
Maximum: ________	Maximum: ________

2. Which gender shows more variability in the ages of the winners? Explain using appropriate measures.

3. What other statistical questions can you think of based on these plots? Is there anything surprising about the differences between genders that could be worth exploring?

LMR_2.7

Lesson 6: Face Off

Objective:

Students will informally compare two or more distributions using their knowledge of shape, center, and spread to answer statistical questions. They will learn how to find the difference between two means and two medians using a histogram or dotplot.

Materials:

1. *Comparing Commute Times with Dotplots* handout (LMR_2.8_Commute Times – Dotplots)
2. *Comparing Exam Scores with Histograms* handout (LMR_2.9_Exam Scores – Histograms)
3. Timer
4. *Comparing Fuel Efficiency with Boxplots* handout (LMR_2.10_Fuel Efficiency – Boxplots)

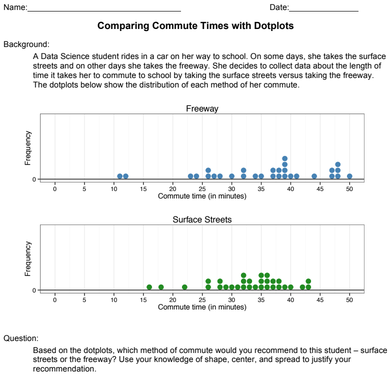
Vocabulary:

rebuttal

Essential Concepts: Writing (and saying) precise comparisons between groups in which variability is present based on the (a) center, (b) spread, (c) shape, and (d) unusual outcomes help to make statements in context of the data. Actual comparison statements should use terms such as "less than," "about the same as," etc.

Lesson:

1. Poll students about the method of transportation they use for their daily school commute. How many of them walk, ride in a car, take the bus, ride a bike, etc.? Record their responses on the board. Ask them to estimate the typical amount of time it takes for them to get to school, in minutes.
2. Inform students that they have learned important features of distributions that will allow them to make decisions when working with data. More specifically, they will be able to use their knowledge of measures of center and measures of spread to compare 2 distributions in order to make a decision.
3. In teams, have students complete the *Comparing Commute Times with Dotplots* handout (LMR_2.8). Allow students time to read the “Background” portion of the handout, and then discuss what statistical question(s) the student in the scenario is trying to answer.



Name:______________________________ Date:______________

Comparing Commute Times with Dotplots

Background:

A Data Science student rides in a car on her way to school. On some days, she takes the surface streets and on other days she takes the freeway. She decides to collect data about the length of time it takes her to commute to school by taking the surface streets versus taking the freeway. The dotplots below show the distribution of each method of her commute.

Question:

Based on the dotplots, which method of commute would you recommend to this student – surface streets or the freeway? Use your knowledge of shape, center, and spread to justify your recommendation.

LMR_2.8

4. Once teams decide on their recommendation, engage half of the class in an *Active Debate*. Half of the students will stand in a debate line and the other half will “fishbowl” the debate. Roles will reverse later in the lesson (see Step14).
5. Of those students standing on the debate line, half will argue the reasons why they recommend street travel and the other half will argue the reasons why they recommend freeway travel.

6. On the debate line, each student will stand face to face with a student who has the opposite recommendation. In other words, a student who recommends street travel will stand facing a student who recommends freeway travel.

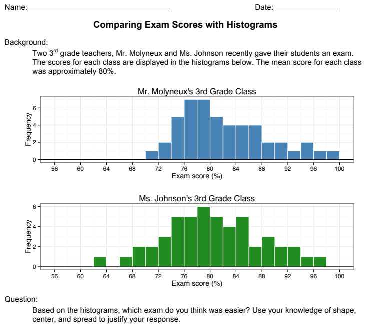
7. Using a timer, allow one minute for students who recommend freeway travel to argue their point to the person they are facing. Then, repeat for students who recommend street travel. Students should not interrupt or respond; they should only listen to the other side.

8. Next, give debaters two minutes to prepare a **rebuttal** of the other person's argument. For example, if one student claimed that freeway travel is better, the other student may ask where the evidence is in the data or show that the data does not support the claim.

9. Allow each debater two minutes to present his/her rebuttal.

10. Finally, ask debaters if any of them changed their recommendations after engaging in the debate.

11. In teams, have students complete the *Comparing Exam Scores with Histograms* handout (LMR_2.9). Allow students time to read the "Background" portion of the handout, and then discuss what statistical question(s) the student in the scenario is trying to answer.



Name:______________________________ Date:_______________

Comparing Exam Scores with Histograms

Background:

Two 3rd grade teachers, Mr. Molyneux and Ms. Johnson recently gave their students an exam. The scores for each class are displayed in the histograms below. The mean score for each class was approximately 80%.

Question:

Based on the histograms, which exam do you think was easier? Use your knowledge of shape, center, and spread to justify your response.

LMR_2.9

12. Repeat debate process (Steps 4 - 10) with the other half of the class.

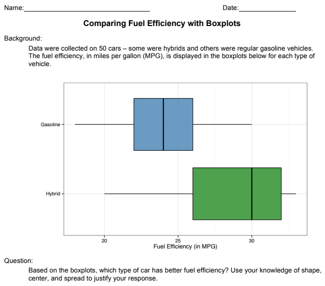
13. Summarize the lesson by conducting a class discussion about what to look for when comparing distributions. Students should be precise when estimating values of means, medians, MAD, and IQR. They should also be able to comment on when it is most appropriate to use each measure of center and spread. ***If a distribution is symmetric, it is best to use the mean as a measure of center and the MAD as a measure of spread. If a distribution is skewed, or has outliers, it is best to use the median as a measure of center and the IQR as a measure of spread.***

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

☑ Similar to the activities they did during class today, for homework, students should complete the *Comparing Fuel Efficiency with Boxplots* handout (LMR_2.10).



Name:______________________________ Date:_______________

Comparing Fuel Efficiency with Boxplots

Background:

Data were collected on 50 cars – some were hybrids and others were regular gasoline vehicles. The fuel efficiency, in miles per gallon (MPG), is displayed in the boxplots below for each type of vehicle.

Question:

Based on the boxplots, which type of car has better fuel efficiency? Use your knowledge of shape, center, and spread to justify your response.

LMR_2.10

Lesson 7: Plot Match

Objective:

Students will learn how to create a boxplot from an already-established dotplot.

Materials:

1. *From Dotplots to Boxplots* handout (LMR_2.11_Dotplots to Boxplots)
2. Sets of plots from *Plot Match* file (LMR_2.12_Plot Match) – one for each team
   Advanced preparation required (see Step 7 below)

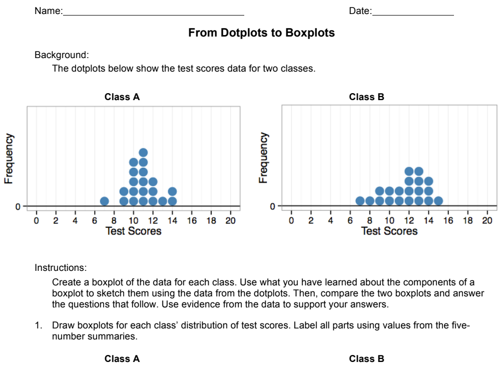
Vocabulary:

representation

> **Essential Concepts:** Boxplots are an alternative visualization of histograms or dotplots. They capture most, but not all, of the features we can see in a dotplot or histogram.

Lesson:

1. Ask students to complete an *Entrance Slip* by recalling the components of the five-number summary that make up a boxplot. ***Five-number summary: minimum, 1st quartile (Q_1), median, 3rd quartile (Q_3), maximum.***
2. Randomly select students to share the components and briefly discuss what each means in a boxplot. If students are missing a component, ask them to add the component to their list.
3. Remind students that during Lesson 5, they created a boxplot from students' heights.
4. Explain that a boxplot is one **representation** of the distribution of a variable in a data set. They have worked with other representations of distributions. Ask students:
   a. What other representations of distributions have we seen? ***Answers may include: dotPlots, bar charts, scatterplots, histograms, and tables.***
5. Distribute the *From Dotplots to Boxplots* handout (LMR_2.11). In teams, students will sketch boxplots from dotplots. They will need to determine the five-number summaries of each plot, and should clearly label each value on their boxplots.



Name:________________________ Date:______________

From Dotplots to Boxplots

Background:

The dotplots below show the test scores data for two classes.

Instructions:

Create a boxplot of the data for each class. Use what you have learned about the components of a boxplot to sketch them using the data from the dotplots. Then, compare the two boxplots and answer the questions that follow. Use evidence from the data to support your answers.

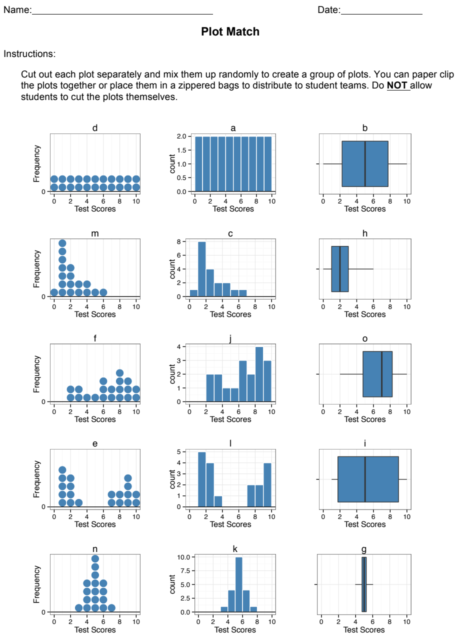
1. Draw boxplots for each class' distribution of test scores. Label all parts using values from the five-number summaries.

Class A **Class B**

LMR_2.11

6. Students should answer the 3 questions included in the handout. They can discuss their answers in pairs, and then have a class share out of the responses.
7. Once the discussion wraps up, inform the students that they will now attempt to find plots that represent the same data but are plotted differently.
8. Distribute one set of plots, from the *Plot Match* file (LMR_2.12), to each student team.

<u>Advanced preparation required:</u> Each student team will receive a set of plots containing all 15 plots from the *Plot Match* file (LMR_2.12). Copies will need to be cut and sorted prior to class time. To keep the plots together, you can either paper clip them or place them in zippered bags. **<u>Note:</u>** Do not distribute the handout for students to cut out the plots!



Name:______________________________ Date:______________

Plot Match

Instructions:

Cut out each plot separately and mix them up randomly to create a group of plots. You can paper clip the plots together or place them in a zippered bags to distribute to student teams. Do **<u>NOT</u>** allow students to cut the plots themselves.

LMR_2.12

9. Inform students that they are now going to gather in their teams and practice matching different representations of distributions. Each group will receive 15 plots (5 dotPlots, 5 histograms, and 5 boxplots). Their task is to determine which dotplots, histograms, and boxplots represent the same data.

10. Once each group has decided upon their 5 groupings, engage the students in a class share out until all students agree. Then, have the students record their responses to the following statements and/or questions in their IDS journals:

    a. What types of data are best for using a histogram? ***Histograms are useful for almost any type of data. They can easily show the shape of a distribution (including skewness and multiple peaks). They are usually best with larger data sets.***
    b. What types of data are best for using a dotplot? ***Dotplots can also easily show the shape of a distribution. They are preferred over histograms when there is a relatively small amount of data.***
    c. What types of data are best for using a boxplot? ***Boxplots are useful when the distribution has one mode (one peak). They are also useful to describe data that are heavily skewed or that contain outliers.***
    d. Describe some characteristics of data that become hidden when a boxplot is used instead of a dotplot or histogram. ***Dotplots and histograms can show the number of modes in a distribution, but a boxplot cannot. If a distribution is bimodal, we will not be able to tell in a boxplot. In general, we lose the ability to talk about the overall shape of the distribution.***

11. Display the uncut version of the *Plot Match* file (LMR_2.12) so that students see the letters that correspond to each set of representations.

Solution key:

Set 1: Plots (d), (a), (b)
Set 2: Plots (m), don't, (h)
Set 3: Plots (f), (j), (o)

Set 4: Plots (e), (l), (i)
Set 5: Plots (n), (k), (g)

12. Have a few students share out their responses. For homework, students will record some pros and cons of using different types of graphical representations to display the same data.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students should reflect on today's class discussion and record their ideas of some pros and cons of using different types of graphical representations to display the same data.

LAB 2B: Oh the Summaries…

Complete Lab 2B prior to the Practicum.

Lab 2B - Oh the Summaries ...

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Just the beginning

- Means, medians, and MAD are just a few examples of *numerical summaries*.
- In this lab, we will learn how to calculate and interpret additional summaries of distributions such as: minimums, maximums, ranges, quartiles and IQRs.
  - We'll also learn how to write our first custom function!
- Start by loading your *Personality Color* data again and name it `colors`.

Extreme values

- Besides looking at *typical* values, sometimes we want to see *extreme* values, like the smallest and largest values.
  - To find these values, we can use the `min`, `max` or `range` functions. These functions use a similar syntax as the `mean` function.
- **Find the `min` value and `max` value for your predominant color.**
- **Apply the `range` function to your predominant color and describe the output.**
  - The *range* of a variable is the difference between available smallest and largest value.
  - Notice, however, that our `range` function calculates the maximum and minimum values for a variable, but not the difference between them.
  - Later in this lab you will create a custom `Range` function that will calculate the difference.

Quartiles (Q1 & Q3)

- The *median* of our data is the value that splits our data in half.
  - Half of our data is smaller than the *median*, half is larger.
- *Q1* and *Q3* are similar.
  - 25% of our data is smaller than *Q1*, 75% are larger.
- Fill in the blanks to compute the value of *Q1* for your predominant color.

```
quantile(~____, data = ____, p = 0.25)
```

- **Use a similar line of code to calculate *Q3*, which is the value that's larger than 75% of our data.**

The Inter-Quartile-Range (IQR)

- Make a `dotPlot` of your *predominant* color's scores.
- Visually (Don't worry about being super-precise):
  - Cut the distribution into quarters so the *number* of *data points* is equal for each piece. (Each piece should contain 25% of the data.)
    - Hint: You might consider using the `add_line(vline = )` to add vertical lines to the quarter marks.
  - **Write down the numbers that split the data up into these 4 pieces.**
  - **How long is the interval of the middle two pieces?**
  - This length is the *IQR*.

Calculating the IQR

- The IQR is another way to describe *spread*.
  - It describes how *wide* or *narrow* the middle 50% of our data are.
- Just like we used the `min` and `max` to compute the `range`, we can also use the *1st* and *3rd* quartiles to compute the *IQR*.
- **Use the values of *Q1* and *Q3* you calculated previously and find the *IQR* by hand**.
  - **Then use the iqr() function to calculate it for you.**
- **Which personality color score has the widest spread according to the *IQR*? Which is narrowest?**

Boxplots

- By using the medians, quartiles, and min/max, we can construct a new single variable plot called the *box and whisker* plot, often shortened to just *boxplot*.
- **By showing someone a `dotPlot`, how would you teach them to make a *boxplot*? Write out your explanation in a series of steps for the person to use.**
  - **Use the steps you write to create a sketch of a *boxplot* for your predominant color's scores in your journal.**
  - **Then use the `bwplot` function to create a *boxplot* using `R`.**

Our favorite summaries

- In the past two labs, we've learned how to calculate numerous *numerical summaries*.
  - Computing lots of different summaries can be tedious.
- Fill in the blanks below to compute some of our *favorite* summaries for your predominant color all at once.

```
favstats(~____, data=colors)
```

Calculating a range value

- We saw in the previous slide that the `range` function calculates the maximum and minimum values for a variable, but not the difference between them.
- We could calculate this difference in two steps:
  - Step 1: Use the `range` function to `assign` the max and min values of a variable the name `values`. This will store the output from the `range` function in the *environment* pane.

```
values <- range(____, data = ____)
```

  - Step 2: Use the `diff` function to calculate the difference of `values`. The input for the `diff` function needs to be a vector containing two numeric values.

```
diff(values)
```

- **Use these two steps to calculate the *range* of your predominant color.**

Introducing custom functions

- Calculating the *range* of many variables can be tedious if we have to keep performing the same two steps over and over.
  - We can combine these two steps into one by writing our own custom `function`.
- Custom functions can be used to combine a task that would normally take many steps to compute and simplify them into one.

- The next slide shows an example of how we can create a custom function called `mm_diff` to calculate the absolute difference between the `mean` and `median` value of a `variable` in our data.

Example function

```
mm_diff <- function(variable, data) {
mean_val <- mean(variable, data = data)
med_val <- median(variable, data = data)
abs(mean_val - med_val)
}
```

- The function takes two *generic* arguments: `variable` and `data`
- It then follows the steps between the curly braces { }
  - Each of the *generic* arguments is used inside the `mean` and `median` functions.
- Copy and paste the code above into an *R script* and *run* it.
- The `mm_diff` function will appear in your Environment pane.

Using `mm_diff()`

- After running the code used to create the function, we can use it just like we would any other numerical summary.
  - In the *console*, fill in the blanks below to calculate the absolute difference between the `mean` and `median` values of your predominant color:

```
____(~____, data = ____)
```

- **Which of the four colors has the largest absolute difference between the `mean` and `median` values?**
  - **By examining a `dotPlot` for this personality color, make an argument why either the `mean` or `median` would be the better description of the *center* of the data.**

Our first function

- Using the previous example as a guide, create a function called Range (*Note the capital 'R'*) that calculates the *range* of a variable by filling in the blanks below:

```
____ <- function (____, ____) {
values <- range(____, data = ____)
diff(___)
}
```

- **Use the `Range` function to find the personality color with the largest difference between the `max` and `min` values.**

On your own

- **Create a function called myIQR that uses the `quantile` function to compute the middle 30% of the data.**

Practicum: The Summaries

Objective:

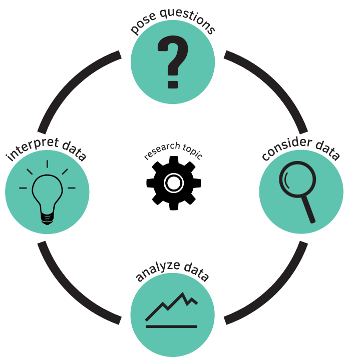
Students will engage in their first statistical investigation using the Data Cycle. They will pose a statistical question based on a data set from a Participatory Sensing campaign, analyze, and interpret the data.

Materials:

1. *The Summaries Practicum* (LMR_U2_Practicum_The Summaries)

Note to Teacher: Before assigning the practicum to your students, engage the class in a discussion about the sample statistical questions below. Guide the discussion so that students identify not only the groups being compared in each question, but also what is being compared about the groups. Remind them of the Data Cycle from Unit 1.

The Data Cycle



Practicum
The Summaries

Using the *Food Habits* campaign data or *Personality Color* survey data, develop a new statistical question that compares two or more groups. Some sample statistical questions (about other data sets) are below:

- Which gender shows a bigger range in age, male or female Oscar winners?
  Grouping variable: gender (male, female)
  Variable: ages
- Do children, teenagers, or adults spend more money on candy?
  Grouping variable: age group (child, teenager, adult)
  Variable: the amount of money spent on candy
- How does the median height of teenage males compare to that of females?
  Grouping variable: gender (male, female)
  Variable: height
- How do the average temperatures of Los Angeles, Las Vegas, and San Francisco compare?
  Grouping variable: city (Los Angeles, Las Vegas, San Francisco)
  Variable: daily maximum temperature

Remember, a statistical question is one that anticipates variability in the question and then addresses the variability in the answer:

Based on the data you chose (*Food Habits* or *Personality Color)*, you need to:

1. Write down your question and think about ways you could answer it using RStudio.
2. Describe the data you are using to answer your question and explain why it is appropriate.
3. Analyze the data to provide evidence that supports the answer to your question. Include plot(s) and numerical summaries (mean, median, MAD, IQR, etc.) related to your plots.

4. Interpret the data to answer your statistical question. You should:
    a. Provide the plot(s) and numerical summaries related to your plot(s).
    b. Describe what the plot shows.
    c. Explain why you chose to make that particular plot(s) and the related numerical summaries.
    d. Explain how the plot and numerical summary answer your statistical question.
5. 5. Write and submit a one-page report.

Note: You may use the scoring guide in Unit 1 to give you an idea of how to score the Practicum.

How Likely Is It?

Instructional Days: 7

Enduring Understandings

Probability measures the long run frequency of occurrence for chance outcomes. Probabilities can be approximated by designing and conducting simulations, and also via mathematical calculation.

Engagement

Students will watch a scene from the movie *Rosencrantz and Guildenstern are Dead* and discuss the likelihood of getting “heads” when tossing a coin 78 times in a row. The scene can be found at: https://www.youtube.com/watch?v=NbInZ5oJ0bc

Learning Objectives

Statistical/Mathematical:

S-CP 2: Understand that two events A and B are independent if the probability of A and B occurring together is the product of their probabilities, and use this characterization to determine if they are independent.

S-CP 9: (+) Use permutations to perform [informal] inference.
*This standard will be addressed in the context of data science.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand how algorithms are used to design simulations.

Applied Computational Thinking using RStudio:

- Design and conduct simulations in RStudio.
- Compare actual data to simulated data using RStudio.
- Re-randomization of permuted data.
- Use estimated probabilities from samples to determine theoretical probabilities

Real-World Connections:

Learn to use simulations to determine expectations of events.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it is used and how to work with it.

2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

Data File or Data Collection Method

Simulated data in RStudio.

Legend for Activity Icons

🎥	💬	📜	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 8: How Likely Is It?

Objective:

Students will understand the basic rules of probability. They will learn that previous outcomes do not give information about future outcomes if the events are independent.

Materials:

1. Video: "Heads" from the movie *Rosencrantz and Guildenstern are Dead* found at: https://www.youtube.com/watch?v=NbInZ5oJ0bc
2. Projector for RStudio functions

Vocabulary:

probability, simulation, model, sample proportion, chance, independence

Essential Concepts: Probability is an area that we humans have poor intuition about. Probability measures a long-run proportion: 50% chance means the event happens 50% of the time *if you repeated it forever*. When we don't repeat forever, we see variability.

Lesson:

1. Ask students to consider possible synonyms to the word **chance**. If someone says, "That just happened by chance," what does that mean? ***Synonyms: possibility, prospect, expectation, unintentional, unplanned. The actual definition of chance is "a possibility of something happening."***

2. Then, ask them which game – chess or the board game, "Sorry" – is more based on chance? Why?

   Note: any game can be chosen. ***"Sorry" is more based on chance because many outcomes are determined by dice rolls. In chess, there are certain strategies and movements that can be planned, so it is more a game of skill. With Sorry, the players roll a die (number cube), so the numbers they roll have an impact on how well they do in the game.***

3. Next ask students if they can think of situations where chance is the only force at play. ***Possible responses: card games, slot machines, the lottery, coin flipping, and rock-paper-scissors.***

4. Play the "Heads" video from the movie *Rosencrantz and Guildenstern are Dead* found at: https://www.youtube.com/watch?v=NbInZ5oJ0bc.

5. In their IDS journals, ask students to write down their initial reactions to the clip by responding to the following questions:

   a. Is it *possible* to get 78 heads in a row when tossing a coin? ***Yes, it is possible to get 78 heads in a row since one coin toss does not determine the next coin toss.***
   b. Do you think it is *likely* to get 78 heads in a row? ***No, although it is possible to get 78 heads in a row.***
   c. How many times should we get heads when tossing a coin? ***1 out of 2 times or 50% of the time.***
   d. On average, how many times out of the 78 tosses should the characters have gotten heads? ***Roughly about 39 times.***

6. Ask students to discuss their findings with their team members and come to an agreement on their responses. Afterwards, conduct a *Whip Around* and ask each team to share its findings. Are there any differences between the teams? Any similarities?

7. As teams share their responses, students should add to or revise their individual findings in their IDS journals.

8. Explain to students that, from the concept of chance, we can start learning about **probability**. Chance is simply the possibility that something will happen, and probability is a measurement for

how often something happens in the "long run." Students may have ideas about how to calculate probabilities based on prior classes or knowledge, but inform them that IDS will be taking a different approach by using simulations (see next step).

9. Since we don't want to actually flip a coin 78 times like the actors did in the video, we can have RStudio simulate them for us. A **simulation** is a way of creating random events that are close to real-life situations without actually doing them. It is a kind of **model**, which is a way of representing real world situations so that predictions can be made.

10. Explain to students that R has a function that does coin flipping for us, and that it assumes an equal probability of heads and tails. Using a projector to display your computer screen to the whole class, demonstrate how to do one simulation of a coin flip in RStudio. Use the following function:

11. `rflip(1)`Explain that the value of 1 in the argument part of the function tells R to flip the coin 1 time. If we want to flip the coin 10 times, we could simply change the function to **`rflip(10)`**.

12. Run the function again using 10 as the number of times to flip the coin. Ask students:

    a. How many heads ("H"s) were there? ***Answers will vary for each sample.***
    b. How many Tails ("T"s) were there? ***Answers will vary for each sample.***
    c. In the output, what does **`Flipping 10 coins [Prob(Heads) = 0.5]`** mean? ***this is RStudio telling us that we are tossing the coin 10 times and that the*** **probability** ***of getting heads should be 0.5 (it is flipping a fair coin).***
    d. In the output, what does **`Number of Heads: 3 [Proportion Heads: 0.3]`** mean? **Note:** This is example of an output. Your sample may have a different value for the number of heads that appeared, and thus a different value for the proportion of heads. ***this is RStudio telling us that in our sample, we got heads 3 out of the 10 times we flipped the coin. The*** **sample proportion** ***is automatically calculated for us by dividing the number of heads by the total number of tosses (in this case, 3/10 = 0.3).***

13. To relate back to the video at the beginning of class, repeat the simulation once more, but use 78 as the number of coin flips **`rflip(78)`**. Ask students:

    a. How many heads ("H"s) were there? Since we know to expect about 39 heads if the coin is fair, does the value seem reasonable? ***Answers will vary for each sample. Most likely, you will see values near 39.***
    b. How many Tails ("T"s) were there? ***Answers will vary for each sample.***
    c. What proportion of the coin flips were heads? ***Answers will vary for each sample.***

14. Using the **`rflip(78)`** command, run the simulation 3–5 more times and have students record the values for the number of heads and the proportion of heads.

    As an example, we ran the function 3 times and saw the following values:

    Sample 1 – amount of heads: 45
    proportion of heads: 0.577

    Sample 2 – amount of heads: 33
    proportion of heads: 0.423

    Sample 3 – amount of heads: 42
    proportion of heads: 0.538

15. Have students answer the questions listed below. The important thing to note is that the values can and (almost always) WILL change each time you run the simulation to create a new sample.

    a. How do the proportions of heads in the samples compare to each other? ***Answers will vary.***
    b. How do the proportions of heads compare to the true probability of heads (1/2 or 50%)? ***Answers will vary, but students should notice that most of the probabilities are close to 50%.***

c. Why is there a 50% chance of getting heads during each coin flip? ***Since there are two sides to a coin, both should be equally likely to come up. So there is a 1 out of 2 chance of getting heads and 1 out of 2 chance of getting tails.***

16. Ask students to engage in a discussion with their group about the statement below, then have a few group reporters share out.

    a. If a coin was flipped 78 times, I would claim that the coin is unfair if I got less than **#** heads or more than **#** heads.

17. Inform students that you are going to perform 500 simulations. Each simulation represents a coin being flipped 78 times. For each simulation, the computer will record the number of heads in the 78 flips. A histogram will be created that represents the number of heads in each of the simulations. The histogram is a model that will display what typically happens when a fair coin is flipped 78 times.

18. Copy and paste the code below in an RScript and run each line of code, one at a time, for the students:

```
set.seed(11) #reproducibility
flips <- do(500)*rflip(78)
View(flips) # 4 variables
histogram(~heads, data = flips)
favstats(~heads, data = flips)
```

19. Engage the students in a discussion about the histogram:

    a. What is the distribution telling us? ***When flipping a fair coin 78 times, what typically happened was that it landed on heads between 36 and 40 times (36/78 = 0.46 to 40/78 = 0.51). It was not uncommon for the coin to land on heads 31-35 (0.40-0.45) times or 41-45 (0.52-0.58) times. Even landing on heads between 46-50 (0.59-0.64) times was not too uncommon. What was very uncommon, however, was landing on heads less than 30 times (less than 38%) or more than 51 times (more than 65%).***

    b. Were the group's cut-offs (item #16) similar to what the chance model displayed? ***Answers will vary. Some groups' intervals might be very wide and others very narrow.***

    c. Use the chance model (histogram) which displays what typically happens when a fair coin is flipped 78 times, make a call for the scenarios below – fair or unfair?

        i. You flip a coin 78 times and get 37 heads. ***Fair. 37 was very common based on the histogram.***

        ii. You flip a coin 78 times and get 46 heads. ***Fair. 46 was less common but not too uncommon.***

        iii. You flip a coin 78 times and get 20 heads. ***Unfair. In the 500 simulations, not once did we see a FAIR coin land on heads 20 times.***

20. Next, pose the following question:

    a. If you get a heads on the first toss of a coin, will you definitely get a heads on the next toss? Will you definitely get a tails on the next toss? ***No. One coin toss should not affect another coin toss. Each time you flip the coin, the chances of getting heads versus tails remains the same.***

21. Introduce the concept of **independence**. Explain that, when tossing a fair coin, there is no relationship between each toss. The second toss does NOT depend on the first toss; therefore, the coin tosses are independent of each other.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students will create a Tweet (they do not have to post it online). Using 280 characters or fewer, write a Tweet about the meaning of probability.

Lesson 9: Bias Detective

Objective:

Students will learn how to use simulations to detect biased probability.

Materials:

1. Poster paper – with 6 columns labeled 1, 2, 3, 4, 5, 6
2. 2 dice (number cubes)
   Note: You can use regular hard dice, or soft foam dice (can be found at dollar stores)
3. Projector for RStudio functions

Vocabulary:

bias

Essential Concepts: In the short-term, actual outcomes of chance experiments vary from what is 'ideal.' An ideal die has equally likely outcomes. But that does not mean we will see exactly the same number of one-dots, two-dots, etc.

Lesson:

1. In pairs, ask students to quickly share their Tweets from the previous lesson's homework. Collect the Tweets and select a few to share with the class. Of the Tweets shared, ask students which one is closest to the definition: probability measures how often something happens in the "long run."

2. Remind students that, during the previous lesson, they were introduced to simulations. The progression following this path: chance → probability → simulations. The motivation for using simulations is that we can use the calculated sample proportions to estimate probabilities of real-life events.

3. During today's lesson, we will be continuing to learn about probability and simulations to determine if an event is not fair (one example: a coin is weighted and lands on heads more often than tails).

4. Ask students what they know about dice (number cubes). If they have never heard of them, show one to the class and explain how it works. ***A die (number cube) is a 6-sided cube. Each face of the cube is labeled with dots to represent a number between 1 and 6. For example, if the face has 3 dots, then it represents the number 3. The cube itself is weighted so that there is an equal probability of rolling each of the 6 numbers.***

5. Have a discussion about what the students would expect the probability of rolling the number 1 should be if a die (number cube) were tossed into the air and allowed to fall back to the ground (or table). ***Since there are 6 numbers on the die, each number should be equally likely to occur, so the probability of rolling a 1 is 1/6.***

6. Display a piece of poster paper on the board with columns labeled 1, 2, 3, 4, 5, 6. Explain that each column represents the numbers on the die (number cube). We will be using poster to tally the results of actual dice (number cube) rolls.

7. Select two students to be dice (number cube) rollers and give each student one die. As noted in the *Materials* section above, you can have the students use either regular hard dice, or softer foam ones (can be found at dollar stores).

8. Tell the class that each student roller will be rolling the dice 6 times (so there will be a total of 12 rolls for our sample). Ask:
   a. If they are rolling the dice 6 times, how often do you think Student 1 will roll a 3? Would you expect it to be the same for Student 2? ***Out of 6 rolls, we would expect to see each of the numbers one time, so we will most likely see about one 3 for Student 1.***

b. Would you expect the Student 2 to roll a 3 just as often? Why? ***Yes, we should expect the same thing from Student 2 because we have independent events. There are actually two ways that independence plays a part here: (1) each student is independent from the other and has no effect on what the other will roll, (2) the 6 die rolls for Student 1 are all independent of each other because each face of the cube has an equal chance of happening on any given roll. So, if Student 1 gets a 3 during his/her first roll, that doesn't give us any information about what he/she will get on the second roll.***

c. Since we will have 12 rolls (and therefore 12 samples), how many tally marks should we expect in each column on the chart? ***We would expect to see 2 tally marks in each column (each number will probably be rolled twice).***

9. Have each student roller toss his/her die one time and share the outcomes with the rest of the class. As they do this, place a tally mark in the corresponding column on the chart. Repeat this process 5 more times so that each student has a total of 6 rolls.

10. As a class, observe the results in the chart and discuss the following:

    a. Do the data from these 12 rolls match what we expected (see responses from Step 8)? Is this surprising? ***Answers will vary by class. Some values may have shown up more than we expected (example: the number 3 was rolled 3 times), and others may not have been rolled at all (example: the number 5 was never an outcome). We only have a small sample of data, so it's not surprising for our results to vary from the expected outcomes.***

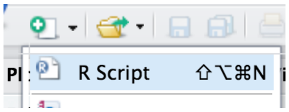
    b. If the data do not match our expectations, does this mean the dice are unfair in some way? ***Even if they don't match our expectations, this does not mean the dice are unfair – we simply don't have enough data yet to know. We would need to roll the dice more.***

    c. If we wanted to purposely create an unfair die, what are some ways we could achieve that? ***Answers will vary by class. Some examples include: (1) We could add tape to one face of the die to give that side more weight. This would increase the chances of the number that is directly opposite of it appearing because the die will land on the heavier side more (and therefore the side facing up will be the number opposite). (2) We could chip the edge of one corner of the die. This would throw off the original balance and favor certain sides.***

11. Similar to the previous day's lesson with coin flipping, we can also simulate dice rolls in RStudio. The function required is called **`roll_die()`**. The arguments for this function are a bit different than the **`rflip()`** function from yesterday. We cannot simply put **`roll_die(1)`** for the computer to roll a die one time. Instead, the function was built with 2 possible dice to choose from – die A and die B.

12. Inform the students that one of the dice in the function is fair and the other has been created with **bias**. Bias is the act of favoring one outcome over another. They will attempt to determine which dice is the biased one by doing multiple simulations.

    Note to Teacher: Many simulations require multiple functions, or code, to perform. This is where RScripts are helpful. An RScript can be used to test code, write notes, and let us easily execute multiple lines of code at one time. This would be a good place to introduce students to RScripts.



13. Using a projector to display your computer screen to the whole class, demonstrate how to open an RScript. Type the following function on your script and click Run. Run simply pastes the function onto the console.

14. `roll_die("A", times = 1)`The output will show one number that represents what value on the die the computer rolled. Go back to your script and modify the function to roll die A 12 Times.

```
roll_die("A", times = 12)
```

15. Compare the results of these 12 simulated rolls to the results of the 12 actual rolls completed by the two students during Step 9. If there is space available on the tally chart, you can add the computer results to it for an easy comparison.

16. Ask students how we could record data from these simulations if we wanted to roll the die 100 times. Would they want to hand count the number of times each value occurred? Is there a function in RStudio that will count them for us? ***It would be difficult to count every individual value in the output on the screen. However, we can use the* `tally()` *function to find out how many times each die value appeared.***

17. To make using the **`tally()`** function easier, we should assign a name to each simulation so we don't have to type the entire function multiple times. We can also have it calculate the proportions for each value. Add the functions below to your RScript and run them one at a time.

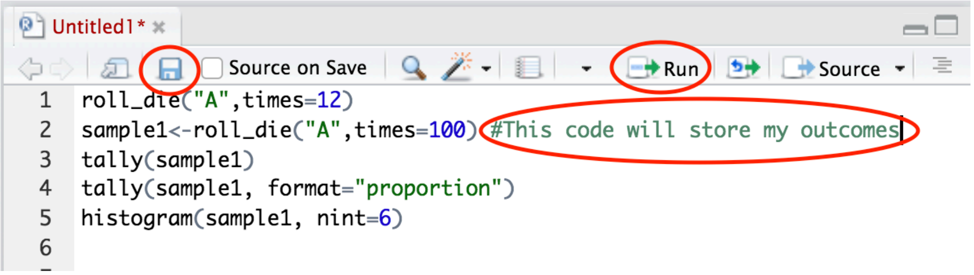
```
sample1 <- roll_die("A", times = 100)
tally(sample1)
tally(sample1, format = "proportion")
```

18. Remind the students that if the die is fair, then each side of the die should appear roughly the same amount of times. Therefore, the proportions should be fairly similar to each other and to the true probability of 1/6.

19. Add the function below to your script, but before running it, ask the students what they think a histogram of the simulated data might look like and then run the command on your screen. Note: Be sure to include the argument **`nint = 6`** so that the resulting histogram has six bars. ***If the die is fair, each of the bars in the histogram should be roughly the same height.***

```
histogram(sample1, nint=6)
```

Note to teacher: Show students how to save an RScript. Inform students that they can take notes on their RScript by including a hashtag (also known as a pound sign or #) at the beginning of the note. Data scientists refer to these types of notes as "code comments" or simply "comments". See image below.



```
roll_die("A",times=12)
sample1<-roll_die("A",times=100)#This code will store my outcomes
tally(sample1)
tally(sample1, format="proportion")
histogram(sample1, nint=6)
```

The Script will be stored in the files tab. To run each function individually, place your cursor on the line and hit the Run button. To run multiple lines of code at once, highlight them and hit Run.

20. Allow the students to access their school computers now to start creating their own simulations in an RScript using die B. Students can pair up, if needed. Have them begin by asking RStudio to roll the die 100 times. They should note their output from both the **`tally()`** function and the histogram. They can then compare the results to those from Step 14. Are they similar? Can they determine which die is unfair yet? ***Answers will vary by class. The results will be similar, but not exact. With the sample sizes of each simulation being fairly small, we cannot see a clear difference between the two dice yet.***

21. Let the students explore by changing the number of times RStudio rolls the dice. Remind them that the goal is to determine which of the two dice is biased. The sample sizes need to be very large in order for them to see a clear difference between the 2 histograms. The pattern becomes more visible when **`times = 2000`**.

    Note: The maximum value for `times` within the `roll_die()` command is 500. Simulations can be combined using the concatenate function `c()`. For example, suppose s1 represents 500 rolls of die A and s2 is a second sample of 500 rolls for die A. To combine these two samples the following can be used `more_rolls <- c(s1, s2)`

22. When students have had enough time to make a decision regarding which dice is biased and how, engage the class in a discussion to verify that everyone agrees. ***Die B is biased; Die A is fair. Die B favors the number 3.***
23. Then, steer the conversation towards why the sample size affected the results. ***The sample size needed to be large because the difference between the probabilities of the die rolls was very small. In order to detect small differences, we must have larger sample sizes.***

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students will consider a four-sided die and imagine rolling it 20 times. They should sketch a histogram of (a) the ideal, expected outcome, (b) an outcome that they think is "realistic," and (c) an outcome they might see if the die were biased to produce more 4's.

Lesson 10: Marbles, Marbles…

Objective:

Students will understand that random events vary, so that the percentage predicted by a probability isn't exact, but varies. Students practice converting percentages to proportions.

Materials:

1. For each student team: 50 marbles – 20 of one color, and 30 of another color
   Note: Marbles can be substituted for other materials, such as small blocks, as long as they are the same size.

Vocabulary:

proportion, percentage, event, with replacement, without replacement

Essential Concepts: There are two ways of sampling data that model real-life sampling situations: with and without replacement. Larger samples tend to be closer to the "true" probability.

Lesson:

1. Remind students that, during the previous two lessons, they learned how to estimate probabilities for a population with the help of simulations to create sample data. Both lessons had nice, prepackaged functions already available in RStudio, which made the simulations fairly quick and easy – in Lesson 8, the **`rflip()`** function was used to simulate flipping a coin; and in Lesson 9, the **`roll_die()`** function was used to simulate rolling one of two dice.
2. But what if we don't have a nice function to perform a simulation for us? Can we create our own method? Yes! We will actually learn to create a simulation from scratch during Lab 2C.
3. Ask students:
   a. If you have a bag of 50 marbles, where 20 of them are blue and 30 of them are red, what is the probability of drawing a red marble? ***30/50 or 60%.***
4. Select a student to answer the question. Ask the class if they agree or disagree. If they agree, ask them to raise their hand. If there are students who disagree, lead a class discussion until a consensus is reached.
5. Ask students to share their strategies on how to convert the **proportion** into a **percentage**. As strategies are being shared, students should take notes in their IDS journals. Review how to turn fractions into percentages, if necessary.
6. Ask students:
   a. What if we actually drew out one marble, recorded its color, then replaced it back in the bag, and did this 10 times? ***Answers will vary by class.***
   b. Would the percentage of red marbles in this sample necessarily be exactly the same as the probability? Identify that each time a marble is drawn, we are creating an **event**. ***Answers will vary by class.***
7. Distribute the bags of marbles to each team. Ask each team to:
   a. Shake the bag of marbles.
   b. Draw one marble out of the bag.
   c. Record the marble's color in their IDS journal.
   d. Replace the marble back into the bag. Inform them that this is called sampling **with replacement**. Ask them to consider what "with replacement" means and discuss with the class. ***"With replacement" means that after you select a marble from the bag, you have to put it back into the bag (replace it) before you select another marble.***

   They should draw 10 marbles from the bag and record the observed colors.

8. Do a *Whip Around* to find out how many times each team drew a red marble out of their 10 draws. Have them calculate the corresponding sample proportions. For example, if one team drew 7 red marbles out of their 10 draws, their sample proportion is 7/10 = 0.70 (which is the sample as a sample percentage of 70%).

9. Ask students why the proportions are perhaps different from each other and from the "true" probability of drawing a red marble?

10. Have the student teams continue drawing marbles, one at a time and with replacement, until they have 50 events recorded. Discuss the following questions:

    a. How many times did they draw a red marble out of these 50? ***Answers will vary by class.***

    b. What's the corresponding sample proportion? Is it closer to the true probability than when you only drew 10 marbles? ***Answers will vary by class. But, they should notice that, as the sample size increases, the sample proportion gets closer to the true population proportion.***

11. Engage students in a discussion about how the sample size affects the sample proportion. They should see that as they draw more marbles, their sample probability gets closer and closer to the true probability. If we were to continue drawing marbles forever, in the long run, our sample proportion should equal our true probability.

12. Have students consider what it might mean to sample **without replacement**. How would they do that with their bag of marbles? ***"Without replacement" means that after you select a marble from the bag, you never put it back into the bag (do not replace it). Instead, you simply select another marble from the bag immediately. Students should recognize that, by not replacing the marble each time, the probabilities will change. This means each draw from the marble bag is NOT independent from another draw because removing one marble impacts the next event.***

13. *Exit Slip*: Based on this lesson, ask students to describe a sample, an event, and replacement.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next Day

LAB 2C: Which Song Plays Next?

Complete Lab 2C prior to Lesson 11.

Lab 2C – Which Song Plays Next?

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

A new direction

- For the past two labs, we've looked at ways that we can summarize data with numbers.
  - Specifically, you learned how to describe the *center*, *shape* and *spread* of variables in our data.
- In this lab, we're going to *estimate the probability* that a rap song will be chosen from a playlist with both rap and rock songs, if the choice is made at random.
  - The playlist we'll work with has 100 songs: 39 are rap and 61 are rock.

Estimate what ... ?

- To *estimate the probability*, we're going to imagine that we select a song at random, write down its genre (*rock* or *rap*), put the song back in the playlist, and repeat 499 more times for a total of 500 times.
- The statistical question we want to address is: *On average, what proportion of our selections will be rap?*
- **Why do we *put a song back* each time we make a selection?**
- **What would happen in our little experiment if we did not do this?**

Calculating probabilities

- Remember that a *probability* is the long-run proportion of time an event occurs.
  - Many probabilities can be answered exactly with just a little math.
  - The probability we draw a single rap song from our playlist of 39 rap and 61 rock songs is `39/100`, `0.39` or `39%`.
- Probabilities can also be answered exactly if we were willing to randomly select a song from the playlist, write down its *genre*, place the song back in the list, and repeatedly do this *forever*.
  - Literally, *forever…*
  - But we don't have that much time. So we're only going to do it 500 times which will give us an *estimate of the probability*.

Estimating probabilities

- You might ask, *Why are we estimating the probability if we know the answer is 39%?*
  - Sometimes, probabilities are too hard to calculate with simple division as we did above. In which case, we can often program a computer to run an experiment to estimate the probability.
  - We refer to these programs as *simulations*.
- The techniques you learn in this lab could be applied to very simple probability calculations or very hard and complex calculations.
  - In both cases, your *estimated* probability would be very close to the *actual* probability.

Getting ready

- Simulations are meant to mimic what happens in real-life using randomness and computers.
  - Before we can start simulating picking songs from a playlist, we need to simulate that playlist in R.

- To simulate our 39 `rap` songs, we'll use the repeat (`Rep`) function.

```
 rap <- rep("rap", times = 39)
```

- Look in the `Environment` pane for the vector containing your rap songs.
- Use a similar line of code to simulate the rock songs in our playlist of 100.

Put the songs in the playlist

- Now that we've got some different songs, we need to combine them together.
  - To do this, we can use the combine function `c()` in R.
- Fill in the blanks to combine your different songs:

```
songs <- __(rap, ____)
```

- And with that, our playlist of songs should be ready to go.
  - Type `songs` into the console and hit enter to see your individual *songs*.

Pick a song, any song

- Data scientists call the act of choosing things randomly from a set, *sampling*.
  - We can randomly choose a song from our playlist by using:

```
sample(songs, size = 1, replace = TRUE)
```

- Run this code 10 times and compute the proportion of `"rap"` songs you drew from the 10.
  - Vocabulary Check: A *proportion* is a fraction of the whole.
    - For example, if 2 rap songs were drawn from the 10, the *proportion* would be 2/10
    - It is more common to express a *proportion* as a decimal, in this case, 0.20
    - It is even more common to express a *proportion* as a percentage, 20%
- **Once everyone in your class has computed their proportions, calculate the *range* of proportions (the largest proportion minus the smallest proportion) for your class and write it down.**

Now `do()` it some more

- Instead of running the same line of code multiple times ourselves we can use R to `do()` multiple repetitions for us.
  - Fill in the blanks below to do the `sample` code from the previous slide *50* times run:

```
do(___) * sample(___, ___ = ___, ___ = ___)
```

- Recall that we need to store our results to be able to perform analysis.
- *Assign* the 50 selected songs the name `draws` and then `View` your file.
- What is the variable name?
  - `R` defaulted to naming the variable based on the function used. You may use the data cleaning skills you learned in lab 6 to `rename` the variable if you wish.
- Fill in the blank below to tally how often each genre was selected:

```
tally(~___, data = draws)
```

- **Compute the proportion of `"rap"` songs for your 50 draws and find out if the *range* for your class' proportions is bigger or smaller than when we drew 10 songs.**

Proportions vs. Probability

- To review, so far in this lab we've:
    - Simulated a "playlist" of songs.
    - Repeatedly simulated drawing a song from the playlist, noting its genre and placing it back in the playlist.
    - Computed the proportion of the draws that were `"rap"`.
- These proportions are all *estimates* of the theoretical probability of choosing a rap song from a playlist.
    - As we increase the number of draws, the *range* of proportions should shrink.

When using simulations to estimate probabilities, using a large number of repeats is better because the estimates have less variability and so we can 'e confident we're closer to the actual value.

Non-rando' Randomness

- We've seen that random simulations can produce many different outcomes.
    - Some estimated probabilities in your class were smaller/larger relative to others.
- There are instances where you might like the same random events to occur for everyone.
    - We can do this by using `set.seed()`.
- For example, the output of this code will always be the same:

```
set.seed(123)
sample(songs, size = 1, replace = TRUE)

## [1] "rap"
```

Playing with seeds

- With a partner, choose a number to include in `set.seed` then redo the simulation of 50 songs.
    - Both partners should run `set.seed(___)` just before simulating the 50 draws.
    - The blank in `set.seed(___)` should be the same number for both partners.
    - Verify that both partners compute the same proportion of `"rap"` songs.
- Redo the 50 simulations one last time but have each partner choose a different number for `set.seed(___)`.
    - **Are the proportions still the same? If so, can you find two different values for `set.seed` that give different answers?**

On your own

- Suppose there are 1,200 students at your school. 400 of them went to the movies last Friday, 600 went to the park and the rest read at home.

If we select a student at random, what is the probability that this student is one of the one's who went to the movies last Friday?

- **Answer this by estimating the probability that a randomly chosen student went to the movies using 500 simulations.**
    - **Write down both the estimated probability and the code you used to compute your estimate. You might find it helpful to write your answer in an R Script *(File -> New File -> R Script)***
    - **Include `set.seed(123)` in your code before you `do` 500 repeated samples.**

Lesson 11: This AND/OR That

Objective:

Students will understand how AND/OR probabilities are defined and will be able to use frequency tables to compute these probabilities.

Materials:

1. *Compound Probabilities* handout (LMR_2.13_Compound Probabilities)
2. Blue sticky notes
3. Gold sticky notes
4. Four signs on the board labeled: *Pickles, Mayonnaise, Both, None* (in that exact order, and equally spaced across the length of the board)

Vocabulary:

compound probabilities

Essential Concepts: What does "A or B" mean versus "A and B" mean? These are compound events and two-way tables can be used to calculate probabilities for them.

Lesson:

1. Remind the students that they have been learning about estimating probabilities of single events based on sample proportions. Inform them that, today, they will learn how to calculate proportions when multiple events happen.
2. Review the basic idea of computing probabilities; in other words, the number of outcomes we are interested in divided by the total number of outcomes possible.
3. Pose the questions below to the class.

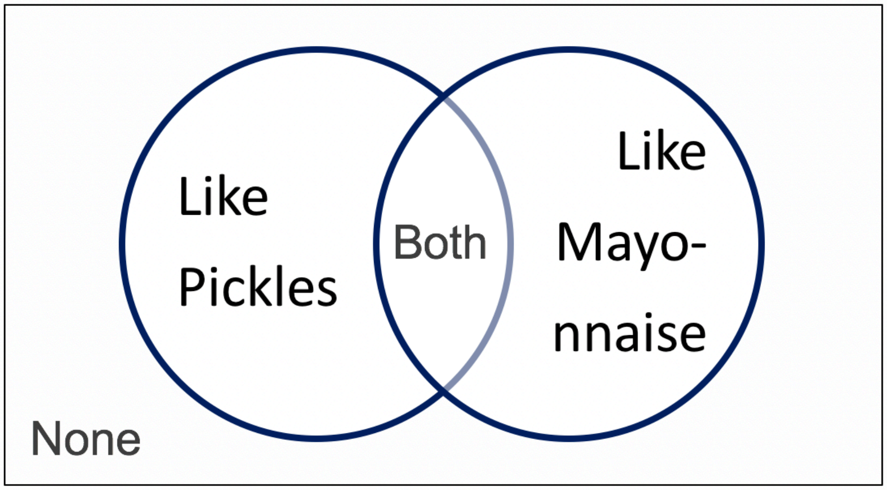
   <u>Note:</u> They do not need to come up with specific answers; this is a time for them to make suggestions.
   a. How would we compute the probability of two outcomes occurring at the same time? For example, what is the probability that a randomly chosen student likes both pickles AND mayonnaise?
   b. How would we compute the probability of either of two outcomes occurring? For example, what is the probability that a randomly chosen student likes either pickles OR mayonnaise?
4. For both questions, steer the students towards using the definition from Step 2. That is, we want the students to realize that they can count the number of people that qualify for the given circumstance and divide by the total number of people to calculate a probability.
5. In order to define AND/OR probabilities, students will participate in an activity where they are grouped by their food preferences.
6. Divide the board into 4 groups and write the words “Pickles,” “Both,” “Mayonnaise,” and “None,” in that order, from left to right.
7. Ask for 10 volunteers to stand by the word that represents their preferences. That is, if they only like pickles, they should stand by the word “Pickles.” If they like both pickles and mayonnaise, they should stand by the word “Both.”

   <u>Note:</u> If all 4 groups do not have at least one student in them, select a few more students to stand at the board.
8. Ask the remaining students (those still seated) to count the total number of people standing by the board and have a student volunteer share the answer with the class. ***Answers will vary by class.***

9. Next, create a 2-way frequency table like the one below to organize the values of student preferences as follows:
   - Counts for students who like both go in the Yes/Like Mayonnaise and Yes/Like Pickles box.
   - Counts for students who like none go in the No/Like Mayonnaise and No/Like Pickles box
   - Counts for students who like only mayonnaise go in the Yes/Like Mayonnaise and No/Like Pickles box.
   - Counts for students who like only pickles go in the No/Like Mayonnaise and Yes/Like Pickles box.

   Note: A Venn diagram like the one below may be used as well, depending on student understanding and at teacher discretion.

		Like Mayonnaise		
		Yes	No	TOTAL
Like Pickles	Yes			
	No			
	TOTAL			



10. Next, ask the students sitting down the following questions:
    a. How many students like both pickles AND mayonnaise? ***Answers will vary by class.***
    b. What is the probability that a randomly selected student at the board likes both pickles AND mayonnaise? ***Answers will vary by class. The probability should be calculated by dividing the number of people who are standing under "Both" (number given in Step 9(a)) by the number of students at the board (number given in Step 8).***

$$\frac{\textbf{\# students under ``Both''}}{\textbf{\# students standing at the board}}$$

11. Now, ask a student from the audience:
    a. How many students like pickles? ***Answers will vary by class.***

    Note: Avoid phrasing the question with "Students that like ONLY pickles." Students need to see that students who like "Both" items also belong to the groups liking the individual items.

If students mistakenly report the number of students who like ONLY pickles, ask the people at the board to raise their hands if they like pickles and then ask the mistaken student to recount.

b. What is the probability that a randomly selected student at the board likes pickles? ***Answers will vary by class. The probability should be calculated by dividing the number of people who are standing under "Pickles" and "Both" by the total number of students at the board.***

$$\frac{\textbf{(\# students under "Pickles") + (\# Students under "Both")}}{\textbf{\# students standing at the board}}$$

12. Finally, ask one more student from the audience:

a. How many students like pickles OR mayonnaise? ***Answers will vary by class.***

Note: Avoid phrasing the question with "Students that like ONLY pickles OR ONLY mayonnaise."

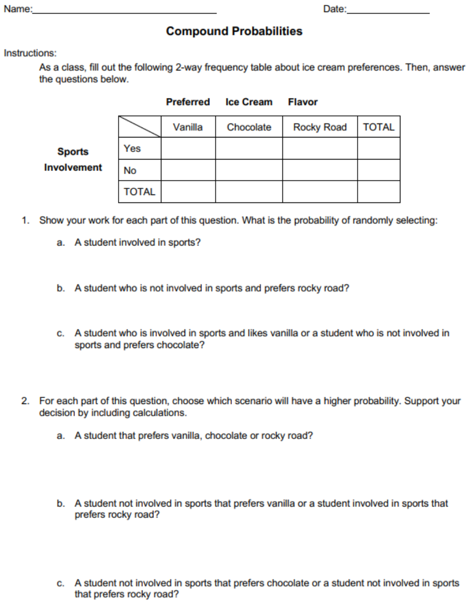
If students mistakenly report the number of students who like ONLY pickles plus the students who like ONLY mayonnaise, ask the people at the board to raise their hands if they like either pickles or mayonnaise (All students at the board should raise their hand except for the students who like "None") and then ask the mistaken student to recount.

b. What is the probability that a randomly selected student at the board likes pickles OR mayonnaise? ***Answers will vary by class. The probability should be calculated by dividing the number of people who are standing under "Pickles," "Mayonnaise," and "Both" by the total number of students at the board.***

$$\frac{\textbf{(\# students under "Pickles") + (\# Students under "Mayonnaise") + (\# Students under "Both")}}{\textbf{\# students standing at the board}}$$

13. Informs students that AND/OR probabilities are called **compound probabilities.** In teams, have students record their own definitions of AND/OR probabilities based on the activity they just completed. ***A compound probability is the probability of some combination of events occurring.***

14. Distribute the *Compound Probabilities* handout (LMR_2.13).



Name:______________________________ Date:______________

Compound Probabilities

Instructions:

As a class, fill out the following 2-way frequency table about ice cream preferences. Then, answer the questions below.

		Preferred	**Ice Cream**	**Flavor**	
		Vanilla	Chocolate	Rocky Road	TOTAL
Sports Involvement	Yes				
	No				
	TOTAL				

1. Show your work for each part of this question. What is the probability of randomly selecting:

a. A student involved in sports?

b. A student who is not involved in sports and prefers rocky road?

c. A student who is involved in sports and likes vanilla or a student who is not involved in sports and prefers chocolate?

2. For each part of this question, choose which scenario will have a higher probability. Support your decision by including calculations.

a. A student that prefers vanilla, chocolate or rocky road?

b. A student not involved in sports that prefers vanilla or a student involved in sports that prefers rocky road?

c. A student not involved in sports that prefers chocolate or a student not involved in sports that prefers rocky road?

LMR_2.13

15. Pass out a blue sticky note to each student who plays a sport and a gold sticky note to each student who does not play a sport.
16. Draw the table from the worksheet on the board (make it large and legible).
17. Have each student who plays a sport hold up their sticky note. Count them and record the number of students who play a sport in the appropriate row of the TOTAL column in the table.
18. Have each student who does not play a sport hold up their sticky note. Count them and record the number of students who do not play a sport in the appropriate row of the TOTAL column in the table.
19. Ask each student which of the following ice cream flavors they most prefer (each student must choose exactly one option): Vanilla, Chocolate or Rocky Road.
    a. Have the students write their ice cream preference on their sticky note.
    b. Fill out the remainder of the table by asking each group of students, those who play a sport and those who do not, to hold up their preference.
    c. Make sure the totals for preferred ice cream flavors and sports involvement add up to the same number.
20. Instruct the students to work in pairs to answer the questions on the *Compound Probabilities* handout (LMR_2.13).

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

If not completed in class, students should finish the *Compound Probabilities* handout (LMR_2.13).

LAB 2D: Queue It Up!

Complete Lab 2D prior to the Practicum.

Lab 2D - Queue it up!

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Where we left off

- In the last lab, we looked at how we can use computer simulations to compute estimates of simple probabilities.
  - Like the probability of drawing a song genre from a playlist.
- We also saw that performing *more* simulations:
  - Took *longer* to finish.
  - Had estimates that *varied less*.
- In this lab, we'll extend our simulation methods to cover situations that are more complex.
  - We'll learn how to estimate their probabilities.
  - We also look at the roll of sampling *with* or *without replacement*.

Back to songs

- In R, simulate a *playlist of songs* containing 30 `"rap"` songs, 23 `"country"` songs and 47 `"rock"` songs.
  - *Assign* the combined playlist the name `songs`.
- Simulate choosing a single song 50 times. Then use your simulated draws to estimate the probability of choosing a *rap* song.
  - The actual (theoretical) probability of choosing a *rap song* in this case is `0.30`.
  - **Write a sentence comparing your estimated probability to the actual probability.**

With or Without?

- So far, you've selected songs *with replacement*.
  - We called it that, because each time you made a selection, you started with the same playlist. That is, you chose a song, wrote down its data, and then placed it back 'n the list.
- It's also possible to select *without replacement* by setting the `replace` option in the `sample` function to `FALSE`.
- Take a sample of `size` 100 from our playlist of songs *without replacement*. Assign this sample the name `without`.
  - **Run `tally(~without)` and describe the output. Does something similar happen if you sample *with replacement*?**
    - Notice that the tilde ~ was not needed with the `tally` function. This is because `without` was not a variable within a data frame but rather a vector which acts like a lone variable.
  - **What happens if `size = 101` and `replace = FALSE`?**

Sample with? Or without?

- Imagine the following two scenarios.
  1. You have a coin with two sides: *Heads* and *Tails*. You're not sure if the coin is fair and so you want to estimate the probability of getting a *Head*.
  2. A child reaches into a candy jar with 10 *strawberry*, 50 *chocolate* and 25 *watermelon* candies. The child is able to grab three candies with their hand and you're interested in probability that all three candies will be chocolate.

- **Which of these scenarios would you sample *with replacement* and which would you sample *without replacement*? Why?**
  - **Write down the line of code you would run to `sample` from the candy jar. Assume the simulated jar is named `candies`.**

Simulations at work

- In reality, songs from a playlist are chosen without replacement.
  - This way, you won't hear the same song several times in a row.
- Let's write a more realistic simulation and estimate the probability that if we select two songs at random, without replacement, that both are rap songs.
  - Use the `do` function to perform 10 simulated `samples` of `size` 2, with replacement and *assign* the simulations the name `draws` and then `View` your file. Use `set.seed(1)`.

What are the variable names? What happened in the first simulation? Did any of your 10 simulations contain two rap songs?

Simulations and probability

- To estimate the probability from our simulations, we need to find the proportion of times that the event we're interested in occurs in the simulations.
- In other words, we need to count the number of times the desired events occurred, divided by the number of attempts we made (the number of simulations).
- The next slides will show you two ways to do this.

Counting similar outcomes

- One way we can estimate the probability of drawing two songs of the *same* genre is to use the following trick to count the number of *rap* songs in each of the 10 simulations:

```
mutate(draws, nrap = rowSums(draws=="rap"))
```

- **Let’s break down the code above by running each part of the code one piece at a time. As you run each line of code below describe the output**

```
draws == "rap"

rowSums(draws == "rap")

mutate(draws, nrap = rowSums(draws=="rap"))
```

- Remember to assign a name to your mutated data set.

Counting other outcomes

- Another method we can use to estimate the probability of complex events is to use the following 2-step procedure:
  1. Subset the rows of the simulations that match our desired outcomes.
  2. Count the number of rows in the subset and divide by the number of simulations.
- The result that you obtain is an estimate of the probability that a specific combination of events ’occurred.
- We'll see an example of this method on the next slide.

Step 1: Creating a subset

- Fill in the blanks below to:

1. Create a subset of our simulations when both draws were `"rap"` songs.
2. Count the number of rows in this subset
3. And divide by the total number of repeated simulations.

```
draws_sub <- filter(draws, ___ == "rap",  ___ == "rap")

nrow(___) / ___
```

Estimating probabilities

- Answer the following questions by performing 500 simulations of sampling 2 songs from a playlist of 30 rap, 23 country and 47 rock songs:
- **Calculate estimated probabilities for the following situations:**
  1. You draw two `"rap"` songs.
  2. You draw a `"rap"` song in the first draw and a `"country"` song in the 2nd.
- **Create a histogram that displays the number of times a `"rap"` song occurred in each simulation. That is, how often were zero rap songs drawn? A single rap song? Two rap songs?**

On Your Own

- Using what you've learned in the previous two labs, answer the following question by performing two computer simulations with 500 repetitions a piece:

If we draw 5 songs from a playlist of 30 rap, 23 country and 47 rock songs, how does the estimated probability of all 5 songs being rap songs change if we draw the songs with or without replacement?

- For each simulation:
  - **Create a histogram for the number of *rap* songs that occurred for each of the 500 repetitions.**
- **Describe how the distribution of the number of *rap* songs changes depending on if we use replacement or not.**

Practicum: Win Win Win

Objective:

Students will create and combine simulations to assess probabilities.

Materials:

1. *Win Win Win Practicum* (LMR_U2_Practicum_Win Win Win)

Practicum
Win Win Win

The California lottery has a game called the *Daily 3*.

- It consists of 3 numbers between 0 - 9 that are drawn daily.
- The numbers are drawn *with replacement*
- Winners are usually awarded a couple hundred dollars.
- To win the maximum amount of money, players must correctly choose the numbers that are drawn, in order.

Based on what you learned in *Lab 2C* and *Lab 2D* (*Which Song Plays Next* and *Queue it Up!)* and using the rules of the *Daily 3*, you need to:

1. Write down the code to correctly simulate the *Daily 3* once.
2. Use your code to simulate the *Daily 3* 500 times.
3. Compute the estimated probability of getting the first 2 numbers of the *Daily 3* correct.
4. Should the estimated probability of correctly guessing the last 2 numbers of the *Daily 3* be less than, the same as, or more than guessing the first 2 numbers? Why?
5. In teams of 4:
    a. Each team member chooses 3 numbers for the *Daily 3*.
    b. Each team member simulates the *Daily 3* game 500 times.
    c. Within your group, combine the team simulation estimates to estimate the probability of winning the *Daily 3*.
6. Write and submit a one-page report. Your report should include the code.

What Are the Chances That You Are Stressing or Chilling?

Instructional Days: 8

Enduring Understandings

Permutations of data provide a model that shows us how the world behaves if chance is the only reason for differences between groups or for associations between variables. If our actual observation is a rare permutation, this suggests that chance is not a good explanation for the difference or association. On the other hand, if the actual observation is a common permutation, this suggests that chance may be a valid explanation. Differences between permuted data and actual data suggest that the chance model can be rejected and there is a dependent relationship between two variables.

Engagement

Students will read the Huffington Post article titled *Don't Take My Stress Away* to set the stage for the Stress/Chill Campaign. High school students who expected, and wanted, to feel stressed out by school wrote this article. The article is found at:

http://www.huffingtonpost.com/jack-cahn/dont-take-my-stress-away_b_2090203.html

Learning Objectives

Statistical/Mathematical:

S-IC 2: Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand that a chance model serves as an indicator of whether or not associations in the actual data are due to chance (understand why a plot might appear to have a trend, but may actually be the result of randomness). Understand that simulations provide a way of comparing expected chance outcomes to real outcomes in order to determine if a model and actual data appear consistent. Learn about merging data sets by understanding the structure of both data sets and the logic of the way they will be combined.

Applied Computational Thinking using RStudio:

- Permutations of data, determining if actual data is similar to permuted data
- Merge multiple data sets together based on a common variable
- Create permutations using a merged data set

Real-World Connections:

In media, citizens read about results and scientific studies in which treatments are applied. In real life, one can ask the question: Does this happen by chance? Understanding chance helps us interpret media reports of scientific and medical findings.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

3. Students will use complex sentences to write informative short reports that use data science concepts and skills.
4. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data Collection Method:

1. **Stress/Chill Participatory Sensing Campaign**: Students will monitor how they feel at different times of the day – whether they are "stressing" or "chilling." Along with how they feel, they will make observations regarding other factors, such as being alone or with others, what they are doing at that moment, and why they are doing that activity.

Data Files:

1. Students' *Personality Color* survey data (`colors`)
2. Students' *Stress/Chill* campaign data
3. Titanic data set (`titanic.rda`)
4. Horror Movie data set (`slasher.rda`)

Legend for Activity Icons

🎥	💬	📜	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 12: Don't Take My Stress Away!

Objective:

Students will read the Huffington Post article titled *Don't Take My Stress Away* to spark their interest about how they spend their time, and will continue to read reports critically to look for claims that may or may not be based on data.

Materials:

1. Article: *Huffington Post's Don't Take My Stress Away* found at: http://www.huffingtonpost.com/jack-cahn/dont-take-my-stress-away_b_2090203.html
2. Data collection devices

Essential Concepts: Generating statistical questions is the first step in a Participatory Sensing campaign. Research and observations help create applicable campaign questions.

Lesson:

1. Become familiar with the *Stress/Chill Campaign Guidelines* (shown at the end of this lesson), particularly the questions, to help guide students during the campaign (see Campaign Guidelines in Teacher Resources).
2. Ask students the following questions and conduct a brief share out of their responses.
   a. Do you know anyone who seems to be always stressed or anyone who seems to be always chilled? ***Answers will vary by class.***
   b. What are some observations you have made that make that person extremely stressed or chilled? ***Answers will vary by class.***
3. Inform students that they will be learning about some high school students who view stress as a part of life in the Huffington Post article titled *Don't Take My Stress Away.*
4. Provide students the link to the article and allow time for them to read it: http://www.huffingtonpost.com/jack-cahn/dont-take-my-stress-away_b_2090203.html
5. As they read the article, students should note whether they agree or disagree with the authors and should write down their comments and/or reactions to the article in their IDS journals.
6. Ask student pairs to share if they agree or disagree with the authors of the article and why. Conduct a *Share Out* of student responses.
7. Inform students that for this unit, we will be investigating how stressed or chilled they are at certain times of the day.
8. Students will collect data using the *Stress/Chill* Participatory Sensing campaign. They will add the *Stress/Chill* campaign to their list of available campaigns either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org
9. Ask students to complete their first survey.
10. After students have completed their first survey, use a random number generator to generate two random times a day for the next 6 days (RStudio example given below). It is recommended that you create 6 sets of random numbers so that students are polled at different times each day.

    Example for RStudio (assuming students are awake between the hours of 7:00 am and 11:00 pm):

    ```
    > sample(7:23, size = 2, replace = FALSE)
    ```

    Note: If a time falls within the school day, it is up to the discretion of the teacher to use this time or not.

11. Based on the times generated, ask students to set reminders on the IDS UCLA App for the next 6 days. Students without a mobile device may set reminders using a method available to them.
12. Focus students’ attention on the *Stress/Chill* survey questions (you may display the questions on the Campaign Guidelines document). Ask students to generate three statistical questions that could be answered using the *Stress/Chill* data.
13. Then, ask them to write down in their IDS journals some predictions about what they think they will see after they collect some data.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

For the next 6 days, students will collect data for the *Stress/Chill* campaign either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org

Campaign Guidelines – Stress/Chill

1. **The Issue:**

   People report being more and more stressed everyday. This trend is extending beyond adults, it is also reported by children and teenagers. The amount of work for which people are responsible has been increasing. To understand what makes us feel stressed, some important questions to ask are:

   a) What factors affect my stress/chill level?
   b) Do different personality types have different things that make them happy/sad?
   c) Do you like to be alone or with people?
   d) Is your stress/chill level a function of the environment in which you are?

2. **Objectives:**

   Upon completing this campaign, students will have compared groups and gained an understanding of variability within and between groups. They will have learned how to conduct and use permutations to model variability, perform informal inference, and how to do simulations to make predictions.

3. **Survey Questions:**

   Use a random number generator to generate two random times a day for the next 6 days, including a weekend if possible. If a time falls within the school day, it is up to the discretion of the teacher to use this time or not.

Prompt	Variable	Data Type
Take a photo of what you are doing right now.	photo	photo
How stressed are you feeling right now (3 is very stressed, 0 is not stressed at all)?	level	integer
How many people are you with (not counting yourself) up to 107,282?	howmanypeople	integer
Who are you with? -alone -friends -family -friends and family -classmates -teacher -teacher and classmates -strangers	who	categorical
Where are you? -school -work -home -public place -others' house -commuting	where	categorical
Why are you here (in one word)?	why	text
AUTOMATIC	location	lat, long
AUTOMATIC	time	time
AUTOMATIC	date	date

When? Surveys are taken two to three times per day at pre-determined randomly selected times.

How Long? About two weeks. Ideally, two of these days include a weekend.

4. **Motivation:**

   Students must understand that they need to keep collecting data. Use the Plot App to look at the data after the first day and have a discussion.

   Ask: Why were most people stressed? Guide students along the way.

   Ask students to predict the following: What is your stress/chill level in the evening versus morning? Does it change everyday? How about during the weekend? What is the difference between groups?

   Data collection: After the first day, use the Campaign Monitoring tool to see who has collected the most data.

5. **Technical Analysis:**

   Students will use RStudio.

6. **Guiding Questions:**

   a) Have students generate predictions and check up on their predictions.
   b) What's the typical stress/chill level of the class across the campaign?
   c) What's my typical stress/chill level and how does it compare to whole class?
   d) Do the stress/chill levels vary by weekday or weekend or the type of people you are with?
   e) Under which conditions is my stress/chill level affected?
   f) Encourage students to generate their own questions.

7. **Report:**

   Students will complete the Stress/Chill Practicum. They will analyze their stress/chill data using data analysis skills and RStudio skills learned in the unit.

Lesson 13: The Horror Movie Shuffle

Objective:

Students will understand that, just by chance, we will see differences between two groups. They will understand that these differences are usually small. Specifically, they will learn that we can determine if outcomes are due to chance for categorical variables by calculating differences in the proportions between two groups.

Materials:

1. 3" x 5" cards (1 per student)

Vocabulary:

chance, simulations, randomness, shuffle

Essential Concepts: We can "shuffle" data based on categorical variables. The statistic we use is the difference in proportions. The distribution we form by shuffling represents what happens if chance were the only factor at play. If the actual observed difference in proportions is near the center of this shuffling distribution, then we would conclude that chance is a good explanation for the difference. But if it is extreme (in the tails or off the charts), then we should conclude that chance is NOT to blame. Sometimes, the apparent difference between groups is caused by chance.

Lesson:

1. **Data Collection Monitoring:** Display the IDS Campaign Monitoring Tool, found at https://portal.idsucla.org. Click on **Campaign Monitor** and sign in.
2. Inform students that you will be monitoring their data collection. Ask:
   a. Who has collected the most data so far? *See User List and sort by Total*.
   b. How many active users are there? How many inactive users are there? *Click on pie chart*.
   c. How many responses were submitted yesterday and today? *See Total Responses*.
   d. How many responses have been shared? How many remain private? *Click on pie chart.*
   e. Using TPS, ask students to think about what they can do to increase their data collection.
3. Conduct a discussion about the data that has been collected.
4. Have students recall what they have learned about **chance** (see Lesson 8). ***Synonyms: possibility, prospect, expectation, unintentional, unplanned. The actual definition of chance is "a possibility of something happening."***
5. To expand on the flow chart from Lesson 9 (chance → probability → simulations), explain that we can use **simulations** to show that sometimes, when we think two groups are different, the difference is really just because of chance, or **randomness**, and does not mean anything. This brings us back to "chance" in the flow chart.
6. Remind students that a simulation is a model for creating random outcomes. Randomness means that something just happens without a specific order.
7. In pairs, ask students to name situations where two groups could be compared, and then have the students record these situations in their IDS journals. Some examples include:
   - ***Men earn more money than women for some work.***
   - ***Basketball players are faster runners than baseball players.***
   - ***Los Angeles students are smarter than _______.***
   - ***UCLA football players are better athletes than USC football players.***
   - ***You and a friend flipped a coin 10 times, and you got more "heads."***

8. Then, ask students to write next to each situation whether they think the differences are either real or due to chance because sometimes differences between two groups are real, but sometimes they might just be due to chance, and they will be learning ways to tell the difference.
9. Explain to the class that we are interested in finding out who will survive by the end of a horror movie. Ask the students:
   a. Do you think men and women have an equal likelihood of surviving by the end of a horror movie? ***Answers will vary by class.***
10. Have a few students share out their opinions along with their reasoning.
11. Inform the students that they will be pretending to be actors from horror movies during today's lesson.
12. Explain that data from horror movies (sometimes called slasher films) were collected of 485 characters from 50 films. For each character, 2 variables were recorded: Gender and Survival. The values for Gender were "Male" and "Female." The values for Survival were "Dies" and "Survives."

	Gender	
Survival	Female	Male
Dies	172	228
Survives	50	35
Total	222	263

Notice that there were more male characters than female characters and that most characters in slasher films do not survive.

13. From this data, the proportion of survivors was calculated for each gender. In other words, for all female characters, the number of female survivors was divided by the total number of females. Similarly, for all male characters, the number of male survivors was divided by the total number of males.

$$\frac{\#\,(\text{“Female” \& “Survives”})}{\#\,(\text{“Female”})} \qquad \text{or} \qquad \frac{\#\,(\text{“Male” \& “Survives”})}{\#\,(\text{“Male”})}$$

14. The percent of females who survived by the end of a horror movie was about **23%**, and the percent of males who survived by the end of a horror movie was about **13%**. Ask the students:
    a. Is this what you expected? (Refer back to the discussion from Step 9.) ***Answers will vary by class. If students thought males would survive more often, then these results would be unexpected. If students thought females would survive more often, then these results would be exactly what they expected. If students thought there was an equal likelihood of survival, these results would also be surprising.***
    b. What is the difference in the proportions of survival rates between genders? What does this mean in the context of surviving a horror movie? ***The difference is 23% - 13% = 10%. This means that 10% more women characters survived than men.***
    c. Is this difference "big" or "small"? How can they define what is "big" and what is "small?" ***Answers will vary by class. Upon first glance, it may seem like 10% is a big difference, but we do not know for sure.***
15. Explain that they will participate in an activity to determine if the 10% difference seen in the actual data set is big or small. This will help them determine if there really is a difference in survival rates for males versus females, or if the 10% difference was just due to chance.
16. Split the class into two groups, 46% of them on one side of the room and the other 54% on the other side of the room. Tell the smaller group they have been assigned to play female characters

in the horror movie (regardless of gender) and tell the larger group that they have been assigned to play male characters in the horror movie (regardless of their gender). Once those groups have been created, have the class calculate the number of students in each group that would have survived a horror film using the actual proportions given in Step 14.

For example: For a class of 30 students:

- ***46% of 30 (0.46x30) ≈ 14 students representing female characters.***
- ***Of those 14 female characters, 23%, or 3 (0.23x14 ≈ 3), are survivors.***
- ***The remaining 16 students (30 – 14 = 16) represent male characters.***
- ***Of those 16 male characters, 13%, or 2 (0.13x16 ≈ 2), are survivors.***

17. Each group should then decide which students will be survivors. Using the 3" x 5" cards, students should write either "dies" or "survives" on their card.

    For example (continued from above):

    Three of the females are survivors; so 3 female characters from the group should write "survives" on their cards. The rest of the group should write "dies" on their cards.

    Two of the male characters are survivors; so 2 males from the group should write "survives" on their cards. The rest of the group should write "dies" on their card.

18. Explain to students that *IF* there really is no difference between genders in horror films, then the characters who survived would only have done so by chance. In other words, males and females would have an equal likelihood of surviving. Have students discuss the following questions:

    a. How many total people in our class are survivors? What is the total proportion of survivors? ***Answers will vary by class. Using the example above, there would be a total of 5 survivors from the class of 30 students. The proportion of survivors would be 5/30 = 0.17 = 17%.***

    b. How many of the survivors would we expect to be male? How many would we expect to be female? ***Answers will vary by class. Using the example above, we would expect to see 17% of males and 17% of females survive since that was the overall proportion of survivors. So, we would expect 0.17*16 ≈ 3 male survivors, and 0.17*14 ≈ 2 female survivors.***

19. Collect all of the 3" x 5" cards from the students and explain that you are going to **shuffle** the cards and redistribute them so that their genders have no influence on whether or not they survive the horror movie.

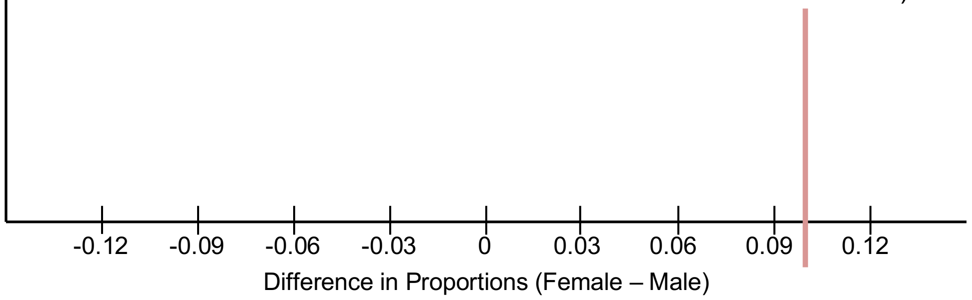
20. Visibly shuffle the survives/dies cards to create a random shuffle. Once the cards have been well-shuffled, pass them back out to the students face down. After all the cards are given out, each group should identify the number of people that are survivors and calculate the corresponding proportion of the survivors.

21. On the board, create a table to display the proportions of survivors for each gender, and include a column for the difference (female survivors – male survivors). Fill in the table with the values the students found in Step 20. **Note:** The first row has been filled in with the example data from above BEFORE the shuffles have taken place. Exact numbers were not used so that the proportions would match the actual horror movie data set.

# of Female Survivors	# of Male Survivors	Proportion of Female Survivors	Proportion of Male Survivors	Difference in Proportions (Female – Male)
3.22	2.08	3.22/14 = 0.23	2.08/16 = 0.13	0.23 – 0.13 = 0.10

22. Note that values in the "Difference in Proportions" column can be positive or negative because sometimes more women will survive, and other times more men will survive.

23. Draw a dotplot on the board labeled “Difference in Proportions.” Include a vertical line at 10% to represent the actual difference in gender survival rates in real horror movies (see example below).



24. Using the information from Steps 20 and 21, place a dot at the corresponding value for the shuffled data’s difference in proportions. Ask the students:
    a. How does this difference compare to the actual data set’s difference of 10%? ***Answers will vary by class. Most likely, the difference in proportions will be much smaller than 10%. In fact, the difference in proportions will be centered around 0.***

25. Repeat Steps 19 – 24 a few more times (depending on how much class time you have available).

26. Ask the students to record their responses to the following questions:
    a. What was the biggest difference we saw from our shuffles? What was the smallest? ***Answers will vary by class.***
    b. What do you think this dotplot would look like if we shuffled our survival cards 1000 times? ***The dotplot would look roughly symmetric and centered around 0, meaning that if there were no relationship between a character’s gender and whether or not they survive, the difference in proportions would typically be 0.***

27. Have a discussion about how the actual difference in gender survival (10%) is rarely seen when we assign “survives” or “dies” just by chance (aka when shuffling). What does this mean in terms of who will die in actual horror movies? ***Since we never (or rarely) saw a 10% difference in the proportions of female survivors versus male survivors, it seems that horror movies actually favor female survivors.***

28. Ask the students:
    a. If you were going to be cast in a horror movie, would you want to be a male character or a female character? ***You would want to be a female character because they are more likely to survive by the end of the film.***

29. Inform the students that they will learn how to shuffle in RStudio in order to determine if an event is real or simply due to chance.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

For the next 5 days, students will collect data for the *Stress/Chill* campaign either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org

LAB 2E: The Horror Movie Shuffle

Complete Lab 2E prior to Lesson 14.

Lab 2E - The Horror Movie Shuffle

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Playing with permutations

- *Slasher* films are notoriously gory and are said to contain recurring biases.
  - One such bias, is that women in slasher films are more likely to survive than men.
- This lab will focus on the statistical question: *Are women in slasher films more likely to survive until the end of the film than men?*
- To answer this question, we'll learn how to use permuted data to gauge how likely an event occurs by chance.
- To begin, use the `data` function to load the `slasher` data file.
  - The data contains information about 485 characters from a random sample of 50 *slasher* horror films.

Initial thoughts...

- To familiarize yourself with the data, answer the following:
  - **How many variables and observations are contained in the data and what are the possible values of the variables?**
  - **Which gender had more survivors? Write down a few sentences as to how you came to your conclusion. Be sure to look at both the *counts* and *percentages* of survivors in each group before deciding.**
  - **Calculate the difference between the percentage of females who survived and the percentage of males who survived. Is the difference large enough to conclude that women tend to survive more often than men?**

Tally whoa ... !

- Something you might have noticed is that these two lines of code aren't equivalent:

```
tally(gender ~ survival, data = slasher)

tally(survival ~ gender, data = slasher)
```

- One of these lines takes the group of *survivors* and tells us how many of them were `Male` or `Female`.
- The other takes the group of *females / males* and tells us how many of them `Dies` or `Survives`.
- **The last question on the previous slide can be answered using the 2nd line of code. Why?**
  - Pro-tip: Include the option `format = "percent"` to obtain a two-way table with percentages.

```
tally(survival ~ gender, format = "percent", data=slasher, margin = TRUE)
```

```
##          gender
## survival    Female      Male
##   Dies      77.47748    86.69202
##   Survives  22.52252    13.30798
##   Total     100.00000   100.00000
```

Examining differences

- When we're comparing the difference between two quantities, such as survival rates of slasher films, it can be difficult to decide how *different* two values need to be before we can conclude that the difference didn't just happen by chance.
  - To help us decide when a difference is not due to chance, we'll use repeated random shuffling.
- By using repeated random shuffling, we'll estimate how often our *actual* difference occurs by *chance*.

Do the shuffle!

- When we shuffle data, we use our original data set as a starting point.
  - Run the following and write down the resulting table on a piece of paper.

```
tally(survival ~ gender, data = slasher)
```

- Now run the following to randomly reassign each `survival` status to each observation. Compare the resulting table to the one you wrote down.

```
tally(shuffle(survival) ~ gender,
  data = slasher)
```

Let's compare ...

- **How many people, in total, survived the slasher film before shuffling? How many people survived after shuffling?**
- **How has shuffling our data changed the percentage of women who survived compared to men who survived?**
  - **Is the difference in proportions from your shuffled data larger or smaller than the difference from the original data? Interpret what this means.**
- **Explain why shuffling our data one time is not enough to decide if the difference seen in our *actual* data occurs by chance or not.**

Detecting differences

- To help us decide if the difference in percentages in our *actual* data occurs by chance or not, we can use the `do()` function to shuffle our data many times and see how often our *actual* difference occurred by chance.
- Run the following lines of code:

```
set.seed(7)
shuffled_outcomes <- do(10) * tally(shuffle(survival) ~ gender,
  format = "percent", data = slasher)
View(shuffled_outcomes)
```

- **In how many simulations did a higher percentage of males survive than females?**
- **What is the largest difference in percentages of survival between males and females?**
- **What patterns are emerging from these simulations?**
- Ten simulations is not enough. Use `do`, `tally` and `shuffle` functions to `shuffle the` `survival` variable and `tally` the percentage of women who survived 500 times. `Assign` your 500 shuffles the name `shuffled_survivors`. Use `set.seed(1)`

Now what?

- The next step to find out how often our *actual* difference occurs by chance is to compare it to the differences in our shuffled data.
- To compute the differences for each shuffle we can use the `mutate` function.
  - Fill in the blanks to add the difference between `Survives.Female` and `Survives.Male` to our `shuffled_survivors` data.

```
shuffled_survivors <- mutate(shuffled_survivors,
          diff = ____ - ____)
```

Time to decide

- Create a `histogram` of the `differences` in our `shuffled_survivors` data. Based on your plot, answer the following
  - **What was the typical difference in percentages between men and women survivors?**
- Include a vertical line in your histogram of the actual difference by running the code below:

```
add_line(vline = 22.52252 - 13.30798)
```

- **Does the actual difference occur very often by chance alone?**
- **Does `gender` play a role in whether or not a character will survive in a horror film? Explain your reasoning.**
- **If you wanted to survive in a horror film, would you want to play a female character or a male character?**

Summary

- By shuffling the `survival` label, we made it so that the proportion of males and females who survived the slasher film was random.
  - The males and females survived by chance alone.
- If surviving the film occurred purely by chance, then most of the time the difference in survival proportions was close to zero.
  - Notice how most values in the histogram occur close to zero.
- When we look to see how often our actual difference occurs in our shuffled data, if the actual difference doesn't occur very often then perhaps there is something more going on than just chance alone ...

On your own

- Carry out another 500 simulations but this time shuffle the `gender` variable instead of the `survival` variable.
  - Include the code `set.seed(1)` before your 500 simulations to make your answer reproducible.
- **Does shuffling the `gender` variable instead of the `survival` variable change your answer to the question: *Does gender play a role in whether or not a character will survive in a horror film?***
  - **Why or why not?**

Lesson 14: The Titanic Shuffle

Objective:

Students will continue to understand that, just by chance, we will see differences between two groups. They will understand that these differences are usually small.

Materials:

1. *LMR_Titanic Strips*
   Advanced preparation required (see Step 8 of lesson)
2. Poster paper
3. Markers

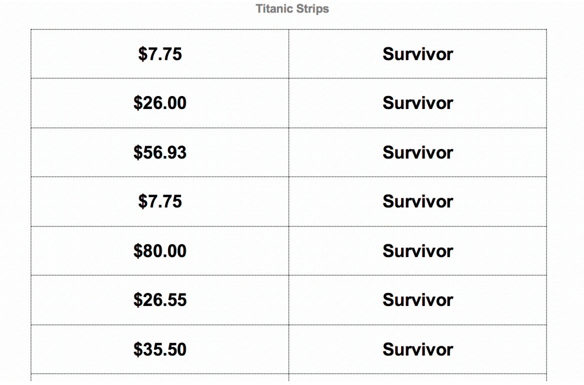
Essential Concepts: We can also "shuffle" data based on numerical variables. The statistic we use is the difference in medians. The distribution we form by this form of shuffling still represents what happens if chance were the only factor at play. When differences are small, we suspect that they might be due to chance. When differences are big, we suspect they might be 'real.'

Lesson:

1. Remind students that they previously learned how to determine if a difference is due to chance by shuffling based on categorical variables (gender and survival).
2. Display the dotplot created during Lesson 13 of the difference in proportions between female and survivors of horror movies. Remind the students that, "by chance," the differences were typically zero. Most of the time, they were pretty small. Sometimes they were bigger, but that was rare and this tells us that if we see "small" differences, we might think they are due to chance. But if we see "big" differences, they are not.
3. Lead a short discussion about what students think small and big differences mean. Make sure they answer in units (which are percentage points for the horror movie data). So, for example, a "big" difference might be 5 percentage points (but don't let them just say "5").
4. Inform students that, during today's lesson, they will learn how to determine if there is a difference between groups when a numerical variable is involved.
5. In particular, they will assume the roles of passengers in the *Titanic* for today's lesson. In case some students may not know about the *Titanic*, ask a volunteer to share what he/she knows.
6. Explain that, at its time, the *Titanic* was the largest cruise ship ever built and was declared to be unsinkable. However, on its first voyage, it sank and was one of the worst maritime disasters in history. About 40% of passengers survived; however, your chances of survival depended very much on your age, gender, and wealth.
7. Inform the students that we are going to look at whether the amount of money a passenger paid for his/her cabin (the fare price) had anything to do with whether or not he/she survived.
8. Each student will need a strip from the *LMR_Titanic Strips* file—see below for instructions.

   Advanced preparation required:

   The Titanic Strips LMR contains data from 40 actual passengers on the titanic. Each strip represents the data from one passenger: the left hand side shows the fare paid and right hand side contains the survival information of that passenger after the collision. Cut the LMR into strips such that the fare price is attached to the survivor status for each of the 40 observations.



Titanic Strips

$7.75	**Survivor**
$26.00	**Survivor**
$56.93	**Survivor**
$7.75	**Survivor**
$80.00	**Survivor**
$26.55	**Survivor**
$35.50	**Survivor**

LMR_Titanic Strips

9. 40 strips were created for large classes. If your class has less than 40 students, assign the students to two groups such that roughly 40% of them are in the survivor group (15/40 = 37.5% ≈ 40%), and the rest are in the victim group. If your class is small (smaller than 10), then put the students in two equal sized groups. The split does not have to be exactly 40%.
10. Inform the smaller group that they are the survivors and distribute a survivor strip with its corresponding fare to each student. Set aside any leftover strips. Tell them that the price on the strip represents the amount of money paid for their ticket to board the *Titanic*. Notify them that $20 in 1912 is worth about $500 today.
11. Divulge to the larger group that they, unfortunately, are the victims and distribute a victim strip with its corresponding fare to each student. Set aside any leftover strips.
12. Ask each group to create a dotplot of their fare prices on a poster. Lead a quick discussion comparing the two dotplots visually. Then, ask each group to calculate the mean fare for their group.
13. As a class, find the difference between the mean fares for the two groups.

    median of “Survivor” fares – median of “Victim” fares

    For example:

    If all 15 survivor cards and all 25 victim cards are used, the difference is medians would be:

    $26.00 – $13.00 = $13.00

14. Explain that one of the controversies of the *Titanic* disaster was that some people felt that the rich people were given better access to the lifeboats than were the poor, so rich people were more likely to survive. Note that the data represented on the fare cards are only a subset of the actual *Titanic* data, which had over 800 passengers. However, the data were randomly selected from the real data and are considered representative of the 800 passengers.
15. In pairs, ask students to discuss the following:
    a. Based on the data from our dotplots, do you think rich people were more likely to survive? In other words, did passengers who paid more for their tickets have a better chance of survival? ***Yes, there is evidence that rich passengers survived more often than poorer passengers. The median difference between the fare prices of the survivors and the victims is $13.00 (see Step 13). Most survivors had higher fare prices than the victims, so the distribution of survivor fares is shifted to the right and is more right-skewed.***
16. Share out a couple of responses with the whole class.
17. Have students tear their strip such that they separate the fare from the outcome (survivor or victim). Collect only the outcomes and randomly shuffle them. Students will keep their fare.

Distribute the shuffled outcome strips face down to the students. Once everyone has a new outcome strip, ask students to turn their outcome strip over and re-group based on their new survival status.

18. Ask the students:

    a. Why do we shuffle the survivor/victim strips and not the fare strips? ***We want to know if the price someone paid for his/her ticket affects whether or not he/she survived. So, when we shuffle, we assume that fare price has nothing to do with survival, so the prices should be irrelevant.***

    b. What do you think the median fare difference of our shuffled groups will be? ***The median fare difference of the shuffled groups should be close to 0, meaning that there should be NO difference in fare price for the survivors and the victims. Everyone would have the same chances of surviving, regardless of their ticket price.***

19. Have each group calculate the median fare price for their new groups. Then, ask:

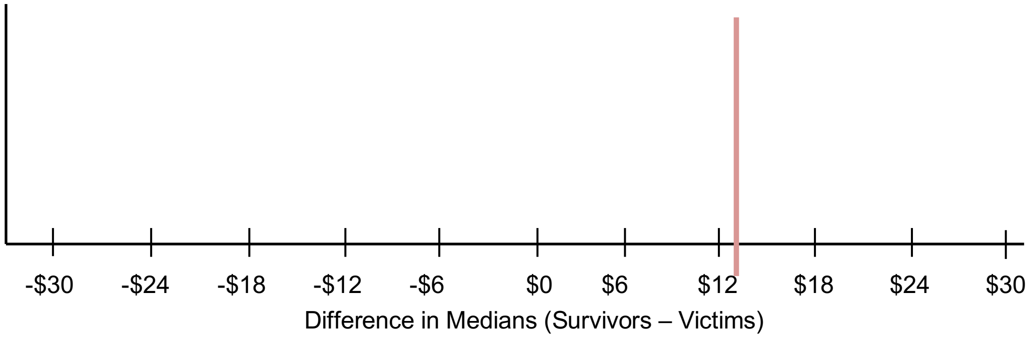
    a. Do you think this difference, of ___ dollars, is real or due to chance? ***Answers will vary by class. Since the data were shuffled, any difference should be due to chance.***

20. On the board, create a table to display the median fare prices for each group, and include a column for the difference (median "Survivor" fare – median "Victim" fare). Fill in the table with the values the students found in Step 13. **Note:** The first row has been filled in with the example data from above BEFORE the shuffles have taken place.

Median Fare Price of Survivors	**Median Fare Price of Victims**	**Difference in Medians (Survivors - Victims)**
$26.00	$13.00	$26.00 - $13.00 = $13.00

21. Note that values in the "Difference in Medians" column can be positive or negative because sometimes the survivors will pay more for their tickets, and other times the victims will pay more for their tickets.

22. Draw a dotplot on the board labeled "Difference in Medians." Include a vertical line at $13.00 (or whatever value was calculated in Step 13 by the class) to represent the actual difference in the median fare prices between the survivors and the victims (see example below).



23. Using the information from Steps 19 and 20, place a dot at the corresponding value for the shuffled data's difference in medians. Ask the students:

    a. How does this difference compare to the actual difference of $13.00 (from Step 13)? ***Answers will vary by class. Most likely, the difference in medians will be much smaller than $13.00. In fact, the difference in medians will be centered around 0.***

24. Remind students that small differences might be due to chance and big differences typically mean that there is a "real" difference between groups. In this case, a big difference might mean that the rich passengers were more likely to survive. And a small difference might mean that survival was just a matter of plain luck.
25. Repeat Steps 17 – 23 a few more times (depending on how much class time you have available).
26. In pairs, ask students to discuss whether they think the real difference in median fare prices they calculated in Step 13 ($13.00 if all cards were used) is small or large. ***Answers will vary by class. Guide students to look at the MAD value of the distribution of differences in median fares.***
27. Explain that one way that we can decide what is "large" or "small" is by creating cut-off values that we think are too far away from the center of the distributions of differences. In general, we can assign a rule that states that any difference in mean fare prices that is greater than 2 MAD values above or below the mean is considered unusual. This means that any value in the outer edges of the plot would indicate that a passenger's ticket price impacted his/her chances of survival.
28. Inform students that they will use RStudio to shuffle the actual *Titanic* data of all 800 passengers during the next class and can decide if the difference in survival rates of rich passengers and poor passengers was real, or just due to chance.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

For the next 3 days, students will collect data for the *Stress/Chill* campaign either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org

LAB 2F: The Titanic Shuffle

Complete Lab 2F prior to Lesson 15.

Lab 2F - The Titanic Shuffle

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Previously ...

- In the previous lab, we learned that by using a `do`-loop and the `shuffle` function, we could simulate randomly shuffling our data many times.
  - This helps us determine how likely it is that a difference between groups is due to chance.
- For this lab, will extend these ideas to *numerical* variables by using random shuffling and numerical summaries.
- The question we will investigate in this lab is:

  Is there any evidence to suggest that wealthier passengers on the Titanic were more likely to survive than poorer passengers?

- We will consider wealthier passengers to be those that paid a higher `fare` for their ticket.

The Titanic

- The Titanic was a ship that sank en route to the U.S.A. from England after hitting an Iceberg in 1912.
  - At the time, it was claimed that the Titanic was *unsinkable* ... it wasn't ... because it did.
- Use the `data` function to load the `titanic` passenger and survival data.
- Create a boxplot of the `fares` paid by passengers and facet the plot based on whether the passenger survived or not.
  - **Based on the plot, do you believe richer passengers were more likely to survive? Explain why and describe how certain you are of being correct.**

The search begins!

- Start your analysis by calculating how much more the *typical* survivor paid than the *typical* non-survivor in our data.
  - Based on the distributions of fares paid, which numerical summary that describes the *typical* value might be preferred?
- **What was the *typical* fare paid by survivors? Non-survivors? How much more did the typical survivor pay?**

Do the shuffle!

- Use the `do` and the `shuffle` functions to shuffle the passenger's survival status 500 times.
  - Use the previous lab if you need some help on how to do this.
  - For each shuffle, compute each group's `median` fare paid.
  - `Assign` your shuffled data the name `shuffled_survival`.
- After shuffling your data, use the `mutate` function to create a variable called `diff` which is the median fare of survivors minus the median fare of non-survivors. (Assign your mutated data the name `shuffled_survival` again).

Put your simulations to use

- **By using your shuffled data, answer the research question we posed at the beginning of the lab.**

Is there any evidence to suggest that wealthier passengers on the Titanic were more likely to survive than poorer passengers?

- **Write up your answer as a statistical analysis. Create a plot and explain how the plot supports your conclusion. Be sure to also explain why shuffling your data is important.**

Comparing Mean Fares

- What about if instead of calculating the median fare price for each group after a shuffle, we calculated the mean fare price and took the difference (mean_survivor – mean_victim).
- **If we did this 500 times, what do you predict the distribution of differences will look like?**
- Use the `do` and the `shuffle` functions to shuffle the passenger's survival status 500 times.
  - For each shuffle, compute each group's mean fare paid.
  - After shuffling your data, use the `mutate` function to create a variable called `diff` which is the mean fare of survivors minus the mean fare of non-survivors.
- **What does the shuffled data reveal? Does the answer to the research question below change when using the mean fares instead of the median fares?**

Is there evidence to suggest that those who survived paid a higher fare than those who died?

Lesson 15: Tangible Data Merging

Objective:

Students will learn how to merge two data sets and ask statistical questions about the merged data.

Materials:

1. *Tangible Data Merging* file (LMR_2.14_Tangible Data Merging)
   Advanced preparation required *(*see Step 4 of lesson*)*
2. Copy paper in two colors
   Advanced preparation required (see Step 4 of lesson)

Vocabulary:

merge

Essential Concepts: We can enhance the context of a statistical problem by merging related data sets together. To merge data, each data set must have a "unique identifier" that tells us how to match up the lines of the data.

Lesson:

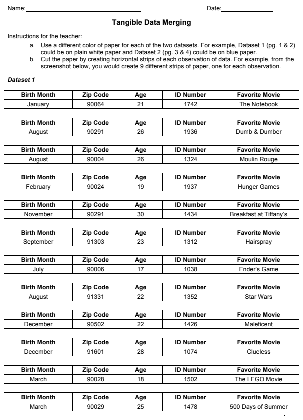
1. Inform students that they are going to examine the research question "Does the personality color test really work?" To answer this, we're going to examine whether the different color groups actually differ on particular beliefs or attitudes, or if these differences might just be due to chance. In particular, we are going to use the *Stress/Chill* data to see if there is evidence that the "colors" actually differ.

2. Show students the variables in each of these data sets. Give students time to brainstorm statistical questions of interest with their teams and record their questions in their IDS journals. Encourage them to think of two- and three-variable questions.

3. Conduct a share out of some of the questions students came up with. Examples include: ***(1) Do people whose predominant color is Gold tend to stress more than people whose predominant color is Blue? (2) Is there a difference between the sorts of things that stress out the different personality colors?***

4. In order to answer the above questions, we will need to merge our class's 2 data sets together (*Personality Color* and *Stress/Chill*). In order to do this, we will be practicing how to merge data sets today.

5. Print out the material from the *Tangible Data Merging* file (LMR_2.14). Use a different color of paper for each of the two data sets. For example, Data set 1 could be on plain white paper and Data set 2 could be on blue paper. Cut the paper by creating horizontal strips of each observation of data. For example, from the screenshot below of the first page of Data Set 1, you would create 12 different strips of paper, one for each observation.



Name:______________________ Date:__________

Tangible Data Merging

Instructions for the teacher:

a. Use a different color of paper for each of the two datasets. For example, Dataset 1 (pg. 1 & 2) could be on plain white paper and Dataset 2 (pg. 3 & 4) could be on blue paper.
b. Cut the paper by creating horizontal strips of each observation of data. For example, from the screenshot below, you would create 9 different strips of paper, one for each observation.

Dataset 1

Birth Month	Zip Code	Age	ID Number	Favorite Movie
January	90064	21	1742	The Notebook

Birth Month	Zip Code	Age	ID Number	Favorite Movie
August	90291	26	1936	Dumb & Dumber

Birth Month	Zip Code	Age	ID Number	Favorite Movie
August	90004	26	1324	Moulin Rouge

Birth Month	Zip Code	Age	ID Number	Favorite Movie
February	90024	19	1937	Hunger Games

Birth Month	Zip Code	Age	ID Number	Favorite Movie
November	90291	30	1434	Breakfast at Tiffany's

Birth Month	Zip Code	Age	ID Number	Favorite Movie
September	91303	23	1312	Hairspray

Birth Month	Zip Code	Age	ID Number	Favorite Movie
July	90006	17	1038	Ender's Game

Birth Month	Zip Code	Age	ID Number	Favorite Movie
August	91331	22	1352	Star Wars

Birth Month	Zip Code	Age	ID Number	Favorite Movie
December	90502	22	1426	Maleficent

Birth Month	Zip Code	Age	ID Number	Favorite Movie
December	91601	28	1074	Clueless

Birth Month	Zip Code	Age	ID Number	Favorite Movie
March	90028	18	1502	The LEGO Movie

Birth Month	Zip Code	Age	ID Number	Favorite Movie
March	90029	25	1478	500 Days of Summer

LMR_2.14

6. Hand each student in the class a strip of paper. Ask them to try to find someone with the other data set (i.e., a person with a different colored strip of paper) that they can "match up," or **merge**, with.
7. For example, a student with the first row of data listed below from Data set 1 might want to match up with the second row of data listed below from Data set 2 because a person who is 21 has probably graduated high school.

Birth Month	Zip Code	Age	ID Number	Favorite Movie
January	90064	21	1742	The Notebook

Zip Code	ID Number	Birth Month	Siblings	Education
91331	1352	August	2	High School

8. However, they should notice that they cannot just make guesses about a person's characteristics in order to match up the data. They should realize that only 3 of the variables are the same in both data sets: *Birth Month*, *Zip Code,* and *ID Number.*
   a. Since multiple people have the same *Birth Month*, discuss why this may not be the best variable to merge with. ***Multiple people are born in January, so we would have no way of differentiating between those people.***
   b. The same is true for the *Zip Codes* variable. Although there are less repeats with *Zip Codes*, we still see some overlap between observations.
   c. So, the only *UNIQUE* identifier in both data sets is *ID Number*. So the students should end up in pairs at the end of the exercise – a student from Data set 1 is matched with the student from Data set 2 that has the same *ID Number*.
9. Have the students write about the experience of tangible data merging in their IDS journals and ask:
   a. Why is it important to have at least one unique identifier for both data sets? ***It is the only way to know which information belongs to which person. We want to make sure we do not match up observations (in this case, people) incorrectly because that will compromise any analysis we do later.***
10. Inform students that they will learn to merge data sets using RStudio during the next lab.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

Students will collect data for one more day for the *Stress/Chill* campaign either through the UCLA IDS UCLA App or via web browser at https://portal.idsucla.org/

LAB 2G: Getting it Together

Complete Lab 2G prior to the Practicum.

Lab 2G - Getting It Together

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Putting data together

- In the labs so far, we've only ever looked at individual data files.
- But often times, we gain additional insights by including additional information from a separate data set.
- In this lab, we will learn how to merge information from our *personality color* data with our *stress/chill* data.
- *Export, upload, import* your *Personality Color* data set and name it `colors`.
- Then, *export, upload, import* your *Stress/Chill* data set and name it `stress`.

Looking at Stress/Chill

- We would like to analyze the research question:

  How do people's personality colors and/or sports participation affect their stress levels?

- We already have data about *personality color* and a separate data set about *stress*.
  - What we don't have is a single data set with information from both ... yet.
- We'll start then by strategizing how to merge our data together.

Deciding how to merge

- Before we merge data, we need to decide *how* we plan to merge it:
- We can *stack* our data sets, that is, take one data set's rows and add them to the bottom of the other data set.
- We can also *join* our data sets horizontally. This is where we take one data set's columns and add them to the end of the other data set's columns based on matching an *ID* variable.
  - The *ID* variable will have entries that we use to *match* observations in both data sets.
- **To answer the statistical question of interest, would it make more sense to *stack* or *join* our `colors` and `stress` data?**

Finding variables in common:

- Look at the `names` of the variables in each data set.
  - To merge different data sets together, we need to find variables they have in common.
- **Which variables do the data sets have in common?**
- **Which variable would make sense to merge the data sets together with? Why not the others?**

Caution required

- Whether *stacking* or *joining*, we need to be careful when we merge data:
- When *stacking* data, we need to be absolutely certain that the variables we're stacking represent the exact same measurements.
  - We wouldn't want to stack `height` in meters and `height` in inches, for instance (without converting one to the other).
- When *joining* data, we need to make sure that the *id* variable in our primary data set matches to *one and only one* observation in the joining data.
  - Otherwise, `R` won't know which observation to match to.

Getting ready

- Our goal is to add the variables from the `colors` data onto the `stress` data.
- Start by ensuring that every `user.id` in the `colors` data is unique.
  - If there's a duplicate, have your teacher remove the duplicate from the IDS *Response Manager* and then re-*export*, *upload*, *import* your `colors` data.
- **After we add the data from *colors* to *stress*, how many rows should our merged data have? Write this number down.**

Putting them together

- We can use the `merge` function to *join* our data sets together using the variables that appear in both sets.
- **Fill in the blanks below to join the information from the `colors` data onto the `stress`.**

```
merge(____, ____, by = "____")
```

- `Assign` this `merged` data set the name `stress_colors`.
  - Make sure your data has the same number of observations that you wrote down on the previous slide.

Saving your data:

- `View` your merged data and make sure nothing appears to be blatantly wrong with it.
- **Why didn't we stack the rows of data instead?**
- **What happens if you swap the order of the data sets in the `merge` function?**
- Fill in the blank below to `save` our `stress_colors` data for later use.

```
save(stress_colors, file = "stress_colors.rda")
```

- Be sure to look in the *Files* tab to make sure your data was saved.

Moving on

- In the next lab, we'll begin analyzing our merged data. In the meantime:
- **Make a few plots using variables from the `stress` data and *facet* or *group* the plots based on variables from the `colors` data.**
  - **Write down the most interesting discovery you make by just exploring your data. Write out how you found your discovery and interpret what it means for the people in your class.**
- **With our *colors* data, we could answer questions about the *typical* color scores in your class. Why can we no longer answer this question in our `stress_colors` data?**

Practicum: What Stresses Us?

Objective:
Students will use RStudio to make graphical representations or numerical summaries of their *Stress/Chill* and *Personality Color* data to answer research questions.

Materials:

1. *What Stresses Us? Practicum* (LMR_U2_Practicum_What Stresses Us)

Practicum
What Stresses Us?

We made a data set that combined our *Stress/Chill* data with our *Personality Color* data. You will use this data to answer the following research questions:

- Do color personalities really predict a person's personality?
- Do people with different personality colors tend to have different stress levels?

Based on the merged data, you need to:

1. Write a one-page report to address these research questions. Use the Data Cycle. Your analysis should include both numerical methods (means, medians, etc.) and graphical methods (plots). The research questions are fairly broad, and you should first think of simpler statistical questions you could ask that would address these research questions.
2. In your report, be sure to:
   a. Provide the plot(s) and numerical summary (or summaries).
   b. Describe what the plot shows.
   c. Explain why you chose to make that particular plot.
   d. Explain how the plot and numerical summary answers your statistical question.
3. Present your report to another member of the class who is not in your team.
   a. Make sure to include any relevant plots or numerical summaries that you use.

   Note: You may use the scoring guide in Unit 1 to give you an idea of how to score the Practicum.

What's Normal?

Instructional Days: 5

Enduring Understandings

Students learn that the Normal curve can be used as a model that describes many real phenomena. Drawing plots of the Normal curve over histograms helps data scientists determine if the distribution represented by the histogram is close to Normal. The Normal curve suggests that one is more likely to obtain values that are close to typical (average), which are found in the center of the curve, and less likely to obtain values that are extreme and farther away from typical.

Engagement

Students will learn about the Normal curve by watching the first 35 seconds the New York Times Video "Bunnies, Dragons, and the Normal World" found at: http://www.nytimes.com/video/science/100000002452709/bunnies-dragons-and-the-normal-world.html.

Learning Objectives

Statistical/Mathematical:

S-ID 4: Use the mean and standard deviation of a data set to fit it to a normal distribution and to estimate population percentages. Understand that there are data sets for which such a procedure is not appropriate. Use calculators and RStudio to estimate areas under the normal curve.

S-IC 6: Evaluate reports based on data.

Data Science:

Learn to eyeball Normal distributions and overlay a Normal curve on a histogram; learn to simulate draws from a Normal distribution, and the impact of sample size; learn that estimating probabilities with a model leads to stable estimates; and estimate probabilities by finding the area under the Normal curve using RStudio.

Applied Computational Thinking Using RStudio:

- Use software to find the area under a Normal curve
- Use software to compare sample distributions (with histograms, for example) with the Normal distribution and make a decision as to whether the distribution appears Normally distributed.
- Draw random samples from a Normal distribution using software.

Real-World Connections:

The Normal curve is used to make inferences about a population. The model makes it possible to estimate the probability of occurrence of any value of a Normally distributed variable. For example, heights are Normally distributed. Using a Normal curve, we can find the probability of that a person would be a height of 6' 2".

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

3. Students will use complex sentences to write informative short reports that use data science concepts and skills.

Data File or Data Collection Method

Data Files:

1. CDC data (`cdc`)
2. Titanic data (`titanic`)

Legend for Activity Icons

🎥	💬	📜	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 16: What is Normal?

Objective:

Students will learn what a Normal distribution is and learn how to identify a Normal distribution.

Materials:

1. Video: *New York Times'* "Bunnies, Dragons, and the Normal World" found at: http://www.nytimes.com/video/science/100000002452709/bunnies-dragons-and-the-normal-world.html

   Note: Show only the first 41 seconds of the video.
2. Graphics from the *Normal Plots* file (LMR_2.15_Normal Plots)
3. Projector to display plots
4. 3" x 5" cards (1 per student)

Vocabulary:

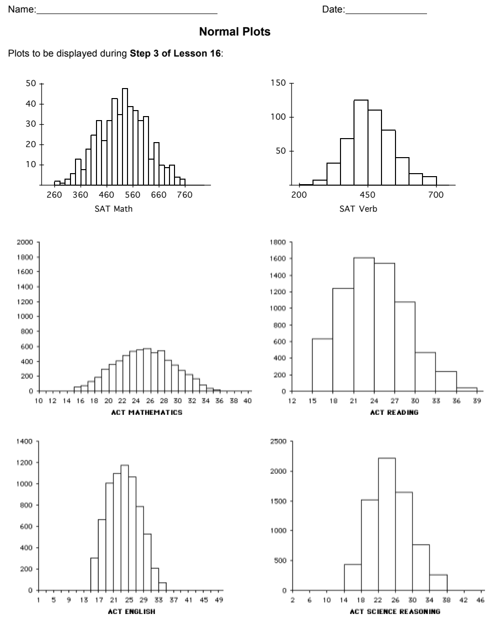
bell-shaped, normal curve, normal distribution

Essential Concepts: The Normal curve, also called the Gaussian distribution and the "bell curve," is a model that describes many real-life distributions and is usually called the Normal Model.

Lesson:

1. Remind students that in Unit 1, Lesson 11 (*What Shape Are You In?*), they sorted histograms into groups based on their shapes.
2. The *Normal Plots* file (LMR_2.15) contains some of the unimodal **bell-shaped** distributions from the original handout of that lesson (*Sorting Histograms* handout (LMR_1.10)).

   Note: You do not need the original handout from Unit 1 – all relevant plots have been compiled in the *Normal Plots* file (LMR_2.15) for accessibility. Six plots are included: SAT Math, SAT Verb, ACT Mathematics, ACT Reading, ACT English, and ACT Science Reasoning.



Name:______________________________ Date:______________

Normal Plots

Plots to be displayed during **Step 3 of Lesson 16**:

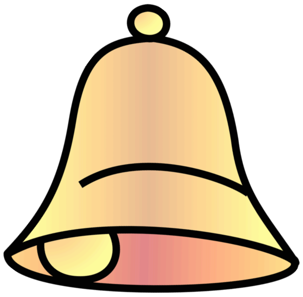
LMR_2.15

3. Display the group of bell-shaped distributions from page 1 of the *Normal Plots* file (LMR_2.15) to the class and ask the students:

    a. What characteristic does this particular group share? ***All of these plots are unimodal (one mode/peak) and symmetric.***

    b. Inform students that these types of distributions are often referred to as bell-shaped. Why might this term be used? ***The histograms look very similar to the shape of a bell.***

4. To show the similarities between the shape of a bell and the shape of these distributions, a clip-art image (shown here) has been included in the *Normal Plots* file (LMR_2.15) on page 2.



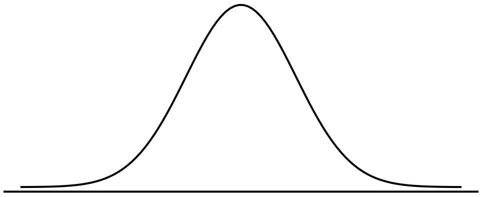
5. Explain that this shape occurs often in real-life. It occurs so often that it's been given its own name: the **normal curve**, or **normal distribution**. Can the students think of distributions where they have seen Normal curves in previous labs?

6. To give some more background on the normal distribution, play the New York Times video titled "Bunnies, Dragons, and the Normal World" found at: http://www.nytimes.com/video/science/100000002452709/bunnies-dragons-and-the-normal-world.html

    Note: Show only the first 41 seconds of the video.

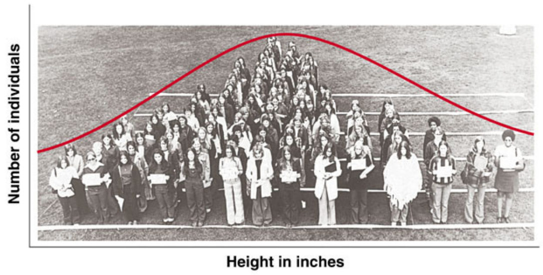
7. Discuss that the normal curve has a very precise mathematical definition, which is pretty complex. But the result is a curve that looks like the one in the "Bunnies" video. In general, the curve looks like the plot shown below.

    Note: You can either draw the diagram below on the board or display it via a projector – the image can be found on page 2 of the *Normal Plots* file (LMR_2.15).



8. Explain that normal distributions are good for describing some populations of people. For example, people's heights are often considered to be normally distributed. Display the famous Frank Anscombe photograph (shown below) via a projector. The graphic can be found on page 2 of the *Normal Plots* file (LMR_2.15). Inform the students that this photo was taken of a group of randomly selected college women who stood in height order.

Tobin/Dusheck, Asking About Life, 2/e
Figure 16.6



9. Next, lead a discussion about why the normal curve is a good fit to the histogram in the above picture. Notice that more people are near the center of the distribution, and fewer are in the outer edges, or tails. Engage the class in a conversation using the following probing questions:
    a. Notice that there is a peak in the center of the distribution. What height do you think is at the center? ***The average height of American women is approximately 5’5” (5 feet, 5 inches) tall. We might therefore expect the average height of this group to be close to 5'5" as well.***
    b. Why are more people in the center, and less people in the edges, or tails, of the distribution? ***The center represents the mean height. Most women will fall somewhere close to the mean and may be a few inches shorter or taller than it. However, less people are likely to be MUCH shorter or MUCH taller than the mean. For example, we would not expect to see many women who are 4’10” tall nor would we expect to see many women who are 6’0” tall.***

10. Explain that the normal curve is a good description of a distribution when it makes sense that there is a single 'typical' value with random deviations above and below that value. Ask students:
    a. Why does this make sense with heights but not with incomes? ***With heights, we expect most people to be near the average, with some deviations above and below the mean (people who are taller or shorter than the mean height). However, with incomes, the distribution will have more deviations that are above the typical value, since there is no upper bound for a maximum income (ex. Bill Gates, Warren Buffett, etc.***
    b. Are there more real-life examples, other than height, that students think might follow a normal distribution? ***Answers will vary by class. Some examples include: (1) scores on standardized math and reading tests (like the SAT and ACT), (2) IQ scores, and (3) body temperatures.***
    c. Does it matter that the curve drawn on the photograph does not match exactly to the women’s heights? ***No. We often refer to the curve as the “normal model” because the curve is a just a model of the true population distribution. So, even though the red curve is not exactly the same as the women’s heights, it is a close enough approximation of the shape of their heights.***

Note to teacher: The main role the normal distribution has historically played has been in modeling errors (most measurements will be close to the actual value while larger errors occur less often) and sample means.

11. Inform the students that, during the next few lessons, they will be learning more about the normal distribution. In particular, they will learn about a new measure of spread used to describe a normal distribution, how to calculate probabilities from this distribution, and how to randomly sample from this distribution.

12. *Cheat Card*: Distribute an index card to students and ask them to create a cheat card that will help them remember information about the normal curve.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students will complete their cheat cards if they were not able to finish in class.

Lesson 17: A Normal Measure of Spread

Objective:

Students will learn that standard deviation is another way to measure variability.

Materials:

1. *How Far Apart?* handout (LMR_2.6_How Far Apart) – completed during Lesson 4
2. *How Far Apart? (with standard deviation – SD)* handout (LMR_2.16_How Far Apart SD)
3. Projector to display visuals using RStudio
4. RScript with the functions in this lesson

Vocabulary:

standard deviation (SD)

> **Essential Concepts:** The standard deviation is another measure of spread. This is commonly used by statisticians because of its role in common models and distributions, such as the Normal Model.

Lesson:

1. In their IDS journals, ask students to create a two-column table and label the left-column as *Measures of Center (Central Tendency)* and the right column as *Measures of Spread (Dispersion).*
2. In pairs, ask students to recall methods they have learned so far for measuring center and measuring spread in distributions.

   Measures of Center: ***mean (average or typical value), median***
   Measures of Spread: ***mean absolute deviation (MAD), interquartile range (IQR)***

3. Share out a pair's explanation and ask the rest of the pairs to agree or disagree. If there is disagreement, hold a class discussion until the lists are correct.
4. Point out that a measure of center or a measure of spread depicts one value for a distribution. Ask student pairs to discuss the following question:
   a. What does the value of each measure tell us about the data in the distribution? ***Possible answer: A measure of center tells us the value that is typical, or in the center. A measure of spread tells us how variable, or how spread apart, the data are.***
5. Next, ask students to add the term **standard deviation (SD)** to their *Measures of Spread* column.
6. Inform students that the standard deviation of a distribution is another way to measure spread, or variability. The standard deviation is similar to the mean absolute deviation (MAD).
7. Ask students to recall the formula for calculating the MAD:

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

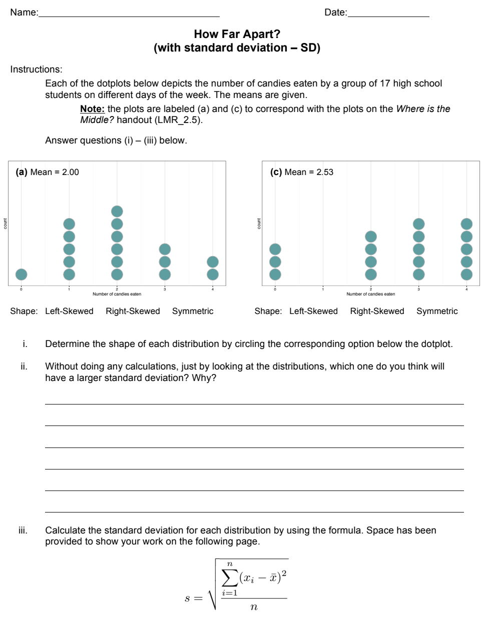
8. While the MAD measures the absolute distance of each data point from the mean, the standard deviation squares the distances of each data point from the mean. Both methods result in positive measurements because distance is always positive.
9. Ask students to recall that they calculated MAD values in the *How Far Apart?* handout (LMR_2.6) during Lesson 4 of this unit.

10. Show and discuss the formula for calculating the standard deviation of a data set:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

<u>Note to teacher:</u> There are different formulas for the standard deviation. We are presenting the simpler one, which divides by *n*. In AP Statistics (or college introductory statistics), students will learn that if they are using a sample of data to estimate the standard deviation for the population, then dividing by $n-1$ is a better estimator than dividing by *n*. But this technically requires a lot of scaffolding and leads to little understanding, and so we will stick with the simpler version. (In some books, this is called the “population value of the standard deviation” and the $n-1$ version is called the “sample estimate of the standard deviation.”)

11. Guide the class to complete the *How Far Apart? (with standard deviation -- SD)* handout (LMR_2.16) to calculate standard deviations of the dotplots using the formula listed above.



Name:______________________________ Date:______________

How Far Apart?
(with standard deviation – SD)

Instructions:

Each of the dotplots below depicts the number of candies eaten by a group of 17 high school students on different days of the week. The means are given.

<u>Note:</u> the plots are labeled (a) and (c) to correspond with the plots on the *Where is the Middle?* handout (LMR_2.5).

Answer questions (i) – (iii) below.

Shape: Left-Skewed Right-Skewed Symmetric

Shape: Left-Skewed Right-Skewed Symmetric

i. Determine the shape of each distribution by circling the corresponding option below the dotplot.

ii. Without doing any calculations, just by looking at the distributions, which one do you think will have a larger standard deviation? Why?

____

____

____

____

____

____

iii. Calculate the standard deviation for each distribution by using the formula. Space has been provided to show your work on the following page.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

LMR_2.16

Answers: Plot (a) – SD = 1.0847 candies; Plot (c) – SD = 1.3770 candies

12. As a whole group, ask students to compare and contrast the standard deviations with the MAD values for the two plots in the handout.

Similarities between SD and MAD:

- ***Measure the same idea: variability, or spread***
- ***Are based on looking at the "deviations" from the mean: the difference between an observation and the mean***
- ***Uses the "typical" deviation***

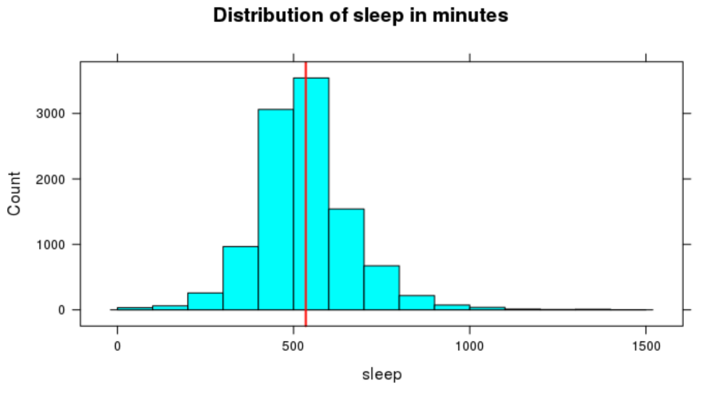
Differences between SD and MAD:

- ***The MAD uses the absolute value and finds the average of the absolute deviations from the mean***
- ***The SD uses the square of each deviation from the mean, and finds the average of the squares***
- ***The SD takes the square root of the average of the squares***

13. Ask students why they think the SD takes the square root of the average of the squares. ***Possible response: Taking the square root of the average of the squares returns the measurements to their original units instead of square units*.**

14. To reinforce students' conceptual understanding of standard deviation, student teams will estimate the standard deviation for a few numerical distributions and explain the reasoning for their estimate. Load and view the atus data, then run the following functions one by one:

```
> histogram(~sleep, data=atus, breaks=seq(0,1500,by=100),
                main = “Distribution of sleep in minutes”)
> sleep_mean<-mean(~sleep, data=atus)
> add_line(vline=sleep_mean)
```



15. Zoom in on the visual and give student teams a few minutes to discuss what they estimate the standard deviation to be. Have the reporter from each team report their estimate using the following sentence frame:

    “The time spent sleeping (in minutes) typically varies from the mean by ____________minutes.”

16. Reveal the actual standard deviation by running the function:

```
> sd(~sleep, data=atus)
```

17. Choose a reporter from a student team that had a good approximation to explain their reasoning.

18. Repeat this process with a few more numerical variables. Functions are provided below.

    Household size

```
> histogram(~household_size, data=atus, nint=13)
> household_mean<-mean(~household_size, data=atus)
> add_line(vline=household_mean)
```

    “Household sizes typically vary from the mean by ____________people.”

```
> sd(~socializing, data=atus)
```

    Socializing

```
> histogram(~socializing, data=atus, breaks=seq(0,2000,by=100))
> social_mean<-mean(~socializing, data=atus)
> add_line(vline=social_mean)
```

    “The time spent socializing (in minutes) typically varies from the mean by ____________minutes.”

```
> sd(~socializing, data=atus)
```

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 18: What's Your Z-Score?

Objective:

Students will understand that a z-score can be used to measure how far away - or how many standard deviations - an observation is away from the mean. Typically z-scores will range between -3 and +3. For simulations involving shuffling, if we compute a z-score that lies far away from the mean, then we might conclude that the outcome was not due to chance. If we see a z-score that lies close to the mean, then we might conclude it was by chance.

Materials:

1. Projector to display RStudio function
2. *RScript with all of the functions in this lesson*
3. *A ruler with centimeter marks on it*

Vocabulary:

z-score, standardized score, Empirical Rule

Essential Concepts: z-scores offer us a way to measure how extreme a value is, regardless of the units of measurement. Typically z-scores will range between -3 and +3, so values that are at or are more extreme than -3 or +3 standard deviations are considered extremely rare.

Lesson:

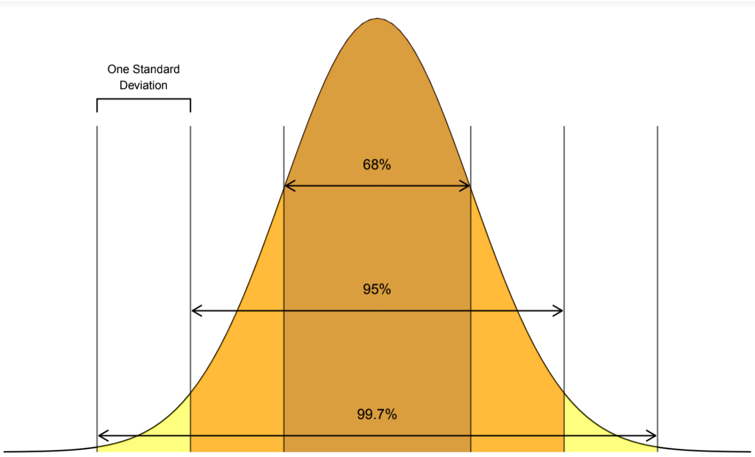
1. Ask students to recall what they remember about normal distributions.

   Answer: Normal distributions are unimodal and symmetric, and are often referred to as bell-shaped. Some real-life examples of variables that produce normal distributions are people's heights, scores on standardized tests, and body temperatures.

2. Display the following statement to students: "All normal distributions are bell-shaped, but not all bell-shaped distributions are normal." Then inform students that normal distributions have special properties.

3. Display the image below and introduce the **Empirical Rule**, which states:

   - Approximately 68% of the observations in a normal distribution fall within one standard deviation of the mean
   - Approximately 95% of the observations in a normal distribution fall within 2 standard deviations of the mean
   - Approximately 99.7% of the observations in a normal distribution fall within 3 standard deviations of the mean



4. Open RStudio and project for students to see. Load the babies dataset, named **Gestation**, by following these steps:

   - Enter **data(Gestation)** in the Console
     - You should see **Gestation** loaded into your Environment
   - Enter **View(Gestation)** in the Console

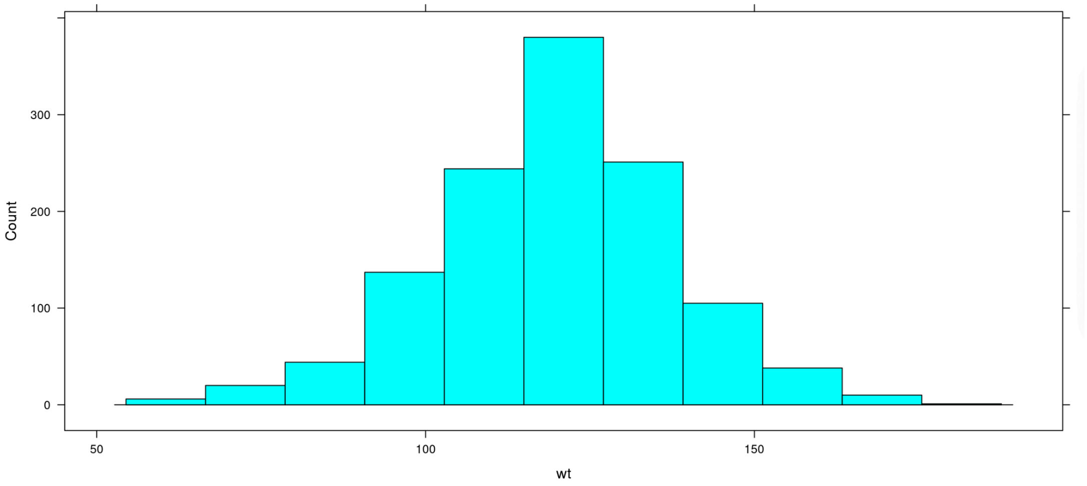
   Display the help documentation by typing **?Gestation** in the Console to consider the variables. Ask student teams to predict which of the variables in the "Gestation" dataset they think might be normally distributed. Choose a couple of teams to share out.

   Description of numerical variables:

   - wt – birth weight (in ounces)
   - gestation – length of pregnancy (in days)
   - parity – 0 if baby was first born, 1-13 otherwise
   - age – mother's age (in years)
   - ht – mother's height (to the last completed inch)
   - wt.1 – mother's weight (in lbs.)
   - dage – father's age (in years)
   - dht – dad's height (to the last completed inch)
   - dwt – father's weight (in lbs.)

5. Create histograms using the variables shared by student teams. There are a few variables that look normally distributed, such as birth mother's heights. We will investigate the babies' birth weights.

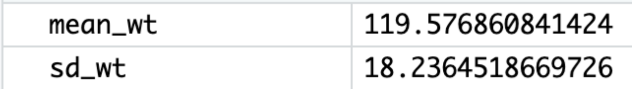
```
histogram(~wt, data = Gestation)
```



6. Ask students:
   a. Does the distribution of baby birth weights look approximately normal? Explain. ***Answer: The distribution of baby birth weights is unimodal, roughly symmetric, and somewhat bell-shaped, so it might be approximately normal.***
   b. What do you approximate the mean weight of the distribution to be? How about the standard deviation? ***Answers will vary. Use this as a check for understanding of standard deviation as well as estimating the mean using the balancing point concept. See next step for calculating the actual mean weight and standard deviation.***

7. Use RStudio to calculate the actual mean and standard deviation.

```
mean_wt <- mean(~wt, data = Gestation)

sd_wt <- sd(~wt, data = Gestation)
```



mean_wt	119.576860841424
sd_wt	18.2364518669726

8. Have students draw a number line with seven equally spaced intervals and label it “Baby birth weight in ounces.” Make sure students leave about 5 centimeters of space above the number line to draw a normal curve. Have students label the middle tick mark with the mean baby weight (round to the nearest tenth of an ounce=119.6 ounces). Then ask students:

   a. What weight is one standard deviation above the mean? ***Answer: A baby whose weight is 137.8 ounces is one standard deviation above the mean baby weight.***
   b. What weight is one standard deviation below the mean? ***Answer: A baby whose weight is 101.4 ounces is one standard deviation below the mean baby weight.***

   Have students label their number line with these values.

9. Have students continue filling their number line with the corresponding weights that are two and three standard deviations from the mean. ***Answer: A baby who weighs 156 ounces is two standard deviations above the mean weight, and a baby who weights 174.2 ounces is three standard deviations above the mean weight. A baby who weighs 83.2 ounces is two standard deviations below the mean weight, and a baby who weighs 65 ounces is three standard deviations below the mean weight.***

10. Ask students: If the distribution of baby weights is approximately normal, what percentage of babies weigh between 101.4 and 137.8 ounces? ***Answer: If the distribution of baby weights is approximately normal, about 68% of babies should be between 101.4 ounces and 137.8 ounces.***

11. Use RStudio to confirm if indeed the distribution of baby weights is approximately normal.

```
> one_sd_wt <- filter(Gestation, wt > 101.4, wt < 137.8)
```

   Answer: In this sample of 1236 observations, there are 861 babies whose weights are one standard deviation from the mean, so 861/1236 = 0.697. This means that around 69.7% of the weights of babies in this sample fall within one standard deviation from the mean baby weight. This is close to 68%, so it seems that the distribution of baby weights is approximately normally distributed.

   Note: If you continue this process for this sample, you will find that the distribution of baby weights is normally distributed as defined by the Empirical Rule. In this sample, 1171/1236 = 94.7% of the baby weights fall within two standard deviations of the mean, and 1229/1236 = 99.4% of the baby weights fall within three standard deviations of the mean.
   .

12. Now that it has been verified that a normal distribution is an appropriate model for this distribution, have students draw a normal curve above the number line. Suggested method to obtain a decent normal curve:
   - Step 1: Draw a dot 4 centimeters above the mean height
   - Step 2: Draw dots 2.4 cm above the heights that are 1 standard deviation from the mean
   - Step 3: Draw dots 0.36 cm above the heights that are 2 standard deviations from the mean

- Step 4: Draw dots right above the number line for the heights that are 3 standard deviations from the mean
- Step 5: Connect the dots with a smooth curve

13. Tell students that we are using this normal curve as a model to represent the distribution of all baby weights. This will allow us to make comparisons, draw conclusions, and make predictions about baby weights. Let's see:
    a. What percentage of babies weigh less than 119.6 ounces? Explain. ***Answer: About 50% of babies weigh less than 119.6 ounces. Since normal distributions are symmetric, the mean and the median are about the same. Since the median divides a distribution into equal halves, in this case so does the mean.***
    b. What percentage of babies weigh between 119.6 and 137.8 ounces? ***Answer: About 34% of babies weigh between 119.6 and 137.8 ounces. According to the Empirical rule, 68% of the observations fall within one standard deviation of the mean, and since normal distributions are symmetric, the area under the curve from the mean to one standard deviation is half of 68% or 34%.***
    c. What percentage of babies weigh more than 137.8 ounces? ***Answer: About 16% of babies weigh more than 137.8 ounces. From part a and b above, we know that 50%+34%=84% of babies weigh less than 137.8 ounces, so 100%-84%=16% weigh more than 137.8 ounces.***

14. Explain that statisticians use something called a **z-score** to compare values. A z-score tells us how many standard deviations away from the mean an observation is. Another name for z-score is a **standardized score**.

15. Introduce the formula for calculating a z-score and discuss what each symbol in the formula means.

$$z = \frac{x - \bar{x}}{s}$$

16. Explain that z-scores answer the question: "How typical is x?" If x is the same as the typical value (the mean), then z = 0. If x is one standard deviation away from the mean, then z = -1 or +1. Remind students from the normal curve that as you move farther from the center (from the mean), there are fewer observations. Therefore, a large z-score is considered an unusual value.

17. Have students calculate the z-score for a baby that weighs 100 ounces:

    z = (100 – 119.6) / 18.2 = -1.08

    Ask the class:

    a. What does a negative z-score mean? ***A negative z-score means the x value is below the mean. This means that the weight is below average.***
    b. What does a positive z-score mean? ***A positive z-score means the x value is above the mean. This means that the weight is above average.***
    c. What is the most negative z-score you think we will find? What is the most positive z-score? ***Typically, values in a normal distribution rarely fall outside two or three standard deviations from the mean. So, if our data is purely by chance, we probably won't see any values that are less than -3 or greater than +3.***

18. Ask students: "Where does a baby that weighs 100 ounces fall within the distribution of baby weights?" Have students find 100 ounces on the x-axis of the normal curve and draw a vertical line from the x-axis until it intersects the normal curve. Have them shade the area under the curve to the left of the vertical line.

19. Tell students that the shaded area represents a percentile in the distribution. A percentile is the exact value in which the desired proportion of observations lie below the specific value in a distribution. Use RStudio to calculate the percentile.

```
pnorm(100, mean = 119.6, sd = 18.2) = 0.140
```

20. Doctors report percentiles to describe a child's development compared to other children their age. For a baby that weighs 100 ounces, a doctor would report the following: "The baby is at the 14th percentile in weight." This means that the baby weighs more than 14% of all babies.

    Note: A z-score can also be used to calculate a percentile, but since a z-score is a standardized score, the mean of the distribution would be zero and the standard deviation would be one.

```
pnorm(-1.08, mean = 0, sd = 1) = 0.140
```

21. Inform the class that they will be using RStudio during the next few days to practice using normal models.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next 2 Days

LAB 2H: Eyeballing Normal

LAB 2I: R's Normal Distribution Alphabet

Complete Labs 2H and 2I prior to the End of Unit Design Project.

Lab 2H - Eyeballing Normal

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

What's normal?

- The *normal distribution* is a curve we often see in real data.
  - We see it in people's blood pressures and in measurement errors.
- When data appears to be *normally distributed*, we can use the *normal model* to:
- Simulate *normally distributed* data.
- Easily compute probabilities.
- In this lab, we'll look at some previous data sets to see if we can find data that are roughly normally distributed.

The normal distribution

- The normal distribution is *symmetric about the mean*:
  - The `mean` is found in the very center of the distribution.
  - And the curve looks the same to the left of the mean as it does on the right.
- Use the following to draw a normal distribution:

```
plotDist('norm', mean = 0, sd = 1)
```

The mean and sd of it

- To draw a normal curve, we need to know exactly 2 things:
  - The `mean` and `sd`.
- The `sd`, or *standard deviation*, is a measure of spread that's similar to the `MAD`.
- **Which part of the normal curve changes when the value of the `mean` changes?**
- **Which part of the normal curve changes when the value of the `sd` changes?**
- *Hint*: Try changing the `mean` and `sd` values in the `plotDist` function.

Finding normal distributions

- Load the `cdc` data and use the `histogram` function to answer the following:
- **Think about the `height` and `weight` variables. Based on what you know about these variables, which of the variables do you think have distributions that will look like the normal distribution?**
  - **Make histograms of these variables. Which ones look like the normal distribution?**
  - *Hint*: To help answer this question, try including the option `fit = "normal"` in the histogram function. You might also try faceting by `gender`.

Using normal models

- Data scientists like using normal models because it often resembles real data.
  - *But not EVERYTHING is normally distributed.*
- As a data scientist in training, you must decide when a normal model seems appropriate.
  - No model is ever perfect 100% of the time.
  - If you choose a model, you should be able to justify why you chose it.

On your own

- **For each of the following, determine which, if any, appear to be normally distributed. Explain your reasoning:**
- Hint: Refer to Lab 2E and 2F
  - **The difference in `percentages` between male and female survivors in a slasher film for 500 random shuffles.**
  - **The difference in `median` fares between survivors and non-survivors on the Titanic for 500 random shuffles.**
  - **The difference in `mean` fares between survivors and non-survivors on the Titanic for 500 random shuffles.**

Lab 2I - R's Normal Distribution Alphabet

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Where we're headed

- In the last lab, you were able to overlay a normal curve on histograms of data to help you decide if the data's distribution is close to a normal distribution.
  - We also saw that calculating the mean of random shuffles also produces differences that are normally distributed.
- In this lab, we'll learn how to use some other R functions to:
  - Simulate random draws from a normal distribution.
  - Calculate probabilities with normal distributions.

Get set up

- Start by loading the `titanic` data and calculate the `mean` `age` of people in the data but `shuffle` their `survival` status 500 times.
  - `Assign` this data the name `shfls`.
- After creating `shfls`, use `mutate` to add a new variable to the data set. This new variable should have the name `diff` and should be the `mean` `age` of those who survived minus those who died.
- Finally, calculate the `mean` and `sd` of the `diff` variable.
  - `Assign` these values the name `diff_mean` and `diff_sd`.

Is it normal?

- Before we proceed, we need to verify that our `diff` variable looks approximately normally distributed.
  - **Is the distribution close to normal? Explain how you determined this. Describe the center and spread of the distribution.**
  - **Compute the mean difference in the age of the *actual* survivors and the actual non-survivors.**

Using the normal model

- Since the distribution of our `diff` variable appears normally distributed, we can use a normal model to estimate the probability of seeing differences that are more extreme than our actual data.
  - **Draw a sketch of a normal curve. Label the mean age difference, based on your shuffles, and the actual age difference of survivors minus non-survivors from the actual data. Then shade in the areas, under the normal curve, that are *smaller* than the actual difference.**
- Fill in the blanks to calculate the probability of an even smaller difference occurring than our actual difference using a normal model.

```
pnorm(____, mean = diff_mean, sd = ____)
```

Extreme probabilities

- The probability you calculated in the previous slide is an estimate for how often we expect to see a difference smaller than the actual one we observed, by chance alone.

- If you wanted to instead calculate the probability that the difference would be larger than the one observed, we could run (fill in the blanks):

```
1 - pnorm(____, mean = diff_mean, sd = ____)
```

Simulating normal draws

- We can simulate random draws from a normal distribution with the `rnorm` function.
  - Fill in the blanks in the following two lines of code to simulate 100 heights of randomly chosen men. Assume the `mean` height is 67 inches and the `standard deviation` is 3 inches.
  - Plot your simulated heights with a `histogram`.

```
draws <- rnorm(____, mean = ____, sd = ____)

histogram(draws, fit = ____)
```

P's and Q's

- We've seen that we can use `pnorm` to calculate *probabilities* based on a specified *quantity*.
  - Hence, why we call it "P" norm.
- Now we'll see how to do the opposite. That is, calculate the *quantity* for a specific *probability*.
  - Hence why we'll call this a "Q" norm.
- How tall can you be and still be in the shortest 25% of heights if the mean height is 67 inches with a standard deviation of 3 inches?

```
qnorm(____, mean = ____, sd = ____)
```

On your own

Conduct one of the statistical investigations below:

- Using the `titanic` data, answer the following statistical question:
  - **Were women on the Titanic typically younger than men?**
  - **Use a histogram, 500 random shuffles and a normal model to answer the question in the bullet above.**
- Using the `cdc` data
  - **Using 500 random shuffles and a normal model, how much taller would the typical male have to be than the typical female in order for the difference to be in the upper 1% by chance alone?**
  - **How can we use this value to justify the claim that the average Male in our data is taller than the average Female?**

<u>End of Unit Design Project and Oral Presentation: Asking and Answering Statistical Questions of Our Own Data</u>

Objective:

Students will apply their learning of the first and second units of the curriculum by completing an end of unit design project.

Materials:

1. *IDS Unit 2 – Design Project and Oral Presentation* (LMR_U2_Design Project)

End of Unit 2 Design Project and Oral Presentation: Asking and Answering Statistical Questions of Our Own Data

Available data sets:

1. Food Habits
2. Time Use
3. Stress/Chill
4. Personality Color

Your mission is to ask and answer a statistical question using at least one data set above.

1. Your question must include a comparison of two distinct groups.
2. Your analysis should address whether any observed differences are real or could be simply due to chance.
3. You should use at least two of the following methods to answer your question with appropriate explanations:
   - ❑ Merge data
   - ❑ Create simulations
   - ❑ Calculate probabilities based on simulations
   - ❑ Use a Normal model
   - ❑ Shuffle/permute data

You will have 5 days to complete this project with your assigned partner. You need to:

- ❑ Prepare an oral presentation (both partners need to participate) that includes:
  - o A 4-slide, 5-minute presentation
  - o An explanation of why you think your statistical question is interesting.
  - o An interpretation of supporting plots and summaries that answer your question.
  - o A reasoning of whether you think the outcome might be due to chance.

- ❑ Submit a 2 -4 page typed, double-spaced summary of your analysis.

<u>Project Assignment Sequence</u>:

- ❑ Day 1: Decide on a statistical question with assigned partner; get approval from teacher
- ❑ Day 2: Working day for analysis – create plots and numerical summaries
- ❑ Day 3: Working day for analysis – create presentation (4 slides maximum)
- ❑ Day 4: Presentations
- ❑ Day 5: Presentations

Introduction to Data Science

Unit 3

Introduction to Data Science
Daily Overview: Unit 3

Theme	Day	Lessons and Labs	Campaign	Topics	Page
Testing, Testing… 1, 2, 3… (7 days)	1	Lesson 1: Anecdotes vs. Data		Reading articles critically, data	227
	2	Lesson 2: What is an Experiment?		Experiments, causation	230
	3	Lesson 3: Let's Try an Experiment!		Random assignments, confounding factors	233
	4	Lesson 4: Predictions, Predictions		Visualizations, predictions	235
	5	Lesson 5: Time Perception Experiment		Elements of an experiment	237
	6	Lab 3A: The results are in!		Analyzing experiment data	239
	7	Practicum: Music to my Ears		Design an experiment	240
Would You Look at That? (4 days)	8	Lesson 6: Observational Studies		Observational study	243
	9	Lesson 7: Observational Studies vs. Experiments		Observational study, experiment	245
	10	Lesson 8: Monsters that Hide in Observational Studies		Observational study, confounding factors	247
	11	Lab 3B: Confound it all!		Confounding factors	251
Are You Asking Me? (9 days)	12	Lesson 9: Survey Says…		Survey	255
	13	Lesson 10: We're So Random		Data collection, random samples	258
	14	Lesson 11: The Gettysburg Address		Sampling bias	262
	15	Lab 3C: Random Sampling		Random sampling	267
	16	Lesson 12: Bias in Survey Sampling		Bias, sampling methods	269
	16	Lesson 13: The Confidence Game		Confidence intervals	272
	17	Lesson 14: How Confident Are You?		Confidence intervals, margin of error	275
	18	Lab 3D: Are You Sure about That?		Bootstrapping	277
	19	Practicum: Let's Build a Survey!		Non-biased survey design	280
What's the Trigger? (5 days)	20	Lesson 15 Ready, Sense, Go!		Sensors, data collection	283
	21	Lesson 16: Does it have a Trigger?		Survey questions, sensor questions	286
	22	Lesson 17: Creating Our Own Participatory Sensing Campaign		Participatory sensing campaign creation	289
	23	Lesson 18: Evaluating Our Own Participatory Sensing Campaign		Statistical questions, evaluate campaign	292
	24^	Lesson 19: Implementing Our Own Participatory Sensing Campaign	Class Campaign—data	Mock-implement campaign, campaign creation, data collection	294
Webpages (6 days)	29	Lesson 20: Online Data-ing	Class Campaign—data	Data on the internet	297
	30	Lab 3E: Scraping web data	Class Campaign—data	Scraping data from the internet	301
	31	Lab 3F: Maps	Class Campaign—data	Making maps with data from the internet	303
	32	Lesson 21: Learning to Love XML	Class Campaign—data	Data storage, XML	305
	33+	Lesson 22: Changing Orientation	Class Campaign—data	Converting XML files	307
	34	Practicum: What Does Our Campaign Data Say?	Class Campaign	Statistical questions, visualizations, numerical summaries	309
End of Unit Project (5 days)	35-40	End of Unit Project: TB or Not TB	Class Campaign	Simulation using experiment data	310

^=Data collection window begins.
+=Data collection window ends.

IDS Unit 3: Essential Concepts

Lesson 1: Anecdotes vs. Data

Data beat anecdotes. In science, we need to closely examine the quality of evidence in order to make sound conclusions. Anecdotes can contain personal bias, might be carefully selected to represent a particular point of view, and, in general, may be completely different from the general trend.

Lesson 2: What is an Experiment?

Science is often concerned with the question "What causes things to happen?" To answer this, controlled experiments are required. Controlled experiments have several key features: (1) there is a treatment variable and a response variable, and we wish to see if the treatment causes a change that we can measure with the response variable; (2) There is a comparison/control group; (3) Subjects are assigned randomly to treatment or control (randomized assignment); (4) Subjects are not aware of which group they are in (a 'blind'). This may require the use of a placebo for those in the control group; and (5) those who measure the response variable do not know which group the subjects were in (if both 4 and 5 are satisfied, this is a 'double blind' experiment).

Lesson 3: Let's Try an Experiment!

Randomized assignment is required to determine cause-and-effect.

Lesson 4: Predictions, Predictions

Designing an experiment requires making many decisions, including what to measure and how to measure it.

Lesson 5: Time Perception Experiment

Designing and carrying out an experiment helps us answer specific statistical questions of interest.

Lesson 6: Observational Studies

Observational studies are those for which there is no intervention applied by researchers.

Lesson 7: Observational Studies vs. Experiments

Experiments are not always possible because of various factors such as ethics, cost limitations, and feasibility.

Lesson 8: Monsters that Hide in Observational Studies

Confounding factors/variables make it difficult to determine a cause-and-effect relation between two variables.

Lesson 9: Survey Says…

Surveys ask simple, straightforward questions in order to collect data that can be used to answer statistical questions. Writing such questions can be hard (but fun)!

Lesson 10: We're So Random

Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to 'center' on the population parameter value.

Lesson 11: The Gettysburg Address

Statistics vary from sample to sample. If the typical value across many samples is equal to the population parameter, the statistic is 'unbiased.' Bias means that we tend to “miss the mark.” If we don't do random sampling, we can get biased estimates.

Lesson 12: Bias in Survey Sampling

Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to ‘center’ on the population parameter value.

Lesson 13: The Confidence Game

We can estimate population parameters. This means that we can give an estimate “plus or minus” some amount that we are confident contains the true value (the population parameter).

Lesson 14: How Confident Are You?

We can estimate population parameters. This means that we can give an estimate “plus or minus” some amount that we are confident contains the true value (the population parameter).

Lesson 15 Ready, Sense, Go!

Sensors are another data collection method. Unlike what we have seen so far, sensors do not involve humans (much). They collect data according to an algorithm.

Lesson 16: Does it have a Trigger?

A key feature that distinguishes the way sensors collect data from more traditional approaches is that sensors collect data when a 'trigger' event occurs. In Participatory Sensing, this event is something we humans agree upon beforehand. Every time that trigger happens, we collect data.

Lesson 17: Creating Our Own Participatory Sensing Campaign

Creating a Participatory Sensing Campaign requires that survey questions must be completed whenever they are “triggered”. Research questions provide an overall direction in Participatory Sensing Campaign.

Lesson 18: Evaluating Our Own Participatory Sensing Campaign

Statistical questions guide a Participatory Sensing Campaign so that we can learn about a community or ourselves. These Campaigns should be evaluated before implementing to make sure they are reasonable and ethically sound.

Lesson 19: Implementing Our Own Participatory Sensing Campaign

Practicing data collection prior to implementation allows optimization of a Participatory Sensing Campaign.

Lesson 20: Online Data-ing

We stretch students' conception of data, to help them see that many web pages present information that can be turned into data.

Lesson 21: Learning to Love XML

XML is a programming language that we use with our campaigns. We create basic XML "tags" in the code, which help us store data in a format we understand.

Lesson 22: Changing Orientation

Converting XML to spreadsheet format helps us better understand and view our data.

Testing, Testing…1, 2, 3…

Instructional Days: 7

Enduring Understandings

An experiment is a data collection method in which the effects of different treatments on an outcome of interest are measured. In an experiment, a treatment is applied to subjects and then observations about the effect of the treatment are made. To isolate the effects from unexplained variation, randomization (or chance) assignment to treatments is applied.

Engagement

Students will view Hans Rosling's video *How Not to Be Ignorant About the World* and will participate in his interactive quiz in order to learn how anecdotes and personal experience can influence what we know and, alternatively, how data provides basis for evidence. The video can be found at:
https://www.ted.com/talks/hans_and_ola_rosling_how_not_to_be_ignorant_about_the_world

Learning Objectives

Statistical/Mathematical:

S-IC 1: Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

S-IC 3: Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

Focus Standards for Mathematical Practice for All of Unit 3:

SMP-1: Make sense of problems and persevere in solving them.

SMP-4: Model with mathematics.

SMP-8: Look for and express regularity in repeated reasoning.

Data Science:

Understand that differences between the measured outcomes of the treatment and control groups in an experiment can be tested. Understand the roles of randomization and of random sampling in statistical inference.

Applied Computational Thinking:

- Test for differences between experimental groups.
- Create graphical representations to compare data between experimental groups.
- Write code to randomly assign subjects to treatment groups

Real-World Connections:

Experiments are used to ensure safety and efficacy of medicines, reliability of electronics and structural materials and find patterns in human behavior.

Language Objectives

5. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
6. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
7. Students will use complex sentences to write informative short reports that use data science concepts and skills.
8. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data Collection Method:

1. Students will gather data generated through a simple experiment.

Data File:

1. Students' *Time Perception* experiment data.

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 1: Anecdotes vs. Data

Objective:

Students will learn the difference between anecdotes and data. They will begin to read articles critically to discern whether the evidence presented is based on anecdotes or data.

Materials:

1. Hans Rosling's video *How Not to Be Ignorant About the World* found at https://www.ted.com/talks/hans_and_ola_rosling_how_not_to_be_ignorant_about_the_world
2. *Article: Miracle at the KK Café* (also available in the LMR folder) https://archives.sfweekly.com/sanfrancisco/miracle-at-the-kk-cafe/Content?oid=2144741
3. *Article: Can Trophy Hunting Actually Help Conservation?* (also available in the LMR folder) https://lastwordwildlife.com/2014/01/21/can-trophy-hunting-actually-help-conservation/

Vocabulary:

anecdote, data

Essential Concepts: Data beat anecdotes. In science, we need to closely examine the quality of evidence in order to make sound conclusions. Anecdotes can contain personal bias, might be carefully selected to represent a particular point of view, and, in general, may be completely different from the general trend.

Lesson:

1. Prepare your video player to show the first 5 minutes and 23 seconds of Hans Rosling's video *How Not to Be Ignorant About the World* found at: https://www.ted.com/talks/hans_and_ola_rosling_how_not_to_be_ignorant_about_the_world
2. Ask students to play along as they watch the video. Each time Hans Rosling asks the audience to choose an answer to each of the three questions, pause the video for about 5 seconds and ask students to write down what they think is the answer to each question.
3. After viewing the video, engage students in *T-I-P-S* (see strategies) with the questions below:

   a. Why did the chimps at the zoo score better than the people? ***Answer: Anyone can select the correct answer just by chance.***

   b. On the second question, Hans Rosling says that "everyone is aware that there are countries and there are areas where girls have great difficulties and they are stopped when they go to school." How could this information influence the answer choice? ***Answer: Personal knowledge and experiences can influence what we think we know.***

   c. Why do you think only a few people know the correct answer to these three questions? ***Answer: People do not know enough about the data that can help them answer these questions.***

4. Display the following statements to students:

   - "My skin glows more…I feel pretty confident." Melissa for Proactiv®
   - "Within four months, I'd lost a grand total of 63 pounds* and was down to my goal weight." Marianne G. for Nutrisystem®
   - "The customer service is obnoxious. The employees are patronizing, smug, and intractable." Seymour773 for Bank of America®

5. Discuss each statement with students by asking the following questions:

   a. Is ________________ a good product? ***For example, is Proactiv® a good skin product?, is Nutrisystem® a good diet program?, is Bank of America® a good bank?***

b. Do you think this person's experience is "typical?" Why? ***Maybe it is typical but maybe not. Their own experience might be very different.***

c. Do you think the company chose this person? How do you know? ***Each company may have chosen the first 2 statements because they were a success. In the case of the Seymour773, a competing company may have chosen his experience to make them appear better.***

d. What about all the other people? How many were successes, how many failures? ***We don't know for sure.***

e. How could we answer such questions? ***Collect data!***

6. Inform students that the statements are called testimonials and they are examples of **anecdotes**. Anecdotes are stories that someone tells about his/her own experience or the experience of someone he/she knows. Anecdotes are good for some things like witness statements in a police report but are not useful for reaching conclusions about groups of people because the assumptions they are based on are not always true. Their claims are easily debunked. Many anecdotes do not equal data.

   <u>Note to teacher about witness statements</u>: Lots of evidence suggests that witness testimony needs to be examined very closely. "As perhaps the single most effective method of proving the elements of a crime, eyewitness testimony has been vital to the trial process for centuries. However, the reliability of eyewitness testimony has recently come into question with the work of organizations such as The Innocence Project, which works to exonerate the wrongfully convicted. This thesis examines previous experiments concerning eyewitness testimony as well as court cases in which eyewitnesses provided vital evidence in order to determine the reliability of eyewitness testimony as well as to determine mitigating or exacerbating factors contributing to a lack of reliability." Information gathered from digitalcommons.liberty.edu

7. On the other hand, **data** are a series of observations, measurements, or facts. Data are information and tell a story.

8. Quickly survey students about whether the video they watched at the beginning of class is based on anecdotes or data. Then inform students that they will analyze two articles to find out if their claims are based on anecdotes or data.

9. Students will read one of two articles, *Miracle at the KK Cafe* or *Can Trophy Hunting Actually Help Conservation?* to analyze whether the claims each makes are based on anecdotes or data. The articles can be found at the following links or in the LMR folder:

   Miracle at KK Café

   https://archives.sfweekly.com/sanfrancisco/miracle-at-the-kk-cafe/Content?oid=2144741

   Can Trophy Hunting Actually Help Conservation?

   https://lastwordwildlife.com/2014/01/21/can-trophy-hunting-actually-help-conservation/

10. Ask students to number themselves off as 1 or 2. Students whose number is 1 will read *Miracle at KK Café* and those whose number is 2 will read *Can Trophy Hunting Actually Help Conservation?*

11. Ask students to find a partner with the same number.

12. Before reading, ask students what they think their article will be about. Have students pair-share their thoughts.

13. During reading, students will take turns reading each paragraph out loud to each other. The student listening will verbally summarize what his/her partner just read.

14. After reading, student pairs will answer the following questions in their DS Journals:

    a. What was the article about?
    b. What claim(s) was/were this article making? Cite examples from the article.
    c. Was this article based on anecdotes or data? Cite examples from the article.
    d. How believable are the claims?

15. After answering the questions, students will find a partner with a different number. Each student will report to their new partner the following information about the article he/she read:

a. The name and publisher of the article.
b. His/her response to the four questions in the DS journal.

16. Quickly survey students about which article was based more on anecdotes and which one was based more on data. Ask a couple of students to explain their choices and give examples.

    Miracle at KK Café makes claims that are anecdotal. Students may cite a customer's claim as an example of an anecdote. Can Trophy Hunting Actually Help Conservation? uses data to make their claims. Students may refer to a statistic used in the article as an example.

17. Class discussion: ***Data Beat Anecdotes!*** Ask students to come up with reasons why this statement is true. Have them come up with situations where you **have** to have an anecdote. For example, if asked what it's like to walk on the moon, only a few people would be able to tell us.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students will do a *Last Word Review* for the words DATA and ANECDOTE.

Last Word Review: Write the word vertically. Students come up with a word or phrase for each letter of the word. Each letter of the word should summarize something about what the students learned about the topic.

Lesson 2: What is an Experiment?

Objective:

Students will learn about the elements of an experiment and the meaning of "causation". Students will learn to distinguish claims of causation from claims of association.

Materials:

1. Video: MythBusters' *Is Yawning Contagious?*
   https://www.discovery.com/tv-shows/mythbusters/videos/is-yawning-contagious

   Note: If video is not found using link, please use a search engine (e.g., Google Video) and type "MythBusters Is Yawning Contagious" to find it. The clip is a little over 5 minutes in length. If you cannot access it, an alternate experiment (MythBusters' *How Does Music Affect Plants*) can be found at https://youtu.be/C5dNhNfGyWQ

Vocabulary:

experiment, subjects, treatment, treatment group, control group, random assignment, representative sample, outcome, research question, confounding factors

Essential Concepts: Science is often concerned with the question "What causes things to happen?" To answer this, controlled experiments are required. Controlled experiments have several key features: (1) there is a treatment variable and a response variable, and we wish to see if the treatment causes a change that we can measure with the response variable; (2) There is a comparison/control group; (3) Subjects are assigned randomly to treatment or control (randomized assignment); (4) Subjects are not aware of which group they are in (a 'blind'). This may require the use of a placebo for those in the control group; and (5) those who measure the response variable do not know which group the subjects were in (if both 4 and 5 are satisfied, this is a 'double blind' experiment).

Lesson:

1. Display the following headlines to students:

   a. Stop Global Warming: Become a Pirate
   b. Lack of sleep may shrink your brain
   c. Early language skills reduce preschool tantrums
   d. Dogs walked by men are more aggressive

2. Discuss each headline by asking the following questions:

   a. What is the headline implying with its wording? ***1a is implying that you can stop global warming by becoming a pirate, 1b is implying that it's possible to shrink your brain if you aren't getting enough sleep, 1c is implying that having early language skills will decrease preschool tantrums, 1d is implying that dogs are more aggressive when they've been walked by men.***
   b. Is it implying causation or association? ***Discuss definitions of causation and association. Causation means there is a cause and effect relationship between variables. For example, heat causes water to boil; whereas association or correlation means that high values of one variable tend to be associated with high values of the other (or high values tend to be with low values). However, this is not necessarily cause-and-effect at play. For example, blanket sales in Canada are associated with brush fires in Australia - not because Canadian blankets cause the fires, but because Canadian winters cause blanket sales, and Canadian winters are Australian summers, which cause fires. 1a, 1c and 1d are implying causation and 1b is implying association.***
   c. How can you tell the difference between causation and correlation? What words stand out in these headlines? ***Answers will vary but some terms for causation include: cause, increase/ decrease, benefits, impacts, effect/ affect, etc.; and for correlation***

include: get, have, linked, more/ less, tied, connected, etc. In 1a, "become" stands out; in 1b, "may" stands out; in 1c, "reduce" stands out; in 1d, "are" stands out.

d. Change each causal version of a headline into a non-causal version and vice versa. ***Answers will vary but an example for 1a is to instead say Global Warming linked to increase of pirates.***

3. Introduce the MythBusters video clip by answering the following questions, in teams, for their headline "Is Yawning Contagious?"

   a. What is the headline implying with its wording? ***That yawning may cause other people to yawn.***
   b. Is it implying causation or correlation? How do you know? ***Causation because "contagious" yawns means that you are yawning because someone else has yawned.***
   c. How can we determine if this is true? ***Split the class into groups and have each team come up with a way to determine if this is true. Each group should assume that they get to examine 50 people.***

4. Show the MythBusters video clip called *Is Yawning Contagious?* The clip can be found at:

   https://www.discovery.com/tv-shows/mythbusters/videos/is-yawning-contagious

5. Focus students on the following guiding questions and ask them to take notes as they watch the video clip:

   a. How did the MythBusters design the investigation?
   b. What steps did they take?
   c. How is this different than your team's headline responses?

6. After viewing the clip, inform students that the MythBusters have just conducted an **experiment**, which is one method of data collection.

7. We begin with a brief introduction into "what is an experiment" but the definition will be developed over the next several lessons.

8. Guide students to identify the elements of an experiment by referring back to the video clip:

   a. **Research Question**—the question to be answered by the experiment (***Is Yawning Contagious?***)
   b. **Subjects** – people or objects that are participating in the experiment (***the 50 adults***)
   c. **Treatment** – the procedure that is assigned to a group of subjects (***Kari yawned before subject entered the room***)
   d. **Treatment group** – the group of subjects that receive the treatment (***two out of every three subjects who were placed into rooms – yawn from Kari***)
   e. **Control group** – the group that does not receive a treatment (***one out of every three subjects who were placed into rooms – no yawn from Kari***)
   f. **Random assignment** – subjects are randomly assigned to either the treatment or control group (***two out of every three subjects received the treatment***) **Note:** In this experiment, random assignment was not used (or if it was, we were not told so.)
   g. **Outcome** – the variable that the treatment is meant to influence. (***whether or not a person yawned***)
   h. **Statistic**—A method for comparing the outcomes of the control and treatment groups is needed. ***In this case, the MythBusters used the difference between the percent of subjects that yawned in the treatment group was 4% higher than the control group***.

Note: In this experiment, and in those found in the IDS curriculum, we use a treatment and a control group. However, a control group is not a *necessary* element of an experiment. Sometimes it is more appropriate to have two treatment groups with no control group (e.g., medical professionals testing different doses of drugs). The effect that is being studied will dictate whether to feature a control group or not.

9. Display the following questions on the board or projector. Using *T-I-P-S*, ask students to discuss them.

a. Why did the MythBusters follow all of these steps to design their experiment? ***In order to determine if watching someone yawn can cause you to yawn.***

b. We don't know how MythBusters chose who would be in the treatment group and who would be in the control group. Suppose that the people who showed up first, early in the morning, were assigned to the treatment group, and the last few people, later in the day, ended up in the control group. Would you believe in the conclusions? ***No, because the two groups were different. The first group might have been sleepier, and so more likely to yawn anyways. Explain that this --another explanation for the cause-and-effect--is caused a confounding variable.***

c. Explain that in order to make the two groups as similar as possible, experimenters usually assign subjects randomly. How might we randomly assign about half of the subjects to the treatment and half to the control? ***We might flip a coin, and those who get Heads go to Treatment.***

d. Why would random assignment improve the MythBusters study? ***Because then the two groups would be more similar. So we wouldn't have a confounding variable to worry about.***

10. **Emphasize that without random assignment, we cannot determine causation because we are not comparing two similar groups.**

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 3: Let's Try an Experiment!

Objective:

Students will explore the importance of randomized assignment in experiments. They will understand that without random assignment, there might be confounding variables and will be able to suggest possible confounding variables.

Materials:

1. Measuring Tape

Essential Concepts: Randomized assignment is required to determine cause and effect.

Lesson:

1. Inform students that they will be exploring the question "Why do we need randomized assignment?" by conducting an experiment. Tell students that you have a treatment that can make people taller. Explain that the class will be divided into two groups, one group will get the treatment, and one group will not. The group that does not receive the treatment will be the control group. After the treatment, they will measure the groups to see which is taller. Now divide the class into two groups by placing the boys in the treatment group and the girls in the control group.

   Remember that in an experiment we typically have a treatment group and a control group. In the MythBusters experiment, they compared number of yawns after treatment, and not any measurements before treatment, because they were comparing the treatment group to the control group (the control group is specifically here because it is a comparable untreated group - this allows us to not need "before" measurements). Therefore, in this case, we will run the experiment and *then* compare average height of the treatment group to the control group.

2. Tell them that after the treatment group takes the treatment, your statistic to compare groups will be to measure the heights. If the treatment group is taller, then the treatment must have worked. There are two possible outcomes to dividing the class this way:

   a. The students will protest (as they should) and you can start a discussion as to why this is not a good way to divide the class.

   b. OR the students don't protest and you continue with the experiment. The treatment should be something silly, like waving a ruler in front of the person's face or by asking them to chant "grow, grow, grow!" three times. After treatment, measure the heights of each group and ask them if they think this is good evidence (**do not say "proves"**) that the treatment is effective.

3. Regardless of the outcome, students should recognize that by putting the boys in one group, the outcome was pre-determined, since boys tend to be taller than girls to begin with. This is an example of a **confounding factor**. Confounding factors are variables that provide an alternative explanation of the effect of the treatment on the outcome variable.

4. Ask students: "How should students be put into groups?"

5. Discuss various other methods of grouping students. Someone will probably say to split the groups into equal numbers of boys and girls. At this suggestion, divide the class into two groups by placing the tallest boys and tallest girls in the treatment group, and the shorter boys and shorter girls in the control group. ***Students should be able to recognize that you shouldn't use any characteristics to decide the groups.***

6. Continue discussion of other ways to decide the groups. Use the following questions as a guide:

   a. What about flipping a coin?
   b. What will the gender balance look like? ***Each group should have about the same balance as the class, though not exactly.***

   c. Why is it important that the groups be similar? ***Because otherwise, something else might be the cause of the response changing.***

7. Inform students that today the class will begin to design their own experiment using what they have learned over the last few lessons. The question they will investigate is:

   How does our perception of time change when exposed to a stimulus?

8. They will be trying to determine the length of one minute without the use of time-aids. In their experiment, they will subject some students to a stimulus and others to no stimulus. They will then analyze the data to determine if subjecting students to a stimulus affects the perception of how long a minute of time lasts.

9. In their DS journals, ask students to answer the following questions about the elements of their experiment:

   a. What is the research question we're interested in addressing?
   b. Who are the subjects that will be participating in the experiment?
   c. How should we randomly assign the subjects into treatment and control groups? (See step 12 for an RStudio method that the teacher can use)
   d. What is the outcome variable that we will be measuring? What unit of measurement should we use?

   Note: Students will decide on a treatment to apply to each group on the following day.

10. As a class, discuss the responses to the questions above (step #9, a-d) and come to a consensus for each question's answer.

11. Inform the class that they will be using the answers they have agreed upon as the final design of the class's experiment.

12. At the end of the class, the students should be assigned to the treatment or control groups using the randomization method they chose as a class in step #9c.

    Note: One method to determine group assignment would be to use the class roster and the sample() function in RStudio. The students have a number that corresponds to their placement on the roster (i.e. student 1's last name most likely starts with an A, and then we move alphabetically through the roster). You can then use RStudio to randomly select which half of the numbers/students will be assigned to the treatment group.

```
> sample(1:30, size = 15, replace = FALSE)
```

13. Students will conduct the experiment in the next lesson.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 4: Predictions, Predictions

Objective:

Students will continue to read articles critically. They will anticipate visualizations about the data that will be collected from the class experiment and make predictions about the outcome.

Materials:

1. Article: PsyBlog's *10 Ways Our Minds Warp Time* found at: http://www.spring.org.uk/2011/06/10-ways-our-minds-warp-time.php
2. *Experiment Predictions* handout (LMR_ 3.1_Experiment Predictions)

Vocabulary:

theory

Essential Concepts: Designing an experiment requires making many decisions, including what to measure and how to measure it.

Lesson:

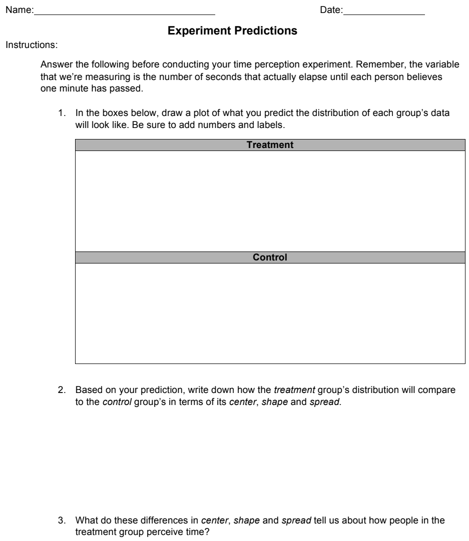
1. Students will read the article *10 Ways Our Minds Warp Time* found at: http://w.spring.org.uk/2011/06/10-ways-our-minds-warp-time.php.
2. They will read the article critically to answer the following questions (displayed or written on the board):
   a. Who was observed and what were the variables measured? ***People and their perceptions of time.***
   b. What statistical questions were the researchers trying to answer? ***How is time perception affected by different stimuli?***
   c. Who collected the data? ***Researchers such as cave expert Michel Siffre collected data.***
   d. How were the data collected? ***Data were collected through various experiments/studies (13 were cited).***
   e. What claim(s) did the article make? ***There were 10 claims made regarding time perception.***
   f. What are some statistics that the article used to make the claim(s)? ***Answers may vary. Article has several percentage statistics.***
3. In their teams, ask students to share their responses from reading the *10 Ways Our Minds Warp Time* article and agree on the responses as a team.
4. Do a quick *Whip Around* of the responses (see step #2 for responses).
5. Remind students that they designed a class experiment during the previous lesson but did not select an actual treatment. As a class, decide on a treatment to use for the experiment. Students can use the methods found in the article for inspiration, or come up with something novel on their own.

   Note: Stimuli examples include music (genres determined by the class), lights off, physical activity (e.g., holding arms out), relaxation/meditation techniques, heads down, eyes closed, etc. Ensure that the experiment can be completed in **one** 50-60 minute class period. Treatments requiring excessive preparation time (e.g., running a mile) are less than ideal.
6. Before they conduct the experiment, students will test their theories by making predictions about the data and the outcomes. A **theory** is an idea used to explain a situation.
7. Display the class experiment's research question:

   How does our perception of time change when exposed to a stimulus?

8. Take a poll of the students who believe that there will be differences in the estimate of the length of a minute between the treatment and control groups. The remaining students, then, do not believe that there will be differences.
9. Then, ask those students who believe there are differences, how small or large they think the difference will be.
10. Distribute the *Experiment Predictions* handout (LMR_ 3.1) and, in pairs, have students discuss and complete the answers for the handout.

    Note: What will the distribution of time perceptions look like? ***The distributions will likely have more points that are closer to 60 seconds, but will also have values that are shorter and longer than 60 seconds. Appropriate plots to use will include histograms, dotplots or boxplots.***



Name:______________________________ Date:______________

Experiment Predictions

Instructions:

Answer the following before conducting your time perception experiment. Remember, the variable that we're measuring is the number of seconds that actually elapse until each person believes one minute has passed.

1. In the boxes below, draw a plot of what you predict the distribution of each group's data will look like. Be sure to add numbers and labels.

Treatment
Control

2. Based on your prediction, write down how the *treatment* group's distribution will compare to the *control* group's in terms of its *center*, *shape* and *spread.*

3. What do these differences in *center*, *shape* and *spread* tell us about how people in the treatment group perceive time?

LMR_3.1

11. Using *Anonymous Author*, select student work to share with the whole class.
12. Give student teams time (about 2 minutes) to discuss each product that is shared/presented.
13. Teams will offer their thoughts using a modified *Two Cents* strategy where, instead of two cents, each team will receive one cent (or a token) and, in order to turn it in, the team will have to make comments or ask questions about the student work that is being shared. Call on teams until you have collected every cent. This ensures that all teams contribute to the discussion.
14. Inform students that they will conduct the experiment in which they will estimate the length of time of one minute during the next lesson.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 5: Time Perception Experiment

Objective:

Students will engage in a collectively designed experiment.

Materials:

1. RStudio's **`stopwatch()`** function
2. IDS UCLA App or Browser-Based survey-taking tool

Essential Concepts: Designing and carrying out an experiment helps us answer specific statistical questions of interest.

Lesson:

1. Begin the lesson by eliciting the elements of an experiment from students (they may refer back to their DS journals for their responses from Lesson 2).
2. Inform students that they will be using RStudio to get a precise measurement of their estimate. Ask for a student volunteer.
3. Demonstrate the stopwatch function using RStudio by typing in the following code:

```
> stopwatch()
```

4. Then, ask the student volunteer to stand in front of your computer and get ready to estimate the length of time of one minute without looking at a clock. Once he/she thinks a minute has passed, ask him/her to press the enter/return key on the keyboard to see the result of the estimate.
5. Inform students that you have just demonstrated how they will measure their one-minute estimates.
6. Begin conducting the experiment by reviewing the research question:

How does our perception of time change when exposed to a stimulus?

7. Refer back to the experiment design.
8. Review the specific treatment that the subjects in the treatment group will receive. If necessary, demonstrate to the treatment group how to do the experiment. For example, if standing with open arms is the stimulus, the estimate begins when the student starts the **`stopwatch()`** function and engages in the stimulus, and ends when the subject presses enter/return in RStudio to stop the timer.
9. For the control group, the students can simply sit at their desks with their eyes closed. Each student will run the **`stopwatch()`** function and stop the timer when they believe a minute has elapsed.
10. Conduct the experiment in its entirety. Use team roles effectively to ensure the experiment is done correctly.
11. Have each student use a computer and the **`stopwatch()`** function to record her/his estimate of one minute. Ensure each student records her/his estimate in the DS journal.
12. When the experiment is completed, have students enter their data in the *Time Perception* survey found in the Survey Taking Tool at https://portal.idsucla.org or by using the IDS UCLA App in their iOS or Android devices.
13. Inform students that they will be analyzing the results from the experiment *in Lab 3.1: The results are in!*

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next Day

LAB 3A: The results are in!

Complete Lab 3A prior to Practicum.

Lab 3A - The results are in!

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Conducting experiments

- Previously in class, you conducted an experiment to gauge how a stimulus affected people's perception of time.
  - Some people were given a treatment, others were not.
- In this lab, we'll use the data cycle to analyze the *research question*:

  Does the stimulus your class chose change people's perception of time?

Coming up with questions

- **Write down two statistical questions that will help you answer the *research question* from the previous slide.**
- Then, *export*, *upload*, *import* your experiment data into RStudio.
  - If you're having trouble coming up with good statistical questions, try loading the data and looking at the variables.
  - Ask yourself, *How would I use these variables to answer the research question?*

Analyzing our data

- Create appropriate plots to answer your statistical questions.
  - **Are there any outliers or unusual observations that require some cleaning before you can interpret your plots?**
- Calculate appropriate numerical summaries to answer your statistical questions.
- Interpret your plots and summaries.
  - **Write down a few sentences with your interpretations.**

Wrapping it up

- Is it possible your initial results occurred by chance alone?
  - **Use repeated shuffling to determine how likely the typical difference between the two groups occurred by chance alone.**
  - **Create a plot and use it to justify your answer.**
- What do you conclude about the *research question*?
  - **Write a report using the plots and analysis you conducted to answer the *research question*.**
  - Be sure to describe how you conducted your experiment.

Practicum: Music to my Ears

Objective: Students will design a simple experiment.

Materials:

1. Practicum: *Music to my ears* (LMR_U3_Practicum_Music to My Ears)

Note to Teacher: Before assigning the practicum to your students, engage the class in a discussion about experiments. Use the following questions as a guide to assess student understanding.

1. When is random assignment used? Why is it important? ***Random assignment is used when you wish to determine whether a treatment causes changes in an outcome variable. It's important because it creates a "balance" of the groups so that the only way the groups differ, on average, is that one gets the treatment and one does not. Thus, if there is a change in the outcome variable, only the treatment could have caused it.***
2. Below are some headlines, determine if they are causal or not. If not causal, re-write so that it is. If causal, state why it's causal.
   - Straight A's in high school may mean better health later in life. ***not causal, re-writing answers will vary***
   - Murder rates affect IQ test scores: Study. ***causal, explanations will vary***
   - Microbe linked to Alzheimer's Disease. ***not causal, re-writing answers will vary***
   - Luckiest people "born in summer" ***causal, explanations will vary***
3. Why is a control group important? ***The control group is important because it allows us to measure the effects of the treatment group with an untreated comparable group. Without the control group, we don't know what would have happened if we had done nothing. [Think of a new vaccine for the flu. If there is no control group, and we see the treatment group improving, we will never know if they would have improved anyways, without the vaccine.]***

Practicum
Music to my Ears

In class, you designed and conducted the *Time Perception* experiment to find out if a person's perception of time changed when exposed to a stimulus. This experiment was designed so that it used random assignment, which is the process of using a chance device (e.g., dice, RStudio, etc.) to determine the placement of subjects into the treatment and control groups. By randomizing, you are removing other possible explanations for why the results happened the way they did.

Now we are asking you to design an experiment to determine whether doing math homework with music playing in the background affects student's test scores. Work with your team to design this experiment.

Submit a paper that clearly lays out your team's design plan. Be sure to include:

1. Descriptions of each element of the experiment by answering the following questions:
   a. What is the research question we are interested in addressing?
   b. Who are the subjects that would be participating in the experiment? How should we select them?
   c. What could be possible treatments? What treatment do you choose and why? What will the control group do in your study?
   d. Describe how to randomize the subjects into the treatment and control groups.
   e. What is the outcome variable that we are measuring? Is it categorical or numerical? What other variables will you measure for each subject?
2. An analysis plan:
   a. What statistical questions will you ask to address your research question?
   b. What analyses (graphical and numerical) will you use to answer these questions?
   c. An explanation of how you will determine whether the treatment affects test scores.

Would You Look at That?

Instructional Days: 4

Enduring Understandings

An observational study is a data collection method in which subjects are observed and outcomes are recorded. Unlike experiments, it may not be possible to assign subjects to treatment and control groups in observational studies, which impedes our ability to control for confounding factors. This means that researchers must rely on existing control and treatment groups to observe the outcomes. Observational studies can show associations in the data, but cause and effect relationships can only be concluded with experiments.

Engagement

Students will participate in the *Observational Studies Activity* described in Lesson 5. They will record information that can be obtained through pictures. The data will then be analyzed to see if there are any variables related to the number of friends a person has on social media.

Learning Objectives

Statistical/Mathematical:

S-IC 1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

S-IC 3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand that data from observational studies can help us find associations among variables. Explain why some variables that are not related in reality might look as though they are due to the presence of confounding factors.

Applied Computational Thinking using RStudio:

- Download data from the Internet that was collected via an observational study.
- Clean data set by adding variable names.
- Create scatterplots of two variables and determine possible relationships between them, as well as identify potential confounding variables.

Real-World Connections:

Economists, psychologists, and biologists conduct observational studies to study human behavior. For example, observational studies are used in epidemiology to study outbreaks of illnesses and people's behavioral patterns.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.

2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will use complex sentences to write informative short reports that use data science concepts and skills.
4. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data Collection Method:

1. Students will record information about a set of high school students by observing characteristics given in a picture.

Data File:

1. *Lung Capacity of Children* data set found at

   https://jse.amstat.org/v13n2/datasets.kahn.html

   NOTE: The raw data set can be found at https://jse.amstat.org/datasets/fev.dat.txt

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 6: Observational Studies

Objective:

Students will learn that an observational study is a data collection method in which subjects are observed and outcomes are recorded. They will learn how to collect this type of data and make informal inferences about the results.

Materials:

1. *Stick Figures Cutouts* (LMR_1.2_ Stick Figures) from Unit 1, Lesson 1
   Note: Advanced preparation required (see step 1 below).
2. *Turning Observations into Data* handout (LMR_3.2_ Observations_to_Data)

Vocabulary:

observational study

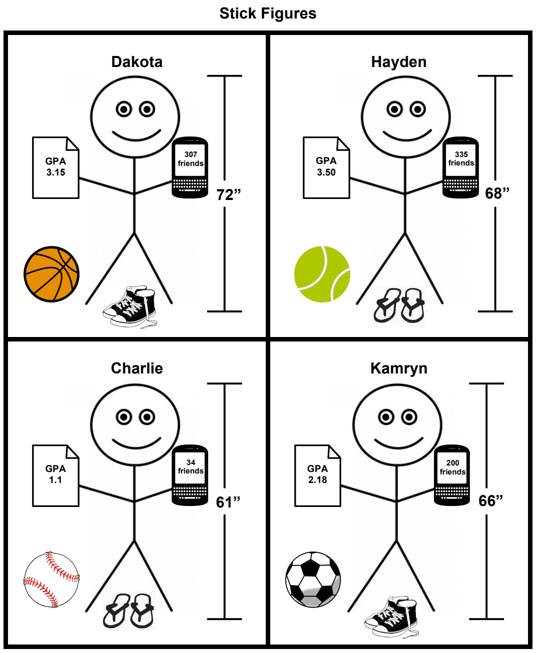
Essential Concepts: Observational studies are those for which there is no intervention applied by researchers.

Lesson:

1. From Unit 1, Lesson 1, redistribute one full set of 8 cards from the *Stick Figures* handout (LMR_1.2) to each student team.

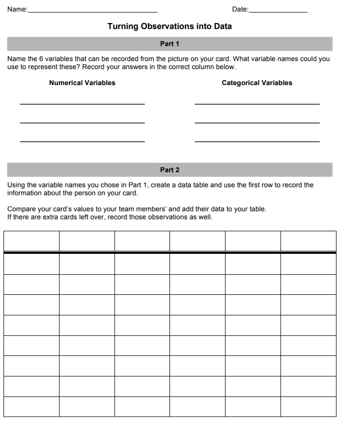
   Advanced preparation required: Print the *Stick Figures* handout (*LMR_1.2*). The handout can then be cut into the 8 cards. You will need enough sets of the cards for each student team to share a full set. For example, if there are 5 student teams in a class, then 5 copies of the file will need to be printed so that each team gets all 8 cards.

2. Have students recall that they used these cards in Unit 1, Lesson 1. When they used them in Lesson 1, the data was collected, recorded, and organized, but without particular structure to it.



LMR_1.2

3. Then, distribute one copy per student of the *Turning Observations into Data* handout (LMR_3.2).



Name:__________________________ Date:_____________

Turning Observations into Data

Part 1

Name the 6 variables that can be recorded from the picture on your card. What variable names could you use to represent these? Record your answers in the correct column below.

Numerical Variables **Categorical Variables**

____________________ ____________________

____________________ ____________________

____________________ ____________________

Part 2

Using the variable names you chose in Part 1, create a data table and use the first row to record the information about the person on your card.

Compare your card's values to your team members' and add their data to your table.
If there are extra cards left over, record those observations as well.

LMR_3.2

4. Every student from the team will then select one of the cards from the team's pile of 8, and should begin working through the *Turning Observations into Data* handout individually.

5. As the students finish each part of the handout, they should compare their responses with their student teams.

6. Go over the names of the variables in Part 1 by doing a quick Whip Around by teams. Then, select a couple of teams to share the information on the first row and one of the columns.

7. Part 3 of the handout asks the students to consider the following research question:

What determines the number of friends a person has on social media?

8. Once the students have completed the handout, discuss the variable that they thought was best associated with the number of friends on social media. ***They should have seen that a person's GPA was related to the number of friends. More specifically, the higher a person's GPA, the more friends he/she had.***

9. Ask a few students to share out their responses to the very last question: "Can you think of another variable (not necessarily given in the pictures) that might impact both the number of friends AND the variable you selected? Give an example and explain how it might impact each of the variables." ***Answers will vary, but one example could be: a person's self-esteem level (if he/she is confident in school, his/her grades might be higher; higher confidence could also be a reason for a person having more friends).***

10. Remind students that in the previous section, they learned about the elements of an experiment. In teams, ask students to discuss how collecting this data is similar or different from experiments. Then have a whole class discussion about this comparison, guiding ***students to realize that there were no assignments to groups and no treatment was applied. The subjects (i.e. the people displayed on the cards) were simply observed, and then information about them was recorded.***

11. Inform students that an **observational stud**y is a data collection method in which subjects are observed and outcomes are recorded. No treatment is applied to the subjects. Instead, researchers are simply watching something happen and have absolutely no control over it.

12. In lesson 7, students will learn more about the differences between experiments and observational studies and what conclusions they can make about each.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 7: Observational Studies vs. Experiments

Objective:

Students will learn how observational studies differ from experiments, and will classify different research scenarios based on which method would be most appropriate. They will also learn about the roles of ethics, cost limitations, and feasibility when deciding between the two data collection methods.

Materials:

1. *What Should We Do?* handout (LMR_3.3_ObsStudies vs Experiments)

Vocabulary:

ethics, cost limitations, feasibility

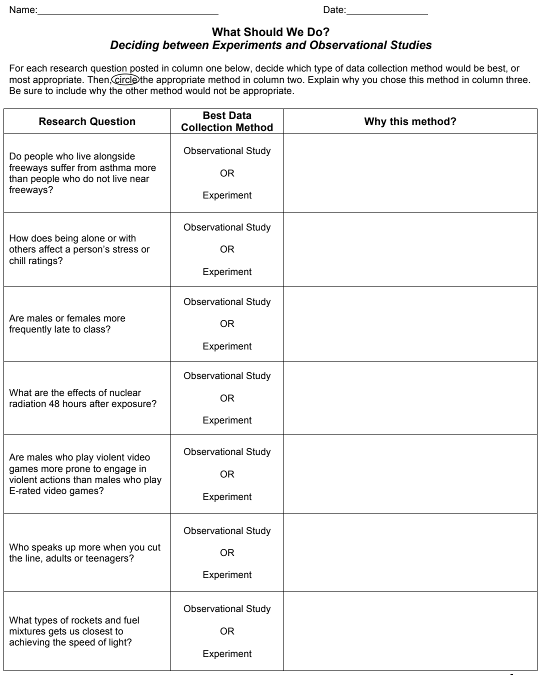
Essential Concepts: Experiments are not always possible because of various factors such as ethics, cost limitations, and feasibility.

Lesson:

1. Remind students that in observational studies, we can never randomly assign subjects to treatment and control groups. Conversely, in experiments, we always need to have random assignment into these groups.
2. Pose the following question to students: Why can't we just always do experiments? Have students discuss this question in their student teams and write down a few responses in their DS journals.
3. Inform students that a researcher wants to perform studies to answer the research questions below. In teams, have students come up with reasons for why an experiment would not be possible for each scenario.
    a. Does smoking cause lung cancer? ***Unethical. You cannot make people smoke cigarettes and then see if they have lung cancer later in life.***
    b. Does drinking water from Mars keep you healthier than drinking water from Earth? ***Cost. It would be incredibly expensive to design a space shuttle that can successfully transport people to Mars and have them live there for an extended period of time and most researchers would not have the funding to do this.***
    c. Do people with higher IQ scores score better on the SAT than people with lower IQ scores? ***Not feasible/not possible. You cannot randomly assign IQ scores to people because it is a measurement based on aptitude.***
4. Select three teams and assign a scenario above to each team. Ask each team to report out on their assigned scenario. As teams share, be sure to discuss the following issues regarding why we cannot always to experiments:
    a. **Ethics:** Sometimes, experiments cannot be performed because it would be unethical to give certain treatments to subjects. For example, we could not inject an HIV infection into participants because the long-term effects might lead to death.
    b. **Cost Limitations:** Sometimes, experiments would be very costly and much too expensive to perform. Some possibilities could be with technology.
    c. **Feasibility**, impossible to randomize: In certain cases, you cannot perform an experiment because it is impossible to randomly assign people to particular groups. For example, you cannot assign a gender to a person.
5. Distribute *What Should We Do?* handout (LMR_3.3). In teams, students will identify whether the research question could best be answered via an experiment or an observational study.

6. Once all student teams have completed the handout, assign one research question to each team to report out. As each response is shared, conduct a whole-class discussion to compare which data collection method was most appropriate for each research question. Ensure everyone understands the reasons each method was chosen before moving on to the next scenario.

Note: Page 2 of the handout is an answer key for teacher reference only!



Name:______________________________ Date:______________

What Should We Do?
Deciding between Experiments and Observational Studies

For each research question posted in column one below, decide which type of data collection method would be best, or most appropriate. Then, circle the appropriate method in column two. Explain why you chose this method in column three. Be sure to include why the other method would not be appropriate.

Research Question	Best Data Collection Method	Why this method?
Do people who live alongside freeways suffer from asthma more than people who do not live near freeways?	Observational Study OR Experiment	
How does being alone or with others affect a person's stress or chill ratings?	Observational Study OR Experiment	
Are males or females more frequently late to class?	Observational Study OR Experiment	
What are the effects of nuclear radiation 48 hours after exposure?	Observational Study OR Experiment	
Are males who play violent video games more prone to engage in violent actions than males who play E-rated video games?	Observational Study OR Experiment	
Who speaks up more when you cut the line, adults or teenagers?	Observational Study OR Experiment	
What types of rockets and fuel mixtures gets us closest to achieving the speed of light?	Observational Study OR Experiment	

LMR_3.3

7. Next, student teams will generate three research questions on their own. They need to identify the best data collection method for answering their question and should provide an explanation. At least one of the three research questions should use an observational study for data collection.

8. Using a share-out strategy, have the reporter of each team share one of their investigation questions with the rest of the class.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 8: Monsters that Hide in Observational Studies

Objective:

Students will learn about confounding factors that may impact the results of an observational study, which is why causation can never be concluded with observational studies, only associations between variables.

Materials:

1. Computers
2. *Spurious Correlations* website (tylervigen.com)

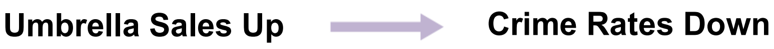
Vocabulary:

cause, confounding factors, associated

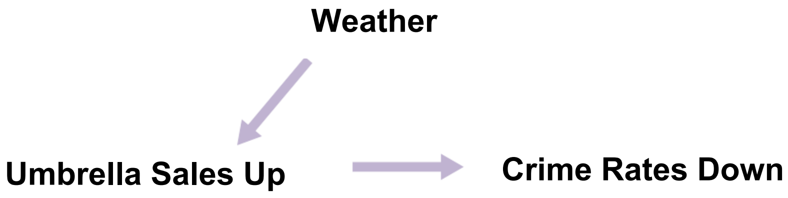
Essential Concepts: Confounding factors/variables make it difficult to determine a cause-and-effect relation between two variables.

Lesson:

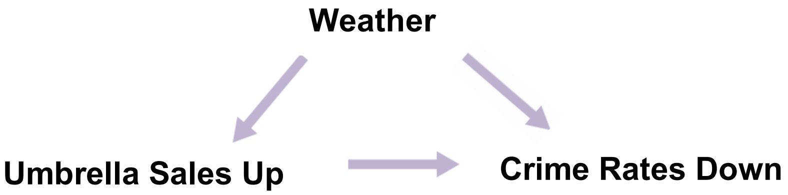
1. Ask students to recall that they looked at the relationship between a student's GPA and the number of friends that person has on social media during lesson 6. It seemed that students with higher GPAs had more friends than students with lower GPAs. But did this mean that the **cause** of a person's GPA is the amount of friends they have? NO!
2. They also identified other variables that could have contributed to the relationship, these outside variables are called **confounding factors**. Confounding factors are variables that are related to both the explanatory variable and the response variable in an observational study.
3. Propose the following statement to students: "Research suggests that a rise in umbrella sales leads to decreased crime rates."
4. Allow the students to work in teams to think about possible confounding factors. They should choose a variable that is related to umbrella sales, and that might lead to decreased crime rates. After they've come up with a few possibilities, use the following diagram progression to further explain the impact of confounding factors.
    a. Step 1: Draw an arrow showing that "a rise in umbrella sales leads to decreased crime rates" since that is what researchers have stated.



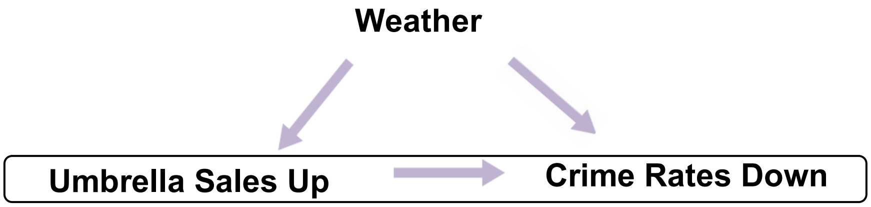
    b. Step 2: Include the variable that might be related to people buying more umbrellas (i.e., the confounding factor). For example, when the weather is rainy, people buy more umbrellas.



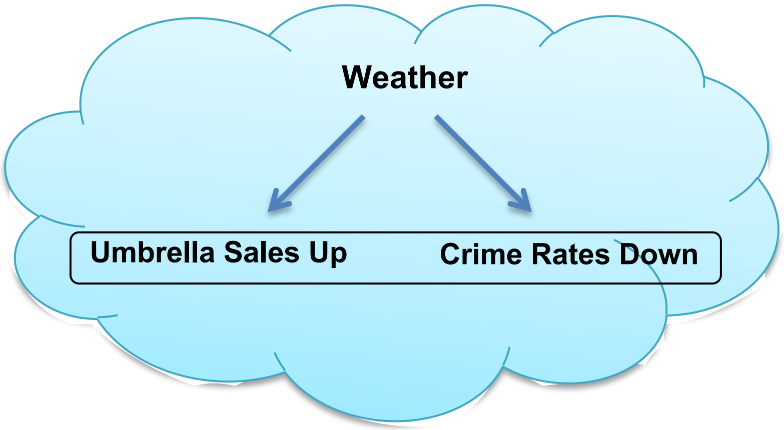
c. Step 3: Draw an additional arrow from "Weather" to "Crime Rates Down" because it is well known that when the weather is bad, people are less likely to be outside committing crimes.



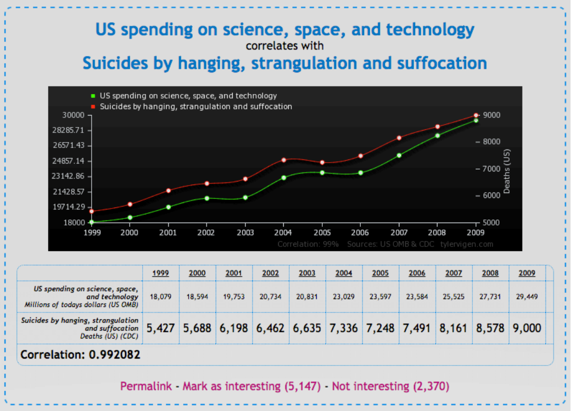
d. Step 4: Remind students that the original claim was that "a rise in umbrella sales leads to decreased crime rates"." However, we've now shown that maybe buying umbrellas is not the only thing that could be contributing to a decrease in crime, which makes us question the link between the two variables.



e. Step 5: Therefore, we have found a confounding factor with the variable "crime rates." This means we can erase the original "link" between a rise in umbrella sales and decreased crime rates since there are outside variables interfering. We can't say buying umbrellas *causes* decreased crime rates, but we can say that a rise in umbrella sales are **associated** with decreased crime rates.



5. Once the students grasp what confounding factors are and how to identify them, introduce them to the website *Spurious Correlations* by Tyler Vigen. This site shows many explanatory and response variables that are randomly associated with each other. Spurious Correlations can be found at: http://www.tylervigen.com/spurious-correlations.



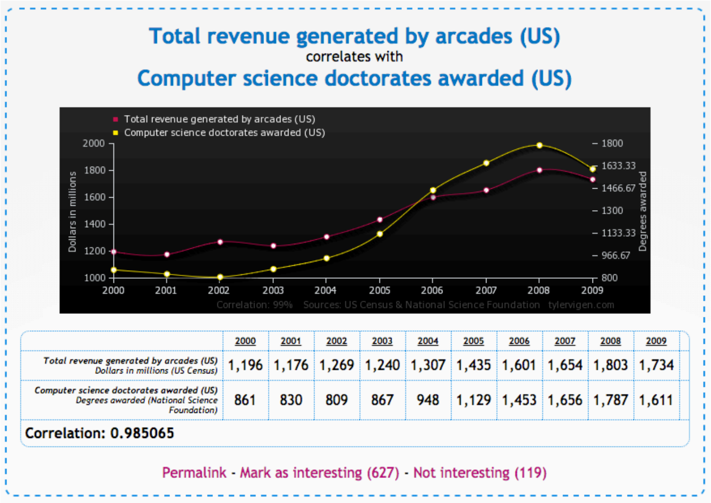
US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

Permalink - Mark as interesting (5,147) - Not interesting (2,370)

6. For the example given above, we see that as the US spends more money on science, space, and technology, more people are dying by way of suicide. Clearly, it does not make sense that if the US keeps spending money on science, then more people are going to commit suicide. It simply happened by chance (or a bizarre chain of confounding factors) that the two variables are related to each other.
7. Allow the students to explore the website on their own (Note: there are multiple pages of graphs, so they are not restricted to simply the homepage). They should choose a graph that interests them and answer the following questions in their DS journals:
    a. What are the two variables shown in your graph?
    b. Is there a positive association or a negative association between the variables?
    c. Write an interpretation of this plot in the context of the data.
    d. Write the data points in a "spreadsheet format" in a form that RStudio could read. Each row should represent a point on the graph, and each column one of the two variables.
    e. By hand, make a scatterplot of the association. Describe whether the association seems strong or weak or moderate to you.
    f. Do you think that the explanatory variable *causes* the response variable? Explain.
    g. If you answered 'no' to f, then draw a diagram like in #4 with possible confounding factors. **Note:** this can be difficult, depending on the graph chosen. Some factors to consider: weather, economy, fashion trends.
8. Example answers to Step 7 are given below:



Total revenue generated by arcades (US)
correlates with
Computer science doctorates awarded (US)

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Total revenue generated by arcades (US) *Dollars in millions (US Census)*	1,196	1,176	1,269	1,240	1,307	1,435	1,601	1,654	1,803	1,734
Computer science doctorates awarded (US) *Degrees awarded (National Science Foundation)*	861	830	809	867	948	1,129	1,453	1,656	1,787	1,611

Correlation: 0.985065

Permalink - Mark as interesting (627) - Not interesting (119)

    a. What are the two variables shown in your graph? ***Total revenue generated by arcades in the US and the number of computer science doctorates awarded.***

b. Is there a positive association or a negative association between the variables? ***There is a direct relationship because the lines have the same shape (they follow the same pattern).***

c. Write an interpretation of this plot in the context of the data. ***It seems that as more doctorates are awarded to computer scientists, arcades are generating more revenue.***

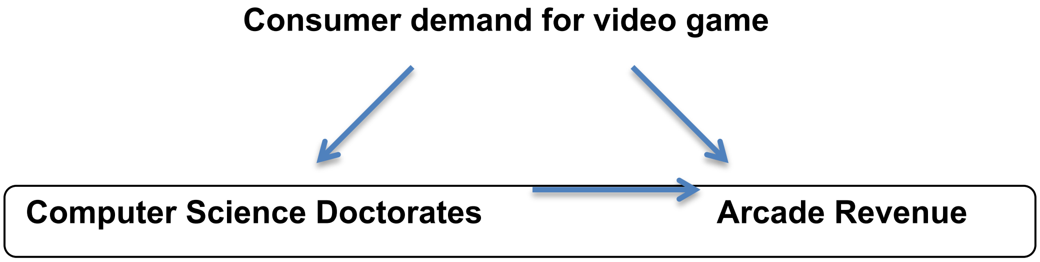
Arcade Revenue	***CS doctorates***
1196	***861***
1176	***830***

etc.

d. Answers will vary.

e. Can you conclude that the one variable *causes* the other? ***No. Although the two variables are associated with one another, we do not have evidence to say that more doctorate awards cause arcades to make more money because the data do not come from a controlled experiment.***

f. Draw a diagram like the one we did together earlier (in step 4 of lesson) with possible confounding factors. ***Student's diagram should look like the one below:***



Consumer demand for video game

Computer Science Doctorates

Arcade Revenue

9. Once all students have selected a graph and have answered the above questions, have them share their responses with a partner. They should explain why they thought their particular graph was interesting, how the two variables are related (directly or inversely), and whether or not there is a causal link between the variables.

10. At the end of this lesson, students should understand that causation can only be concluded when an experiment is performed, but associations can be concluded for observational studies.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next Day

LAB 3B: Confound it all!

Complete Lab 3B prior to Lesson 9.

Lab 3B - Confound it all!

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Finding data in new places

- Since your first forays into doing data science, you've used data from two-sources:
  - Built-in datasets from RStudio.
  - Campaign data from IDS Campaign Manager.
- Data can be found in many other places though, especially online.
- In this lab, we'll read an *observational study* dataset from a website.
  - We'll use this data to then explore what factors are associated with a person's lung capacity.

Our new data

- You can find the data online here:
  - (Right-click and select *Open in New Window*) https://raw.githubusercontent.com/IDSUCLA/dataset/main/fev.csv
- Variables that were measured include:
  - Age in years.
  - Lung capacity, measured in liters.
  - The youth's heights, in inches
  - Genders; `"1"` for males, `"0"` for females.
  - Whether the participant was a smoker, `"1"`, or non-smoker `"0"`.

Importing our data

- Rather than *export*-ing the data and then *upload*-ing and *importing*-ing it, we'll pull the data straight from the webpage into R.
- Click on the *Import Dataset* button under the *Environment* tab.
  - Then click on the *From CSV* option.
  - Type or copy/paste the URL into the box and then hit *Update*.
- Before importing, change the following *Import Options*:
  - Name: `lungs`
  - *Uncheck* the *First Row as Data* box
  - Change *Delimiter* to *Whitespace*

About the data

- The data come from the *Forced Expiratory Volume (FEV)* study that took place in the late 1970's.
  - The observations come from a sample of 654 youths, aged 3 to 19, in/around East Boston.
  - Researchers were interested in answering the *research question*:

  What is the effect of childhood smoking on lung health?

Cleaning your data

- Now that we've got the data loaded, we need to clean it to get it ready for use *(Look at lab 1F for help)*. Specifically:
  - We want to name the variables: `"age"`, `"lung_cap"`, `"height"`, `"gender"`,`"smoker"`, in that order.
  - Change the type of variable for `gender` and `smoker` from *numeric* to *character*.
- After changing the variable types for `gender and smoker`:
  - For `gender`, use `recode` to change `"1"` to `"Male"` and `"0"` to `"Female"`.
  - For `smoker`, use `recode` to change `"1"` to `"Yes"` and `"0"` to `"No"`.

Analyzing our data

- Our `lungs` data is from an observational study.
- **Write down a reason the researchers couldn't use an experiment to test the effects of smoking on children's lungs.**
- Observational studies are often helpful for analyzing how variables are related:
- **Do you think that a person's age affects their lung capacity? Make a sketch of what you think a scatterplot of the two variables would look like and explain.**
- Use the `lungs` data to create an `xyplot` of age and `lung_cap`.
  - **Interpret the plot and describe why the relationship between the two variables makes sense.**

Smoking and lung capacity

- Make a plot that can be used to answer the statistical question:

  Do people who smoke tend to have lower lung capacity than those who do not smoke?

- **Use your plot to answer the question**.
  - **Were you surprised by the answer? Why?**
  - **Can you suggest a possible confounding factor that might be affecting the result?**

Let's compare

- Create three subsets of the data:
  - One that includes *only* 13 year olds ...
  - One that includes *only* 15 year olds ...
  - and one that includes *only* 17 year olds.
- Make a plot that compares the lung capacity of smokers and non-smokers for each subset.
- **How does the relationship between smoking and lung capacity change as we increase the age from 13 to 15 to 17?**

Sum it up!

- **Does smoking affect lung capacity? If so, how?**
  - Support your answers with appropriate plots.
  - Explain why you included the variables you used in your plots.

Are You Asking Me?

Instructional Days: 9

Enduring Understandings

A survey is a data collection method that is administered to a sample. The sample is fraction of the target population. The sample must be representative of the population and random sampling is used to ensure an equal chance of being selected. A census is a special survey that collects data from the entire population. Sampling error and bias cause problems in analysis made from survey data.

Engagement

Students will view a video clip from the game show *Family Feud* to begin to think about survey components. The video can be found at:
https://www.youtube.com/watch?v=-3Nk9t7-rCs

Learning Objectives

Statistical/Mathematical:

S-IC 1: Understand statistics as a process for making inferences about population parameters based on a random sample from that population.

S-IC 3: Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand that bias and sampling error should be minimized when conducting surveys. The wording of questions, as well as who is asked to participate in a survey, can lead to bias. Learn that sampling error can be minimized when larger random samples are collected from a population.

Applied Computational Thinking Using RStudio:

- Create random samples of different sizes to make estimates about a population.
- Create informal confidence intervals based on sample medians.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it is used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will write questions that emphasize differences in data science concepts and skills.

Data File or Data Collection Method

Data File:

1. American Time-Use Survey (ATUS) data

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 9: Survey Says…

Objective:

Students will learn that a survey is another data collection method. They will learn what a survey is, what types of questions are used in a survey, and how a survey is conducted.

Materials:

1. Video: *Family Feud*'s "Shocking Fast Money" found at:
   https://www.youtube.com/watch?v=-3Nk9t7-rCs (good quality, but sad ending)
   OR
   Video: *Family Feud* video clip titled "Family Feud – Comeback of the Century" found at:
   https://www.youtube.com/watch?v=ofQkOfeg60g (bad quality, but happy ending)

2. *Designing a Survey* handout (LMR_3.4_Designing a Survey)

Vocabulary:

survey, self-reported, open-ended questions, closed-ended questions

Essential Concepts: Surveys ask simple, straightforward questions in order to collect data that can be used to answer statistical questions. Writing such questions can be hard (but fun)!

Lesson:

1. Introduce one of the videos listed above by informing students that they will be watching a clip from the television game show *Family Feud*. This segment of the show is called *Fast Money*, where the winning family plays for an additional $20,000. Two family members are chosen to play and must reach a combined score of 200 points to win the money. The goal is to guess the most common responses to five questions. For example, if the question "What animal is a common pet?" were asked, each family member might answer with "dog" or "cat" since these are popular household pets. The first person accumulates as many points as possible during the 20-second first round. The second person is given 25 seconds to earn points with different answers.

2. As students watch the video, have them answer the following questions in their DS journals:

   a. How many people were surveyed? ***100***
   b. Who was represented in the survey? ***Single men***
   c. How many survey questions were asked? ***5***
   d. When the host says "survey said" and we see the response, what does it mean? ***It means that X number of people out of the 100 gave that response to the survey question.***

3. *Family Feud* uses surveys as its main data collection tool. In their DS journals, students should write down what they know about surveys individually.

4. Then, with a partner, students will share what each one of them knows about surveys.

5. Select a couple of students to either share their response or their partner's response with the whole class.

6. Inform students that a **survey** is a data collection method where the data are **self-reported**, meaning that participants answer questions themselves. Surveys are composed of:

   a. Questionnaires or a series of questions
   b. A representative sample of the population of interest
   c. Carefully worded questions

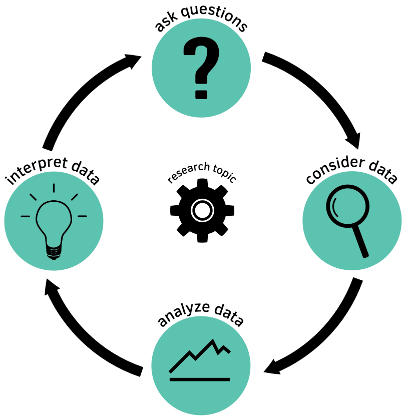
7. Surveys rely on questions. There are two types of questions that can be asked in a survey: **open-ended questions** and **closed-ended questions**. Open-ended questions offer a free-response/text approach, whereas closed-ended questions give a fixed set of choices.

8. Display the following list to the class. With a partner, have the students categorize the following types of questions as either open-ended or closed-ended:

   (a.) Multiple choice *(closed)*
   (b.) Write a paragraph *(open)*
   (c.) Yes/No *(closed)*
   (d.) Comments *(open)*
   (e.) Essays *(open)*
   (f.) On a scale from 1 to 5 *(closed)*
   (g.) Choose from a list *(closed)*
   (h.) Write a sentence *(open)*
   (i.) Check a box *(closed)*

9. Do a quick *Whip Around* to share the categorization for each type of question. Be sure that students make corrections to the list if any items were miscategorized.

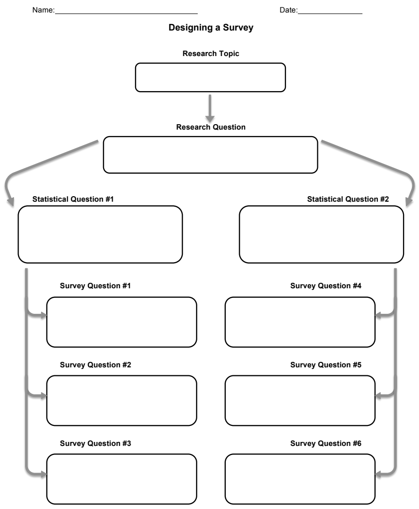
10. Quickly review the Data Cycle.



11. To give students an introduction to conducting surveys, they will first go through a practice scenario as a class to try to answer the following research question:

What are 'families' in the United States?

12. Distribute the *Designing a Survey* handout (LMR_3.4) and let students fill in the boxes for "Research Topic" ***(Families)*** and "Research Question" ***(What are 'families' in the United States?)***.


Name:______________________________ Date:______________

Designing a Survey

Research Topic

Research Question

Statistical Question #1

Statistical Question #2

Survey Question #1

Survey Question #4

Survey Question #2

Survey Question #5

Survey Question #3

Survey Question #6

LMR_3.4

13. Inform students that the left side of the handout will be completed as a class, and then student teams will work together to complete the right side.

14. Using the Data Cycle as a guide, students should brainstorm a statistical question that is related to the research question. One might be: ***What is the typical family size in the United States?***

    Note: This requires a definition of "family," which can have a variety of meanings to different people. Different definitions will likely guide the discussion of possible survey questions in the following step.

15. Next, students need to determine 3 survey questions to help answer the statistical question. The goal in creating survey questions is to make sure they (1) are unambiguous, and (2) address the statistical question. Some examples are listed below (which come from different definitions of "family"):

    Note: Survey questions MUST match the statistical question.

    (a.) How many siblings do you have?
    (b.) How many people live with you?

16. It may help to actually collect data once the first survey question has been created. For example: "How many siblings do you have?" – each student would give a response and the values could be recorded in a dotplot (if desired). If the question is too vague (do we include half-siblings, step-siblings, etc.?), students can revise the question.

17. Once the class has agreed upon 3 survey questions for the first statistical question, allow students to join their student teams for the remainder of the activity.

18. Each team should come up with a statistical question that might answer the research question, then determine 3 survey questions that match their statistical question. Have the students create both open- and closed-ended questions in the handout. Each survey question should be a different type (see Step 8).

19. Have student teams share out their statistical questions and related survey questions with the rest of the class.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

For homework, students should choose one of their team's survey questions and rewrite it 3 ways, using 3 different question types (see Step 8). Example rewrites for the statistical question "How many siblings do you have?" are given below for reference.

(a.) Multiple choice:

How many siblings do you have? Select one option.

(a) 0 siblings
(b) 1 sibling
(c) 2 siblings
(d) 3 siblings
(e) more than 3 siblings

(b.) Write a paragraph:

In your own words, describe your siblings.

(c.) Yes/No:

Do you have any siblings?

Lesson 10: We're So Random

Objective:

Students will learn how to collect random samples from a population in order to estimate a parameter.

Materials:

1. *Populations & Samples* handout (LMR_3.5_Populations and Samples)
2. RStudio
3. Projector for RStudio functions
4. Dotplot titled "Percent of Students Who Have Met Friends Online"
5. *Parameters & Statistics* handout (LMR_3.6_Parameters and Statistics)

Vocabulary:

population, sample, representative, random sample, parameter, statistic

Essential Concepts: Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to 'center' on the population parameter value.

Lesson:

1. Introduce today's lesson by displaying the following statement made by the Pew Research Center in their August 2015 report titled *Teens, Technology & Friendships*:

   > "For today's teens, friendships can start digitally: 57% of teens have met new friends online. The margin of error is plus or minus 3.7 percentage points. Social media and online gameplay are the most common digital venues for meeting friends."

   Note: The data for this report were collected via interviews of 1,060 teenagers between the ages of 13 and 17.

2. Discuss the results of the Pew poll with the following prompting questions:

   a. The report says that 57% of teens have met new friends online. Since the report was based on a sample of 1,060 teens, how many of the teens reported that they have met friends online? ***0.57(1060) = 604.2. This means that approximately 604 teens in the sample have met friends online.***

   b. Do you think 57% of students in our class have met friends online? Why or why not? ***Answers will vary by class. The discussion should include points about how similar and different samples can be. The sample of students in the Pew poll may not represent the students in our class.***

   c. What percent of students in our class have met friends online while a teenager? ***Answers will vary by class. Calculate the percentage by dividing the number of students who have met friends online by the total number of students who came to class today.***

   d. What if [absent person] were in class today? Would that change our percentage? What effect would it have on the percentage if [absent person] answered "yes?" What effect would it have if [absent person] answered "no?" ***Answers will vary by class. If a student who was absent for today's lesson had actually come to class, we would have a different sample of students. It would change the percentage because our sample size now includes 1 more person. If the person answered "yes," the values in the numerator and denominator of the percentage would change. If the person answered "no," the value in the denominator would change. (Students should calculate these values.)***

   e. If we were able to interview every single teenager in the United States, would exactly 57% of them say they have met friends online? ***Probably not because there are many more teenagers in the US than the 1,060 they interviewed. It would be unlikely for a group of 1,060 teenagers to exactly represent all teenagers in the entire country.***

    f. Why do you think the Pew Research Center only interviewed 1,060 teenagers, and not all teenagers in the US? ***It would be impossible to talk to all teenagers in the US in a short period of time, or even a fairly long period of time.***

3. Introduce students to the terms **population** and **sample**. Explain that a population consists of all of the people we want to learn something about. A sample consists of people (or objects) that are selected *from* the population. In pairs, ask students to discuss and record answers to the following two questions:

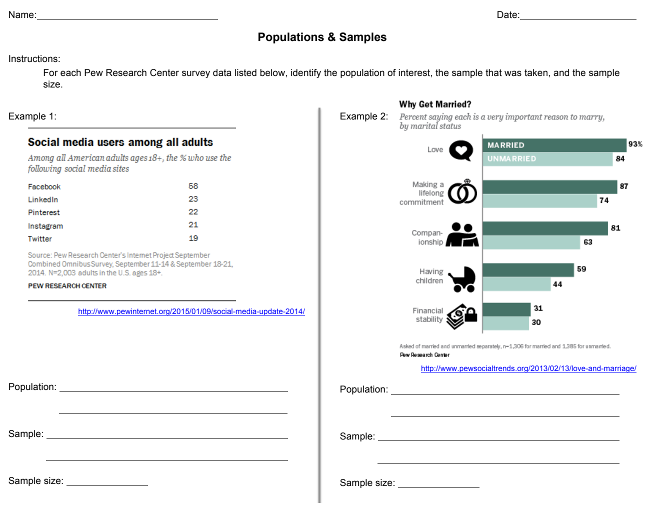
    a. What was the population of interest to the researchers for the Pew poll above? ***All teenagers in the US right now.***

    b. Based on your answer in (a), what characteristics should people have in order to be included in a sample for this poll? ***People would need to be in the US and be between the ages of 13 and 17. People could be from many states, but we would not want to sample only people from California, or only people from Los Angeles. It would be impossible to survey every single person in the US; this is why we create a random*** **representative** ***sample of the population.***

    Note: Steer the discussion so that students see that a sample has to be "like" or "similar to" or "representative of" the population.

4. Select pairs to share their responses to the questions and let students revise their responses.

5. Distribute the *Populations & Samples* handout (LMR_3.5), which contains survey results from other Pew Research Center reports. Give students time to determine the population and sample for each scenario, and then have them verify their results with a partner.



Name:______________________________ Date:____________________

Populations & Samples

Instructions:

For each Pew Research Center survey data listed below, identify the population of interest, the sample that was taken, and the sample size.

Example 1:

Social media users among all adults

Among all American adults ages 18+, the % who use the following social media sites

Facebook	58
LinkedIn	23
Pinterest	22
Instagram	21
Twitter	19

Source: Pew Research Center's Internet Project September Combined Omnibus Survey, September 11-14 & September 18-21, 2014. N=2,003 adults in the U.S. ages 18+.

PEW RESEARCH CENTER

http://www.pewinternet.org/2015/01/09/social-media-update-2014/

Population: ______________________________

Sample: ______________________________

Sample size: ____________

Example 2:

Why Get Married?

Percent saying each is a very important reason to marry, by marital status

	MARRIED	UNMARRIED
Love	93%	84
Making a lifelong commitment	87	74
Compan-ionship	81	63
Having children	59	44
Financial stability	31	30

Asked of married and unmarried separately, n=1,306 for married and 1,385 for unmarried.

Pew Research Center

http://www.pewsocialtrends.org/2013/02/13/love-and-marriage/

Population: ______________________________

Sample: ______________________________

Sample size: ____________

LMR_3.5

6. State that we want to know the percentage of students in our class that have made friends online, but we don't want to ask every single student. Instead, we would like to ask only 10 students and then make some guesses about our class from those 10. Ask:

    a. What is the population of interest? ***The students in our classroom.***

    b. How can we select 10 students to be part of our sample? ***Answers will vary by class. There may be a variety of suggestions; here are some examples of what may be given: (1) put every student's name in a hat and pick out 10; (2) select the 10 students sitting closest to the teacher's desk; (3) have 10 students volunteer to be in the sample.***

7. Inform students that, in general, we want samples to "look like" the population. One way to get a representative sample like this is to take a **random sample**. Ask the students:

a. Would the selection techniques we came up with in Step 6 result in random samples of our class. ***Answers will vary by class. Using the examples from Step 6: (1) putting each student's name in a hat and then picking out 10 would be a random sample as long as each piece of paper is the same size; (2) selecting the 10 students sitting closest to the teacher's desk would not be a random sample because those students might not represent the whole class; (3) if 10 students volunteer, we would not have a random sample because those students selected themselves to be part of the group and may not represent everyone in the class.***

8. Next, assign each student in the class a number by having them count off from 1 to *N* (*N* being the total number of students in the class). Show students that we can use RStudio to create random samples with the following function:

```
> sample(1:N, size = 10, replace = FALSE)
```

Note: we use `replace = FALSE` because we only want each student to be selected once.

9. Using the results given in the output of RStudio, ask the students whose numbers were chosen to stand. Inform them that they are "in" the sample. Then, determine what percent of the sample (these 10 students) have made friends online. How does this percentage compare to the overall class percentage we found in Step 2(c)? ***Answers will vary by class.***

10. Create a dotplot on the board titled "Percent of Students Who Have Met Friends Online." Record the sample percentage from Step 9 on this dotplot.

11. Have the students return to their seats so that we can select a new sample. Before we do this, ask:

    a. What do you predict the percentage to be for the next sample of 10 students? ***Answers will vary by class. They might say the expected percentage will be close to the class's overall percentage.***

12. Using RStudio, create a new sample, calculate the percentage of those 10 students who have met friends online, and record the value in the dotplot.

13. Repeat Step 11 for a few more rounds (at least 5 samples should be taken). Be sure that the students give a prediction before finding each new sample.

14. Display the following questions. Refer to the dotplot of sample percentages. Ask students to discuss the questions in teams:

    a. What do you notice about the sample percentages? What is the "typical" value? ***Answers will vary by class. The typical value should be close to the class's overall percent calculated in Step 2(c) since it is the population percent.***
    b. What is the smallest value? What is the biggest value? ***Answers will vary by class. There might be a lot of variability in the dotplot based on the selected samples. Most sample percentages will be within 30% of the population value, so that really gives a wide variety of possible sample values.***
    c. If we took a larger sample – maybe of size 15 or 20 – would there be more or less variability in the dotplot? ***There will be less variability because adding more people gets us closer to the population size. Be sure to point out that if the sample were exactly the same as the population, then there would be no variability in the plot.***

15. Select teams at random to share their responses to the questions above with the whole class. The rest of the teams should be in full agreement with the responses before moving on to the next question.

16. Explain that the population percentage (the percentage of all students in the class who met friends online) we have been using is called a **parameter**. A parameter is any number that summarizes a population. So, our class has been the population, and the percentage of students that have met friends online is the parameter.

17. Similarly, the term **statistics** is used for numbers that summarize a sample. Ask students what sample statistics they have seen today? ***Each value we included in the dotplot is a statistic.***

18. Be sure to point out that there can be multiple values for a sample statistic (i.e. “We had 5 sample percentages in our dotplot.”), but there is always only one parameter value.

19. Using these new definitions, ask students to consider the original Pew poll data, which found that 57% of teens have met friends online. Ask the students:

    a. Is 57% a parameter or statistic? ***This is a statistic because it is based on a sample. Remind them that the population was ALL US teenagers and the sample included 1,060 teens.***

    b. What is the population parameter then? ***We don’t know! We would have to interview every teenager in the US to determine the parameter, and that is not possible.***

20. Conclude the lesson by telling the students: although we cannot determine the actual population parameter for the percent of teens that have made friends online, we can estimate it using random samples.

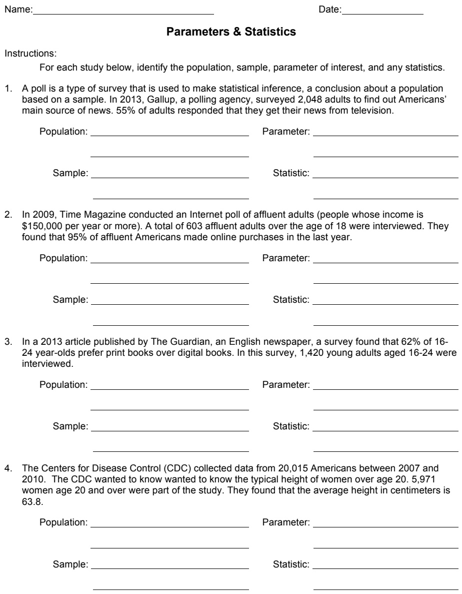
Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students should complete the *Parameters & Statistics* handout (LMR_3.6) for homework.

<u>Note:</u> Page 2 of the handout is an answer key for teacher reference only!



Name:______________________________ Date:______________

Parameters & Statistics

Instructions:

For each study below, identify the population, sample, parameter of interest, and any statistics.

1. A poll is a type of survey that is used to make statistical inference, a conclusion about a population based on a sample. In 2013, Gallup, a polling agency, surveyed 2,048 adults to find out Americans’ main source of news. 55% of adults responded that they get their news from television.

   Population: ________________________ Parameter: ________________________
   ________________________ ________________________
   Sample: ________________________ Statistic: ________________________
   ________________________ ________________________

2. In 2009, Time Magazine conducted an Internet poll of affluent adults (people whose income is $150,000 per year or more). A total of 603 affluent adults over the age of 18 were interviewed. They found that 95% of affluent Americans made online purchases in the last year.

   Population: ________________________ Parameter: ________________________
   ________________________ ________________________
   Sample: ________________________ Statistic: ________________________
   ________________________ ________________________

3. In a 2013 article published by The Guardian, an English newspaper, a survey found that 62% of 16-24 year-olds prefer print books over digital books. In this survey, 1,420 young adults aged 16-24 were interviewed.

   Population: ________________________ Parameter: ________________________
   ________________________ ________________________
   Sample: ________________________ Statistic: ________________________
   ________________________ ________________________

4. The Centers for Disease Control (CDC) collected data from 20,015 Americans between 2007 and 2010. The CDC wanted to know wanted to know the typical height of women over age 20. 5,971 women age 20 and over were part of the study. They found that the average height in centimeters is 63.8.

   Population: ________________________ Parameter: ________________________
   ________________________ ________________________
   Sample: ________________________ Statistic: ________________________
   ________________________ ________________________

LMR_3.6

Lesson 11: The Gettysburg Address

Objective:

Students will learn the definition of sampling bias and will learn that random samples reduce bias when estimating a population parameter. They will gain practice collecting a random sample from a small population and estimating the population parameter.

Materials:

1. *Gettysburg Address* handout (LMR_3.7_Gettysburg_Address)
2. *Sampling the Gettysburg Address* handout (LMR_3.8_Sampling the Gettysburg Address)
3. Dotplot titled "Mean Word Length, Self-Selected Sample, Size = 10" – on board or poster paper
4. *Gettysburg Address – Word Length Histogram* file (LMR_3.9_Gettysburg Histogram)
5. *Gettysburg Word Lengths* handout (LMR_3.10_Gettysburg_Words)
6. RStudio
7. Projector for RStudio functions
8. Dotplot titled "Mean Word Length, Random Sample, Size = 10" – on board or poster paper

   Note: This dotplot will be used again during Lesson 13, so the results need to be kept in some way (this can be either on poster paper or by simply taking a photo).

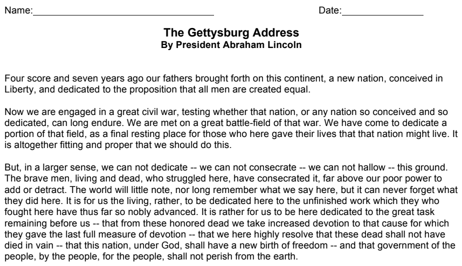
Vocabulary:

sampling bias

Essential Concepts: Statistics vary from sample to sample. If the typical value across many samples is equal to the population parameter, the statistic is 'unbiased.' Bias means that we tend to "miss the mark." If we don't do random sampling, we can get biased estimates.

Lesson:

1. Introduce the lesson by describing the Gettysburg Address:
   a. President Lincoln delivered the Gettysburg Address in November 1863.
   b. It is one of the most famous speeches in the United States.
   c. In it, Lincoln invoked the principles of human equality contained in the U.S. Constitution and Declaration of Independence.
2. Read the Gettysburg Address aloud to the class OR have students read it aloud. The text of the speech can be found in the *Gettysburg Address* handout (LMR_3.7).



Name:______________________________ Date:______________

The Gettysburg Address
By President Abraham Lincoln

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

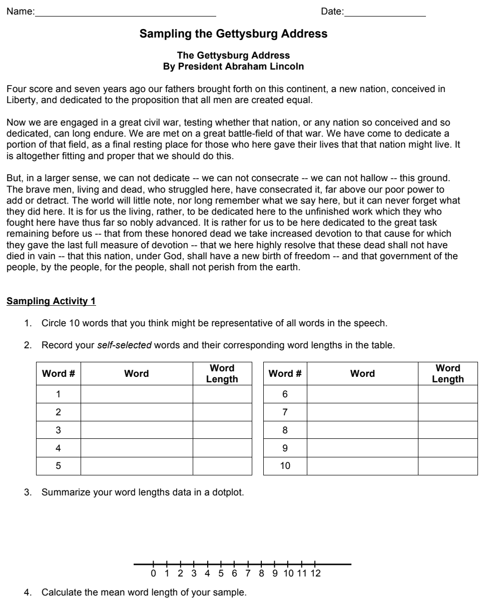
But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

LMR_3.7

3. Today we will use the Gettysburg Address to learn about different sampling techniques.
4. Inform students that the Gettysburg Address contains 272 words. We can consider these 272 words to be our population because it includes all words in the entire speech. From the population, we can sample 10 words that we think represent the speech. It is ok for this step to be vague – students can come up with their own concept for what they think "representative" means in this case.

5. Distribute the *Sampling the Gettysburg Address* handout (LMR_3.8), which includes the actual speech, as well as 2 sampling activities. For this part of the lesson, we will only be looking at Sampling Activity 1 on page 1 of the handout.

   Note: This activity was originally created by Allan Rossman and Beth Chance, and has been modified for the IDS curriculum.



Name:______________________________ Date:______________

Sampling the Gettysburg Address

The Gettysburg Address
By President Abraham Lincoln

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

Sampling Activity 1

1. Circle 10 words that you think might be representative of all words in the speech.
2. Record your *self-selected* words and their corresponding word lengths in the table.

Word #	Word	Word Length
1		
2		
3		
4		
5		

Word #	Word	Word Length
6		
7		
8		
9		
10		

3. Summarize your word lengths data in a dotplot.

4. Calculate the mean word length of your sample.

LMR_3.8

6. Inform students that they will get 30 seconds to select 10 words that they think are representative of all words in the speech.

   Note: It is important that students work fast so they are forced to choose based on first impressions and don't have time to reflect. Also, this activity tends to not work well if students are informed of the punch line (that random samples are unbiased) before they begin.

7. At this point, explain to students that we are actually interested in answering a specific question:

   What is the typical <u>word length</u> in the Gettysburg Address?

8. Next, students should record each circled word, as well as the number of letters each word has (this is the word length) in the table on the handout. Then, they should summarize the data in a dotplot and calculate the mean word length of the sample.

9. On the board, create a class dotplot (may also be done on poster paper) titled “Mean Word Length, Self-Selected Sample, Size = 10.” Once all students have completed the first page of the *Gettysburg Address* handout (LMR_3.8), ask each student to record the mean word length of his or her sample on the class’s dotplot.

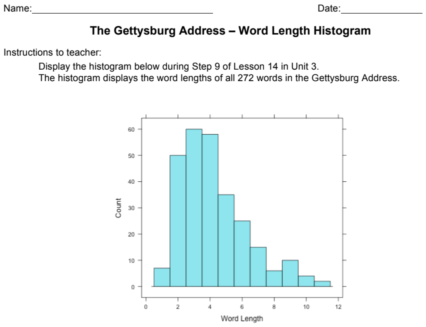
10. When all students have recorded their sample statistics in the dotplot, conduct a class discussion based on the questions listed below.

    Note: You might need to do a reality check. Students will often make mistakes when adding the word lengths and when dividing. Be suspicious (and double-check) extreme values.

    a. What does each point on the plot represent? ***Each point represents one student's estimate of the mean length of all of words in the Gettysburg address.***
    b. What is the typical value represented in the dotplot? ***Answers will vary by class. You should indicate the approximate location of the mean of the distribution (the balancing point, on the dotplot. Remind students that when we ask for the 'typical' value we mean the value in the center of the distribution.***

    c. How much variability is there in the distribution? ***Answers will vary by class. One reasonable approach is for students to give the range (the difference between the largest and smallest values).***
    d. What is the shape of the distribution? ***Answers will vary by class. Often, the shape is right-skewed, but it might not be for you. Note that outliers here will often be arithmetic errors, but not always.***

11. Next, display the histogram from the *Gettysburg Address – Word Length Histogram* file (LMR_3.9), which shows the distribution of word lengths for the entire population of words in the Gettysburg Address.



Name:______________________________ Date:______________

The Gettysburg Address – Word Length Histogram

Instructions to teacher:

Display the histogram below during Step 9 of Lesson 14 in Unit 3.
The histogram displays the word lengths of all 272 words in the Gettysburg Address.

LMR_3.9

12. Remind students that the population is the 272 words from the speech, and inform them that the mean word length of the population, or the population parameter, is 4.22. Using *Think-Pair-Share*, ask:

    a. How does the typical value of our class's sample means compare to the actual population mean of 4.22? ***Almost always, the class's typical mean will be higher (sometimes much higher) than 4.22. Some students will be close to 4.22. But point out that we are talking about the "trend" or typical outcome. The typical outcome is usually too high.***

13. Explain that self-selected samples often produce biased results. **Sampling bias** is a description of the process, or the sampling plan, that is used to collect data. If the resulting samples tend to produce results that are influenced in one particular direction, we say that the sampling plan is biased.

    Note: Bias is NOT the same as prejudice. Bias is a tendency to lean towards a certain belief or viewpoint, and is mostly unconscious. Prejudice is a very conscious phenomenon though, where a person actively makes a decision to dislike something based on unfounded facts.

14. Refer back to the dotplot of sample means and point out how it is biased. Ask:

    a. Why was our original sampling procedure biased? ***When we look for 'representative' words, and do so quickly, our eyes are drawn by the larger, more unusual words, and we tend to overlook small words such as "in," "a," "we," etc.***

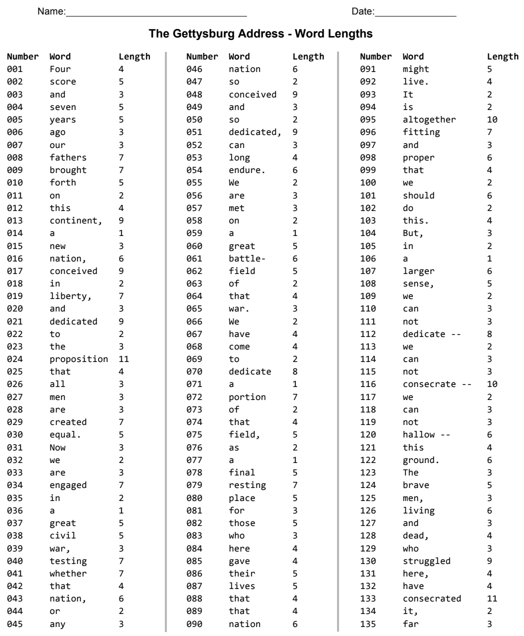
15. Go back to the *Gettysburg Address* handout (LMR_3.8), and direct students to page 2 for Sampling Activity 2. Inform students that they will now do a sampling procedure that results in a better representation of the population of words in the speech.

16. Explain that a random sample tends to produce unbiased sample results.

17. Before students begin the activity, demonstrate how to generate 10 random numbers from a possible 272 using RStudio.

```
> sample((1:272), size = 10, replace = FALSE)
```

18. Each student should generate his or her own set of 10 random numbers. Once students have created their random numbers, distribute the *Gettysburg Address Word Lengths* handout (LMR_3.10).



Name:______________________________ Date:______________

The Gettysburg Address - Word Lengths

Number	Word	Length	Number	Word	Length	Number	Word	Length
001	Four	4	046	nation	6	091	might	5
002	score	5	047	so	2	092	live.	4
003	and	3	048	conceived	9	093	It	2
004	seven	5	049	and	3	094	is	2
005	years	5	050	so	2	095	altogether	10
006	ago	3	051	dedicated,	9	096	fitting	7
007	our	3	052	can	3	097	and	3
008	fathers	7	053	long	4	098	proper	6
009	brought	7	054	endure.	6	099	that	4
010	forth	5	055	We	2	100	we	2
011	on	2	056	are	3	101	should	6
012	this	4	057	met	3	102	do	2
013	continent,	9	058	on	2	103	this.	4
014	a	1	059	a	1	104	But,	3
015	new	3	060	great	5	105	in	2
016	nation,	6	061	battle-	6	106	a	1
017	conceived	9	062	field	5	107	larger	6
018	in	2	063	of	2	108	sense,	5
019	liberty,	7	064	that	4	109	we	2
020	and	3	065	war.	3	110	can	3
021	dedicated	9	066	We	2	111	not	3
022	to	2	067	have	4	112	dedicate --	8
023	the	3	068	come	4	113	we	2
024	proposition	11	069	to	2	114	can	3
025	that	4	070	dedicate	8	115	not	3
026	all	3	071	a	1	116	consecrate --	10
027	men	3	072	portion	7	117	we	2
028	are	3	073	of	2	118	can	3
029	created	7	074	that	4	119	not	3
030	equal.	5	075	field,	5	120	hallow --	6
031	Now	3	076	as	2	121	this	4
032	we	2	077	a	1	122	ground.	6
033	are	3	078	final	5	123	The	3
034	engaged	7	079	resting	7	124	brave	5
035	in	2	080	place	5	125	men,	3
036	a	1	081	for	3	126	living	6
037	great	5	082	those	5	127	and	3
038	civil	5	083	who	3	128	dead,	4
039	war,	3	084	here	4	129	who	3
040	testing	7	085	gave	4	130	struggled	9
041	whether	7	086	their	5	131	here,	4
042	that	4	087	lives	5	132	have	4
043	nation,	6	088	that	4	133	consecrated	11
044	or	2	089	that	4	134	it,	2
045	any	3	090	nation	6	135	far	3

LMR_3.10

19. Explain that the table contains the word number, word, and length of each word in the Gettysburg Address. Demonstrate how to find a word that corresponds to one of the random numbers generated by RStudio, and explain that this word is now part of our random sample.

20. Then, each student will complete the handout by creating a dotplot and calculating the mean of their random sample.

21. On the board, near the first dotplot, create another class dotplot (may also be done on poster paper) titled "Mean Word Length, Random Sample, Size = 10." Once all students have completed the second page of the *Gettysburg Address* handout (LMR_3.8), ask each student to record the mean word length of his or her random sample on the class's dotplot.

    Note: As in Step 9, be sure to check arithmetic for outliers!

22. When all students have recorded their sample statistics in the dotplot, conduct a class discussion based on the following questions:

    a. What does each point in the dotplot represent? ***Each dot represents one student's estimate of the mean word length. But this time, the estimates are based on a random sample of 10 words.***
    b. What do you notice about the typical value in this distribution? ***Answers will vary by class. They should notice that the means of the random samples are fairly close to the population mean of 4.22. (Again, you might have to discard or correct outliers.)***
    c. What shape does this distribution have? What does that tell us? ***Typically, the distribution of sample means for random samples will be symmetric and unimodal.***
    d. What does this distribution tell us about the benefits of random samples? ***We can reduce bias by collecting random samples instead of self-selected samples.***
    e. Why do we need sampling? Why can't we just get the information from the actual population? ***It is usually impossible to include every person or object from a population. Even for the population of size 272 words in the Gettysburg Address, it would take a long time to calculate the word lengths of every single word.***

23. Conclusions and takeaways:
    a. It turns out that there are approximately 5.17•10^{17} different possible samples of 10 words from the Gettysburg Address.
    b. If we could determine the mean for each of these samples and produce a dotplot for all of these means, then the center of the dotplot would lie exactly at 4.22.
    c. The resulting distribution of the means from all possible samples is called the sampling distribution for the sample mean (for samples of size 10 from this population).
    d. The above dotplot is an approximation to the actual sampling distribution.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

Students should write a reflection about why random sampling is better at reducing bias than other sampling procedures.

Lab 3C: Random Sampling

Complete Lab 3C prior to Lesson 12.

<u>Lab 3C - Random Sampling</u>

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Learning by sampling

- In many circumstances, there's simply no feasible way to gather data about everyone in a *population*.
  - For example, the Department of Water & Power (DWP) wants to determine how much water people in Los Angeles use to take a shower. They've created a survey to pass out to collect this information.
  - **Write down two reasons why getting *everyone* in Los Angeles to fill out the survey would be difficult. Also, write a sentence why the DWP might consider using a sample of households instead.**
- In this lab, we'll learn how *sampling methods* affect how *representative* a sample is of a *population*.

Loading a population

- In previous labs, we used the `cdc` data as a sample for young people in the United States.
  - In this lab, we'll consider these survey respondents to be our population.
- Load the `cdc` data into `R` and fill in the blanks to take a *convenience* sample of the first 50 people in the data:

```
s1 <- slice(____, 1:____)
```

- **Why do you think we call this method a *convenience* sample?**

Comparing your convenience sample

- A convenience sample is a sample from a population where we collect data on subjects because they're easy-to-find.
- Using your convenience sample, create a `bargraph` for the number of people in each `grade`.
  - **Do you think the distribution of `grade` for your sample would look similar when compared to the whole `cdc` data?**
  - **Which groups of people do you think are over or under represented in your convenience sample? Why?**
- Create a `bargraph` for `grade` using the `cdc` data.
  - **Compare the distributions of the `cdc` data and your convenience sample and write down how they differ.**

Using randomness

- Fill in the blanks below to create a sample by randomly selecting 50 people in the `cdc` data, without replacement. Call this new sample `s2`:

```
___ <- sample(___, size = ___, replace = ___)
```

- **Write a sentence that explains why you think the distribution of `grade` for this *random sample* will look more or less similar to the distribution from the whole `cdc` data.**
  - Create a `bargraph` for `grade` based on this *random sample* to check your prediction.

Increasing sample size

- Create `bargraphs` for `grade` based on each of the following sample sizes: 10, 100, 1,000, 10,000.
  - Compare each distribution to that of the population.
- **How do the distributions change as the size of the sample increases? Why do you think this occurs?**
- `tally()` the proportion of `grades` for your *convenience* sample and all your *random* samples.
  - **Which set of proportions looks most similar to the proportions of the population?**

Lessons learned

- The mean, or proportion, from a *random* sample might not always be closer to that of the true population when compared to a *convenience* sample.
- However, as sample sizes get larger:
  - *Random* samples will tend to be better estimates for the population.
  - With *convenience* samples, this might not be the case.
- **Write down a reason why estimates based on *convenience* samples might not improve even as sample size increases.**

Lesson 12: Bias in Survey Sampling

Objective:

Students will learn about bias in relation to survey sampling. More specifically, they will learn what types of sampling methods could result in a biased sample, who might be over/under-represented in the sample, and how to select a better sample.

Materials:

1. *Identifying Biased Samples* handout (LMR_3.11_Identifying Biased Samples)
2. Poster paper

Vocabulary:

survey sample, over-represented, under-represented, random sampling

Essential Concepts: Another popular data collection method involves collecting data from a random sample of people or objects. Percentages based on random samples tend to 'center' on the population parameter value.

Lesson:

1. Remind students that they learned about biased samples during the last few lessons. Today, they will continue with this topic and discuss how people are selected to be in a sample.
2. The people who are asked to participate in a survey are known as the **survey sample**. Ideally, the people who are included in the survey sample are a representative group of the target population, or the population we would like to make inferences about.
3. Propose the following scenario to the class: "An elementary school is going to start serving ice cream in the cafeteria every Friday during lunch, and needs to know the favorite flavor of its students."
4. In pairs, ask students to come up with two examples of samples that might be biased. For instance, one biased sample might include only the four 3rd grade classes at the school. For each biased sample, the students should answer the following questions in their DS journals:
    a. Who is the target population? ***All students at the elementary school.***
    b. Who is included in your biased sample? ***Only 3rd grade students. These students are*** **overrepresented** ***in the sample.***
    c. Who is not included in your biased sample? ***All other students in the school (Kindergartners, 1st, 2nd, 4th, and 5th graders). These students are*** **underrepresented** ***in the sample.***
    d. Is your sample representative of the target population? ***No! We're only including 3rd graders, and they may not have the same preferences as other students.***
5. Once pairs have come up with their biased samples and answered the questions in Step 4, they should share out with their student teams and answer the following questions:
    a. How is your biased sample different from the samples created by other pairs in your team? ***Answers will vary by class. An example might be that one pair sampled only 3rd graders and the other pair sampled only girls.***
    b. Which do you think is more representative of the target population? Why? ***Answers will vary by class. Using the example above, we could argue that either one is more representative. We could maybe say that, since 3rd graders include both boys and girls, we have a more representative sample than if we just sampled girls. Or, we could say that since girls come from all grade levels, they're more representative of the entire school than just 3rd graders.***

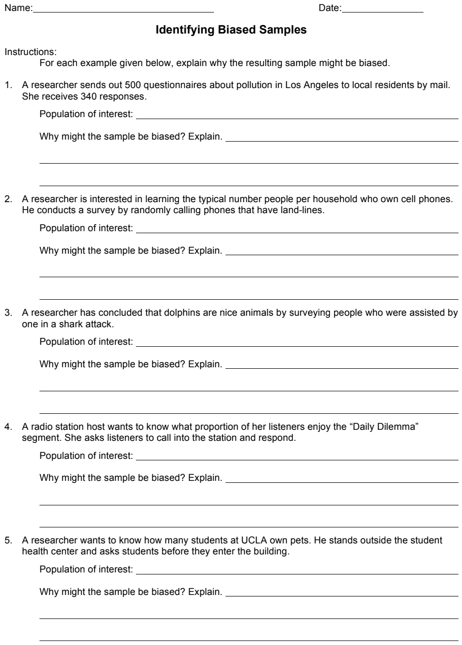
6. After the teams have discussed their samples, they should select one pair's biased sample to share with the rest of the class. Record each team's biased sample on a sheet of poster paper with the following layout:

Biased Sample	Who is overrepresented?	Who is underrepresented?
All 3rd grade students	*3rd grade students*	*All other students (kindergartners, 1st, 2nd, 4th, and 5th graders)*

7. Distribute the *Identifying Biased Samples* handout (LMR_3.11). In this activity, students will explain why a particular sampling method might result in a biased sample – a sample that is not representative of the population of interest.

   Note: It is NOT enough for students to say that the "sample is not random." They need to explain how the sample is biased.

   Note: Page 2 of the handout provides sample answers for teacher reference ONLY. Do NOT distribute page 2 to students.



Name:_______________________________ Date:_______________

Identifying Biased Samples

Instructions:
For each example given below, explain why the resulting sample might be biased.

1. A researcher sends out 500 questionnaires about pollution in Los Angeles to local residents by mail. She receives 340 responses.

   Population of interest: ______________________________

   Why might the sample be biased? Explain. ______________________________

2. A researcher is interested in learning the typical number people per household who own cell phones. He conducts a survey by randomly calling phones that have land-lines.

   Population of interest: ______________________________

   Why might the sample be biased? Explain. ______________________________

3. A researcher has concluded that dolphins are nice animals by surveying people who were assisted by one in a shark attack.

   Population of interest: ______________________________

   Why might the sample be biased? Explain. ______________________________

4. A radio station host wants to know what proportion of her listeners enjoy the "Daily Dilemma" segment. She asks listeners to call into the station and respond.

   Population of interest: ______________________________

   Why might the sample be biased? Explain. ______________________________

5. A researcher wants to know how many students at UCLA own pets. He stands outside the student health center and asks students before they enter the building.

   Population of interest: ______________________________

   Why might the sample be biased? Explain. ______________________________

LMR_3.11

8. Each student should complete the handout independently. Afterwards, conduct a whole-class discussion to compare and contrast different students' explanations of how the samples might be biased. For each example given in the handout, discuss who is most likely over-represented and who is most likely under-represented in the sample.

9. Ask the students:

   a. Now that we have learned about sampling biases, how can we eliminate this type of bias? ***Answers will vary by class.***

   Note: Allow students to collaborate and come up with a few ideas on their own of how to eliminate sampling bias. If desired, ideas can be written on the board for discussion and comparison.

10. Conclude with the actual answer: **random sampling**.

    a. If we randomly sample people from our population of interest, we can reduce the bias of any sample statistics obtained from the survey responses.

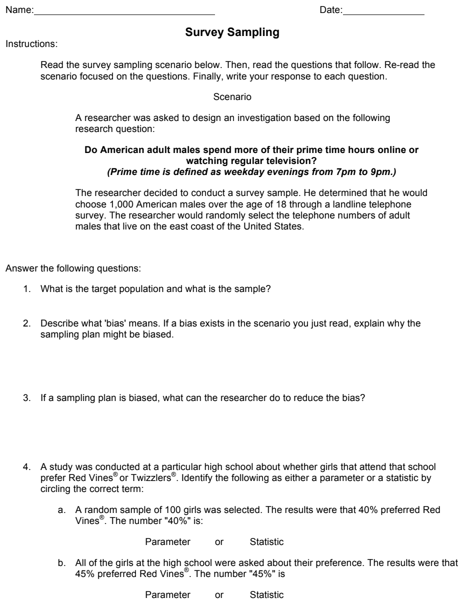
b. If we have a biased sample, we can only give descriptions about that particular sample; we CANNOT generalize to the population of interest.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

☑ Students will complete the *Survey Sampling* handout (LMR_3.12) for homework.



Name:______________________________ Date:______________

Survey Sampling

Instructions:

Read the survey sampling scenario below. Then, read the questions that follow. Re-read the scenario focused on the questions. Finally, write your response to each question.

Scenario

A researcher was asked to design an investigation based on the following research question:

Do American adult males spend more of their prime time hours online or watching regular television?
(Prime time is defined as weekday evenings from 7pm to 9pm.)

The researcher decided to conduct a survey sample. He determined that he would choose 1,000 American males over the age of 18 through a landline telephone survey. The researcher would randomly select the telephone numbers of adult males that live on the east coast of the United States.

Answer the following questions:

1. What is the target population and what is the sample?

2. Describe what 'bias' means. If a bias exists in the scenario you just read, explain why the sampling plan might be biased.

3. If a sampling plan is biased, what can the researcher do to reduce the bias?

4. A study was conducted at a particular high school about whether girls that attend that school prefer Red Vines® or Twizzlers®. Identify the following as either a parameter or a statistic by circling the correct term:

   a. A random sample of 100 girls was selected. The results were that 40% preferred Red Vines®. The number "40%" is:

      Parameter or Statistic

   b. All of the girls at the high school were asked about their preference. The results were that 45% preferred Red Vines®. The number "45%" is

      Parameter or Statistic

LMR_3.12

Lesson 13: The Confidence Game

Objective:

Students will learn about informal confidence intervals for making estimates about population parameters based on statistics from random samples.

Materials:

1. *The Confidence Game* handout (LMR_3.13_Confidence Game)
2. Dotplot titled "Number Correct" displayed on the board (or on poster paper)

Vocabulary:

inferences, interval, confidence interval

Essential Concepts: There is uncertainty when we estimate population parameters. Because of this, it is better to give a range of plausible values, rather than a single value.

Lesson:

1. Remind students that they have been learning about why sampling allows us to make **inferences** about a population. Some methods of sampling produce biased sample statistics, which does not allow us to generalize the results from a sample to the population of interest. To obtain unbiased statistics, random sampling methods need to be used.
2. Conduct a class brainstorm about what it means to "estimate" something. Have students come up with possible synonyms for the word "estimate." ***Some example synonyms include: guess, approximation, projection, opinion, impression, etc.***
3. Inform students that, in statistics, to provide an estimate means that we can give a range of values that we are confident include the population parameter value.
4. In today's lesson, explain that the students will be playing a game, called *The Confidence Game*, in which they will be asked a series of questions that each have one numerical answer. However, instead of guessing what the exact answer is, the students will create a range of possible values that they think might include the real answer. They should be 90% confident that the true value is within their interval.
5. Introduce *The Confidence Game* to students by first going through an example using the question:

   How tall is the Empire State Building, in feet (including the spire at the very top)?

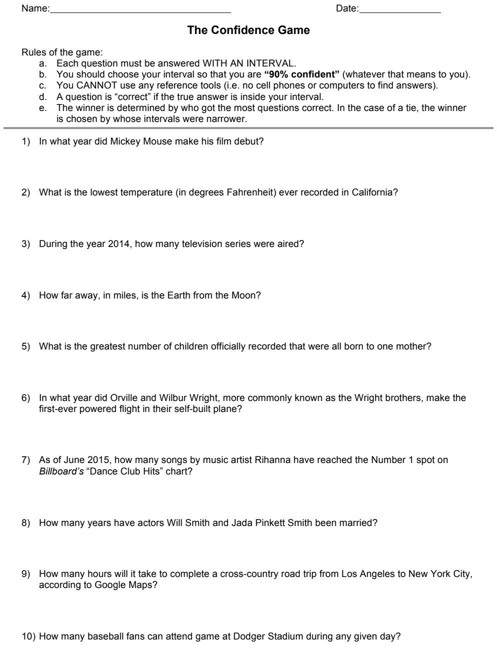
   a. Ask the students to write down an **interval**, or range, of values that they think contains the true height of the building.
   b. Have a few students share their intervals with the class and discuss any obvious similarities or differences between them.

      <u>For example:</u> If Student A gives an interval from 500 to 2000 feet and Student B gives an interval from 1100 to 1400 feet, one discussion could stem from asking Student A why he or she isn't as sure of the answer as Student B is (since Student B gave a narrower interval). Then see if Student A wants to change his or her interval.

   c. After the discussion, tell the students that the actual height of the Empire State Building is 1,454 feet tall. Take a poll to see how many students' intervals contained this value. We will learn what it means to have the true value in our intervals after we play the game.
6. Now, we can actually play the game! Distribute *The Confidence Game* handout (LMR_3.13) and explain the rules. Students will have about 5 minutes to complete the handout, which gives them approximately 30 seconds per question.

   <u>Note:</u> The rules are printed at the beginning of the handout. They are included here for your convenience.

a. Each question must be answered WITH AN INTERVAL.
b. You should choose your interval so that you are “90% confident” (whatever that means to you).
c. You CANNOT use any reference tools (i.e. no cell phones or computers to find answers).
d. A question is “correct” if the true answer is inside your interval.
e. The winner is determined by who got the most questions correct. In the case of a tie, the winner is chosen by whose intervals were narrower.



Name:______________________________ Date:______________

The Confidence Game

Rules of the game:

a. Each question must be answered WITH AN INTERVAL.
b. You should choose your interval so that you are **“90% confident”** (whatever that means to you).
c. You CANNOT use any reference tools (i.e. no cell phones or computers to find answers).
d. A question is “correct” if the true answer is inside your interval.
e. The winner is determined by who got the most questions correct. In the case of a tie, the winner is chosen by whose intervals were narrower.

1) In what year did Mickey Mouse make his film debut?

2) What is the lowest temperature (in degrees Fahrenheit) ever recorded in California?

3) During the year 2014, how many television series were aired?

4) How far away, in miles, is the Earth from the Moon?

5) What is the greatest number of children officially recorded that were all born to one mother?

6) In what year did Orville and Wilbur Wright, more commonly known as the Wright brothers, make the first-ever powered flight in their self-built plane?

7) As of June 2015, how many songs by music artist Rihanna have reached the Number 1 spot on *Billboard's* “Dance Club Hits” chart?

8) How many years have actors Will Smith and Jada Pinkett Smith been married?

9) How many hours will it take to complete a cross-country road trip from Los Angeles to New York City, according to Google Maps?

10) How many baseball fans can attend game at Dodger Stadium during any given day?

LMR_3.13

7. Once each student has completed *The Confidence Game* handout (LMR_3.13), have students choose partners and exchange handouts so that they can grade each other. Remind them that a question is marked as “correct” if the actual value (see answers in Step 8) falls within the interval.

8. Display the answers for each of the 10 questions from the handout found below:

   1) In what year did Mickey Mouse make his film debut? ***1928***
   2) What is the lowest temperature (in degrees Fahrenheit) ever recorded in California? ***-45 degrees Fahrenheit***
   3) During the year 2014, how many television series were aired? ***1,715 TV shows***
   4) How far away, in miles, is Earth from the moon? ***238,900 miles***
   5) What is the greatest number of children officially recorded that were all born to one mother? ***69 children***
   6) In what year did Orville and Wilbur Wright, more commonly known as the Wright brothers, make the first-ever powered flight in a plane? ***1903***
   7) As of June 2015, how many of Rihanna’s songs have reached the Number 1 spot on *Billboard’s* “Dance Club Hits” chart? ***23 songs***
   8) How many years have actors Will Smith and Jada Pinkett-Smith been married? ***18 years***
   9) How many hours will it take to complete a cross-country road trip from Los Angeles to New York City according to Google Maps? ***41 hours (2,789.5 miles)***
   10) How many baseball fans can attend game at Dodger Stadium during any given day? ***56,000 fans***

9. Each student should write the total number of “correct” responses at the top of his or her partner’s handout, and then return it.

10. Engage the students in a discussion about how well they did at estimating the true values with their intervals. The following questions can be used to steer the discussion:
    a. Remember that we were aiming to be 90% confident for each question. Based on this, how many of the 10 questions should we each have gotten correct? ***If we are 90% confident, then we would expect 90% of the 10 intervals to include the true value, which is 9 intervals.***
    b. Did anyone in the class get exactly 9 correct? Did anyone get all 10 correct? ***Answers will vary by class. However, it is very unlikely that many students will have gotten 9 or 10 correct responses on this first round.***

11. Create a dotplot on the board (or on poster paper) titled "Number Correct" and have each student record his or her value. Then, ask:
    a. How many students got 9 correct? In other words, how many students were actually 90% confident of their intervals? ***Answers will vary by class.***
    b. What is the typical number of correct responses for our class? Does it seem too high or too low? Explain. ***Answers will vary by class. Most likely, the typical number of correct responses will be fairly low (maybe even 4 or less).***
    c. Why is our typical score so much lower than 9? ***We tend to be more confident than we should be, so we create narrower intervals.***
    d. It looks like, even though we thought we were 90% confident, most of us (or all of us) did not succeed 90% of the time. How could we increase our level of confidence? ***We could use wider intervals.***

12. Recall from Step 3 that, in statistics, to estimate something means that we can give a range of values that we are confident include the population parameter value. This range of values, like the ones the students created during *The Confidence Game* activity, is known as a **confidence interval**.

13. Students will continue to learn about confidence intervals during the next lesson.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

☑ In your own words, write a description of what a confidence interval is and why it is used in statistics.

Lesson 14: How Confident Are You?

Objective:

Students will learn about informal confidence intervals and estimates for the margin of error.

Materials:

1. Dotplot titled "Mean Word Length, Random Sample, Size = 10" – from Lesson 11

Vocabulary:

margin of error, bootstrapping

Essential Concepts: The margin of error expresses our uncertainty in an estimate. The estimate, plus or minus the margin of error, gives us an interval in which we are very confident the true value lies.

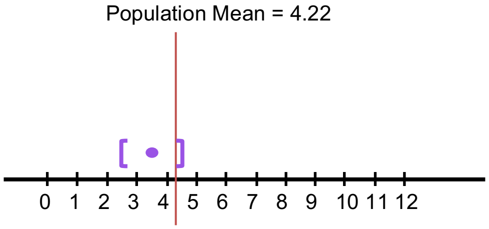
Lesson:

1. In this lesson, students will learn about confidence intervals in more detail.
2. Display the dotplot the class created during Lesson 11 (The Gettysburg Address) titled "Mean Word Length, Random Sample, Size = 10."
3. Have students recall that each dot represents one student's calculation of the mean word length of a sample of 10 randomly selected words from the Gettysburg Address.
4. Also remind them that the population parameter, which is the mean word length of all words in the speech, was 4.22. There should already be a vertical line on the dotplot to indicate this value, but if it is not present, please add it during this step. Ask:
    a. What vocabulary word was used to describe each of the sample means we each created during Lesson 11? ***The sample statistic. Every dot on the graph represents one sample statistic, more specifically each dot corresponds to a different sample mean.***
    b. How many of us got exactly the right value? ***Probably none.***
    c. Thinking back on The Confidence Game we played yesterday, what approach could we do so that 90% of us would be correct? ***We could give an interval.***
5. Show students that they can give an interval in the form:

    Your sample statistic plus or minus AMOUNT

6. Ask them to calculate what their AMOUNT must be so that their interval includes the parameter value. Ask them to write this as an interval.
7. Choose one student to illustrate what is to be done. Ask them for their AMOUNT. On the dotplot, find their value, and use bars to go out plus and minus the AMOUNT. Confirm that it includes the parameter value.

For example: The purple dot represents a sample mean of 3.5. The AMOUNT we have chosen for this particular case is 0.8, so the lower bracket is 0.8 below the sample mean, and the upper bracket is 0.8 above the sample mean. Notice that the population parameter is included within the brackets.



8. Now, convert this to an interval of the form [lowest value, highest value] by subtracting the amount from the sample statistic to get the lowest value, and adding to get the highest.
9. Inform students that this AMOUNT is called the **margin of error**. Explain that the students all now have different margins of error because in this unusual 'game' they know the population value. But in real life we do not, and so we have to choose one single margin of error that will work 90% of the time.
10. Ask the students what margin of error they should use so that 90% of the estimates will have a 'successful' interval. You might want to tell them how many estimates that is for your class. A ballpark figure for the margin of error is 1.3.
11. Explain: If we were to start all over, we could imagine picking one of these sample statistics at random.
    a. What's the probability that the sample statistic plus or minus the margin of error would include the parameter value? ***90%.***
12. Because of this, we call these 'confidence intervals.' When we report an interval, for example 2.7 to 4.3, we say "We are 90% confident that the population parameter value is between 2.7 and 4.3." This is another way of saying "We don't know what the exact true value is, but we're confident it is somewhere in this interval."
13. Remind students of the Pew Poll they discussed during Lesson 10. For reference, the Pew Research Center made the following statement in their August 2015 report titled *Teens, Technology & Friendships*: **Pew Poll**

    > "For today's teens, friendships can start digitally: 57% of teens have met new friends online. The margin of error is plus or minus 3.7 percentage points. Social media and online gameplay are the most common digital venues for meeting friends."

    Note: The data for this report were collected via interviews of 1,060 teenagers between the ages of 13 and 17.
14. Now that students have learned about the margin of error, have them write an *Exit Slip* about what the margin of error means in context of the Pew Poll.
15. Conclusions and takeaways:
    a. Estimates that are based on random samples vary.
    b. We can measure this variation.
    c. The margin of error can tell us how much estimates vary.
    d. We can use the estimate from our random sample, along with the margin of error, to give us a range of plausible values for the population parameter. This is called a confidence interval.
16. If time allows, introduce students to the idea of **bootstrapping**, which is where we take random samples of really large samples. For example, if we were looking at Twitter data, it would be almost impossible to compile every single tweet that exists in the population. Instead, we might be able to access 500,000 tweets, which is a very large sample. From this sample, we could create smaller random samples of size 100 and make inferences about the overall population of tweets from these samples. This will be discussed further in Lab 3D.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next Day

LAB 3D: Are You Sure about That?

Complete Lab 3D prior to the Let's Build a Survey! Practicum.

Lab 3D - Are you sure about that?

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Confidence and intervals

- Throughout the year, we've seen that:
  - Means are used for describing the typical value in a sample or population, but we usually don't know what they are, because we can't see the entire population.
  - Means of samples can be used to *estimate* means of populations.
  - By including a margin of error with our estimate, we create an interval that increases our confidence that we've located the correct value of the population mean.
- Today, we'll learn how we can calculate *margins of error* by using a method called the *bootstrap*.
  - Which comes from the phrase, *Picking yourself up by your own bootstraps*.

In this lab

- Load the built-in `atus` (*American Time Use Survey*) data set, which is a survey of how a sample of Americans spent their day.
  - **The United States has an estimated population of 327,350,075. How many people were surveyed for this particular data set?**
- The statistical question we wish to investigate is:

  What is the mean age of people older than 15 living in the United States?

- **Why is it important that the ATUS is a random sample?**
- **Use our `atus` data to calculate an estimate for the average age of people older than 15 living in the U.S.**

One bootstrap

- A *bootstrapped* sample is when we take a random `sample()` of our original data (`atus`) *WITH* replacement.
  - The `size` of the sample should be the same size as the original data.
- We can create a single *bootstrapped* sample for the `mean` in 3 steps:
  1. Sample the number of the rows to use in our *bootstrap*.
  2. `slice` those rows from our original data into our *bootstrap* data.
  3. Calculate the `mean` of our *bootstrapped* data.

Our first bootstrap

- Fill in the blanks to `sample` the row numbers we'll use in our *bootstrapped* sample.
  - Be sure to re-read what a *bootstrapped* sample is from the previous slide to help you fill in the blanks.
  - Use `set.seed(123)` before taking the sample.

```
bs_rows <- ____(1:____, size = ____, replace = ____)
```

- We can use the `slice` function to create a new data set that includes each row from our `sample`

```
bs_atus <- slice(atus, bs_rows)
```

Take a look

- Look at the values of `bs_rows` and `bs_atus`.
  - **Write a paragraph that explains to someone that's not familiar with `R` how you created `bs_rows` and `bs_atus`. Be sure to include an explanation of what the *values* of `bs_rows` mean and how those values are used to create `bs_atus`. Also, be sure to explain what each argument of each function does.**

One strap, two strap

- Calculate the `mean` of the `age` variable in your `bootstrapped` data, then use a different value of `set.seed()` to create your own, personal *bootstrapped* sample. Then calculate its `mean`.
  - Compare this second *bootstrapped* sample with three other classmates and write a sentence about how similar or different the *bootstrapped* sample means were.

Many bootstraps

- To use *bootstrapped* samples to create *confidence intervals*, we need to create many *bootstrapped* samples.
  - Normally, the more *bootstrapped* samples we use, the better the *confidence interval*.
  - In this lab, we'll `do()` 500 *bootstrapped* samples.
- To make `do()`-ing 500 *bootstraps* easier, we'll code our 3-step bootstrap method into a function.
  - Open a new R script (File -> New File -> R Script) to write your function into.

Bootstrap function

- Fill in the blank space below with the 3-steps needed to create a *bootstrapped* sample `mean` for our `atus` data.
  - Each step should be written on its own line between the curly braces.

```
bs_func <- function() {



 }
```

- Highlight and *Run* the code you write.

Visualizing our bootstraps

- Once your function is created, fill in the blanks to create 500 *bootstrapped* sample means:

```
bs_means <- do(____) * bs_func()
```

- **Create a histogram for your bootstrapped samples and describe the *center*, *shape* and *spread* of its distribution.**
  - These bootstrapped estimates no longer estimate the average age of people in the U.S.
  - Instead, they estimate how much the estimate of the average age of people in the U.S. varies.
- In the next slide, we'll look at how we can use these bootstrapped means to create *90% confidence intervals*.

Bootstrapped confidence intervals

- To create a 90% confidence interval, we need to decide between which two *ages* the middle 90% of our bootstrapped estimates are contained.
- **Using your histogram, fill in the statement below:**

  The lowest 5% of our estimates are below _______ years and the highest 5% of our estimates are above_______ years.

- Use the `quantile()` function to check your estimates.
- **Based on your bootstrapped estimates, between which two ages are we 90% confident the actual mean age of people living in the U.S. is contained?**

On your own

- Using your *bootstrapped* sample means, create a 95% confidence interval for the mean age of people living in the U.S.
  - **Why is the 95% confidence interval wider than the 90% interval?**
  - **Write down how you would explain what a 95% confidence interval means to someone not taking *Introduction to Data Science*.**

<u>Practicum: Let's Build a Survey!</u>

Objective: Students will design a non-biased survey.

Materials:

1. Practicum: *Let's Build a Survey!* (LMR_U3_Practicum_Build a Survey)

Practicum
Let's Build a Survey!

Based on what you have learned in Lessons 9 through 14, you will now design a survey. You and your team members must do all of the following:

1. Select a topic from the list below:
   a. Social Media
   b. Entertainment
   c. Sports
   d. The Environment
   e. Health
   f. Education
   g. Other topic of interest
2. Create a research question about your topic of interest.
3. Create a statistical question that is related to the research question.
4. Identify the population of interest.
5. Describe how you will select your sample from the population so that you'll be able to make generalizations about your population of interest.
6. Identify the number of people who will be in your sample.
7. Create five survey questions that will try to answer your statistical question and describe how you have made sure that they are non-leading questions.
8. Identify a statistic that can be used to summarize the responses from this survey. Can you identify a parameter?
9. Submit a typed paper that details the survey you just designed.

What's the Trigger?

Instructional Days: 5

Enduring Understandings

Sensors are data collection devices that collect data either continuously or whenever they are triggered. A sensor is a converter that measures a physical quantity and converts it into a signal, which can be read by an observer or by an instrument. Participatory Sensing is a specific data collection method that uses sensor technology. This method emphasizes the involvement of citizens and community groups in the process of sensing and documenting where they live, work, and play. Triggers play an important role in the Participatory Sensing data collection process. The response to the triggers may or may not be the same each time.

Engagement

Students will view and discuss a video clip called *Play Like Nadal With a Smart Tennis Racket* to begin to think about the sensors as data collection devices found ubiquitously in today's world. The video can be found at: https://youtu.be/lcBnzddQECc

Learning Objectives

Statistical/Mathematical:

S-IC 3: Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

Data Science:

Understand that sensors provide a continuous stream of data. Participatory Sensing provides real-time data from a user who is willingly providing the data. What differentiates a sensor as a data gathering method is the use of a trigger that signals a data collection session.

Applied Computational Thinking:

- Create a Participatory Sensing campaign using a campaign Authoring Tool.

Real-World Connections:

Sensors are found everywhere in today's world. They can provide data about environmental conditions as well as personal habits. More and more, sensors are being used for personal tracking, especially in the medical field, to inform people about what they do.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

3. Students will use write questions that use emphasize differences in data science concepts and skills.

Data File or Data Collection Method

Data Collection Method:

1. Students will gather data generated through a class-generated Participatory Sensing campaign.

Data File:

1. Students’ Participatory Sensing campaign data

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 15: Ready, Sense, Go!

Objective:

Students will learn what sensors are and how they are used to collect data.

Materials:

1. Video: *Play Like Nadal With a Smart Tennis Racket*
   https://youtu.be/lcBnzddQECc
2. Computers (see Step 5)
3. Poster paper
4. Flags in 3 different colors
   Advanced preparation required (see Step 10 below)

Vocabulary:

sensor, trigger, algorithm

Essential Concepts: Sensors are another data collection method. Unlike what we have seen so far, sensors do not involve humans (much). They collect data according to an algorithm.

Lesson:

1. *Entrance Ticket:* What are some of the data collection methods we have learned about so far in this unit? ***We have learned about experiments, observational studies, surveys, and getting data from a URL (in Lab 3B).***
2. Inform students that, in this lesson, they will be introduced to another data collection method known as sensors.
3. With a partner, ask students to discuss what they think a sensor is. Ask each pair to write down their ideas.
4. Show the *Play Like Nadal With a Smart Tennis Racket* video found at: https://youtu.be/lcBnzddQECc. As students watch the video, they should think about other sensors they may have come across, particularly ones used with smartphones. After watching the video, ask students to add to their definition of a sensor.
5. Now, inform students that they will work in teams to compile a list of data-collecting sensors. They may use computers to conduct online research for this part of the lesson. Challenge each team to generate the longest list in the class.
6. After students have had time to research and create their lists, ask students in each team to number off one through four (or five, depending on team sizes).
7. Share out in rounds. First, ask students in each team whose number is one to share one sensor from their list. On the poster paper, create a class list of sensors as shared by the students. Repeat with the rest of the numbers.
8. Score keeping: Each person gets five seconds to respond. You may hold up your hand with the palm facing the students and count down. The rules for teams are as follows:

   a. add a sensor to the list, get 1 point
   b. repeat an answer, lose 1 point
   c. do not answer in five seconds, lose a turn
   d. do not have an answer to contribute, may pass

   Note: You may reward the winning team with extra credit points, if desired.
9. Next, students will engage in an activity to see sensors in action.
10. Create 3 groups of students:

    a. Group 1 – Triggers (3 students)

Provide each Trigger a different colored flag (for example: **Pink**, **Purple**, **Green**). The teacher will call out a color, at random, and the Trigger assigned to that color will raise his or her flag. Each flag corresponds to a research question of interest.

Pink – Who is in our class?
Purple – What is on our classroom walls?
Green – What do we like to do after school?

b. Group 2 – Sensors (2 students)
Each Sensor should be assigned to one Trigger, or colored flag (**Pink** = Sensor A, **Purple** = Sensor B, **Green** = no sensor assigned to it). When the Sensors see their assigned Trigger, they send a signal to the Collector (see below) telling him or her to collect data from another student in the class. The Sensors are basically go-betweens for the Triggers and the Collector.

c. Group 3 – Collector (1 student)
One student is the Collector of all the data. The Collector is in charge of asking survey questions related to the research question of the original Trigger. Survey questions are provided here:

Survey questions related to the **Pink** trigger:

(1) How did you get to school today (bus, car, walking, etc.)?
(2) What size shoe do you wear?
(3) What is your favorite pizza topping?

Survey questions related to the **Purple** trigger:

(1) What is your favorite wall decoration?
(2) What type of poster is it (motivational, reference, class work, etc.)?
(3) What color is most prevalent in the poster?

Survey questions related to the **Green** trigger:

(1) Do you have a sports team practice or club meeting today?
(2) Are you hanging out with friends today after school?
(3) Will you be working today after school?

Each time a sensor is active, the Collector must ask a new student in the class the appropriate survey questions.

11. Explain the activity to your students. Then, call out a flag color at random. Repeat several times. Make sure you call out the flag that has no assignment at least once so that students see that no action took place. Reflect on the activity with the following discussion questions:

a. What data were missed? Why? ***Data about what our class likes to do after school. They were missed because there was no Sensor connected to the Green trigger, so the Collector never knew to collect this type of data.***

b. Grocery stores keep track of customer data when purchases are made with a loyalty card. What is the trigger in this case? What data are being collected? ***The trigger is checking out at a grocery store. There are lots of data that are collected, including: items bought, cost of items, number of items on sale, etc.***

12. After they engage in the sensor activity, ask students to revisit their definition of a sensor (see Step 3). Have them revise their definition based on the following concepts:

a. A **sensor** is a converter that measures a physical quantity and converts it into a signal, which can be read by an observer or by an instrument.
b. A sensor collects data continuously, or whenever a **trigger** is activated. A trigger is a something that responds to an event so that an action can occur.
c. Sensors collect data according to an **algorithm**. An algorithm is a process or set of rules that are followed (just like the rules followed during the activity).

d. Sensors may also collect data automatically, without anyone's knowledge or input. Examples include GPS location, time, and date.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

☑ Now that students learned what sensors are, ask them what data they would they like to see collected on a sensor that they couldn't collect in an experiment or survey. They must explain why it is difficult to collect that data in an experiment or survey, and how a sensor would make it easier to collect that data.

Lesson 16: Does It Have a Trigger?

Objective:

Students will learn to identify and categorize survey questions versus sensor questions, and will practice writing sensor questions.

Materials:

1. Poster paper (one per student team)
2. Sticky notes
3. *Sensor or Survey?* handout (LMR_3.14_Sensor or Survey)

Vocabulary:

Participatory Sensing

Essential Concepts: A key feature that distinguishes the way sensors collect data from more traditional approaches is that sensors collect data when a 'trigger' event occurs. In Participatory Sensing, this event is something we humans agree upon beforehand. Every time that trigger happens, we collect data.

Lesson:

1. Refer back to the list of sensors the class created during the previous lesson. Distribute a piece of poster paper to each student team, and have them create the following table:

Sensor	How is it triggered?
1.	1.
2.	2.
3.	3.

2. Assign each student team 3 sensors from the class's list.

3. Then, each team should complete the table using their knowledge of triggers discussed during the previous lesson. Remind students that when a trigger occurs, a sensor reacts to it and sends a signal to a data collector.

4. Conduct a *Gallery Walk* of the posters. Each team will get to write one reaction or question about what they see on each poster.

5. After the Gallery Walk, ask each team to return to their posters. If the posters include questions, have teams take turns responding to the questions.

6. *Quickwrite*: In their DS Journals, ask student to respond to the following questions. They will have two minutes to write as much as they can:

   a. When you learned about survey questions, what were the two categories of questions you learned about? ***Answer: Open-ended and Closed-ended are the categories.***

   b. What are some examples of these types of questions? ***Open-ended: write a paragraph, comments, essays, write a sentence, single answer. Closed-ended: multiple or single choice, yes/no, scales (e.g. 1-5), choose from a list, check a box.***

7. In teams, ask students to share their responses using the *Give One/Get One* strategy. You may use a timer to keep track of time.

8. Remind students that one of the most important things they learned about sensors is that there is a trigger that reminds either a device or a person to answer a question or to collect data.

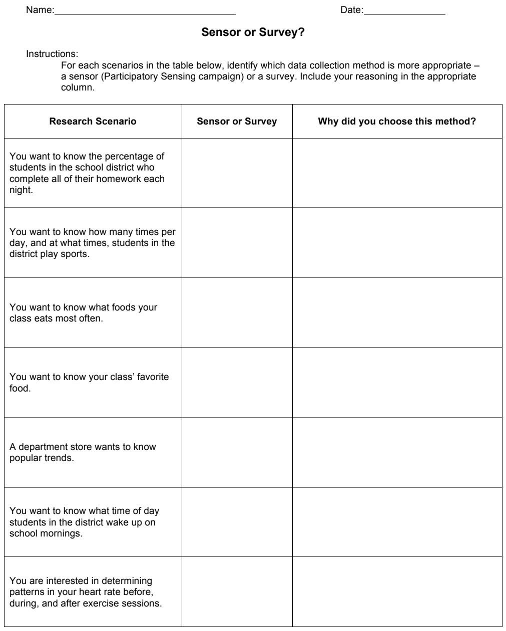
9. For this class, students have already had experience with using sensors as a data collection tool – all the **Participatory Sensing** campaigns.

10. Explain that survey questions are asked in Participatory Sensing campaigns. There is no difference in the type of questions that are asked when collecting data via surveys and when collecting data via PS campaigns.
11. When deciding whether to use a survey or a PS campaign for data collection, we have to look at the research question of interest. Some questions are better answered with survey data, while others with PS campaigns. Research questions that include variation across time or across locations are good candidates for PS. Some questions might be answered by both.

    For example:

    Consider the research question: How does my sense of safety and security change as I go about my daily routine? This question would best be answered via a PS campaign because students could collect data in real time about their sense of security. A possible trigger could be "whenever you change locations" or "once at the start of every hour" or perhaps whenever a random alarm goes off.

    Consider the research question: What proportion of high school students are superstitious? This question could be done with a survey based on a random sample from the population of all high school students.
12. Distribute the *Sensor or Survey?* (LMR_3.14) handout. In teams, students will determine whether a sensor or survey is better for a given research scenario.



Name:______________________________ Date:______________

Sensor or Survey?

Instructions:
For each scenarios in the table below, identify which data collection method is more appropriate – a sensor (Participatory Sensing campaign) or a survey. Include your reasoning in the appropriate column.

Research Scenario	Sensor or Survey	Why did you choose this method?
You want to know the percentage of students in the school district who complete all of their homework each night.		
You want to know how many times per day, and at what times, students in the district play sports.		
You want to know what foods your class eats most often.		
You want to know your class' favorite food.		
A department store wants to know popular trends.		
You want to know what time of day students in the district wake up on school mornings.		
You are interested in determining patterns in your heart rate before, during, and after exercise sessions.		

LMR_3.14

13. Once the teams have completed the handout, assign each team one research scenario from the *Sensor or Survey* activity.
14. Conduct a *Whip Around* and have each team share their responses with the class. Allow students time to revise any incorrect responses.
15. Summarize the lesson by highlighting that PS campaigns and surveys use similar questions. However, depending on the research topic of interest, the decision to use one or the other relies on whether or not a trigger is involved.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Suppose we wish to know more about whether people behave superstitiously. Write two research scenarios, using the following questions as a guide:

a. How would you collect data to address this using PS? Include the trigger event you would use, and the data you would like to collect when the trigger happens.
b. How would you collect data to collect this using a survey based on a random sample of people in California?
c. Describe the differences between these two approaches. What can you learn in one approach that you can't in the other?

Lesson 17: Creating Our Own Participatory Sensing Campaign

Objective: Students will be guided through the creation of a new Participatory Sensing campaign and survey on a topic of interest chosen by the class.

Materials:

1. *Food Habits Campaign Questions* handout (LMR_3.15_Food Habits Qs)
2. *Campaign Creation Brainstorm* handout (LMR_3.16_Campaign Creation)

Essential Concepts: Creating a Participatory Sensing Campaign requires that survey questions must be completed whenever they are "triggered". Research questions provide an overall direction in Participatory Sensing Campaign.

Lesson:

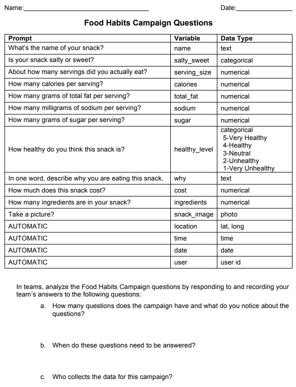
1. Review homework questions by asking a couple of students to share their responses. The rest of students will engage in *Agree/Disagree* as the questions are shared.
2. Display the following definition of Participatory Sensing that some computer scientists have agreed to and ask students to read and record this definition in their DS journals:

   > At its heart, Participatory Sensing is data collection and interpretation. Participatory Sensing emphasizes the involvement of citizens and community groups in the process of sensing and documenting where they live, work, and play. It can range from private personal observations to the combination of data from hundreds, or even thousands, of individuals that reveals patterns across an entire city. Most important, Participatory Sensing begins and ends with people, both as individuals and members of communities. The type of information collected, how it is organized, and how it is ultimately used, may be determined in a traditional manner by a centrally organized body, or in a deliberative manner by the collection of participants themselves. The latter case, in particular, emphasizes the novelty of Participatory Sensing as an approach and underscores the importance of using widely available and familiar technology. [Source: "Participatory Sensing: A citizen-powered approach to illuminating the patterns that shape our world."]

3. Activate prior knowledge: Based on this definition, ask students to recall the Participatory Sensing campaigns in which they have engaged thus far. ***Answer: Food Habits, Time Use, Stress/Chill.***

   Note: *Personality Color* and *Time Perception* were surveys, not Participatory Sensing campaigns because they were only completed once. Their data was not collected over time.

4. Inform students that they will be creating a new, whole class Participatory Sensing campaign, but before they do that, they will analyze the *Food Habits Campaign Questions* handout (LMR_3.15).



Name:______________________________ Date:______________

Food Habits Campaign Questions

Prompt	Variable	Data Type
What's the name of your snack?	name	text
Is your snack salty or sweet?	salty_sweet	categorical
About how many servings did you actually eat?	serving_size	numerical
How many calories per serving?	calories	numerical
How many grams of total fat per serving?	total_fat	numerical
How many milligrams of sodium per serving?	sodium	numerical
How many grams of sugar per serving?	sugar	numerical
How healthy do you think this snack is?	healthy_level	categorical 5-Very Healthy 4-Healthy 3-Neutral 2-Unhealthy 1-Very Unhealthy
In one word, describe why you are eating this snack.	why	text
How much does this snack cost?	cost	numerical
How many ingredients are in your snack?	ingredients	numerical
Take a picture?	snack_image	photo
AUTOMATIC	location	lat, long
AUTOMATIC	time	time
AUTOMATIC	date	date
AUTOMATIC	user	user id

In teams, analyze the Food Habits Campaign questions by responding to and recording your team's answers to the following questions:

a. How many questions does the campaign have and what do you notice about the questions?

b. When do these questions need to be answered?

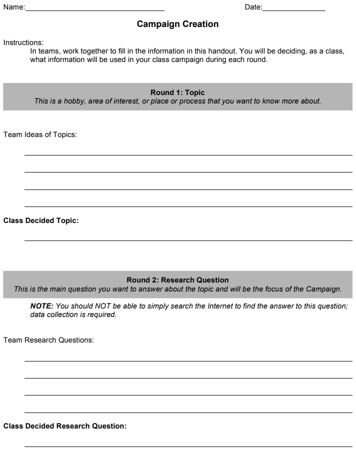
c. Who collects the data for this campaign?

LMR_3.15

5. In teams, allow students two minutes to discuss the following as they analyze the Food Habits Campaign questions:
    a. How many questions does the campaign have and what do they notice about the questions? ***Answers will vary. Students may notice that they are survey type of questions and may identify the type of questions such as open-ended, single-choice, etc.***
    b. When do these questions need to be answered? ***Each time they eat a snack.***
    c. Who collects the data for this campaign? ***The participants collect their own data.***

6. Ask a few teams to share their insights about the discussion. In the share-out, guide students to see that the questions are in fact survey questions. Although survey questions are answered once, when we collect data every time a 'trigger' event occurs, then we are engaging in Participatory Sensing.

7. Ensure that team roles have defined duties to keep teams on task for the rest of this lesson. Creating this class campaign will follow a process in which consensus (or a majority rule) will be reached in each step of the campaign development within each team. Inform students that they will be creating a Participatory Sensing campaign on a topic of their interest using LMR_3.16.



Name:______________________ Date:____________

Campaign Creation

Instructions:
In teams, work together to fill in the information in this handout. You will be deciding, as a class, what information will be used in your class campaign during each round.

Round 1: Topic
This is a hobby, area of interest, or place or process that you want to know more about.

Team Ideas of Topics:

____

____

____

____

Class Decided Topic:

____

Round 2: Research Question
This is the main question you want to answer about the topic and will be the focus of the Campaign.

NOTE: *You should NOT be able to simply search the Internet to find the answer to this question; data collection is required.*

Team Research Questions:

____

____

____

____

Class Decided Research Question:

____

LMR 3.16

8. <u>Round 1:</u> First, teams will discuss their hobbies, areas of interest, or places or processes they want to know more about. Prompt students to think about whether they want to learn about "where they live, work, or play." All students within the group must agree on a hobby or area of interest to be their topic of interest to create a campaign for. ***An example of a hobby is practicing cello. An area of interest might be 'the environment.' A place of interest might be "our school" or "my church" or "Disneyland."***

9. Once teams have decided on a topic for their group, have teams share out their topic of interest. As a class, decide on one topic that will be used for creating a new Campaign.

10. <u>Round 2:</u> Now have teams consider what research questions you might ask about this topic of interest. ***An example of a research question for practicing cello is “How can I improve my playing?” or "How can I practice more effectively?"***

11. Once teams have decided on a research question for their group, have teams share out their research question. As a class, decide on one research question that will be used for creating a new Campaign.

12. <u>Round 3:</u> Next, they will examine what kind of data needs to be collected in order to answer this research question. They should discuss possible triggers that will determine when data should be collected. Allow teams to engage in a discussion about when is the best time to trigger the data collection/completion of the survey. For example: every day at 8am; whenever they practice the cello; whenever they see an advertisement; etc. They should record it in their DS Journals. ***An***

example of a trigger for practicing cello is whenever you play the cello. In this case, it could be any time of day or even multiple times of the day.

13. Once teams have decided on a trigger for their group, have teams share out their possible trigger. As a class, decide on one trigger that will be used for creating a new Campaign.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Using the class topic, research question, trigger event, and discussion of the data they plan to collect. Classify our class campaign under the appropriate category with your justification:

(A) Individual; (B) Groups of people; (C) Community

Note to teacher: To determine which category a campaign should be placed under, consider the question "Who or what will we learn about?" If the answer is "only one person", then place in the Individual category. The cello campaign is an example of this. If we might learn about lots of people, put it in the Groups of People category. The Food Habits, Stress/Chill, and Time Use campaigns fit into this category (they learn about the students in the class). A campaign that wanted to know where all of the churches in the neighborhood were located, or wanted to try to keep people from littering, or wasting water, these should go into the "community" category.

Lesson 18: Evaluating Our Own Participatory Sensing Campaign

Objective: Students will create statistical questions and evaluate their Participatory Sensing Campaign.

Materials:

1. *Campaign Creation* handout (LMR_3.16_Campaign Creation) from previous lesson
2. Class Campaign Information from Lesson 16

Essential Concepts: Statistical questions guide a Participatory Sensing Campaign so that we can learn about a community or ourselves. These Campaigns should be evaluated before implementing to make sure they are reasonable and ethically sound.

Lesson:

1. Review homework by giving students about five minutes to share their classifications in their teams. They will decide as a team which classification is the most fitting.
2. Once the five minutes have passed, have a class discussion of classifications and their justifications. Explain to the class that the campaign must be carried out by the whole class so if it has been classified in the Individual category, it must be revised. Also discuss whether the campaign is feasible. (For example, is the trigger so rare that no one will collect data? Are the questions too intrusive?).
3. Inform students that one of the promises of PS is its potential for helping people bring about social and civic change. Ask teams to consider the following questions and report back:
    a. Does our campaign try to do this?
    b. Could it be changed or modified to do this?

    Note: Feasible campaigns fall under the groups of people or community categories. If a campaign is in the individual category, it should be modified to fall under the other categories before moving to round 4.
4. Display the campaign information students generated (and selected as a class) the previous day or revised today: Topic, Research question, Trigger, and Type of Data needed.
5. Now they will continue the rounds using the Campaign Creation handout LMR 3.16 from the previous lesson.
6. Round 4: Now that the class has decided on a trigger and the type of data needed, they will create survey questions to ask when the trigger is set. The questions should consider all of the possible data they might collect at this trigger event. It's ok if the list is long; the goal is to be creative and think of lots of different ideas.

    Examples of survey questions for practicing cello are:
    "How long did you practice?"
    "What did you play?"
    "How would you rate your practice session: 1 to 5?"
    "Any thoughts or comments about your practice?"
7. Once teams have created 4 survey questions for their group, have teams share out their survey questions. As a class, decide on no more than 10 survey questions that will be used for creating a new Campaign.
    a. Then, evaluate each survey question. For each question they should consider:
        i. What type of data will this question collect? (Numerical, discrete numerical, text, categories, photos, location).
        ii. How does this question help address the research question?

        iii. Does the question need to be reworded? (Is it clear what is being asked for? Do they know how to answer it?)

    b. If the survey questions need to be rewritten, assign teams to rewrite survey questions. Then, as a class, decide on the changes.

    c. Once finalized, write the survey question that goes along with that data variable, being cognizant of question bias.

8. <u>Round 5:</u> In teams, now generate two to three statistical questions that they might answer with these data. Make sure your statistical questions are interesting and relevant to the class topic of interest. They may keep a record in their DS Journals. Remind students that they will also have data about the date, time, and place of data collection.

    Examples of statistical questions that can be answered for practicing cello are:

    "How frequently do I practice?"

    "When I practice more frequently, do I rate my sessions higher?"

    "Are higher-rated sessions associated with time of day?"

9. Once teams have generated their statistical questions, have them share out with the class. Confirm that the questions are statistical and that they can be answered with the data the students propose to collect. As a class, decide on no more than 3 statistical questions to guide your campaign.

10. Now that they have all the pieces of the campaign, evaluate whether it's a reasonable and ethically sound campaign. Engage the class in a whole group discussion on the following questions:

    a. Are answers to your survey questions likely to *vary* when the trigger occurs? (If not, you'll get bored entering the same data again and again)
    b. Can the entire class carry out the campaign?
    c. Do triggers occur so rarely that you'll have very little data? Do they occur so often that you'll get frustrated entering too much data?
    d. Ethics: Would sharing these data with strangers or friends be embarrassing or undermine someone's privacy?
    e. Can you change your trigger or survey questions to improve your evaluation?

    f. Will you be able to gather enough relevant data from your survey questions to be able to answer your statistical questions?

11. Students have collaboratively created their first Participatory Sensing campaign. Inform them that you will be demonstrating one tool used to create the campaigns that they see on their smart devices or the computer. Students should take notes in their DS journals, as they will be using the tool later.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 19: Implementing Our Own Participatory Sensing Campaign

Objective: Students will mock-implement, create their Participatory Sensing campaign, survey on their topic of interest, then begin data collection.

Materials:

1. *Campaign Creation* handout (LMR_3.16_Campaign Creation) from previous lesson
2. Campaign Authoring Tool (https://portal.idsucla.org)

Essential Concepts: Practicing data collection prior to implementation allows optimization of a Participatory Sensing Campaign.

Lesson:

1. Display the class generated campaign information for the class to clearly see.
2. In teams, have students mock-implement the campaign they have created. They can do this by asking each other the survey questions to make sure they make sense/ will generate relevant data to their research question and statistical questions. They can use the evaluative questions from Lesson 17 step #10.
3. If there are suggestions for improvement, have teams propose them to the class and make final changes to the campaign.
4. Inform students that you will now demonstrate the tool used to create the campaigns that is displayed on their mobile devices or computers.
5. Login to the **IDS Home Page** found at https://portal.idsucla.org. Click on the **Campaigns tab** on the navigation bar at the top of the page. Then, follow the steps in the tool:
    a. **Campaign Info Window:**
        i. **Campaign Name:** Give your campaign a name. A name related to the topic is recommended.
        ii. **Select your class/period.**
        iii. **Description:** Provide a one-sentence description of your campaign.
        iv. **Data Sharing:** Select Disabled in order to monitor for improper responses.
        v. **Campaign Status:** Select Running.
        vi. **Click the +Add Survey button.**

    b. **Survey Window**:
        i. **Title:** Give the survey a title (again, it may or may not be the same as the campaign name). Users see the title and the all the prompts that follow.
        ii. **ID:** Give the survey a name (it may or may not be the same as the campaign name). Users do not see the survey ID.
        iii. **Description:** Provide a short description of the survey for display.
        iv. **Submit Text:** Provide a brief message to be displayed after survey submission.
        v. **Anytime:** Select the checkbox if you want the survey to be available at anytime.
        vi. **Click the +Add Prompt button and select the prompt type for your first survey question. Note:** You should only select from the following choices: Single choice, number, photo, and text. Multiple-choice does not mean select one choice; it means select many choices. It is not recommended that multiple-choice be used at this point.

    c. **Prompt Information:**
        i. **Click the new prompt bar.**
        ii. **Prompt ID:** This will be your first variable. A short one-word name or short two-word name separated by an underscore is recommended.

        iii. **Prompt Label:** This is the variable name that will be displayed (it may be the same as the prompt ID without the underscore, if used).
        iv. **Question Text:** Type the survey question about which you want to collect data.
        v. **Additional Prompt Information:** Depending on the prompt type, you will be asked to enter additional information. For example, if your prompt is Text, you will be asked a minimum and a maximum value for the number of characters the participant can enter.
        vi. **Skippable:** Select the checkbox if you would like the prompt to be skipped. It is recommended that photo prompts be skippable, since some users will submit their responses via a browser.

    d. **Repeat step c for the remaining survey questions by clicking the +Add Prompt button.**

    e. **XML Code:** As you create the campaign, the code that creates it will be displayed. You may select the checkbox titled **Enable Syntax Highlighting** so that students can keep track of where the information you are adding is embedded in the code. Inform students that they will be learning about XML syntax in the next several lessons.

    f. **Click the Submit Campaign button on the top, right hand side of the page once all prompts have been added.** This action will send the campaign to the server for users to see.

6. Once all prompts have been created, students may use their smart devices or login to the IDS Home Page to view the new campaign. Remember to **Refresh Campaigns**.

7. Students should go through the entire participatory sensing survey to see how their questions are displayed. They do not have to upload the survey.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

For the next 5 days, students will collect data using their newly created Participatory Sensing campaign.

Webpages

Instructional Days: 6

Enduring Understandings

Data takes on a variety of forms online and requires a different style of representation.

Engagement

Students will view a video clip about a data farm, specifically, Google's Street View Data Center to begin thinking about data formats and accessing data online. The video can be found at: https://www.engadget.com/2012-10-17-google-inside-data-centers.html

Learning Objectives

Statistical/Mathematical:

S-IC 3: Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

S-IC 6: Evaluate reports based on data.

DS: Use different techniques to access data from the web and understand why different data representations are useful for different software platforms.

Applied Computational Thinking using RStudio:

- Read data from xml and html table and convert to R data frames
- Use latitude and longitude coordinates of mountain data and overlay it on a map

Real-World Connections:

Data from the web has been used to predict outbreaks of the flu and is a source of extremely rich data sets.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will engage in discussions regarding internet research as it applies to data science.

Data File or Data Collection Method

Data Collection Method:

1. Students will scrape data from online HTML and XML sources.

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 20: Online Data-ing

Objective:

Students will discover that data exists on the Internet in a variety of areas, formats, and for a variety of purposes.

Materials:

1. *Video: Explore a Google Data Center with Street View* found at: https://www.engadget.com/2012-10-17-google-inside-data-centers.html
2. *Wikipedia – Video Games* handout (LMR_3.17_Wikipedia – Video Games)
3. *Wikipedia – Video Games – CSV Format* handout (LMR_3.18_Video Games – CSV)
4. *Online Data-ing* handout (LMR_3.19_Online Data-ing)

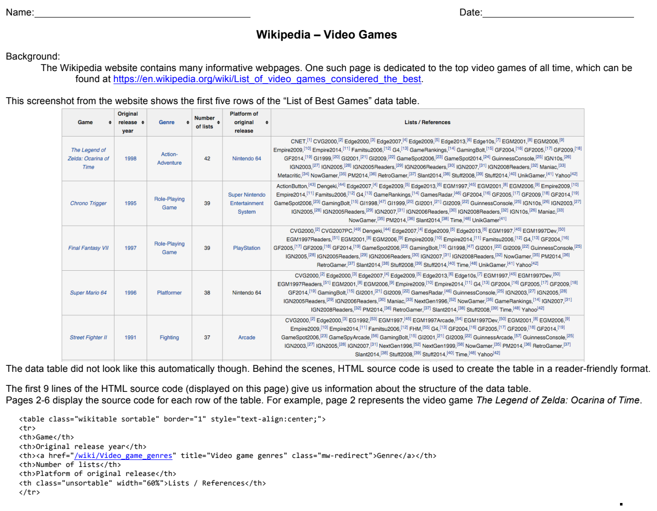
Vocabulary:

data farm, tags, HTML

Essential Concepts: We stretch students' conception of data, to help them see that many web pages present information that can be turned into data.

Lesson:

1. By a show of hands, ask students if they have ever heard of the term **data farm**. If any of them have, ask him or her to share what they know about it.
2. Inform students that a data farm is a physical space where high capacity servers are placed to store large amounts of data.
3. Introduce the video titled *Explore a Google data center with Street View* found at https://www.engadget.com/2012-10-17-google-inside-data-centers.html by explaining that the data center they are about to see is one of these large data farms used to store vast amounts of data.
4. After students watch the video, have a class discussion using the following questions:
   a. We have been talking about data for a few months now. How would you respond if someone asked you, "What are data?" ***Answers will vary by class.***
   b. What are some ways that we have stored data? ***Data frames in R, Excel spreadsheets, .csv files.***
5. Explain that one of the main ways data are distributed is through the Internet. Storing and sharing data on the Internet requires a different format than what we have seen. For example, Wikipedia has a page dedicated to the top video games.
6. Distribute the *Wikipedia – Video Games* handout (LMR_3.17), and have students explain the information that the data table provides.



Name:_________________________________ Date:________________________

Wikipedia – Video Games

Background:

The Wikipedia website contains many informative webpages. One such page is dedicated to the top video games of all time, which can be found at https://en.wikipedia.org/wiki/List_of_video_games_considered_the_best.

This screenshot from the website shows the first five rows of the "List of Best Games" data table.

Game	Original release year	Genre	Number of lists	Platform of original release	Lists / References
The Legend of Zelda: Ocarina of Time	1998	Action-Adventure	42	Nintendo 64	CNET,[1] CVG2000,[2] Edge2000,[3] Edge2007,[4] Edge2009,[5] Edge2013,[6] Edge10s,[7] EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12] G4,[13] GameRankings,[14] GamingBolt,[15] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19] GI1999,[20] GI2001,[21] GI2009,[22] GameSpot2006,[23] GameSpot2014,[24] GuinnessConsole,[25] IGN10s,[26] IGN2003,[27] IGN2005,[28] IGN2005Readers,[29] IGN2006Readers,[30] IGN2007,[31] IGN2008Readers,[32] Maniac,[33] Metacritic,[34] NowGamer,[35] PM2014,[36] RetroGamer,[37] Slant2014,[38] Stuff2008,[39] Stuff2014,[40] UnikGamer,[41] Yahoo[42]
Chrono Trigger	1995	Role-Playing Game	39	Super Nintendo Entertainment System	ActionButton,[43] Dengeki,[44] Edge2007,[4] Edge2009,[5] Edge2013,[6] EGM1997,[45] EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12] G4,[13] GameRankings,[14] GamesRadar,[46] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19] GameSpot2006,[23] GamingBolt,[15] GI1998,[47] GI1999,[20] GI2001,[21] GI2009,[22] GuinnessConsole,[25] IGN10s,[26] IGN2003,[27] IGN2005,[28] IGN2005Readers,[29] IGN2007,[31] IGN2006Readers,[30] IGN2008Readers,[32] IGN10s,[26] Maniac,[33] NowGamer,[35] PM2014,[36] Slant2014,[38] Time,[48] UnikGamer[41]
Final Fantasy VII	1997	Role-Playing Game	39	PlayStation	CVG2000,[2] CVG2007PC,[49] Dengeki,[44] Edge2007,[4] Edge2009,[5] Edge2013,[6] EGM1997,[45] EGM1997Dev,[50] EGM1997Readers,[51] EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12] G4,[13] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19] GameSpot2006,[23] GamingBolt,[15] GI1998,[47] GI2001,[22] GI2009,[22] GuinnessConsole,[25] IGN2005,[28] IGN2005Readers,[29] IGN2006Readers,[30] IGN2007,[31] IGN2008Readers,[32] NowGamer,[35] PM2014,[36] RetroGamer,[37] Slant2014,[38] Stuff2008,[39] Stuff2014,[40] Time,[48] UnikGamer,[41] Yahoo[42]
Super Mario 64	1996	Platformer	38	Nintendo 64	CVG2000,[2] Edge2000,[3] Edge2007,[4] Edge2009,[5] Edge2013,[6] Edge10s,[7] EGM1997,[45] EGM1997Dev,[50] EGM1997Readers,[51] EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] G4,[13] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19] GamingBolt,[15] GI2001,[21] GI2009,[22] GamesRadar,[46] GuinnessConsole,[25] IGN2003,[27] IGN2005,[28] IGN2005Readers,[29] IGN2006Readers,[30] Maniac,[33] NextGen1996,[52] NowGamer,[35] GameRankings,[14] IGN2007,[31] IGN2008Readers,[32] PM2014,[36] RetroGamer,[37] Slant2014,[38] Stuff2008,[39] Time,[48] Yahoo[42]
Street Fighter II	1991	Fighting	37	Arcade	CVG2000,[2] Edge2000,[3] EG1992,[53] EGM1997,[45] EGM1997Arcade,[54] EGM1997Dev,[50] EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12] FHM,[55] G4,[13] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19] GameSpot2006,[23] GameSpyArcade,[56] GamingBolt,[15] GI2001,[21] GI2009,[22] GuinnessArcade,[57] GuinnessConsole,[25] IGN2003,[27] IGN2005,[28] IGN2007,[31] NextGen1996,[52] NextGen1999,[58] NowGamer,[35] PM2014,[36] RetroGamer,[37] Slant2014,[38] Stuff2008,[39] Stuff2014,[40] Time,[48] Yahoo[42]

The data table did not look like this automatically though. Behind the scenes, HTML source code is used to create the table in a reader-friendly format.

The first 9 lines of the HTML source code (displayed on this page) give us information about the structure of the data table.
Pages 2-6 display the source code for each row of the table. For example, page 2 represents the video game *The Legend of Zelda: Ocarina of Time*.

```
<table class="wikitable sortable" border="1" style="text-align:center;">
<tr>
<th>Game</th>
<th>Original release year</th>
<th><a href="/wiki/Video_game_genres" title="Video game genres" class="mw-redirect">Genre</a></th>
<th>Number of lists</th>
<th>Platform of original release</th>
<th class="unsortable" width="60%">Lists / References</th>
</tr>
```

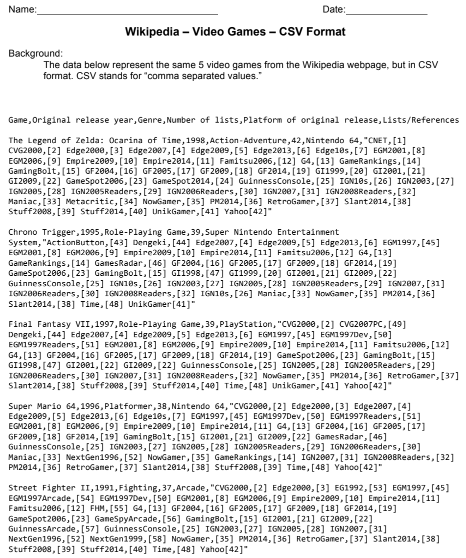
LMR_3.17

7. Once the students understand what the data table describes, walk them through the first portion of the **HTML**, or Hypertext Markup Language, source code (on page 1). Notice that the first header on the table is denoted as "Game." Ask:
    a. How is "Game" represented in the source code? ***<th>Game</th>***
    b. What do you think the <th> code represents? ***The <th> is a tag for "table header"***
    c. If this were in RStudio, what would we call this header? ***A variable.***

8. Assign each student team one video game from the Wikipedia data table. Each team will compare how the information is stored in the table with its corresponding HTML source code. Each team should answer the following questions in the DS journals.
    a. Where are the variable names stored? ***The variable names are stored at the beginning of the code, in between <th> and </th>. <th> and </th> are called* tags *– they tell the browser to represent the information between them as a header in the table.***
    b. How are different values of the variables stored? ***Values are stored between the <td> and </td> tags.***
    c. Why do you think the data are stored in such a complex way? Why can't we just put them in a spreadsheet? ***Answers may vary by class. One reason is that the data must be displayed in a way that allows a browser to make it look pretty (and readable) on a computer screen.***
    d. How could we get this into an R dataframe so we can analyze it? ***In its current form, this would be very difficult. We would need to represent the data in a different format in order for R to understand it.***

9. Distribute the *Wikipedia – Video Games – CSV Format* handout (LMR_3.18) and explain that this is yet another way to represent the same video game data.

    <u>Note:</u> The handout only provides information on the first 5 rows of the Wikipedia table. A full version of the file (including all video games in the table) is located on the server with the title bestgames.csv.



Name:______________________________ Date:______________

Wikipedia – Video Games – CSV Format

Background:

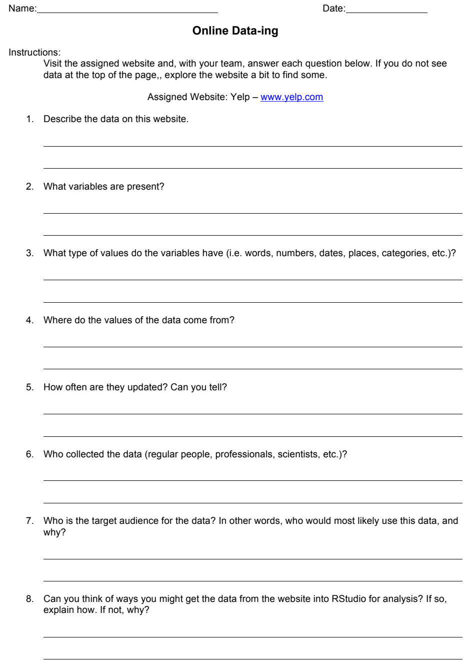
The data below represent the same 5 video games from the Wikipedia webpage, but in CSV format. CSV stands for “comma separated values.”

```
Game,Original release year,Genre,Number of lists,Platform of original release,Lists/References

The Legend of Zelda: Ocarina of Time,1998,Action-Adventure,42,Nintendo 64,"CNET,[1]
CVG2000,[2] Edge2000,[3] Edge2007,[4] Edge2009,[5] Edge2013,[6] Edge10s,[7] EGM2001,[8]
EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12] G4,[13] GameRankings,[14]
GamingBolt,[15] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19] GI1999,[20] GI2001,[21]
GI2009,[22] GameSpot2006,[23] GameSpot2014,[24] GuinnessConsole,[25] IGN10s,[26] IGN2003,[27]
IGN2005,[28] IGN2005Readers,[29] IGN2006Readers,[30] IGN2007,[31] IGN2008Readers,[32]
Maniac,[33] Metacritic,[34] NowGamer,[35] PM2014,[36] RetroGamer,[37] Slant2014,[38]
Stuff2008,[39] Stuff2014,[40] UnikGamer,[41] Yahoo[42]"

Chrono Trigger,1995,Role-Playing Game,39,Super Nintendo Entertainment
System,"ActionButton,[43] Dengeki,[44] Edge2007,[4] Edge2009,[5] Edge2013,[6] EGM1997,[45]
EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12] G4,[13]
GameRankings,[14] GamesRadar,[46] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19]
GameSpot2006,[23] GamingBolt,[15] GI1998,[47] GI1999,[20] GI2001,[21] GI2009,[22]
GuinnessConsole,[25] IGN10s,[26] IGN2003,[27] IGN2005,[28] IGN2005Readers,[29] IGN2007,[31]
IGN2006Readers,[30] IGN2008Readers,[32] IGN10s,[26] Maniac,[33] NowGamer,[35] PM2014,[36]
Slant2014,[38] Time,[48] UnikGamer[41]"

Final Fantasy VII,1997,Role-Playing Game,39,PlayStation,"CVG2000,[2] CVG2007PC,[49]
Dengeki,[44] Edge2007,[4] Edge2009,[5] Edge2013,[6] EGM1997,[45] EGM1997Dev,[50]
EGM1997Readers,[51] EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] Famitsu2006,[12]
G4,[13] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19] GameSpot2006,[23] GamingBolt,[15]
GI1998,[47] GI2001,[22] GI2009,[22] GuinnessConsole,[25] IGN2005,[28] IGN2005Readers,[29]
IGN2006Readers,[30] IGN2007,[31] IGN2008Readers,[32] NowGamer,[35] PM2014,[36] RetroGamer,[37]
Slant2014,[38] Stuff2008,[39] Stuff2014,[40] Time,[48] UnikGamer,[41] Yahoo[42]"

Super Mario 64,1996,Platformer,38,Nintendo 64,"CVG2000,[2] Edge2000,[3] Edge2007,[4]
Edge2009,[5] Edge2013,[6] Edge10s,[7] EGM1997,[45] EGM1997Dev,[50] EGM1997Readers,[51]
EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11] G4,[13] GF2004,[16] GF2005,[17]
GF2009,[18] GF2014,[19] GamingBolt,[15] GI2001,[21] GI2009,[22] GamesRadar,[46]
GuinnessConsole,[25] IGN2003,[27] IGN2005,[28] IGN2005Readers,[29] IGN2006Readers,[30]
Maniac,[33] NextGen1996,[52] NowGamer,[35] GameRankings,[14] IGN2007,[31] IGN2008Readers,[32]
PM2014,[36] RetroGamer,[37] Slant2014,[38] Stuff2008,[39] Time,[48] Yahoo[42]"

Street Fighter II,1991,Fighting,37,Arcade,"CVG2000,[2] Edge2000,[3] EG1992,[53] EGM1997,[45]
EGM1997Arcade,[54] EGM1997Dev,[50] EGM2001,[8] EGM2006,[9] Empire2009,[10] Empire2014,[11]
Famitsu2006,[12] FHM,[55] G4,[13] GF2004,[16] GF2005,[17] GF2009,[18] GF2014,[19]
GameSpot2006,[23] GameSpyArcade,[56] GamingBolt,[15] GI2001,[21] GI2009,[22]
GuinnessArcade,[57] GuinnessConsole,[25] IGN2003,[27] IGN2005,[28] IGN2007,[31]
NextGen1996,[52] NextGen1999,[58] NowGamer,[35] PM2014,[36] RetroGamer,[37] Slant2014,[38]
Stuff2008,[39] Stuff2014,[40] Time,[48] Yahoo[42]"
```

LMR_3.18

10. Inform students that a file with the CSV format is easily readable by R. Then ask:
    a. Where are the variable names stored? ***The variable names are stored in the first row***
    b. How are values of the variables separated? ***The values are separated by commas.***
    c. If we were interested in using the online data, how would we obtain it? ***This is a challenging problem – one which students may not know how to answer at this point. The objective is for them to struggle with how they would obtain data and recognize that it is not always as simple as “export, upload, import.”***

☑ 11. Split the class into their student teams and distribute the *Online Data-ing* handout (LMR_3.19). Assign each team a different website (each page of the handout lists a different site) and have them use this site to complete the questions in the handout.



Name:______________________________ Date:______________

Online Data-ing

Instructions:

Visit the assigned website and, with your team, answer each question below. If you do not see data at the top of the page,, explore the website a bit to find some.

Assigned Website: Yelp – www.yelp.com

1. Describe the data on this website.

   ____

   ____

2. What variables are present?

   ____

   ____

3. What type of values do the variables have (i.e. words, numbers, dates, places, categories, etc.)?

   ____

   ____

4. Where do the values of the data come from?

   ____

   ____

5. How often are they updated? Can you tell?

   ____

   ____

6. Who collected the data (regular people, professionals, scientists, etc.)?

   ____

   ____

7. Who is the target audience for the data? In other words, who would most likely use this data, and why?

   ____

   ____

8. Can you think of ways you might get the data from the website into RStudio for analysis? If so, explain how. If not, why?

   ____

   ____

LMR_3.19

12. Have each student team share their findings with one other team. They should have their website displayed while discussing their results.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next 2 Days

For the next 4 days, students will collect data using their newly created Participatory Sensing campaign.

Lab 3E: Scraping Web Data

Lab 3F: Maps

Complete Labs 3E and 3F prior to Lesson 21.

Lab 3E - Scraping web data

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

The web as a data source

- The internet contains huge amounts of information.
  - Using computers to gather this information in an automated fashion is referred to as *scraping web data*.
  - Scraping data from the web can be difficult because each website displays & stores data differently.
- In this lab, we'll learn how to scrape data in two steps:
  - Step 1: Gather information from the web.
  - Step 2: Clean it up and turn it into a usable data frame for `Lab 3F`.

Our first web scraper

- Copy and paste the link below into a web browser to view the website of data we'd like to *scrape* and analyze.

  https://labs.idsucla.org/extras/webdata/mountains.html

- **Briefly describe what the data on the website is about.**
  - **Then write down 3 questions you'd be interested in answering by analyzing this data.**

HTML

- `HTML` is the code that's used to render every website you've ever visited.
- The following slide shows the `HTML` code used to create the first two rows of the web data.
  - **How is the data table in `HTML` different than the data tables we're used to seeing in `R`, for example, when we use the `View()` function?**
  - **What do you think the *tags* `<TABLE>`, `<TR>`, `<TH>`, `<TD>` mean? How does `HTML` use these *tags* to display the table?**

```
<TABLE>
  <TR>
    <TH>peak</TH>
    <TH>range</TH>
    <TH>state</TH>
    <TH>long</TH>
    <TH>lat</TH>
    <TH>elev_ft</TH>
    <TH>elev_m</TH>
    <TH>prominence_ft</TH>
    <TH>prominence_m</TH>
    <TH>rank</TH>
  </TR>
  <TR>
    <TD>Denali (Mount McKinley)</TD>
    <TD>Alaska Range</TD>
    <TD>Alaska</TD>
    <TD>-151.0063</TD>
    <TD>63.0690</TD>
    <TD>20236</TD>
    <TD>6168</TD>
    <TD>20174</TD>
    <TD>6149</TD>
    <TD>1</TD>
  </TR>
</TABLE>
```

Get to scraping!

- Use your browser to go back to the website with the data we're interested in scraping.
- Find the URL address for the site and assign it the name `data_url` in `R`.
  - Then fill in the blanks below to have `R` scrape *every* web table available on the site:

```
         tables <- readHTMLTable(____)
```

Find our data

- Since `readHTMLTable()` scrapes *every* table that is on a particular web URL, we need to find out which table has the data we're interested in.
  - For example, `wikipedia.org` often has articles with 3 or more tables.
  - This means we need to check all 3 tables to find the one we're interested in.
- Use the `length()` function to find out how many tables of data were scraped in our set of `tables`.

Saving tables

- Now that we know how many tables we've scraped, we can go back and scrape individual tables by adding the `which` argument to the `readHTMLTable()` function.
  - Use `readHTMLTable()` to re-scrape the data from the web but this time use the `which` argument to scrape just the individual table.
  - The `which` argument should be the integer denoting which table you want scraped.
  - Assign the scraped data the name `mtns`

Check, save and use!

- After scraping the data, the only thing left to do is to save it and use it.
- Fill in the blanks to save the data and give it a file name

```
save(____, file = "____.Rda")
```

- **What is the mean and standard deviation of `elev_ft`?**
- **Which `state` has the most mountains in our data?**

Lab 3F - Maps

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Informative and Fun!

- Maps are some of the most interesting plots to make because the info represents:
  - Where we live.
  - Where we go.
  - Places that interest us.
- Maps are also helpful to display geographic information.
  - John Snow (the physician, not the character from *Game of Thrones*...) once famously used a map to discover how cholera was transmitted.
- In this lab, we'll use `R` to create an interactive map of the `mtns` data we scraped in Lab 3E.

Getting ready to map

- The map we'll be creating will end up in RStudio's *Viewer* pane.
  - Which means you'll need to alternate between building the map and loading the lab.
- You'll find it very helpful, for this lab, to write all of the commands, including the `load_lab(23)` command, as an `R` script.
  - This way you can edit the code that builds the map and quickly reload the lab.

Load your data!

- In Lab 3E you created a dataset. Load it into RStudio now by filling in the blank with the file name of the data.

```
load("___.Rda")
```

- Didn't finish the lab or save the data file? Ask a friend to share it!

Build a Basic Map

- Let's start by building a basic map!
- Use the `leaflet()` function and the mtns data to create the `leaf` that we can use for mapping.

```
mtns_leaf <- leaflet(____)
```

- Then, insert `mtns_leaf` into the `addTiles()` function and assign the output the name `mtns_map`
- Run `mtns_map` in the console to look at your basic map with no data displayed.
  - Be sure to try clicking on the map to pan and zoom.

Including our data

- Now we can add markers for the locations of the mountains using the `addMarkers()` function.
  - Fill in the blanks below with the basic map we've created and the values for latitude and longitude.

```
addMarkers(map = ____, lng = ~____, lat = ~____)
```

- Supply the `peak` variable, in a similar way as we supplied the `lat` and `long` variables, to the popup argument and include it in the code above.

- **Click on a marker within California and write down the name of the mountain you clicked on.**

Colorize

- Our current map looks pretty good, but what if we wanted to add some colors to our plot?
- Fill in the blanks below to create a new variable that assigns a color to each mountain based on the `state` its located.

```
mtns <- mutate(____, state_colors = colorize(____))
```

- Now that we've added a new variable, we need to re-build `mtns_leaf` and `mtns_map` to use it.
  - Create `mtns_leaf` and `mtns_map` as you did before.
  - Then change `addMarkers` to `addCircleMarkers` and keep all of the arguments the same.

Showing off our colors

- To add the colors to our plot, use the `addCircleMarkers` like before but this time include `color = ~state_colors` as an argument.
- It's hard to know just what the different colors mean so let's add a legend.
  - First, assign the map with the circle markers as `mtns_map`.
  - Then, fill in the blanks below to place a legend in the top-right hand corner.

```
addLegend(____, colors = ~unique(____), labels = ~unique(____))
```

Lesson 21: Learning to Love XML

Objective:

Students will understand the need for data to be stored in different ways - specifically, why it makes sense for web data to be formatted as XML.

Materials:

1. *Online Data-ing* handout (LMR_3.19_Online Data-ing)
   Note: This should have been completed during the previous class.
2. Mountain Peak XML data found at:
   https://labs.idsucla.org/extras/webdata/mountains.html

   Note: Open with Google Chrome or Firefox browsers, NOT with Safari.
3. Projector
4. *Mountains – HTML vs. XML* handout (LMR_3.20_Mountins – HTML vs. XML)

Vocabulary:

XML

Essential Concepts: XML is a programming language that we use with our campaigns. We create basic XML "tags" in the code, which help us store data in a format we understand.

Lesson:

1. Allow time for student teams to present their findings from the *Online Data-ing* handout (LMR_3.19) if there was not sufficient time during the previous lesson.

2. Remind students that in the previous lesson they learned about a variety of ways that data can be presented online.

3. They've been working with comma separated (CSV) files and R data frames. Last time and in the lab, they worked with HTML tables. Today they are going to learn how HTML can be displayed as an XML table.

4. **XML**, or Extensible Mark up Language, is a popular format for storing data on the Internet. It is useful because it creates readable web pages, and also because it allows programmers to easily update values in the data table if those values change.

5. In pairs, ask students to brainstorm ways in which data that is found online is different than the way we see data in RStudio. Then, create a class brainstorm from the student pair responses.

6. After the brainstorm, emphasize the following:

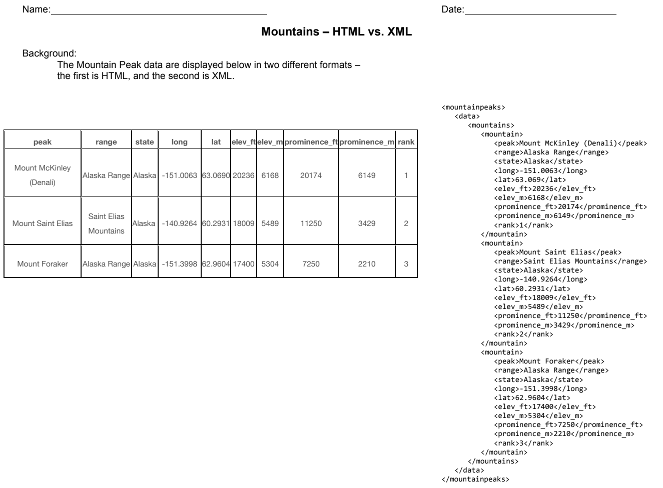
   a. RStudio's default way to work with data is as large data frames (tables) where rows represent observations and columns represent variables.
   b. Data that is viewed online often has a different structure.
   c. Data structures found on the web might be displayed in tables, such as those on Wikipedia, or streams, such as Twitter, and might even include data spread across multiple sections of a web page, such as Yelp.

   Show students, on a projector, the Mountain Peak XML data found at https://labs.idsucla.org/extras/webdata/mountains.html

   Ask students to look at the data and determine if they have seen it before. Hint: They have! It was the data they scraped during Lab 3E.

7. Once students figure out that the XML is just the same data as the website they scraped during Lab 3E, distribute the *Mountains – HTML vs. XML* handout (LMR_3.20), which displays both HTML and XML versions of the data.

   Note: The handout only includes the first 3 mountains.



Name:___________________________________ Date:__________________________

Mountains – HTML vs. XML

Background:

The Mountain Peak data are displayed below in two different formats – the first is HTML, and the second is XML.

peak	range	state	long	lat	elev_ft	elev_m	prominence_ft	prominence_m	rank
Mount McKinley (Denali)	Alaska Range	Alaska	-151.0063	63.0690	20236	6168	20174	6149	1
Mount Saint Elias	Saint Elias Mountains	Alaska	-140.9264	60.2931	18009	5489	11250	3429	2
Mount Foraker	Alaska Range	Alaska	-151.3998	62.9604	17400	5304	7250	2210	3

```
<mountainpeaks>
   <data>
      <mountains>
         <mountain>
            <peak>Mount McKinley (Denali)</peak>
            <range>Alaska Range</range>
            <state>Alaska</state>
            <long>-151.0063</long>
            <lat>63.069</lat>
            <elev_ft>20236</elev_ft>
            <elev_m>6168</elev_m>
            <prominence_ft>20174</prominence_ft>
            <prominence_m>6149</prominence_m>
            <rank>1</rank>
         </mountain>
         <mountain>
            <peak>Mount Saint Elias</peak>
            <range>Saint Elias Mountains</range>
            <state>Alaska</state>
            <long>-140.9264</long>
            <lat>60.2931</lat>
            <elev_ft>18009</elev_ft>
            <elev_m>5489</elev_m>
            <prominence_ft>11250</prominence_ft>
            <prominence_m>3429</prominence_m>
            <rank>2</rank>
         </mountain>
         <mountain>
            <peak>Mount Foraker</peak>
            <range>Alaska Range</range>
            <state>Alaska</state>
            <long>-151.3998</long>
            <lat>62.9604</lat>
            <elev_ft>17400</elev_ft>
            <elev_m>5304</elev_m>
            <prominence_ft>7250</prominence_ft>
            <prominence_m>2210</prominence_m>
            <rank>3</rank>
         </mountain>
      </mountains>
   </data>
</mountainpeaks>
```

LMR_3.20

8. Ask student pairs to answer the following:

   a. Why are certain XML tags indented in the XML version of the data? ***The indentations tell us how to structure the HTML table. For example, all the mountains are contained in the <data> section, but are further tagged by each particular mountain within the <mountain> and </mountain> tags. All information stored between those two tags will be displayed as one row of the HTML table.***

   b. What are the role of tags (ex. <state>) and end tags (ex. </state>) in the XML code? ***Tags tell us when a certain type of data begins, and end tags tell us when the data should end. In other words, it tells us where to find the specific values of a variable (ex. Alaska would be the value of the "state" variable since it is between the <state> and </state> tags.***

   c. Where are the variable names? ***The variable names can be found between each <mountain> and </mountain> tags. Specifically, the first variable is "peak" and the last variable is "rank."***

   d. Where are the observations? ***The observations are located within each of the variable tags. For example, the observation "Mount McKinley (Denali)" is found between the <peak> and </peak> tags.***

9. Assign student pairs one of the above questions to share out with the class. Student pairs that did not receive an assignment must participate using the *Agree/Disagree* strategy.

10. As a class, discuss the answers to the questions above.

11. XML formats make it easier to display data on the web in a pleasant matter and make it easier for programmers to find and alter data if the values change or if, for example, they wish to add a new row to a table.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

For the next 3 days, students will collect data using the class's newly created Participatory Sensing campaign (see Lessons 16-18).

For homework, students should reflect about how XML and HTML data are displayed. They should discuss when each format is appropriate.

Lesson 22: Changing Format

Objective:

Students will learn how to convert XML files to the more familiar data table format and vice versa.

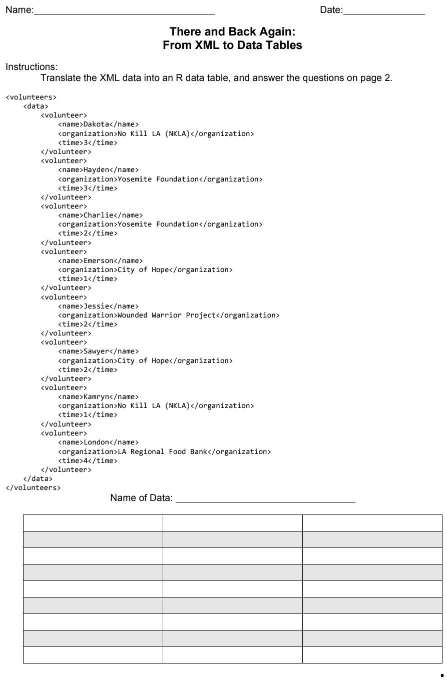
Materials:

1. *There and Back Again: From XML to Data Tables* handout (LMR_3.21_From XML to Data Tables)
2. *There and Back Again: From Data Tables to XML* handout (LMR_3.22_From Data Tables to XML)

Essential Concepts: Converting XML to spreadsheet format helps us better understand and view our data.

Lesson:

1. Take a few minutes to compare the structure of XML code to HTML data tables (refer to Step 7 from Lesson 21).
2. Inform students that in today's lesson, they will learn how to translate information from XML code into a data table.
3. Distribute the *There and Back Again: From XML to Data Tables* handout (LMR_3.21) to students.



Name:______________________________ Date:______________

There and Back Again:
From XML to Data Tables

Instructions:

Translate the XML data into an R data table, and answer the questions on page 2.

```
<volunteers>
    <data>
        <volunteer>
            <name>Dakota</name>
            <organization>No Kill LA (NKLA)</organization>
            <time>3</time>
        </volunteer>
        <volunteer>
            <name>Hayden</name>
            <organization>Yosemite Foundation</organization>
            <time>3</time>
        </volunteer>
        <volunteer>
            <name>Charlie</name>
            <organization>Yosemite Foundation</organization>
            <time>2</time>
        </volunteer>
        <volunteer>
            <name>Emerson</name>
            <organization>City of Hope</organization>
            <time>1</time>
        </volunteer>
        <volunteer>
            <name>Jessie</name>
            <organization>Wounded Warrior Project</organization>
            <time>2</time>
        </volunteer>
        <volunteer>
            <name>Sawyer</name>
            <organization>City of Hope</organization>
            <time>2</time>
        </volunteer>
        <volunteer>
            <name>Kamryn</name>
            <organization>No Kill LA (NKLA)</organization>
            <time>1</time>
        </volunteer>
        <volunteer>
            <name>London</name>
            <organization>LA Regional Food Bank</organization>
            <time>4</time>
        </volunteer>
    </data>
</volunteers>
```

Name of Data: ____________________________

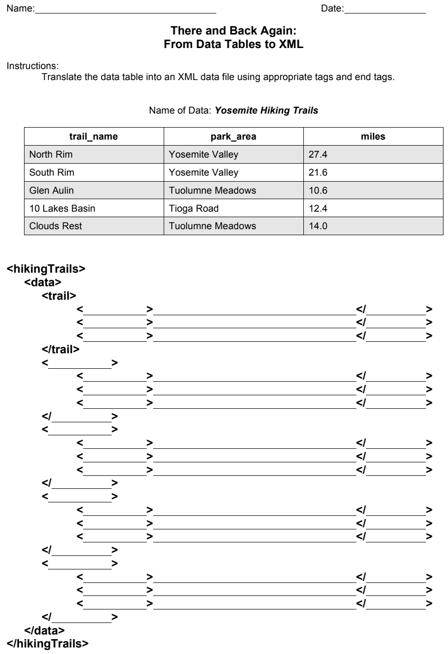
LMR_3.21

4. Inform the students that XML code is provided on page 1 of the handout, and their goal is to transfer all the information to the empty data table.
5. As a guide, ask a volunteer to find and name one of the variables in the XML code and then have all the students write the name of the variable in the first column of the top row in the data table.
6. Next, ask another student to find the first value of the variable named in Step 5. This value should be placed in the correct column and row of the data table.
7. Provide time for students to complete the handout individually.

8. Using the Anonymous Author strategy, share a couple of the completed data tables. Ask teams to discuss how they are alike and how they are different.

   Note: Most tables will probably be the same, but could vary slightly based on which columns each variable name was placed in, and in what order the observations were listed in the rows. Ultimately, the information contained in the data tables is the same.

9. Then, conduct a whole class discussion regarding student responses to the questions on page 2 of the handout.
10. Distribute the *There and Back Again: From Data Tables to XML* (LMR_3.22) to student teams and allow them time to complete it.



Name:______________________________ Date:______________

There and Back Again:
From Data Tables to XML

Instructions:
Translate the data table into an XML data file using appropriate tags and end tags.

Name of Data: ***Yosemite Hiking Trails***

trail_name	park_area	miles
North Rim	Yosemite Valley	27.4
South Rim	Yosemite Valley	21.6
Glen Aulin	Tuolumne Meadows	10.6
10 Lakes Basin	Tioga Road	12.4
Clouds Rest	Tuolumne Meadows	14.0

```
<hikingTrails>
    <data>
        <trail>
            <________>__________________________</________>
            <________>__________________________</________>
            <________>__________________________</________>
        </trail>
        <_________>
            <________>__________________________</________>
            <________>__________________________</________>
            <________>__________________________</________>
        </_________>
        <_________>
            <________>__________________________</________>
            <________>__________________________</________>
            <________>__________________________</________>
        </_________>
        <_________>
            <________>__________________________</________>
            <________>__________________________</________>
            <________>__________________________</________>
        </_________>
        <_________>
            <________>__________________________</________>
            <________>__________________________</________>
            <________>__________________________</________>
        </_________>
    </data>
</hikingTrails>
```

LMR_3.22

11. Once teams have finished, teams will guide you to write the correct XML code.
12. Using a *Whip Around*, teams will tell you the first line of the XML code you need to write. Teams waiting their turn will check if the team is guiding you correctly. If not, they need to stop you and propose their line of code. You may not continue writing the lines of code until all teams are in agreement.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

Students will continue to collect data using the class's Participatory Sensing campaign (see Lessons 17-19). They will analyze the data the next day during the practicum.

Practicum: What Does Our Campaign Data Say?

Objective:

Students will answer the statistical question they generated at the beginning of the Participatory Sensing campaign creation lesson. They will use RStudio to make graphical representations or numerical summaries of their data to answer their question.

Materials:

1. *Our Own Campaign* (LMR_U3_Practicum_Our Own Campaign)

Practicum
Our Own Campaign

At the start of the Participatory Sensing campaign creation in lesson 16, the class developed a research question about your class's topic of interest.

It is now time to analyze and interpret your class campaign data. You will use the data from your class-created campaign only. Based on the analysis, you can also wonder about what other data would be necessary to better answer your question, if any.

Based on the class's campaign data collected:

1. Refer back to the statistical questions your class generated in lessons 16-18 that address the research question.

2. Choose one of these statistical questions and determine which variables will answer this question.

3. Analyze the data to answer the question you've chosen. Your analysis should include graphs and numerical summaries. You should:
   a. Provide the plot and numerical summary.
   b. Describe what the plot shows.
   c. Explain why you chose to make that particular plot.
   d. Explain how the plot and numerical summary answers your statistical question.
   e. Include the code you used in RStudio to make your plot.

4. After analyzing your data, determine if additional data would better answer your statistical question. If so, propose what that data would be. Different variables? Different data collection approach? Same variables, but more people? Same variables and people but more time?

5. Now, choose two more statistical questions that address the research question.

6. Analyze and interpret the data to answer these questions.

7. Sometimes, when analyzing data, we think of new statistical questions to ask, or we realize that the data need to be cleaned before we can answer. Explain whether this is the case with any of your statistical questions.

8. Write a one-page report and present it to another member of the class who is not in your team.

End of Unit Project and Oral Presentation: TB or Not TB?

Objective:
Students will apply what they have learned in the unit.

Materials:

1. Computers
2. *IDS Unit 3 – Project and Oral Presentation* (LMR_U3_End of Unit Project)

IDS Unit 3 – End of Unit Project
TB or Not TB

Experiments in the medical field that involve new treatments (new medications) are called clinical trials. You have received a data set that shows the results from Sir Austin Bradford Hill's first randomized study in 1948 examining the effects of the antibiotic Streptomycin on 107 tuberculosis patients. You and a partner will use this data set to find out if Streptomycin is an effective treatment for tuberculosis.

A short article about tuberculosis facts can be found at:

http://www.cdc.gov/tb/publications/factsheets/general/tb.htm

Since this is an experiment, answer the following questions below. You may need to research the answer to some of the questions.

a. What is the research question?
b. Who are the subjects that participated in the experiment?
c. What is the treatment?
d. Who is in the treatment group?
e. Who is in the control group?
f. How were the subjects assigned to each group?
g. What population is this experiment representative of?
h. What is the variable that we will be measuring?
i. What is the outcome of this experiment?

To answer your research question, you and a partner will compare the outcome of the data with the outcomes given by a chance model (in which Streptomycin has no effect on TB).

1. First, scrape the data. Refer to the web scraping lab if you need to recall how to scrape data. To access Sir Hill's data, go to: https://labs.idsucla.org/extras/webdata/tb.html

2. Second, determine the percentages of subjects in the study that died and the percentages of the subjects that recovered for each group.

3. Third, assuming that the treatment had no effect, use the data to:

   a. Calculate the percentage of people with tuberculosis we would expect to die.
   b. Use the *expected* percentage for (a), above, to calculate the number of people we expect to die from the treatment group.
   c. Compare the percentage from (b) to the percentage from the treatment group *actually* died.

4. Then, if we assume that the outcome does not depend on the treatment, design and complete an appropriate simulation in RStudio using a chance model to replicate Sir Hill's study:

   a. Shuffle the treatment and control labels 300 times; each time, calculate the percentage of treatment patients who "died". Plot the distribution of the 300 percentages. Refer to the simulation labs if you need to recall how to create a simulation.
   b. Use the results from the chance model (shuffling) to determine whether (i.) or (ii.) below is the most reasonable explanation for the actual data in Sir Hill's study and state why:

   i. Streptomycin is a much better treatment for tuberculosis than bed rest. So, the outcome depends on the treatment.
   ii. The actual difference between treatments is due to chance; Streptomycin may not be effective on tuberculosis. So, it is possible that treatment and outcome are independent.

5. Can we say that Streptomycin ***causes*** the recovery of tuberculosis patients? Explain your answer.

Create a 4-5 slide, 5-minute presentation that shows your results. Be sure to include a detailed explanation of how you and your partner decided to conduct your simulation. Each person must participate in the presentation. In addition to the presentation, submit a 2-4 page, double-spaced summary of your analysis.

Introduction to Data Science

Unit 4

Introduction to Data Science

Daily Overview: Unit 4

Theme	Day	Lessons and Labs	Campaign	Topics	Page
Predictions and Models (15 days)	1	Lesson 1: Water Usage		Data cycle, official data sets	318
	2	Lesson 2: Exploring Water Usage		Exploratory data analysis, campaign creation	321
	3	Lesson 3: Evaluating and Implementing a Water Campaign	Water Campaign—data	Statistical questions, evaluate & mock implement campaign	323
	4^	Lesson 4: Refining the Water Campaign	Water Campaign—data	Revise and edit campaign, data collection	325
	5	Lesson 5: Statistical Predictions Using One Variable	Water Campaign—data	One-variable predictions using a rule	327
	6	Lesson 6: Statistical Predictions by Applying the Rule	Water Campaign—data	Predictions applying mean square deviation, mean absolute error	330
	7	Lesson 7: Statistical Predictions Using Two Variables	Water Campaign—data	Two-variable statistical predictions, scatterplots	334
	8	*LAB 4A: If the Line Fits…*	Water Campaign—data	Estimate line of best fit	337
	9	*LAB 4B: What's the Score?*	Water Campaign—data	Comparing predictions to real data	339
	10	Lesson 8: What's the Trend?	Water Campaign—data	Trend, associations, linear model	341
	11	Lesson 9: Spaghetti Line	Water Campaign—data	Estimate line of best fit, single linear regression	345
	12	*LAB 4C: Cross-Validation*	Water Campaign—data	Use training and testing data for predictions	348
	13	Lesson 10: Predicting Values	Water Campaign—data	Predictions based on linear models	351
	14	Lesson 11: How Strong Is It?	Water Campaign—data	Correlation coefficient, strength of trend	353
	15	*LAB 4D: Interpreting Correlations*	Water Campaign—data	Use correlation coefficient to determine best model	355
Piecing it Together (6 days)	16	Lesson 12: More Variables to Make Better Predictions	Water Campaign—data	Multiple linear regression	359
	17	Lesson 13: Combination of Variables	Water Campaign—data	Multiple linear regression	362
	18	*LAB 4E: This Model Is Big Enough for All of Us*	Water Campaign—data	Multiple linear regression	365
	19	Practicum: Predictions	Water Campaign—data	Linear regression	366
	20	Lesson 14: Improving Your Model	Water Campaign—data	Non-linear regression	367
	21	*LAB 4F: Some Models Have Curves*	Water Campaign—data	Non-linear regression	369
The Growth of Landfills (5 days)	22	Lesson 15: The Growth of Landfills	Water Campaign—data	Modeling to answer real-world problems	373
	23	Lesson 16: Exploring Trash via the Dashboard	Water Campaign—data	Analyze data to improve models	376
	24	Lesson 17: Exploring Trash via RStudio	Water Campaign—data	Analyze data to improve models	377
	25	Prepare Team Presentations	Water Campaign—data	Modeling with statistics	-
	26	Present Team Recommendations	Water Campaign—data	Modeling with statistics	-
Decisions, Decisions! (3 days)	27	Lesson 18: Grow Your Own Classification Tree	Water Campaign—data	Multiple predictors, classifying into groups, decision trees	379
	28	Lesson 19: Data Scientists or Doctors?	Water Campaign—data	Decision trees based on training and testing data	384
	29	*LAB 4G: Growing Trees*	Water Campaign—data	Decision trees to classify observations	387
Ties that Bind (3 days)	30	Lesson 20: Where Do I Belong?	Water Campaign—data	Clustering, k-means	390
	31	*LAB 4H: Finding Clusters*	Water Campaign—data	Clustering, k-means	395
	32+	Lesson 21: Our Class Network	Water Campaign—data	Clustering, networks	397
End of Unit Project (7 days)	33-40	End of Unit 3 and 4 Design Project and Oral Presentations: Water Usage	Water Campaign	Synthesis of above	400

^=Data collection window begins.
+=Data collection window ends.

IDS Unit 4: Essential Concepts

Lesson 1: Water Usage

Data can be used to make predictions. Official data sets rely on censuses or random samples and can be used to make generalizations. On the other hand, data from Participatory Sensing campaigns are not random and rely on the sensors, in our case, humans, to be gathered and limits the ability to generalize.

Lesson 2: Exploring Water Usage

Exploring different data sets can give us insight about the same processes. Information from an official data set compared with a Participatory Sensing data set can yield more information than one data set alone. Research questions provide an overall direction to make comparisons between data sets.

Lesson 3: Evaluating and Implementing a Water Campaign

Statistical questions guide a Participatory Sensing campaign so that we can learn about a community or ourselves. These campaigns should be evaluated before implementing to make sure they are reasonable and ethically sound.

Lesson 4: Learning About Our Water Campaign

Statistical questions guide a Participatory Sensing campaign so that we can learn about a community or ourselves. These campaigns should be tried before implementing to make sure they are collecting the data they are meant to collect and refined accordingly.

Lesson 5: Statistical Predictions using One Variable

Anyone can make a prediction. But statisticians measure the success of their predictions. This lesson encourages the classroom to consider different measures of success.

Lesson 6: Statistical Predictions by Applying the Rule

If we use the squared residuals rule, then the mean of our current data is the best prediction of future values. If we use the mean absolute error rule, then the median of the current data is the best prediction of future values.

Lesson 7: Statistical Predictions Using Two Variables

When predicting values of a variable *y*, and *if y* is associated with *x*, then we can get improved predictions by using our knowledge about *x*. Basically, we “subset” the data for a given value of *x*, and use the mean *y* for those subset values. If the resulting means follow a trend, we can model this trend to generalize to as-yet unseen values of *x*.

Lesson 8: What’s the Trend?

Associations are important because they help us make better predictions; the stronger the trend, the better the prediction we can make. “Better” in this case means that our mean squared residuals can be made smaller.

Lesson 9: Spaghetti Line

We can often use a straight line to summarize a trend. “Eye balling” a straight line to a scatterplot is one way to do this.

Lesson 10: Predicting Values

The regression line can be used to make good predictions about values of *y* for any given value of *x*. This works for exactly the same reason the mean works well for one variable: the predictions will make your score on the mean squared residuals as small as possible.

Lesson 11: How Strong Is It?

A high absolute value for correlation means a strong linear trend. A value close to 0 means a weak linear trend.

Lesson 12: More Variables to Make Better Predictions

We can use scatterplots to assess which variables might lead to strong predictive models. Sometimes using several predictors in one model can produce stronger models.

Lesson 13: Combination of Variables

If multiple predictors are associated with the response variable, a better predictive model will be produced, as measured by the mean absolute error.

Lesson 14: Improving your Model

If a linear model is fit to a non-linear trend, it will not do a good job of predicting. For this reason, we need to identify non-linear trends by looking at a scatterplot or the model needs to match the trend.

Lesson 15: The Growth of Landfills

Modeling does not always have to produce an equation. Instead, we can create models to answer real-world problems related to our community.

Lesson 16: Exploring Trash via the Dashboard

Exploring the IDS Dashboard provides a visual approach to data analysis.

Lesson 17: Exploring Trash via RStudio

RStudio can be used to verify initial results/findings from data analysis done via the IDS Dashboard.

Lesson 18: Grow Your Own Classification Tree

Many data sets have multiple predictors and are very non-linear. We can still use this data, but need to model it differently, such as in a decision tree. Decision trees are a useful tool for classifying observations into groups.

Lesson 19: Data Scientists or Doctors?

We can determine the usefulness of decision trees by comparing the number of misclassifications in each.

Lesson 20: Where Do I Belong?

We can identify groups, or “clusters,” in data based on a few characteristics. For example, it is easy to classify a classroom into males and females, but what if you only knew each student’s arm span? How well could you classify their genders now?

Lesson 21: Our Class Network

Networks are made when observations are interconnected. In a social setting, we can examine how different people are connected by finding relationships between other people in a network.

Predictions and Models

Instructional Days: 16

Enduring Understandings

The regression line is a prediction machine. We give it an x-value, it gives us a predicted y-value. The regression line summarizes the trend in the data, but there may still remain variability in the dependent variable that is not explained by the independent variable. Although the regression line provides optimal predictions when the association is linear, other models are needed for when it is not linear.

Engagement

Students will analyze a map from the Medical Daily website. The map and its article called *How Twitter Can Predict Heart Disease: Negative Tweets Associated With Stress, Higher Risk Of Disease*, shows a side-by-side comparison of CDC heart attack deaths data and Twitter's predicted data. They will engage in a discussion comparing and contrasting the visualization. The map can be found at:

http://www.medicaldaily.com/how-twitter-can-predict-heart-disease-negative-tweets-associated-stress-higher-risk-318830

Learning Objectives

Statistical/Mathematical:

S-ID 6: Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. *Use given functions or choose a function suggested by the context. Emphasize linear models*.
b. Informally assess the fit of a function by plotting and analyzing residuals.
c. Fit a linear function for a scatter plot that suggests a linear association.

S-ID 7: Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

S-ID 8: Compute (using technology) and interpret the correlation coefficient of a linear fit.

S-IC 6: Evaluate reports based on data. *

*This standard is woven throughout the course. It is a recurring standard for every unit.

Focus Standards for Mathematical Practice for All of Unit 4:

SMP-2: Reason abstractly and quantitatively.

SMP-4: Model with mathematics.

SMP-7: Look for and make use of structure.

Data Science:

Judge whether or not the linear model is appropriate. Learn to interpret a correlation coefficient in a linear model and interpret slope and intercept. Evaluate the strength of a linear association. Evaluate the potential error in a linear model.

Applied Computational Thinking using RStudio:

- Use linear regression models to predict response values based on sets of predictors.
- Fit a regression line to data and predict outcomes.
- Compute the correlation coefficient of a linear model.
- Create a Participatory Sensing campaign using a campaign Authoring Tool.

Real-World Connections:

Many studies are published in which predictions are made, and media reports often cite data that make predictions. They involve one or more explanatory variable and a response variable, such as income vs. education, weight vs. exercise, and cost of insurance vs. age. Understanding linear regression helps evaluate these studies and reports.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will use complex sentences to write informative short reports that use data science concepts and skills.
4. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data File:

1. LA DWP (`dwp_2010`)
2. Movies (`movie`)

Data Collection:

Students will collect data for their water usage campaign.

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 1: Water Usage

Objective:

Students will compare and contrast an official data set versus a Participatory Sensing data set. They will begin to analyze an official data set from 2010 provided by the Los Angeles Department of Water and Power (DWP) to help them understand how water was used in the Los Angeles area in the recent past, before the drought.

Materials:

1. *Video: California Drought Crisis Reaches Worst Level as It Spreads North* http://www.nbcnews.com/storyline/california-drought/california-drought-crisis-reaches-worst-level-it-spreads-north-n169516
2. *Webpage: Twitter vs. Heart Disease* Webpage (Found at: http://www.medicaldaily.com/how-twitter-can-predict-heart-disease-negative-tweets-associated-stress-higher-risk-318830)
3. Class Created Campaign Information (from Unit 3, Lessons 17-19)

Vocabulary:

census

Essential Concepts: Data can be used to make predictions. Official data sets rely on censuses or random samples and can be used to make generalizations. On the other hand, data from Participatory Sensing campaigns are not random and rely on the sensors, in our case, humans, to be gathered and limits the ability to generalize.

Lesson:

1. Ask students to recall that statistics are used to make predictions about population parameters.

2. Project the map found on the Medical Daily website. Inform students that Twitter data was compared to CDC heart disease deaths data on side-by-side maps. Using a *Think, Pair, Share* ask students to discuss:

    a. What is the source of the data on each map? ***A: Tweets as predicted by Twitter and heart attacks as listed on a death certificate and recorded by the CDC***.
    b. What do the colors on the map mean? ***A: On the spectrum from green to red, green means fewer deaths by heart attack and red means a greater number of deaths by heart attacks***.
    c. How are the maps the same? How are they different?
    d. How reliable are the methods used to report these data? ***A: In the case of the CDC data, we have verifiability (death certificates). On the other hand, the Twitter data predicts based on a person's word.***
    e. How scalable are the methods and can they be generalized? ***A: Official data sets are usually censuses or random samples; they address things at a high level. Participatory Sensing or, in this case the Twitter data, is not random, but addresses things at a personal or local level; however, because it is not a census nor a random sample, it is difficult to be precise about uncertainty or ability to generalize***.

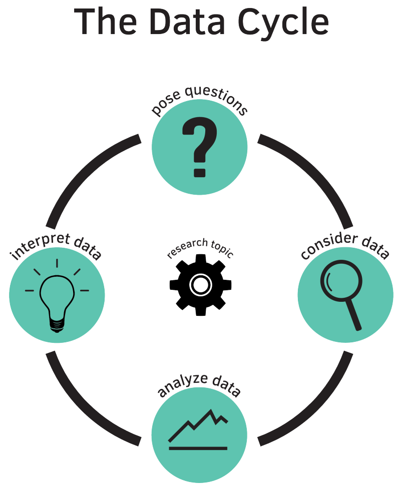
3. Quickly share student responses to the discussion. Then, inform them that this unit focuses on data to make predictions.

4. Set the context for the next three lessons. Inform students that they will be delving into the topic of water usage. In California, water usage is extremely important, given that the state has been in an exceptional drought in the second decade of the 21st century.

5. Using the K-L-W strategy in their DS journals, give students a couple of minutes to write what they *Know* about droughts. Then, students will write what they *Learned* about the California drought as they watch the brief NB News video clip titled *California Drought Crisis Reaches Worst Level as It Spreads North*. Finally, they will write 2-3 questions about what they *Want to*

know/learn about droughts. Video clip is found at: http://www.nbcnews.com/storyline/california-drought/california-drought-crisis-reaches-worst-level-it-spreads-north-n169516.

6. Do a quick *Whip Around* to share some of the students' responses to the *K-L-W*.
7. Inform students that they will be learning about water usage in their own neighborhoods. The LADWP data provides information at a high level about some (most) neighborhoods in L.A. Students will investigate how they can learn more using participatory sensing. In statistics, the Data Cycle provides a process by which we can learn about or investigate a particular topic of interest. To review the components of the Data Cycle, give students two minutes to *Quick Sketch* each phase of the cycle in their DS Journals.
8. After students have had an opportunity to do their sketches, display the Data Cycle graphic below to review it:



9. Next, review their class created campaign from Unit 3. Using a Pair-Share strategy, ask students to discuss when a Participatory Sensing campaign should be used rather than a survey. ***A: Answers will vary. Research questions that include variation across time or across locations are good candidates for Participatory Sensing campaigns; therefore, a trigger is necessary in order to record observations at multiple time points and locations. If a question needs to be answered only once, then a survey is a better method.***
10. Remind students that in the last unit, they created one campaign for the entire class. In this unit, each student team will be creating and implementing a campaign on the topic of water usage.
11. Before they start creating their campaign, they are going to explore an official data set provided by the Los Angeles Department of Water and Power (DWP) to learn more about water usage in the Los Angeles area. Some students may receive services from the DWP.
12. Explain how the data were collected:
    - The data you will see reflects the water usage in Los Angeles in the fiscal year that began in 2010 (July 2010-June 2011). At this time, L.A. was entering a drought, but water conservation efforts had not yet begun.
    - The DWP supplies water to businesses and addresses within its boundaries. It records the amount of water delivered to each address each month. For privacy purposes, it doesn't report how much water a single address uses.
    - Instead, it combines these into neighborhoods. These neighborhoods are defined by the U.S. Census and called Census Blocks. A census block is usually one, sometimes two, square blocks.
    - The DWP reports separate water usage figures for businesses, government structures (such as schools), and residences. For privacy purposes, this data set eliminated any Census Block that had fewer than 15 addresses.

- Water use is reported in Hundreds of Cubic Feet (HCF) per month (one HCF is about 748 gallons). Display the picture below, which shows a truck that holds about 6 HCF, so students can get a sense of the amount of water as reported.

13. Load and display the DWP data set in RStudio using the command `data(dwp_2010)`. Then, expand the spreadsheet (⧉) and explain what each variable in the data set means (they may want to record these in their DS journals):
    a. census = census block
    b. sector_type = category of facility
    c. longitude, latitude = GPS coordinates for center of census block location
    d. census_pop = population of census block (number of water users)
    e. total = total number of HFCs used by sector type per block in 2010
    f. july through june = number of HFCs used by sector type per month
    g. count = number of facilities per census block for that sector type
14. Next, load an interactive map of the DWP 2010 data by visiting: https://labs.idsucla.org/extras/animations/watermap/watermap.html
15. Lead a discussion about what is on the page. Ask:
    a. What do the colors and percentages on the legend mean?
    b. What trends do you see?
16. Then click on a marker (circle) to show the popup. Ask:
    a. What information is the popup displaying?
    b. How is the popup displaying the information?
17. Then, click on the Size by census_pop circle under the legend. Ask:
    a. What do you notice about the markers?
    b. What is the size of each marker telling us?
    c. What else do you see?
18. Now that they know what the variables mean, ask student teams to generate two statistical questions about the data. Below are three examples of possible statistical questions:
    a. What month uses the most water?
    b. Typically, how much water do residences consume during that month?
    c. Does this change if you factor in the number of people living in that census block?
19. If time permits, conduct a share-out of the teams' statistical questions.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 2: Exploring Water Usage

Objective:

Students will engage in exploratory data analysis with a Los Angeles Department of Water and Power (DWP) data set and begin the creation of a water usage Participatory Sensing campaign to observe patterns of water use in their neighborhoods.

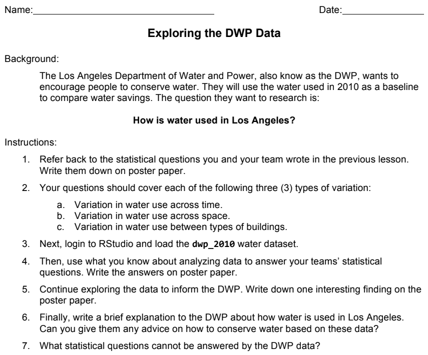
Materials:

1. *Exploring the DWP Data* (LMR_4.1_Exploring DWP Data)
2. *Water Campaign* (LMR_4.2_Water Campaign)
3. Poster paper
4. Markers
5. Class Created Campaign Information (from Unit 3, Lessons 17-19)

Essential Concepts: Exploring different datasets can give us insight about the same processes. Information from an official dataset compared with a participatory sensing dataset can yield more information than one dataset alone. Research questions provide an overall direction to make comparisons between datasets.

Lesson:

1. Display the DWP data using RStudio. In pairs, ask students to recall what each of the variables mean.
2. Next, ask student teams to refer back to the statistical questions they generated in the previous lesson - they will need it for the data exploration.
3. Distribute the *Exploring the DWP Data* handout (LMR_4.1). In their teams, allow students about 20-30 minutes to explore the DWP data set and complete the handout.



Name:______________________________ Date:______________

Exploring the DWP Data

Background:

The Los Angeles Department of Water and Power, also know as the DWP, wants to encourage people to conserve water. They will use the water used in 2010 as a baseline to compare water savings. The question they want to research is:

How is water used in Los Angeles?

Instructions:

1. Refer back to the statistical questions you and your team wrote in the previous lesson. Write them down on poster paper.
2. Your questions should cover each of the following three (3) types of variation:
   a. Variation in water use across time.
   b. Variation in water use across space.
   c. Variation in water use between types of buildings.
3. Next, login to RStudio and load the `dwp_2010` water dataset.
4. Then, use what you know about analyzing data to answer your teams' statistical questions. Write the answers on poster paper.
5. Continue exploring the data to inform the DWP. Write down one interesting finding on the poster paper.
6. Finally, write a brief explanation to the DWP about how water is used in Los Angeles. Can you give them any advice on how to conserve water based on these data?
7. What statistical questions cannot be answered by the DWP data?

LMR_4.1

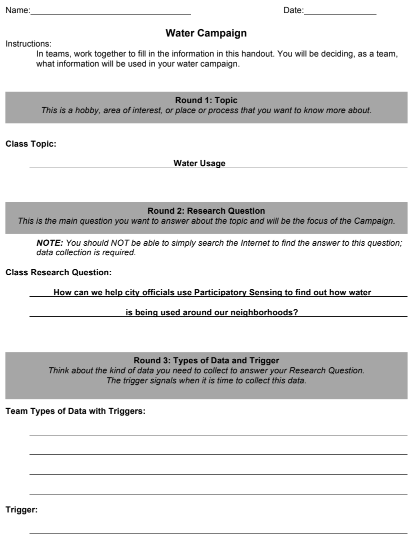
4. After students have had time to explore the DWP data, conduct a whole class discussion based on the following (answers will vary based on student teams' data exploration):
   1. What were some interesting findings?
   2. Which number statistics provided you information to help answer the research question?
   3. Which plots gave you some insights into Los Angeles' water usage?
   4. Based on your findings and by citing evidence from your analysis, what would you say about how water is being used in Los Angeles and who is using it?
5. To prepare for the creation of the Participatory Sensing campaign, ask students to discuss the following in their teams:
   a. How do you think water usage has changed since 2010?

b. What sectors do you think have changed the most?

6. Now that students have an idea about water usage in Los Angeles based on the DWP 2010 data exploration, inform them that they will create a water usage campaign. The research question fro this campaign is:

How can we save water in our neighborhood?

7. Quickly review their class campaign from Unit 3; placing emphasis on the trigger and at how the data they decided to collect answers the research question.
8. Distribute the *Water Campaign* handout (LMR_4.2). Ask students to notice that Rounds 1 and 2 are completed. Their task is to design the rest of the campaign by completing the remaining rounds.



Name:______________________________ Date:______________

Water Campaign

Instructions:

In teams, work together to fill in the information in this handout. You will be deciding, as a team, what information will be used in your water campaign.

Round 1: Topic
This is a hobby, area of interest, or place or process that you want to know more about.

Class Topic:

Water Usage

Round 2: Research Question
This is the main question you want to answer about the topic and will be the focus of the Campaign.

NOTE: *You should NOT be able to simply search the Internet to find the answer to this question; data collection is required.*

Class Research Question:

How can we help city officials use Participatory Sensing to find out how water
is being used around our neighborhoods?

Round 3: Types of Data and Trigger
Think about the kind of data you need to collect to answer your Research Question.
The trigger signals when it is time to collect this data.

Team Types of Data with Triggers:

____
____
____
____

Trigger:

____

LMR_4.2

9. Round 3: Allow student teams a reasonable amount of time to engage in a *Brainstorm,* in which they will discuss what kind of data needs to be collected in order to answer this research question and when is the best time to trigger the data collection/completion of the survey. Before they begin, ask students to keep the following question in mind: Which of these data will give us information that addresses our research question?
10. Facilitate the student teams' *Brainstorm* session by circulating around the room to check for understanding. If teams need help with deciding which data they should collect, you may ask them to ponder the following:
    a. What are some water sources?
    b. What do we use water for?
    c. Where might we see water as we walk around our neighborhoods?
    d. What would you consider wasted water?
    e. What are some uses of water that cannot be avoided?
11. Ask students to record the information from each round on poster paper - in this lesson, Rounds 1-3.
12. Inform students that they will be completing Rounds 4 and 5 during the next lesson.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 3: Evaluating and Implementing a Water Campaign

Objective:

Students will complete the design of their water usage Participatory Sensing campaign, create the campaign using the campaign authoring tool, and implement a mock campaign to evaluate the feasibility of the campaign.

Materials:

1. *Water Campaign* (LMR_4.2_Water Campaign) from previous lesson
2. Posters from previous lesson
3. Markers
4. *Campaign Authoring Instructions* handout (LMR_4.3_Campaign Authoring)

Essential Concepts: Statistical questions guide a participatory sensing campaign so that we can learn about a community or ourselves. These campaigns should be evaluated before implementing to make sure they are reasonable and ethically sound.

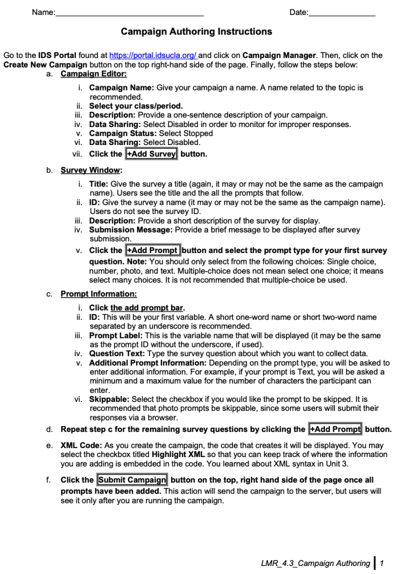
Lesson:

1. Student teams will continue designing their water usage Participatory Sensing campaign. Allow them a couple of minutes to review the information on their posters before moving on to round 4.
2. Round 4: Now that the teams have decided on a trigger and the type of data needed, they will discuss and create survey questions/prompts to ask when the trigger is set. The questions should consider all of the possible data they might collect at this trigger event.
3. Once teams have created their survey questions/prompts, they will evaluate each survey question. For each question they should consider:
    a. What type of survey question/prompt will this be (e.g. single choice, text, photo, numerical, discrete numerical, categorical, location)?
    b. How does this question/prompt help address the research question?
    c. Does the question/prompt need to be reworded? (Is it clear what is being asked for? Do they know how to answer it?) One way to do this is to pair teams and take turns asking each other prompts. The team that is being asked may explain what information they think the question is asking for.
4. If survey questions need to be rewritten, students will decide as a team on the changes.
5. Once finalized, they will record the survey question/prompt that goes along with each data variable on their *Water Campaign* handout (LMR_4.2), being cognizant of question bias.
6. Round 5: In teams, students will now generate three statistical questions that they might answer with the data they will collect and to guide their campaign. They need to make sure that their statistical questions are interesting and relevant to the water usage topic. They will record these statistical questions on their posters. Remind students that they will also have data about the date, time, and place of data collection.
7. Confirm that the questions are statistical and that they can be answered with the data the students propose to collect by circulating around the room to check on each team. Each team will decide on no more than 3 statistical questions to guide their campaign.
8. Now that they have all the pieces of the campaign, teams will evaluate whether their campaign is reasonable and ethically sound. Each team will hold a discussion on the following questions:
    a. Are answers to your survey questions likely to *vary* when the trigger occurs? (If not, you'll get bored entering the same data again and again)
    b. Can the team carry out the campaign?

c. Do triggers occur so rarely that you'll have very little data? Do they occur so often that you'll get frustrated entering too much data?
d. Ethics: Would sharing these data with strangers or friends be embarrassing or undermine someone's privacy?
e. Can you change your trigger or survey questions to improve your evaluation?
f. Will you be able to gather enough relevant data from your survey questions to be able to answer your statistical questions?

9. Students have collaboratively created their Water Usage Participatory Sensing campaign. They will now use the Campaign Authoring tool to create a campaign like the ones they see on their smart devices or the computer.

10. Distribute the *Campaign Authoring Instructions* handout (LMR_4.3). Each team will select a member to type the information required to create their campaign. Then, they will follow the instructions on the handout.



Name:______________________________ Date:_______________

Campaign Authoring Instructions

Go to the **IDS Portal** found at https://portal.idsucla.org/ and click on **Campaign Manager**. Then, click on the **Create New Campaign** button on the top right-hand side of the page. Finally, follow the steps below:

a. **Campaign Editor:**
   i. **Campaign Name:** Give your campaign a name. A name related to the topic is recommended.
   ii. **Select your class/period.**
   iii. **Description:** Provide a one-sentence description of your campaign.
   iv. **Data Sharing:** Select Disabled in order to monitor for improper responses.
   v. **Campaign Status:** Select Stopped
   vi. **Data Sharing:** Select Disabled.
   vii. **Click the +Add Survey button.**

b. **Survey Window**:
   i. **Title:** Give the survey a title (again, it may or may not be the same as the campaign name). Users see the title and the all the prompts that follow.
   ii. **ID:** Give the survey a name (it may or may not be the same as the campaign name). Users do not see the survey ID.
   iii. **Description:** Provide a short description of the survey for display.
   iv. **Submission Message:** Provide a brief message to be displayed after survey submission.
   v. **Click the +Add Prompt button and select the prompt type for your first survey question. Note:** You should only select from the following choices: Single choice, number, photo, and text. Multiple-choice does not mean select one choice; it means select many choices. It is not recommended that multiple-choice be used.

c. **Prompt Information:**
   i. **Click the add prompt bar.**
   ii. **ID:** This will be your first variable. A short one-word name or short two-word name separated by an underscore is recommended.
   iii. **Prompt Label:** This is the variable name that will be displayed (it may be the same as the prompt ID without the underscore, if used).
   iv. **Question Text:** Type the survey question about which you want to collect data.
   v. **Additional Prompt Information:** Depending on the prompt type, you will be asked to enter additional information. For example, if your prompt is Text, you will be asked a minimum and a maximum value for the number of characters the participant can enter.
   vi. **Skippable:** Select the checkbox if you would like the prompt to be skipped. It is recommended that photo prompts be skippable, since some users will submit their responses via a browser.

d. **Repeat step c for the remaining survey questions by clicking the +Add Prompt button.**

e. **XML Code:** As you create the campaign, the code that creates it will be displayed. You may select the checkbox titled **Highlight XML** so that you can keep track of where the information you are adding is embedded in the code. You learned about XML syntax in Unit 3.

f. **Click the Submit Campaign button on the top, right hand side of the page once all prompts have been added.** This action will send the campaign to the server, but users will see it only after you are running the campaign.

LMR_4.3_Campaign Authoring | *1*

LMR_4.3

11. To name their campaign, a naming convention is suggested. Otherwise, you will have multiple campaigns with the same name. For example, teams may include their team name or number in order to easily identify their campaigns.

12. Once their campaign is authored, students will save their work and make edits after they mock implement the campaign for a few days.

13. They will collect data during the mock implementation of their campaign using the information they recorded on the *Water Campaign* handout (LMR_4.2) and record the answers to the survey/prompts on paper. They will make observations about how well their campaign worked and what improvements or changes need to be made.

14. Round 6 will be completed once students have mock-implemented their campaigns.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students will collect data by mock implementing their Water Usage Participatory Sensing campaign.

Lesson 4: Refining the Water Campaign

Objective:

Students will revise their water usage Participatory Sensing campaign according to the finding from the mock implementation of their campaign to refine it. Student teams will then share their final campaigns with the rest of the class.

Materials:

1. Posters from previous lesson
2. Markers

Essential Concepts: Statistical questions guide a participatory sensing campaign so that we can learn about a community or ourselves. These campaigns should be tried before implementing to make sure they are collecting the data they are meant to collect and refined accordingly.

Lesson:

1. Student teams will come together to discuss their findings regarding the mock implementation of their campaign. Allow them time to share their findings with their team members.
2. Next, ask student teams to discuss the revisions they need to make according to their findings.
3. Once they have made a decision on the revisions, they will reflect these changes on their posters.
4. Now they will edit their campaigns. The Recorder/Reporter will type for his/her team and will login by going to the IDS Portal and clicking on Campaign Manager.
5. Then, ask students to find their team's campaign. If the campaign is not visible, they may do a **Search** by typing in their campaign name.
6. Once the campaign is found, the Recorder/Reporter will click on the drop down menu to the right of the campaign information and select **Edit Campaign**.
7. On the **Campaign Editor**, students may scroll down to see their prompts. Students may do the following actions to the prompts:
    - **Edit**: Click on the prompt's name to expand and make changes as needed. To close the prompt, they may click on the **X** that appears on the expanded prompt.
    - **Delete:** Click on the **X** that appears on the prompt's name (non-expanded prompt).
    - **Add:** Scroll down to the **+Add Prompt** button.
8. Once finished making the edits/revisions to the campaign, the Reporter/Recorder will change the **Campaign Status** to **Running** (green).
9. To run the campaign and begin collecting data via the mobile app or web browser, the Reporter/Recorder will click on **Update Campaign** on the top right hand side of the Campaign Editor.
10. Ask teams to refresh their campaigns on their smartphones or the web browser to verify that their campaign appears as one of the choices.
11. Now that they are finished with their campaigns, student teams will share out their campaigns with the rest of the class by engaging in a Gallery Walk of the posters that show their work.
12. Encourage teams to ask questions or make comments as they visit each poster.
13. Before moving to the next round (from poster to poster), ask the teams if they have questions or need clarification or would like to make a comment to the team that created the poster. Repeat until teams have visited all the posters.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students may begin collecting data by implementing their Water Usage Participatory Sensing campaign. They will have a longer period of time to collect these data—about a month—to ensure they collect a sufficient amount of data since only the members of their team will be collecting data for their campaign and not the entire class.

Lesson 5: Statistical Predictions in One Variable

Objective:

Students will devise a rule to determine how to choose a winner when predicting the typical height of all students in a large high school and measure the success of their prediction. They will consider different measures of success.

Materials:

1. *Heights of Students at a Large High School* handout (LMR_4.4_HS Student Heights)

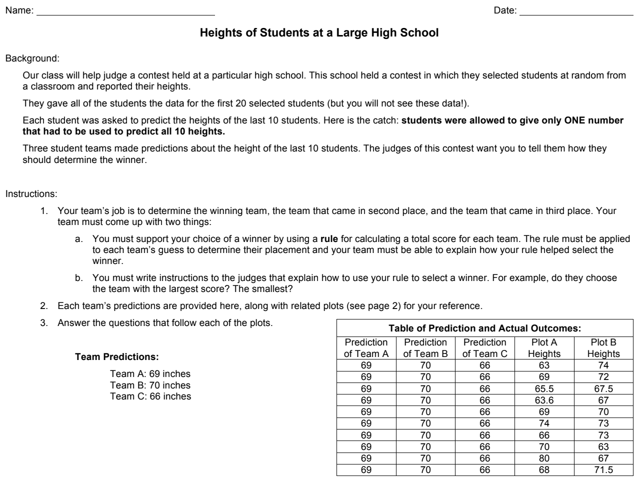
Vocabulary:

rule

Essential Concepts: Anyone can make a prediction. But statisticians measure the success of their predictions. This lesson encourages the classroom to consider different measures of success.

Lesson:

1. Inform the class that for this lesson, our class will help judge a contest held at a particular high school. This school held a contest in which they selected students at random from a classroom and reported their height.
2. The information in Steps 3 – 7 is included in the *Heights of Students at a Large High School* handout (LMR_4.4).



Name: ______________________________ Date: ____________________

Heights of Students at a Large High School

Background:

Our class will help judge a contest held at a particular high school. This school held a contest in which they selected students at random from a classroom and reported their heights.

They gave all of the students the data for the first 20 selected students (but you will not see these data!).

Each student was asked to predict the heights of the last 10 students. Here is the catch: **students were allowed to give only ONE number that had to be used to predict all 10 heights.**

Three student teams made predictions about the height of the last 10 students. The judges of this contest want you to tell them how they should determine the winner.

Instructions:

1. Your team's job is to determine the winning team, the team that came in second place, and the team that came in third place. Your team must come up with two things:
   a. You must support your choice of a winner by using a **rule** for calculating a total score for each team. The rule must be applied to each team's guess to determine their placement and your team must be able to explain how your rule helped select the winner.
   b. You must write instructions to the judges that explain how to use your rule to select a winner. For example, do they choose the team with the largest score? The smallest?
2. Each team's predictions are provided here, along with related plots (see page 2) for your reference.
3. Answer the questions that follow each of the plots.

Team Predictions:

Team A: 69 inches
Team B: 70 inches
Team C: 66 inches

Table of Prediction and Actual Outcomes:

Prediction of Team A	Prediction of Team B	Prediction of Team C	Plot A Heights	Plot B Heights
69	70	66	63	74
69	70	66	69	72
69	70	66	65.5	67.5
69	70	66	63.6	67
69	70	66	69	70
69	70	66	74	73
69	70	66	66	73
69	70	66	70	63
69	70	66	80	67
69	70	66	68	71.5

LMR_4.4

3. They gave all of the students the data for the first 20 selected students, but you will not see these data! Each student was asked to predict the heights of the last 10 students. Here is the catch: **students were allowed to give only ONE number that had to be used to predict all 10 heights.** .
4. Three student teams made predictions about the height of the last 10 students. The judges of this contest want you to tell them how they should determine the winner.
5. Your team's job is to determine the winning team, the team that came in second place, and the team that came in third place. Your team must come up with two things:
   a. You must support your choice of a winner by using a **rule** for calculating a total score for each team. The rule must be applied to each team's guess to determine their placement and your team must be able to explain how your rule helped select the winner.

b. You must write instructions to the judges that explain how to use your rule to select a winner. For example, do they choose the team with the largest score? The smallest?

6. Here are the predictions of the three teams:

   Team A: 69 inches
   Team B: 70 inches
   Team C: 66 inches

7. Display Plot A, found on page 2 of the *Heights of Students at a High School* handout (LMR_4.4). This dotplot displays the heights shown in the Actual Outcome column of the table. Inform students that this dotplot and table is provided to help them come up with a method to determine a winner. The table is to visualize the predicted heights side-by-side with the randomly selected heights.

   Notes to teacher:

   a. Students may have to be reminded that negative values with large absolute value are larger than positive values with small absolute values (e.g., 10 is larger than 3).

   b. Let students struggle for a little bit. A prompt to get them started: Look at the difference between a team's prediction and the actual outcomes (e.g., for the first height, Team A predicted 69, actual outcome was 63, so 69-63=6). They might also need to be nudged towards the *sum* of these differences – they need to produce a single score, not 10 separate scores.

   c. Here are some rules you can "feed" to the class to move them along. Ask them (a) Describe this rule in words. (b) Is it better to get a high score or a low score or some other score? (c) Which teams win for each? (Note, some of these rules produce ties).

      i. Rule 1: sum(heights-predicted.value == 0) ***words: the number of exactly correct predictions***

      ii. Rule 2: sum(heights-predicted.value) ***words: the sum of the differences between predicted value and the actual heights***

      iii. Rule 3: sum(abs(heights-estimate)) ***words: the sum of the absolute values of the deviations***

      iv. Rule 4: sum((heights-estimate)^2) ***words: the sum of the squared deviations***

      Note: It is unlikely that students will think of the last two. That's okay, because we will introduce them in a future lesson, but you might want to present one (or both) to see what they think about these rules.

8. Allow student teams time to discuss and complete the task for Plot A.

9. Do not share their responses to Plot A. Instead, display the following questions:

   a. What if we had a different set of 10 randomly selected students and plotted their heights?

   b. Would the same team win?

10. Allow teams to discuss the questions, then share a couple of responses to the questions in the previous step.

11. Display Plot B, found on page 2 of the *Heights of Students at a High School* handout (LMR_4.4), then have them find the winner using this new sample. Is it the same as they chose before?

    Note: We do NOT know the value of the true population mean/typical value. This is what we are really trying to predict.

12. Teams will take turns to share their work as follows:

    a. Which team did you select as the winner using Plot A?
    b. Explain the method, or **rule**, your team used to declare the winner.

c. Which team did you select as the winner using Plot B? Is the winner the same?
d. Did you use the same rule to select a winner or did it change? If it changed, explain.

13. During the share out, students will take notes about the other teams’ rules in their DS journals.

14. Teams may continue to share at the start of the next lesson, if they run out of time.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 6: Statistical Predictions Applying the Rule

Objective:

Students will apply the rule statisticians use to determine the best method for predicting heights for students at a high school.

Materials:

1. Each team's rule for determining a winner (from previous lesson)
2. *Prediction Games* handout (LMR_4.5_Prediction Games)

Vocabulary:

mean squared deviation, mean absolute error

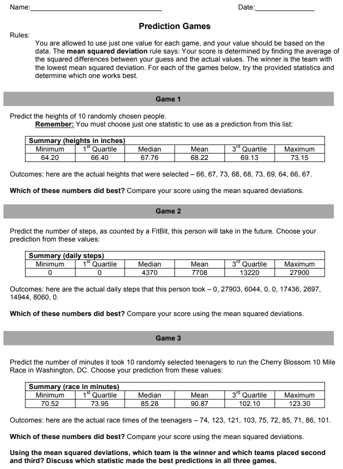
Essential Concepts: If we use the squared residuals rule, then the mean of our current data is the best prediction of future values. If we use the mean absolute error rule, then the median of the current data is the best prediction of future values

Lesson:

1. Ask students to recall that in the previous lesson, each student team created a rule to determine a winner. Which team's rules worked well for determining a winner?
2. Remind them that in their DS Journals, they took notes about each team's rule as they presented. This time, they will be switching roles – instead of creating a rule to judge the given predictions, they will be given a rule and it's their job to find the best procedure to win the contest.
3. Inform students that the question we are trying to answer is:

   How can we create a general rule that will always select the BEST guess to win no matter what 10 data points we are given?
4. Explain to students that data scientists use the "mean squared deviation" rule (also called the "mean squared error" or "mean squared residual" rule or "residual sums of squares" rule, the latter term being the most common). A "*deviation*" is the difference between our prediction and the actual outcome (as in MAD) and is sometimes called a "residual."

   Note to Teacher: Basically, students are being asked to determine which of these predicted values is "closest" to the data. One issue that comes up is dealing with positive and negative differences.
5. Distribute the *Prediction Games* handout (LMR_4.5).



Name:______________________________ Date:______________

Prediction Games

Rules:

You are allowed to use just one value for each game, and your value should be based on the data. The **mean squared deviation** rule says: Your score is determined by finding the average of the squared differences between your guess and the actual values. The winner is the team with the lowest mean squared deviation. For each of the games below, try the provided statistics and determine which one works best.

Game 1

Predict the heights of 10 randomly chosen people.

Remember: You must choose just one statistic to use as a prediction from this list:

Summary (heights in inches)					
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
64.20	66.40	67.76	68.22	69.13	73.15

Outcomes: here are the actual heights that were selected – 66, 67, 73, 68, 68, 73, 69, 64, 66, 67.

Which of these numbers did best? Compare your score using the mean squared deviations.

Game 2

Predict the number of steps, as counted by a FitBit, this person will take in the future. Choose your prediction from these values:

Summary (daily steps)					
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0	0	4370	7708	13220	27900

Outcomes: here are the actual daily steps that this person took – 0, 27903, 6044, 0, 0, 17436, 2697, 14944, 8060, 0.

Which of these numbers did best? Compare your score using the mean squared deviations.

Game 3

Predict the number of minutes it took 10 randomly selected teenagers to run the Cherry Blossom 10 Mile Race in Washington, DC. Choose your prediction from these values:

Summary (race in minutes)					
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
70.52	73.95	85.28	90.87	102.10	123.30

Outcomes: here are the actual race times of the teenagers – 74, 123, 121, 103, 75, 72, 85, 71, 86, 101.

Which of these numbers did best? Compare your score using the mean squared deviations.

Using the mean squared deviations, which team is the winner and which teams placed second and third? Discuss which statistic made the best predictions in all three games.

LMR_4.5

6. Explain the rules of the game as follows:

   You are allowed to use just one value for each game, and your value should be based on the data. The **mean squared deviation** rule says: Your score is determined by finding the average of the squared differences between your guess and the actual values. The winner is the team with the lowest mean squared deviation. For each of the games below, try the provided statistics and determine which one works best.

$$MSD = \frac{\sum_{i=1}^{n}(x_i - \hat{x})^2}{n}$$

   a. Game 1: Predict the heights of 10 randomly chosen people.
   Remember: You must choose just one statistic to use as a prediction from this list:

	Summary (heights in inches)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	64.20	66.40	67.76	68.22	69.13	73.15
MSD	**22.9**	**10.58**	**7.8056**	**7.7044**	**8.7509**	**33.1925**

	Summary (heights in inches)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	64.20	66.40	67.76	68.22	69.13	73.15
MAE	**3.94**	**2.34**	**2.1**	**2.188**	**2.578**	**5.05**

   Outcomes: here are the actual heights that were selected – 66, 67, 73, 68, 68, 73, 69, 64, 66, 67. Which of these numbers did best? Compare your score using the mean squared deviations.

   For example, using the minimum and outcomes above, gives you a mean squared deviation of:

$$\frac{\sum(66-64.20)^2+(67-64.20)^2+\ldots+(67-64.20)^2}{10}=\frac{229}{10}=22.9$$ ***square in.***

   Note to teacher: The value of the mean squared deviation will always be in square units. In order to convert back to the original units, simply take the square root of the mean squared deviation.

   Interpretation: When using the minimum height to make predictions about all heights, our predictions will typically be off by $\sqrt{22.9} = 4.79$ inches.

   b. Game 2: Predict the number of steps, as counted by a FitBit, this person will take in the future. Choose your prediction from these values:

	Summary (daily steps)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	0	0	4370	7708	13220	27900
MSD	**141,468,199**	**141,468,199**	**93,193,683**	**82,048,768**	**112,426,503**	**489,749,479**

	Summary (daily steps)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	0	0	4370	7708	13220	27900
MAE	**7708.4**	**7708.4**	**7169**	**7501.8**	**9636.2**	**20192.2**

Outcomes: here are the actual daily steps that this person took – 0, 27903, 6044, 0, 0, 17436, 2697, 14944, 8060, 0. Which of these numbers did best? Compare your score using the mean squared deviations.

Note to teacher: For Game 2, you might consider allowing students to utilize RStudio to calculate the mean squared deviation. The example below can be used to calculate the mean squared deviation for predicting daily steps using the minimum. Before revealing the codes, elicit a class discussion about how RStudio can be used to calculate the MSD.

Step 1: Create a vector of the given daily steps

```
> steps<-c(0,27903,6044,0,0,17436,2697,14944,8060,0)
```

Step 2: Store the squared deviations

```
> sqr_dev<-((steps-0)^2)
```

Step 3: Find the mean of the squared deviations

```
> mean(sqr_dev)
```

The code can be shortened to two steps if you apply a composition of the last two functions

```
> mean((steps-0)^2)
```

c. Game 3: Predict the number of minutes it took 10 randomly selected teenagers to run the Cherry Blossom 10 Mile Race in Washington, D.C.

	Summary (race in minutes)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	70.52	73.95	85.28	90.87	102.10	123.30
MSD	**777.0264**	**647.6125**	**387.3624**	**353.5429**	**474.49**	**1390.33**

	Summary (race in minutes)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	70.52	73.95	85.28	90.87	102.10	123.30
MAE	**20.58**	**18.13**	**15.7**	**16.674**	**19.14**	**32.2**

Outcomes: here are the actual race times of the teenagers – 74, 123, 121, 103, 75, 72, 85, 71, 86, 101. Which of these numbers did best? Compare your score using the mean squared deviations.

7. Using the mean squared deviations, which statistic is the winner and which statistics placed second and third? Discuss which statistic made the best predictions in all three games.

Note to teacher: Explain that the mean worked best for all three contests. Data scientists (and mathematicians) can prove that the mean will **always** work best (except in a few weird cases from time to time). So if you want to predict the future, the mean is the best single guess you can make.

8. Ask: What if another data science class has a best rule that is different from ours?
9. Another agreed upon method that data scientists and statisticians often use is the **mean absolute error**. It's unlikely that students will figure this out on their own. The reasons why we do it in statistics date back to the 18th century, so it won't make a lot of sense; but it's what statisticians do. The mean absolute error is expressed as (where $\hat{x}$ stands for the predicted value):

$$MAE = \frac{\sum_{i=1}^{n} |x_i - \hat{x}|}{n}$$

10. Explain that each team will now use the statisticians' method for declaring a winner. Display the mean absolute error formula and discuss what each symbol means.
11. Using our previous examples, recalculate your predictions using the MAE.
12. Using the mean absolute error, which statistic is the winner and which statistics placed second and third?

Answers:

	Summary (heights in inches)					
	Minimum	*1st Quartile*	*Median*	*Mean*	*3rd Quartile*	*Maximum*
	64.20	*66.40*	*67.76*	*68.22*	*69.13*	*73.15*
MAE	*3.94*	*2.34*	*2.1*	*2.188*	*2.578*	*5.05*

	Summary (daily steps)					
	Minimum	*1st Quartile*	*Median*	*Mean*	*3rd Quartile*	*Maximum*
	0	*0*	*4370*	*7708*	*13220*	*27900*
MAE	*7708.4*	*7708.4*	*7169*	*7501.8*	*9636.2*	*20192.2*

	Summary (race in minutes)					
	Minimum	*1st Quartile*	*Median*	*Mean*	*3rd Quartile*	*Maximum*
	70.52	*73.95*	*85.28*	*90.87*	*102.10*	*123.30*
MAE	*20.58*	*18.13*	*15.7*	*16.674*	*19.14*	*32.2*

Note to teacher: Explain that in this instance, the median is the "winner." This means that the way you play the game depends on the rules of the game. If we used squared deviations, play with the mean. If we use the mean absolute error (MAE), play with the median.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 7: Statistical Predictions Using Two Variables

Objective:

Students will learn how to predict height using arm span data - and vice versa - visually on a scatterplot.

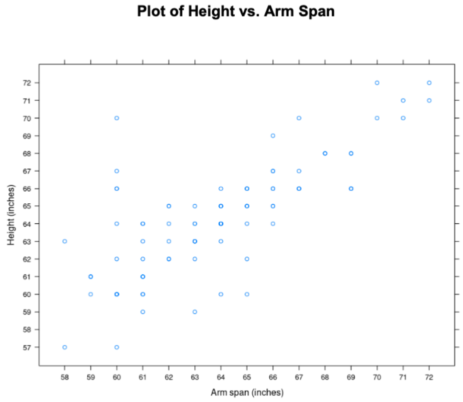
Materials:

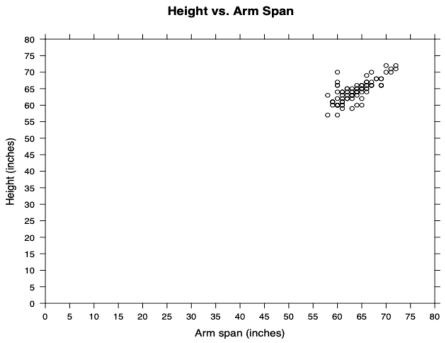
1. *Arm span vs. Height Scatterplot* (LMR_4.6_Arm Span vs Height)
   Note: This handout will be referenced in subsequent lessons.
2. Assorted color markers (dry erase or overhead)—See step 3 of lesson.
3. Overhead or LCD projector

Essential Concepts: When predicting values of a variable *y* - and if *y* is associated with *x* - then we can get improved predictions by using our knowledge about *x*. We essentially "subset" the data for a given value of *x* and use the mean *y* for those subset values. If the resulting means follow a trend, we can model this trend to generalize to as-yet unseen values of *x*.

Lesson:

1. Remind students that in the previous lessons they were working with height data to predict the typical height of all the students at a large high school, implementing a method used by statisticians to help them make good predictions.
2. In addition to the height data, it turns out that each student's arm span data was also collected and recorded.
3. Display the *Arm Span vs. Height Scatterplot* (LMR_4.6) on a white board or overhead projector (you will write on the board or the transparency later in the lesson—see step 9).





LMR_4.6

4. Distribute the *Arm Span vs. Height* handout (LMR_4.6). Students will refer to this handout again later in a subsequent lesson.

5. In teams, ask students to analyze the plot and discuss the following questions:

   - What kind of plot is this? ***Scatterplot.***
   - How many variables are displayed in this plot? ***Two variables.***
   - Which variable is shown on the x-axis? On the y-axis? ***Arm span is shown on the x-axis and height is shown on the y-axis.***
   - What is this plot showing? ***It is showing the relationship between a person's height and the person's corresponding arm span measurement.***
   - How can I find out the height of the person whose arm span measures 68 inches? ***Find 68 on the x-axis. Then find the data point located at 68. Place finger on the data point and track its location on the y-axis. The height is also 68 inches.***

6. Using Talk Moves, conduct a class discussion of the questions in step 5.

7. Remind students that we've learned that the mean is the best way of predicting heights. The mean heights of these people is 64 inches.

8. Ask the students: Do you think we can do better? Is 64 a good prediction for someone whose arm span is 72"? What about 60"? How can you come up with a rule for determining the best predicted height *if you know the person's arm span*?

   Note to teacher: Lead students to realize that they can do this by "subsetting" the data for the fixed *x* value. For example, if arm span is 60, they should consider only the heights of people whose arm span is 60 and find the mean.

9. In teams, ask students to approximate the mean height for people whose arm span is 60, 64, 68, and 72.

   Note: Because the plot does not clearly show duplicate ordered pairs, an approximation is sufficient at this point. You may have students use RStudio to calculate the mean height for the specific armspans. Refer to the OPTIONAL section at the end of this lesson.

10. Then plot these points on the graph. We'll use this later – the points should be roughly along a straight line. ***These arm spans have a range of height values associated with them. Students may take a mean of the heights, but answers may vary.***

11. Ask students if they see any patterns or rules they can use from this to help with predictions. Because there were multiple height values associated with each arm span length, you will likely get multiple answers from students. The goal now is to come up with a rule that suggests a plausible height value for anyone with a particular arm span.

12. A sentence starter to guide students: If a person has a bigger arm span, then we should predict____ [a bigger height]. If time permits, you might push them to be more precise. Let's take someone who has a 60 inch arm span. You predicted a height of ____. How much should we increase our prediction for people with a 62 inch arm span? Can you do this without subsetting the data and re-calculating?

13. Conceptually, students are wrestling with the notion of the slope of the regression line but there's no need to point this out just yet. Important: The equation of the line of best fit will be revealed in lesson 10.

    OPTIONAL FOR ITEM 9 If you want to obtain the exact mean height for each arm span value in step 9, copy the code below and run it in an RScript.

```
xyplot(height~armspan, data = arm_span,
       scales = list(x = list(at = seq(58, 72, 1)),
       y = list(at = seq(52, 72, 1))),
       xlab = "Arm span (inches)", ylab = "Height (inches)")

armspan_60 <- filter(arm_span, armspan==60)
```

```
mean(~height, data = armspan_60)
#62.66667

armspan_64 <- filter(arm_span, armspan==64)
mean(~height, data = armspan_64)
#64

armspan_68 <- filter(arm_span, armspan==68)
mean(~height, data = armspan_68)
#68

armspan_72 <- filter(arm_span, armspan==72)
mean(~height, data = armspan_72)
#71.5

#Base R Code
#syntax to create a scatterplot using base R
plot(arm_span$height, arm_span$armspan)

#Points function in base R is more user friendly
points(60, 62.66667, col = "red", cex = 2)
points(64, 64, col = "red", cex = 2)
points(68, 68, col = "red", cex = 2)
points(72, 71.5, col = "red", cex = 2)
```

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next 2 Days

LAB 4A: If the Line Fits…

LAB 4B: What's the Score?

Complete Labs 4A and 4B prior to Lesson 8.

Lab 4A - If the line fits ...

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

How to make predictions

- Anyone can make predictions.
  - Data scientists use data to inform their predictions by using the information learned from the sample to make predictions for the whole population.
- In this lab, we'll learn how to make predictions by finding the *line of best-fit*.
  - You will also learn how to use the information from one variable to make predictions about another variable.

Predicting heights

- Use the `data()` function to load the `arm_span` data.
- This data comes from a sample of 90 people in the Los Angeles area.
  - The measurements of `height` and `armspan` are in inches.
  - A person's `armspan` is the maximum distance between their fingertips when they spread their arms out wide.
- Make a plot of the `height` variable.
  - **If you had to predict the height of someone in the LA area, what single height would you choose and why?**
  - **Would you describe this as a *good* guess? What might you try to improve your predictions?**

Predicting heights knowing arm spans

- Create two subsets of our `arm_span` data:
  - One for `armspan >= 61 & armspan <= 63`.
  - A second for `armspan >= 64 & armspan <= 66`.
- Create a histogram for the `height` of people in each subset. Answer the following based on the data:
  - **What `height` would you predict if you knew a person had an `armspan` around 62 inches?**
  - **What `height` would you predict if you knew a person had an armspan around 65 inches?**
  - **Does knowing someone's `armspan` help you predict their height. Why or why not?**

Fitting lines

- Notice that there is a trend that people with a larger `armspan` also tend to have a larger mean `height`.
  - One way of describing this sort of trend is with a line.
- Data scientists often *fit* lines to their data to make predictions.
  - What we mean by *fit* is to come up with a line that's close to as many of the data points as possible.
- Create a scatterplot for `height` and `armspan`. Then run the following code. Draw a line by clicking twice on the *Plot* pane.

```
add_line()
```

Predicting with lines

- Draw a line that you think is a good *fit* and write down its equation. Using this equation:
  - **Predict how tall a person with a 62 and a person with a 65 inch `armspan` would be.**
- Using a line to make predictions also lets us make predictions for `armspans` that aren't in our data.
  - **How tall would you predict a person with a 63.5 inch `armspan` to be?**
- **Compare your answers with a neighbor's. Did both of you come up with the same equation for a line? If not, can you tell which line fits the data best?**

Regression lines

- If you were to go around your class, each student would have created a different line that they feel *fit* the data best.
  - Which is a problem because everyone's line will make slightly different predictions.
- To avoid this variation in predictions, data scientists will use *regression lines*.
  - This line connects the mean `height` of people with similar `arm_spans`.
  - Fill in the blanks below to create the a *regression line* using an `lm`, or *linear model*:

```
lm(____ ~ ____, data = arm_span)
```

Predicting with regression lines

- Use the output of the code from the previous slide to write down the equation of the *regression line* in the form

  `y = a + bx`.
- Add this line to a scatterplot by filling in the blanks below:

```
add_line(intercept = ____, slope = ____)
```

- Predict the height of a person with a 63.5 inch `armspan` and compare it with a neighbor. Ensure you both arrive at the same predicted value.
- **Measure your `armspan` and use the regression line to predict your height. How close was the prediction?**

Lab 4B - What's the score?

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Previously

- In the previous lab, we learned we could make predictions about one variable by utilizing the information of another.
- In this lab, we will learn how to measure the accuracy of our predictions.
  - This in turn will let us evaluate how well a model performs at making predictions.
  - We'll also use this information later to compare different models to find which model makes the best predictions.

Predictions using a line

- Load the `arm_span` data again.
  - Create an `xyplot` with `height` on the y-axis and `armspan` on the x-axis.
  - Type `add_line()` to run the `add_line` function; you'll be prompted to click twice in the plot window to create a line that you think fits the data well.
- Fill in the blanks below to create a function that will make predictions of people's `heights` based on their `armspan`:

```
predict_height <- function(armspan) {
  ____ * armspan + ____
}
```

Make your predictions

- Fill in the blanks to include your predictions in the `arm_span` data.

```
____ <- mutate(____, predicted_height = ____(____))
```

- Now that we've made our predictions, we'll need to figure out a way to decide how accurate our predictions are.
  - We'll want to compare our *predicted heights* to the *actual heights*.
  - At the end, we'll want to come up with a single number summary that describes our model's accuracy.

Sums of differences

- A residual is the difference between the actual and predicted value of a quantity of interest.
- Fill in the blanks below to create a function which calculates the sum of differences:
- **What do residuals measure?**

```
____ <- mutate(____, residual = ____(____))
```

- **What do residuals measure?**
- One method we might consider to measure our model’s accuracy is to sum the residuals.
- Fill in the blanks below to calculate our accuracy summary.

```
summarize(____, sum(____))
```

- Hint: Like `mutate`, the firs argument of `summarize` is a dataframe, and the second argument is the action to perform on a column of the dataframe. Whereas the output of `mutate` is a column, the output of `summarize` is (usually) a single number summary.

Checking our work

- **Describe and interpret, in words, what the output of your accuracy summary means.**
  - **Compare your accuracy summary with a neighbor's. Whose line was more accurate and why?**
- **Write down why adding positive and negative errors together is problematic for assessing prediction accuracy.**
  - **Why does calculating the squared values for the differences solve this problem?**
- The *mean squared error* (MSE) is calculated by squaring all of the residuals, and then taking the mean of the squared residuals.
- Fill in the blanks below to calculate the MSE of your line.

```
summarize(____, mean((____)^2)
```

On your own

- Create a *regression line* as you did in the previous lab, for `height` and `armspan`.
  - We also refer to *regression lines* as *linear models*.
  - Assign this model the name `best_fit`.
- Making predictions with models `R` is familiar with is simpler than with lines, or models, we come up with ourselves.
  - Fill in the blanks to make predictions using `best_fit`:

```
____ <- mutate(____, predicted_height = predict(____))
```

- Hint: the `predict` function takes a linear model as input, and outputs the predictions of that model.
- Calculate the MSE for these new predicted values.

The magic of lm()

- The `lm()` function creates the *line of best fit* equation by finding the line that minimizes the *mean squared error*. Meaning, it's the *best fitting line possible*.
  - Compare the MSE value you calculated using the line you fitted with `add_line()` to the same value you calculated using the `lm` function.
  - Ask your neighbors if any of their lines beat the `lm` line in terms of the MSE. Were any of them successful?
- To see how the `lm` line fits your data, create a scatterplot and then run:

```
add_line(intercept = ____, slope = ____)
```

<u>Lesson 8: What's the Trend?</u>

Objective:

Students will understand that the regression line is a model for a linear association (trend). They will learn to identify the direction and strength of trends.

Materials:

1. *What's the Trend?* handout (LMR_4.7_What's the Trend)
   <u>Note:</u> This handout will be referenced and used in subsequent lessons.
2. *Strength of Association* handout (LMR_4.8_Strength of Association)

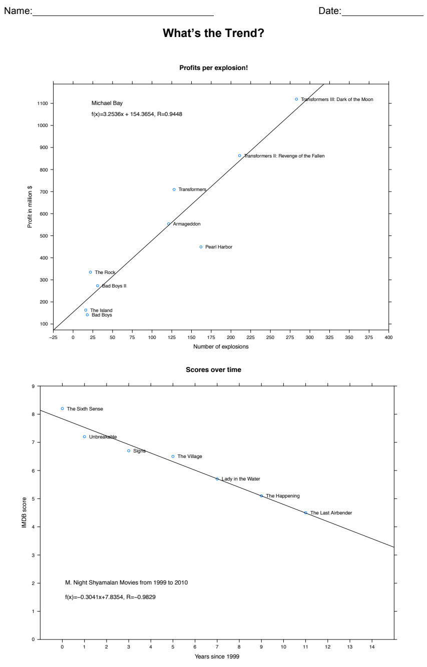
Vocabulary:

trend, positive association, negative association, no association, shape, linear, model, strength of association

Essential Concepts: Associations are important because they help us make better predictions; the stronger the trend, the better the prediction we can make. "Better" in this case means that our mean squared residuals can be made smaller.

Lesson:

1. Distribute *What's the Trend?* (LMR_4.7). Students will analyze the two scatterplots on the handout. The *Profits per Explosion* plot shows the relationship between the number of explosions in Michael Bay's movies and the profit earned by each movie. The *Scores Over Time* plot shows the relationship between M. Night Shyamalan movies made since *The Sixth Sense* was released in 1999 and their Internet Movie Database (IMBD) scores.



Name:______________________________ Date:______________

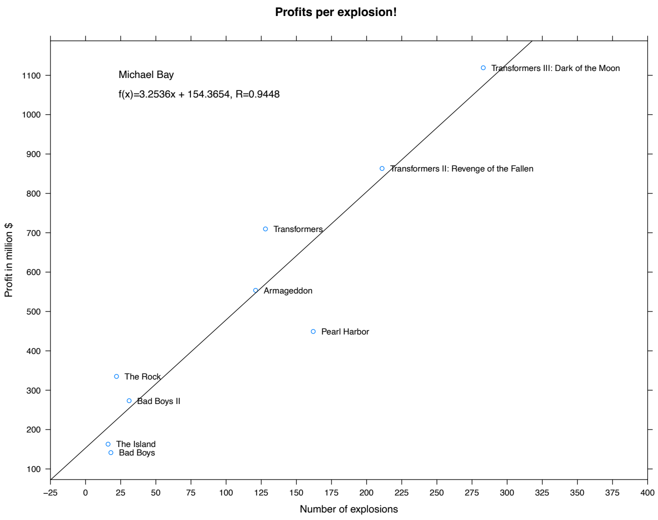
What's the Trend?

Profits per explosion!

Scores over time

LMR_4.7

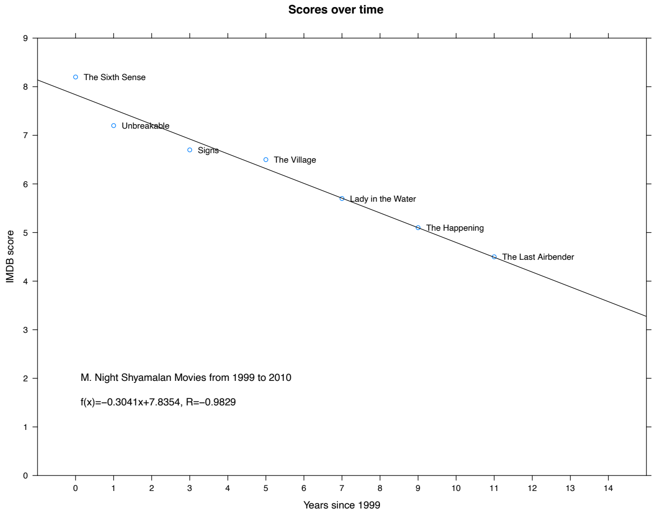
2. In teams, students will discuss and record their responses to the following questions for each plot:



a. What kind of plot is this? ***Scatterplot.***

b. What do the numbers on the x-axis represent? What do the numbers on the y-axis represent? ***The x-axis shows number of explosions and y-axis shows profit in millions of dollars.***

c. What is this plot telling us? ***Answers will vary. One example could be that if there are more explosions in a movie, then the movie will earn a greater profit.***



d. What kind of plot is this? ***Scatterplot.***

e. What do the numbers on the x-axis represent? What do the numbers on the y-axis represent? ***The x-axis shows the number of years since 1999 and the y-axis shows the movie's IMDB score.***

f. What is this plot telling us? ***Answers will vary. One example could be that as M. Night Shyamalan has produced more movies, their IMDB ratings have gone down.***

3. Allow students time to discuss and record their answers to the questions.

4. Display both plots, if possible (students may also refer to the plots in their own handout). Discuss the following questions with the whole class:

- What is happening in each plot? What seems to be the trend? ***Guide students to understand that the Profits per Explosion plot shows an increasing trend, while the Scores Over Time plot shows a decreasing trend. An increasing trend is called a positive association and a decreasing trend is called a negative association.***

- What does it mean to have an increasing trend and a positive association? ***In Profits per Explosion, it means that as the number of explosions increase, the movie profits also increase.***
- What does it mean to have a decreasing trend and a negative association? ***In Scores Over Time, it means that as the years after 1999 pass, the movie IMBD ratings decrease.***

5. Quickwrite: What if we had a plot with **no association**? Ask students to sketch what they think a scatterplot that shows no association looks like. ***A correct sketch will show a scatterplot with data points that show no positive or negative association; no trend or pattern. There would be no association or a very weak one. The data would be scattered.***

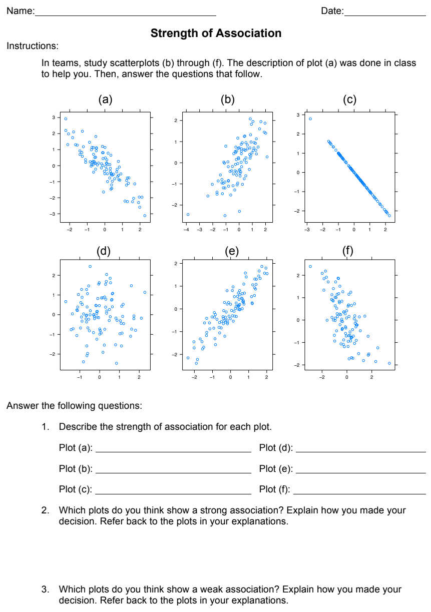
6. Select a couple of sketches to share with the whole class. Discuss why the sketches show no association.

7. Ask students to discuss their thoughts about why a line was drawn through the points of the two plots and why there are equations for each plot.

8. Conduct a share out of their observations. Guide students to the understanding that the **shape** of both plots is **linear**. The line represents a *model* for the relationship between two variables. The equation shown in the plots above represents the line through the points. It provides a description of the data and the relationship between the variables.

9. Distribute *Strength of Association* (LMR_4.8_Strength of Association). In teams, students will examine the scatterplots (b) through (e). Their task is to discuss the **strength of the association** for each plot. They will determine which plots they think show strong associations and which ones they believe show weak associations. They must explain how they made their decision. Reasons must reference the plots.

10. As an example, demonstrate how to describe plot (a) in the *Strength of Association* handout. ***Possible description: Plot (a) shows a negative association, or decreasing trend. The association appears to be fairly strong because the points are relatively close together, forming a moderate linear pattern***.



Name:______________________________ Date:______________

Strength of Association

Instructions:

In teams, study scatterplots (b) through (f). The description of plot (a) was done in class to help you. Then, answer the questions that follow.

Answer the following questions:

1. Describe the strength of association for each plot.

   Plot (a): ________________________ Plot (d): ________________________

   Plot (b): ________________________ Plot (e): ________________________

   Plot (c): ________________________ Plot (f): ________________________

2. Which plots do you think show a strong association? Explain how you made your decision. Refer back to the plots in your explanations.

3. Which plots do you think show a weak association? Explain how you made your decision. Refer back to the plots in your explanations.

LMR_4.8

11. Once all teams have completed the handout, assign one plot to each team for a share out. If two teams have the same plot, one team will share its explanation first and the second team can agree, disagree, or add to the first team's description.

12. Guide students to understand that a strong association has points closer to each other and a weak association has points more scattered.
13. If students run out of time, they will complete the remainder of the activity for homework.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

☑ Complete the *Strength of Association* handout (LMR_4.8_Strength of Association).

Lesson 9: The Spaghetti Line

Objective:

Students will estimate the line of best fit for a height and arm span data set using a strand of spaghetti as a modeling tool.

Materials:

1. *The Spaghetti Line* (LMR_4.9_The Spaghetti Line)
   Note: Advance preparation is required. Cut out plots prior to beginning the lesson.

2. *What's the Trend?* (LMR_4.7_What's the Trend) from lesson 8
3. *Arm Span vs. Height Scatterplot* (LMR_4.6_Arm Span vs Height) from Lesson 7
4. 1 lb. of Uncooked Spaghetti
5. Grid Paper
6. Tape or Glue
7. Poster paper

Vocabulary:

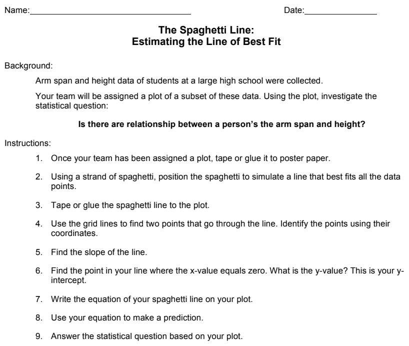
line of best fit, regression line

Essential Concepts: We can often use a straight line to summarize a trend. "Eye balling" a straight line to a scatterplot is one way to do this.

Lesson:

1. If necessary, begin by sharing out the descriptions for the plots in the *Strength of Association* (LMR_4.8_Strength of Association) handout from the previous lesson.
2. Inform students that in this lesson, they will estimate the equation of the **line of best fit** for a height and arm span data set.
3. Refer students back to the plots in the *What's the Trend?* handout (LMR_4.7_What's the Trend). The line in each of the plots is known as the **line of best fit**, or the **regression line**. This is a trend line that best represents or models the data in each scatterplot. Ask students:
   a. Why do you think this line is called "best fit"? ***Some possible answers are that it is a line that is closest to all data points or that it "fits" evenly among the data points. This is a good time to refer back to the discussion about height versus arm span in lesson 7.***
4. Distribute *The Spaghetti Line* (LMR_4.9_The Spaghetti Line) to each student and a couple of spaghetti strands per team. Students will estimate the line of best fit as outlined in the handout. Team solutions should be recorded on poster paper. They will glue their assigned plot on the poster and record their responses to the questions on the poster paper.

   Note to teacher: If necessary, review how to find the slope of a line using two points and how to write an equation using the slope and y-intercept.



Name:______________________________ Date:______________

The Spaghetti Line:
Estimating the Line of Best Fit

Background:

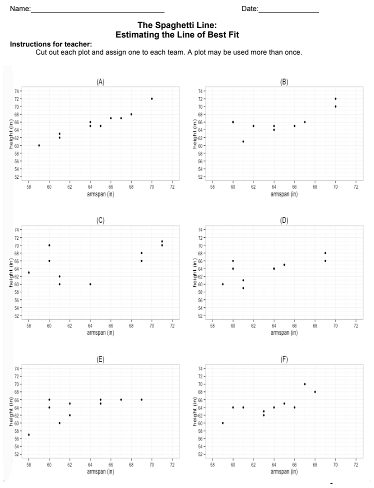
Arm span and height data of students at a large high school were collected.

Your team will be assigned a plot of a subset of these data. Using the plot, investigate the statistical question:

Is there are relationship between a person's the arm span and height?

Instructions:

1. Once your team has been assigned a plot, tape or glue it to poster paper.
2. Using a strand of spaghetti, position the spaghetti to simulate a line that best fits all the data points.
3. Tape or glue the spaghetti line to the plot.
4. Use the grid lines to find two points that go through the line. Identify the points using their coordinates.
5. Find the slope of the line.
6. Find the point in your line where the x-value equals zero. What is the y-value? This is your y-intercept.
7. Write the equation of your spaghetti line on your plot.
8. Use your equation to make a prediction.
9. Answer the statistical question based on your plot.



Name:______________________________ Date:______________

The Spaghetti Line:
Estimating the Line of Best Fit

Instructions for teacher:

Cut out each plot and assign one to each team. A plot may be used more than once.

LMR_4.9

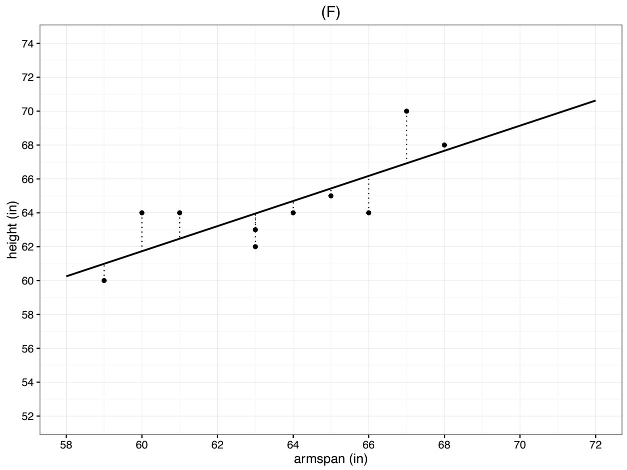
5. Ask teams to post their work around the room. Conduct a *Gallery Walk* so that teams can see each other's work.

6. Lead a discussion about the teams' lines. Ask: Which team has the best line? Why?

   Note to teacher: Push the students a bit by adding an obviously bad line to the graph and asking why their line is better than this one. Push them to come to an understanding that the "best" line comes close to the *most* points.

7. Inform students that data scientists have a way of finding the best line. They choose the line so that the mean squared distances between the points and the line is as small as possible. Discuss with students:
   a. What methods have we used so far? ***We've used Mean Squared Deviations and Mean Absolute Error (Lesson 6).***
   b. How did we use these methods? ***It was best to use Mean Squared Deviations when we are looking at mean and Mean Absolute Error when we are looking at median.***
   c. Which method do you think data scientists use most often? ***Data scientists often use MAE.***

8. [See graphic below] If time permits, ask students to calculate the distances and squares of two different lines so that they can understand what it means. This is the 2D version of the game they played in Lesson 6.



9. Inform students that they will see the equation of the arm span vs. height data in lesson 10.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

Students will use a straight edge to draw a line of best fit for the scatter plot in the *Arm Span vs. Height* handout (LMR_4.6_Arm Span vs. Height) from lesson 7. They will use their knowledge of slope and y-intercept to determine the equation for the line of best fit that they drew.

LAB 4C: Cross-Validation

Complete Lab 4C prior to Lesson 10.

Lab 4C - Cross-Validation

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Predictions

- In the previous lab, we learned how to:
  - Create a linear model predicting `height` from the `arm_span` data (4A).
  - See how well our model predicts `height` on the `arm_span` data by computing mean squared error (MSE) (4B).
- In this lab, we will see how our model predicts heights of people *we haven't yet measured*.
- To do this, we will use a method called *cross-validation*.
- Cross-validation consists of three steps:
  - Step 1: Split the data into *training* and *test* sets.
  - Step 2: Create a model using the *training* set.
  - Step 3: Use this model to make predictions on the *test* set.

Step 1: train-test split

- Waiting for new observations can take a long time. The U.S. takes a census of its population once every 10 years, for example.
- Instead of waiting for new observations, data scientists will take their current data and divide it into two distinct sets.
- Split the `arm_span` data into `training` and `testing` data sets using the following steps.
- First, fill in the blanks below to randomly select which rows of `arm_span` will go into the `training` set.

```
set.seed(123)
train_rows <- sample(1:____, size = 85)
```

- Second, use the `slice` function to create two dataframes: one called `train` consisting of the `train_rows`, and another called `test` consisting of the remaining rows of `arm_span`.

```
train <- slice(arm_span, ____)
test <- slice(____, - ____)
```

- **Explain these lines of code and describe the `train` and `test` data sets.**

Aside: set.seed()

- When we split data, we're randomly separating our observations into *training* and *testing* sets.
  - It's important to notice that no single observation will be placed in both sets.
- Because we're splitting the data sets randomly, our models can also vary slightly, person-to-person.
  - This is why it's important to use `set.seed`.
- By using `set.seed`, we're able to reproduce the random splitting so that each person's model outputs the same results.
  Whenever you split data into training and testing, always use `set.seed` first.

Aside: train-test ratio

- When splitting data into *training* and *testing* sets, we need to have enough observations in our data so that we can build a good model.
  - This is why we kept 85 observations in our `training` data.
- As data sets grow larger, we can use a larger proportion of the data to *test* with.

Step 2: train the model

- Step 2 is to create a linear model relating `height` and `armspan` using the `training` data.
- Fit a line of best fit model to our `training` data and assign it the name `best_train`.
- Recall that the slope and intercept of our linear model are chosen to minimize MSE.
- Since the MSE being minimized is from the training data, we can call it *training MSE*.

Step 3: test the model

- Step 3 is to use the model we built on the `training` data to make predictions on the `test` data.
- Note that we are NOT recomputing the slope and intercept to fit the test data best. We use the same slope and intercept that were computed in step 2.
- Because we're using the *line of best fit*, we can use the `predict()` function we introduced in the last lab to make predictions.
  - Fill in the blanks below to add predicted heights to our `test` data:

```
test <- mutate(test, ____ = predict(best_train, newdata = ____))
```

- Hint: the `predict` function without the argument `newdata` will output predictions on the `training` data. To output predictions on the `test` data, supply the `test` data to the `newdata` argument.
- **Calculate the MSE in the same way as you did in the previous lab (test MSE is simply MSE of the predictions on the test data).**

Recap

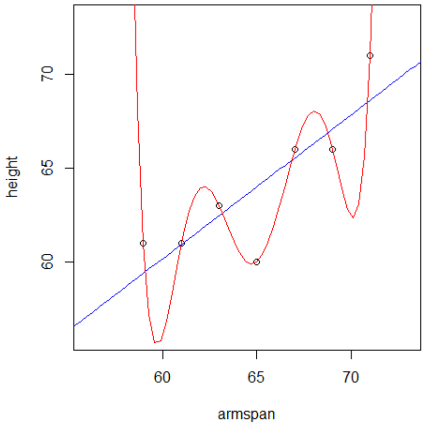
- Another way to describe the three steps is
- Step 1: Split the data into `training` and `test` sets.
- Step 2: Choose a slope and intercept that minimize training MSE.
- Step 3: Using the same slope and intercept from step 2, make predictions on the `test` set, and use those predictions to compute test MSE.
- This begs the question, why do we care about test MSE?

Why cross-validate?

- Why go to all this trouble to compute test MSE when we could just compute MSE on the original dataset?
- When we compute MSE on the original dataset, we are measuring the ability of a model to make predictions on *the current batch of data*.
- Relying on a single dataset can lead to models that are so specific to the current batch of data that they're unable to make good predictions for future observations.
  - This phenomenon is known as *overfitting*.
- By splitting the data into a training and test set, we are *hiding a proportion of the data* from the model. This emulates future observations, which are unseen.
- Test MSE estimates the ability of a model to make predictions of *future observations*.

Example of overfitting

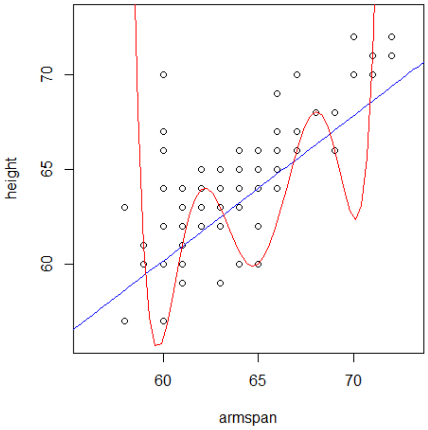
- The following example motivates cross-validation by illustrating the dangers of overfitting.
- We randomly select 7 points from the `arm_span` dataset and fit two models: a linear model, and a polynomial model.
    - You will learn how to fit a polynomial model in lab 4F.
- Below is a plot of these 7 `training` points, and two cures representing the value of height each model would predict given a value of armspan.



- **Which model does a better job of predicting the 7 `training` points?**
- **Which model do you think will do a better job of predicting the rest of the data?**

Example of overfitting, continued

- Below is a plot of the rest of the `arm_span` dataset, along with the predictions each model would make.



- **Which model does a better job of generalizing to the rest of the `arm_span` dataset?**

Lesson 10: Predicting Values

Objective:

Students will learn how to make predictions based on linear models.

Materials:

1. *What's the Trend?* handout (LMR_4.7_What's the Trend) from lesson 8
2. *Predicting Values* handout (LMR_4.10_Predicting Values)

Vocabulary:

observed value, predicted value

Essential Concepts: The regression line can be used to make good predictions about values of *y* for any given value of *x*. This works for exactly the same reason the mean works well for one variable: the predictions will make your score on the mean squared residuals as small as possible.

Lesson:

1. Entrance ticket: How do you find the equation of a line using two points?

   Note to Teacher: Students may share their responses with a partner or you may choose to use this ticket as an assessment.

2. Reveal the equation of the line of best fit for the Arm Span vs. Height data and ask students to check their equations from the homework assignment:

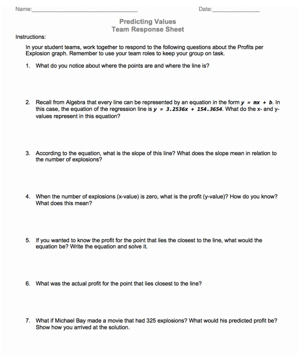
$$\widehat{\text{height}} = 0.7328(\text{armspan}) + 17.4957$$

   Note: Any time a *hat* is on top of a variable, this means we are making "predicted values" of that variable.

3. Whose equation came closest to the equation of the regression line? Ask the student whose equation came closest to share how he/she came up with the equation.
4. Inform students that the equation of the line is a rule that predicts the height based on a second variable, in this case, arm span.
5. Team discussion question:

   Using the equation of the line of best fit provided, how can we predict the height of a student whose arm span is 67 inches?

6. Remind students that lines of best fit are also known as regression lines and they are models that can be used to make predictions. Today, they will explore more about this line.
7. Ask student teams to refer back to *What's the Trend?* Handout (LMR_4.7). They should discuss the following questions and record their responses on the *Predicting Values* handout (LMR_4.10):



Name:____________________ Date:__________

Predicting Values
Team Response Sheet

Instructions:

In your student teams, work together to respond to the following questions about the Profits per Explosion graph. Remember to use your team roles to keep your group on task.

1. What do you notice about where the points are and where the line is?
2. Recall from Algebra that every line can be represented by an equation in the form $y = mx + b$. In this case, the equation of the regression line is $y = 3.2536x + 154.3654$. What do the x- and y-values represent in this equation?
3. According to the equation, what is the slope of this line? What does the slope mean in relation to the number of explosions?
4. When the number of explosions (x-value) is zero, what is the profit (y-value)? How do you know? What does this mean?
5. If you wanted to know the profit for the point that lies the closest to the line, what would the equation be? Write the equation and solve it.
6. What was the actual profit for the point that lies closest to the line?
7. What if Michael Bay made a movie that had 325 explosions? What would his predicted profit be? Show how you arrived at the solution.

LMR_4.10

a. What do you notice about where the points are and where the line is? ***Some points are near the line, others are further away, and one point is exactly on the line. Data points are*** **observed values** ***and points on the line are*** **predicted values.**

b. Recall from Algebra that every line can be represented by an equation in the form $y = mx + b$. In this case, the equation of the regression line is $y = 3.2536x + 154.3654$. What do the x- and y-values represent in this equation? ***The x-values represent the number of explosions and the y-values represent the predicted profit.***

c. According to the equation, what is the slope of this line? What does the slope mean in relation to the number of explosions? ***The slope is 3.2536. It is the rate of change between the number of explosions and the profit. It means that for every explosion increase the profit increases by 3.2536 dollars.***

d. When the number of explosions (x-value) is zero, what is the profit (y-value)? How do you know? What does this mean? ***The profit is 154.3654 million dollars. Students may use the equation to show that they substituted zero for x, so the y-intercept is the profit. It means that if Michael Bay were to make a movie with NO explosions, this would be his projected profit.***

e. If you wanted to know the profit for the point that lies the closest to the line, what would the equation be? Write the equation and solve it. ***Profit=3.2536(211)+154.3654. Profit=840.875 or 840,875,000 million dollars.***

f. What was the actual profit for the point that lies closest to the line? ***The actual profit was 836,303,693 million dollars.***

g. What if Michael Bay made a movie that had 325 explosions? What would his predicted profit be? Show how you arrived at the solution. ***By substituting 325 in the value of x in the equation, predicted profit will be $1,211,785,400 or $ 1,211.7854, or by finding the point on the line or both.***

8. Assign one question to each team for a share out. If two teams have the same question, one team will share its explanation first and the second team can agree, disagree, or add to the first team's explanation.

   Note: If students ask/wonder about the meaning of the R^2, inform them that it is related to R, also known as the correlation coefficient. They will learn about R (not R^2)in lesson 11.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students will answer the following questions about the *Scores Over Time* plot (LMR_4.7):

- What do you notice about where the points are and where the line is?
- What do the y-and x-values represent in this equation?
- According to the equation, what is the slope of this line? What does the slope mean?
- When the x-value is zero, what is the y-value? How do you know? What does this mean?
- What would the predicted value of the score be if M. Night Shyamalan released a movie in 2015? How do you know?

Lesson 11: How Strong Is It?

Objective:

Students will learn that the correlation coefficient is a value that measures the strength in linear associations only.

Materials:

1. *Correlation Coefficient* handout (LMR_4.11_Correlation Coefficient)
   Note: Advance preparation required. This handout is the resource for the plot cutouts. DO NOT distribute as-is to students.

Vocabulary:

correlation coefficient

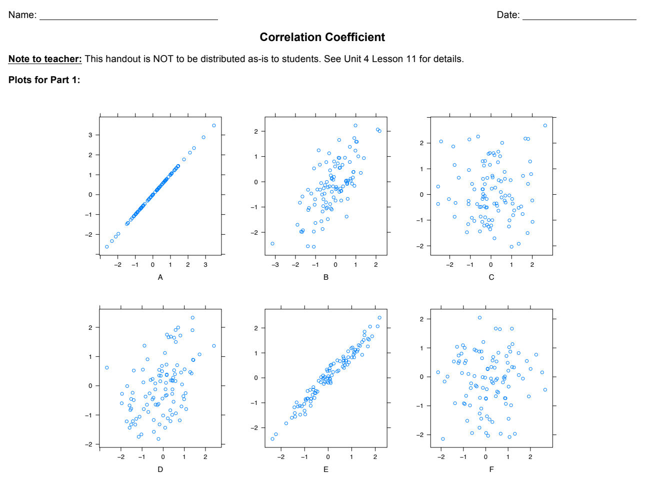
> **Essential Concept**: A high absolute value for correlation means a strong linear trend. A value close to 0 means a weak linear trend.

Lesson:

1. Inform students that, so far, they have been labeling associations as strong, very strong, or weak. A number called the **correlation coefficient** measures strength of association. The correlation coefficient only applies to linear relationships, which must be checked visually with a scatterplot. Later we will learn how to calculate this number using RStudio.

   Note to teacher: Advance preparation is needed for this lesson. Each team needs one envelope with cutouts of plots A-F in LMR_4.11 (Part A). Make envelopes according to the number of teams in the class. This process will be repeated for LMR_4.11 (Part B).

2. Distribute the envelopes to the teams. Students will examine the strength of association in each plot. Their task is to assign the correlation coefficient that corresponds to each plot and to explain why they assigned that correlation coefficient to that particular plot. The only piece of information they will receive is that a correlation coefficient equal to 1 has the strongest linear association and a correlation coefficient equal to 0 has the weakest association.



Name: ____________________________ Date: __________________

Correlation Coefficient

Note to teacher: This handout is NOT to be distributed as-is to students. See Unit 4 Lesson 11 for details.

Plots for Part 1:

LMR_4.11

3. Assign each team one plot. If there are more teams than plots, these teams will be assigned a plot in the next round. Each team will share the correlation coefficient they assigned to their plot and the explanation that goes with it.

4. Using the *Voting Cards* strategy (see Instructional Strategies), the rest of the teams will show whether they approve, disapprove, or are uncertain about the teams' assignment and/or explanation. Repeat for each plot. The correlation coefficients for each plot are:

- ***Plot A: r = 1.00***
- ***Plot B: r = 0.72***
- ***Plot C: r = 0.19***
- ***Plot D: r = 0.48***
- ***Plot E: r = 0.98***
- ***Plot F: r = 0.00***

5. The last set of plots showed positive associations. Now students will assign the correlation coefficients for plots G-L for LMR_4.11 (Part 2).
6. Distribute the envelopes to the teams. Students will examine the strength of association in each plot. Their task is to assign the correlation coefficient that corresponds to each plot and to explain why they assigned that correlation coefficient to that particular plot. The only piece of information they will receive is that a correlation coefficient equal to -1 has the strongest linear association and a correlation coefficient equal to 0 has the weakest association.
7. Teams previously not assigned a plot are now assigned one. Each team will share the correlation coefficient they assigned to their plot and the explanation that goes with it.
8. Using the Voting Cards strategy, the rest of the teams will show whether they approve, disapprove, or are uncertain about the teams' assignment and/or explanation. Lead a class discussion whenever there is disapproval or uncertainty. Repeat for each plot. The correlation coefficients for each plot are:
   - ***Plot G: r = -1.00***
   - ***Plot H: r = 0.72***
   - ***Plot I: r = -0.19***
   - ***Plot J: r = -0.48***
   - ***Plot K: r = 0.98***
   - ***Plot L: r = 0.00***
9. Journal Entry: What is a correlation coefficient, what does it do, and what does it tell us about a scatterplot?

Homework & Next Day

Students will complete journal entry for homework if not completed in class.

LAB 4D: Interpreting Correlations

Complete Lab 4D prior to Lesson 12.

Lab 4D - Interpreting correlations

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

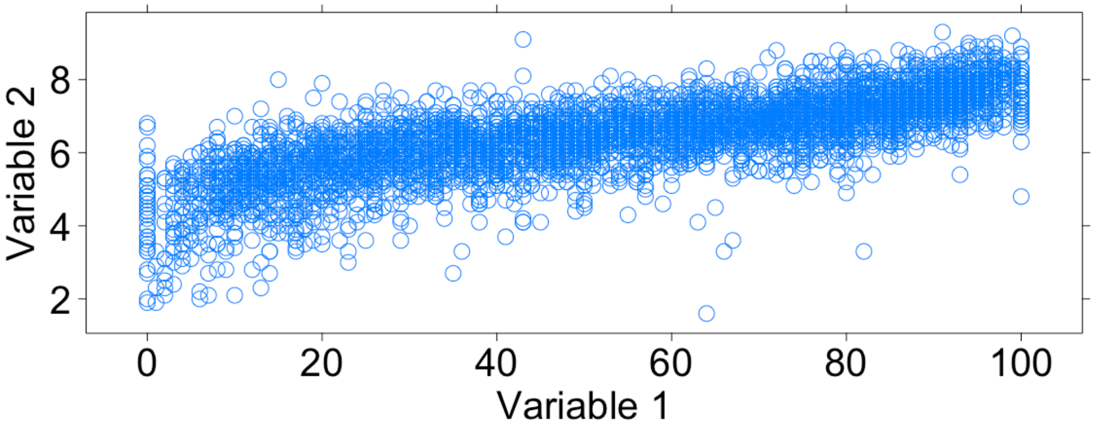
Some background...

- So far, we've learned about measuring the success of a model based on how close its predictions come to the actual observations.
- The *correlation coefficient* is a tool that gives us a fairly good idea of how these predictions will turn out without having to make predictions on future observations.
- For this lab, we will be using the `movie` data set to investigate the following questions:

  Which variables are better predictors of a movie's `audience_rating` when the predictions are made using a line of best fit?

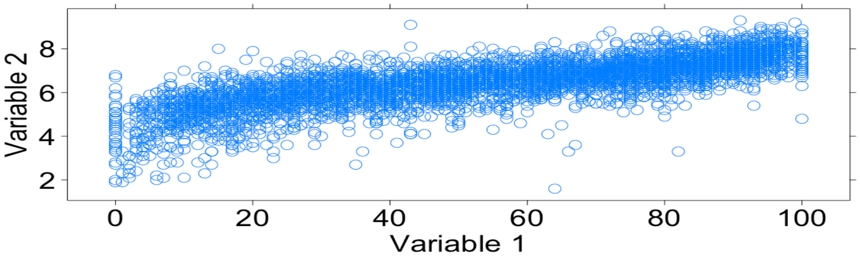
Correlation coefficients

- The *correlation coefficient* describes the *strength* and *direction* of the linear trend.
- It's only useful when the trend is linear and both variables are numeric.



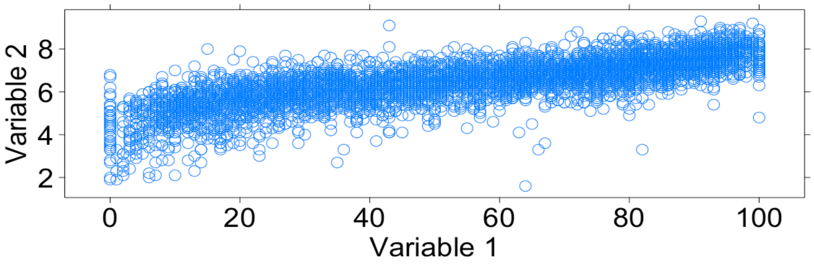
- **Are these variables linearly related? Why or why not?**

Correlation review I



- Correlation coefficients with values close to 1 are very strong with a positive slope. Values close to -1 means the correlation is very strong with a negative slope.
  - **Does this plot have a positive or negative correlation?**

Correlation review II



- **Recall that if there is no linear relation between two numerical variables, the correlation coefficient is close to 0. What do you guess the correlation coefficient will be for these two variables?**

The movie data

- Load the `movie` data using the `data` command.
- The data comes from a variety of sources like *IMDB* and *Rotten Tomatoes*.
  - The `critics_rating` contains values between 0 and 100, 100 being the best.
  - The `audience_rating` contains values that range between 0 and 10, 10 being the best.
  - `n_critics` and `n_audience` describe the number of reviews used for the ratings.
  - `gross` and `budget` describes the amount of money the film made and took to make.

Calculating Correlation Coefficients!

- We can use the `cor()` function to find the particular correlation coefficient of the variables from the previous plot, which happen to be `audience_rating` and `critics_rating`.
  - But note, the `cor()` function removes any observations which contains an `NA` value in either variable.
  - **Calculate the correlation coefficient for these variables using the `cor` function. The inputs to the functions work just like the inputs of the `xyplot` function.**

Now answer the following.

- **What was the value of the correlation coefficient you calculated?**
- **How does this actual value compare with the one you estimated previously?**
- **Does this indicate a strong, weak, or moderate association? Why?**
- **How would the scatter plot need to change in order for the correlation to be stronger?**
- **How would it need to change in order for the correlation to be weaker?**

Correlation and Predictions

- Find the two variables that look to have the strongest correlation with `critics_rating`.
  - Compute the correlation coefficients for `critics_rating` and each of the two variables.
  - **Use the correlation coefficient to determine which variable has a stronger linear relationship with `critics_rating`.**
- Fit two `lm` models to predict `critics_rating` with each variable and compute the MSE for each.

   - **Use the MSE to determine which variable is a better predictor of `critics_rating`.**
- **How are the correlation coefficient and the MSE related?**

On your own

- Select two different numerical variables from the `movie` data.
- Plot the variables using the `xyplot()` function.
   - **Would calculating a correlation coefficient for the two variables be appropriate? Justify your answer.**
   - **Predict what value you think the correlation coefficient will be. Compare this value to the actual value. Finally, interpret what the actual correlation coefficient means.**
- Work with your classmates to determine which two variables have the strongest correlation coefficient.
- **Why do you think these variables are so strongly related? Is using the correlation coefficient to describe the relationship appropriate and why/why not?**

Piecing it Together

Instructional Days: 5

Enduring Understandings

Real-life phenomena are often complex. Data scientists use multiple regression models to create simple equations to help explain and predict these phenomena. Data scientists can also use polynomial transformations to add flexibility to rigid linear models.

Engagement

Students will read the article titled *How Long Can a Spinoff Like Better Call Saul Last?* that will set the context for students to begin thinking about more than one explanatory variable to make better predictions. The article can be found at:
http://fivethirtyeight.com/features/how-long-can-a-spinoff-like-better-call-saul-last/

Learning Objectives

Statistical/Mathematical:

S-ID 6: Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. *Use given functions or choose a function suggested by the context. Emphasize linear models*.

Data Science:

Understand that multiple regression can be a better tool for predicting that simple linear regression and know when it is appropriate to use multiple regression versus simple linear regression. Understand when linear models are not appropriate based on the shape of the scatterplot.

Applied Computational Thinking using RStudio:

- Use multiple linear regression models with other predictor variables
- Fit regression lines to data and predict outcomes.
- Create non-linear models to look for relationships.
- Fit polynomials functions to data.

Real-World Connections:

Economists and marketing firms use multiple regression to predict changes in the market and adjust strategies to fit the demands of changes in the marketplace.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data Set:

1. NFL data set
2. USMNT data set

Data File:

3. Movies: `data(movie)`

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 12: More Variables to Make Better Predictions

Objective:

Students will see that information from different variables can be used together to create linear models that make more accurate predictions.

Materials:

1. *Advertising Plots Part 1* handout (LMR_4.12_Advertising Plots 1)
2. *Advertising Plots Part 2* handout (LMR_4.13_Advertising Plots 2)
3. *Article: How Long Can a Spinoff Like 'Better Call Saul' Last?* http://fivethirtyeight.com/features/how-long-can-a-spinoff-like-better-call-saul-last/

Vocabulary:

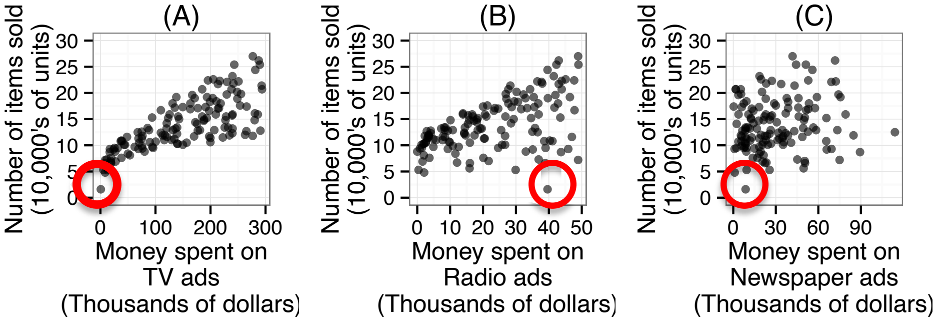
market

Essential Concepts: We can use scatterplots to assess which variables might lead to strong predictive models. Sometimes using several predictors in one model can produce stronger models.

Lesson:

1. Remind students that models are used to make predictions. Ask a volunteer to think of a TV show that had a "spinoff" and to name both of the shows. Ask if he/she knows whether or not the original was more or less successful than the spinoff. Then, ask the class: Is there a way to predict spinoff success?
2. Next, using the *Talking to the Text* instructional strategy, ask students to read the article titled: *How Long Can a Spinoff Like Better Call Saul Last?*

   Note: If this is the first time using this strategy with your students, make sure you model/explain it before they begin reading it. See Instructional Strategies in Teacher Resources for a description.
3. After reading the article, ask students to discuss three *Talking to the Text* responses with a partner. You may set a time limit for each student to share with his/her partner.
4. Then, in teams, students will answer the following questions pertaining to the article:
   a. What is the article trying to predict?
   b. How many variables are used?
   c. What other variables might affect a spinoff?
   d. The dotted line in the plot is not a regression line. How would you draw a regression line to make predictions?
   e. What other information would you like to know to predict a spinoff's success?
5. Allow students time to discuss and record their answers. Then conduct a share out of their responses to the discussion questions.
6. Discuss the following questions with the class:
   a. What effect does advertising have on retail sales?
   b. Where do stores advertise (What mediums do they use)? Does each method of advertisement reach the same people?
   c. Does each method of advertisement have a similar effect? Or are some methods more effective than others?
7. Distribute the 3 plots from the *Advertising Plots Part 1* handout (LMR_4.12) and inform the students about the data using the details below:



LMR_4.12
(Plots are presented separately in the LMR)

a. These 3 plots show the number of items sold by a retailer (in 200 different markets) and the amount of money the company spent on *TV*, *Radio* and *Newspaper* advertisements.

b. The data has 200 observations, one for each different market. A **market** is simply a location where an item is sold. For example, Los Angeles and San Francisco are two different markets.

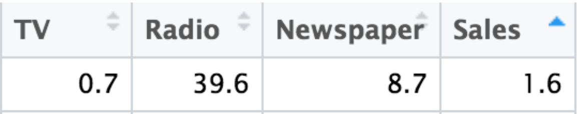
c. Each observation has 4 variables: (1) The number of items sold (in 10's of thousands of units), (2) the money spent on TV ads (in thousands of dollars), (3) the money spent on radio ads (in thousands of dollars), and (4) the money spent on newspaper ads (in thousands of dollars).

d. The data were collected using an observational study.

8. To illustrate a-d above, ask students to refer to plot A (TV ads) and circle the market in which this retailer sold the least number of items (see circles in plots above). Ask: How many items did this market sell? ***About 20,000 items. The actual number of items sold was 1.6 (in 10,000's of units) which is 16,000 items***. How much money did this retailer spend on TV ads in this market? ***This retailer spent zero dollars on TV ads***. ***The actual amount the retailer spent on TV ads was 0.7 thousands of dollars, which is $700.***

9. Students should then refer to plot B (Radio ads), find the same market (the one in which the retailer sold about 20,000 items) and circle it. Ask: How much money did the retailer spend on Radio ads in the same market? ***About 40 thousand dollars***. ***The actual amount spent on Radio ads was 39.6 thousands of dollars, which is $39,600***.

10. Finally, ask students to refer to plot C (Newspaper ads), find the same market (the one in which the retailer sold about 20,000 items), and circle it. Ask: How much money did the retailer spend on Newspaper ads in the same market? ***About 10 thousand dollars***. ***About The actual amount spent on Newspaper ads is 8.7 thousands of dollars, which is $8,700***



TV	Radio	Newspaper	Sales
0.7	39.6	8.7	1.6

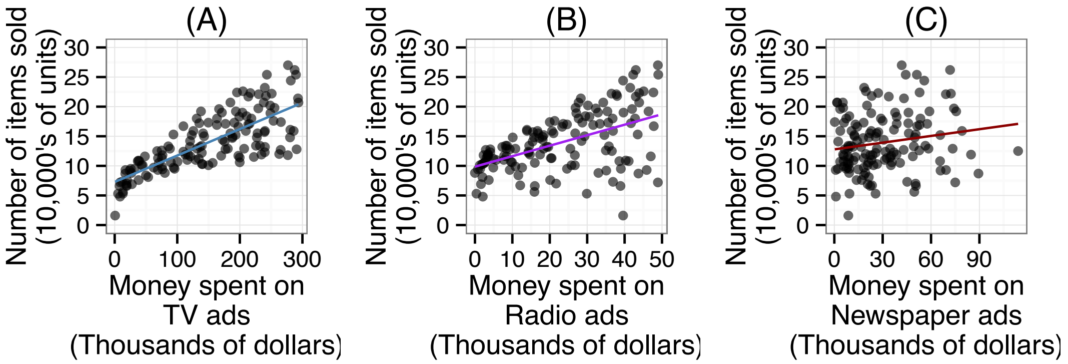
.

11. Based on the above plots, use a Pair-Share to discuss the following:

    a. Describe the relationship between advertisements and the number of items sold.

    b. Which type of advertisement is the most strongly correlated with the number of units sold? How can you tell?

12. Distribute the *Advertising Plots Part 2* handout (LMR_4.13_Advertising Plots 2), which contains plots A-C, but now include the line of best fit.



LMR_4.13

(Plots are presented separately in the LMR)

13. Ask students to recall from Lesson 6 that a method statisticians use to figure out which predicted values is closest to the actual data is the mean absolute error (MAE).

    Note to teacher: In the labs, students will use the mean squared error (MSE) - also learned in Lesson 6 - which calculates the regression line. In the lessons, we discuss the issue more generally using the mean absolute error (MAE).

14. In teams, ask students to discuss the following:

    a. How would you use the mean absolute error to determine which plot would make the most accurate predictions? ***Answers will vary, but you would expect to hear something like: "the prediction line that has the least amount of distance to all the points on the plot would make the most accurate prediction because the predicted values will be closer to the actual data."***

15. Next, have students select a statement they think is best (a or b), then write a justification for their selection based on what they learned in this lesson. This may be completed as homework.

    a. Combining multiple variables (e.g., TV and Newspaper ads, TV and Radio ads, TV, Radio, and Newspaper ads, etc.) into one model will lead to worse predictions because the variables that make poor predictions will contaminate those that make good predictions.

    b. Combining multiple variables (e.g., TV and Newspaper ads, TV and Radio ads, TV, Radio, and Newspaper ads, etc.) into one model will lead to better predictions because the model can use more information to make predictions.

16. Inform students that RStudio has the capability of creating models that combine multiple variables to make predictions about another variable. For example, it can make a model to predict number of items sold using both money spent on TV and money spent on Newspaper ads. Students will learn more about it during the next lesson.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students may continue writing their justifications for the selected statement in item 15 if they were unable to finish.

Lesson 13: Combination of Variables

Objective:

Students will learn that we can make better predictions by including more variables. Then they will wrestle with how the information should be combined.

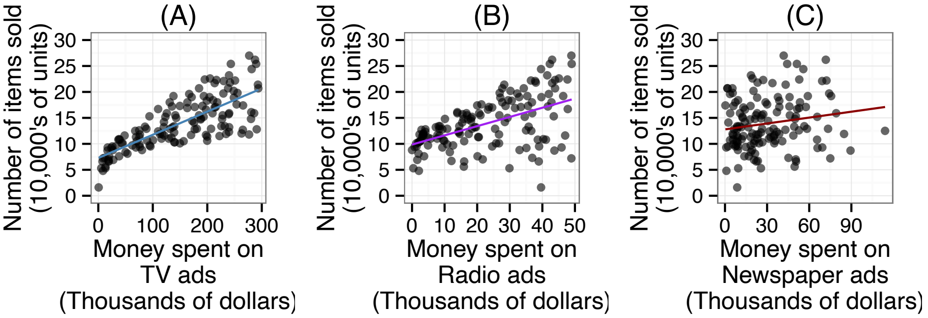
Materials:

1. *Advertising Plots Part 2* handout (LMR_4.13_AdvertisingPlots2) from Lesson 12

Essential Concepts: If multiple predictors are associated with the response variable, a better predictive model will be produced, as measured by the mean absolute error.

Lesson:

1. Display the plots and statements from the previous day:



LMR_4.13

(Plots are presented separately in the LMR)

   a. Combining multiple variables (e.g., money spent on TV and Newspaper ads, TV and Radio ads, TV, Radio, and Newspaper ads, etc.) into one model will lead to worse predictions because the variables that make poor predictions will contaminate those that make good predictions.

   b. Combining multiple variables (e.g., TV and Newspaper ads, TV and Radio ads, TV, Radio, and Newspaper ads, etc.) into one model will lead to better predictions because the model can use more information to make predictions.

2. Ask the students to share out their opinions in an Active Debate (see Unit 2 Lesson 6 as an example).

3. Next, inform teams that they will have 2 minutes to come up with as many combinations of ads (variables) as they can think of (e.g., TV + Newspaper ads, TV+ Radio ads, TV + Radio + Newspaper ads, etc.)

4. After 2 minutes, list all the different combinations by conducting a Whip Around and eliciting a combination from each team.

5. By a show of hands, ask students to select which combination or single model will be the best predictor for the number of items sold by the retailer.

6. Then inform students that we will determine which of the statements is true by comparing the mean absolute error (MAE) of single models (like the ones we showed in the previous lesson) vs. combined models. But first, use the line of best fit for the combined variables:

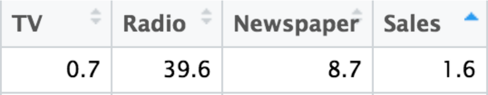
$$\widehat{sales} = 0.045449(tv) + 0.186570(radio) - 0.004952(newspaper) + 3.029878$$

Note: The function that produced the line of best fit using RStudio was

```
lm(Sales ~ TV + Radio + Newspaper, data= retail)
```

a. Use this equation to predict the amount of sales for the same market they circled in the previous lesson. ***Students' calculation should yield the predicted value in (b), below***.

Note: Remind students that they need to substitute the values as they appear in the x-axis of the plots without converting to thousands of dollars. For example, the circled market spent about 10 thousand dollars on newspaper ads, so students should substitute 10 instead of the expanded value in the equation.



TV	Radio	Newspaper	Sales
0.7	39.6	8.7	1.6

b. Does the predicted value (10.407) seem like a plausible number of sales? Why? ***It is not a plausible number of sales because the prediction is too high. The prediction says the retailer will sell about 104,070 units, when the actual sales were about 16,000 units. Although the model did not make a very good prediction for this market, it is not surprising because as LMR_4.13 displays, that market did not fit the overall pattern in any of the scatterplots.***

7. Reveal that RStudio calculated the mean absolute error for different combinations plus the single models, and the results are displayed on the table below. This means that, for example, when using the TV model to predict number of items sold, our predictions will typically be off by about 2.337808 (in 10,000s) of units or 23,378 units. Then ask students:

Model	Mean Absolute Error
TV	2.337808
Radio	3.565113
Newspaper	4.538444
TV-Radio	1.160937
TV-Newspaper	2.344971
Radio-Newspaper	2.93832
TV-Radio-Newspaper	1.161068

a. Which model is the best predictor of number of items sold? ***Answer: The TV-Radio model is the best predictor of number of items sold because it had the least amount of error, on average. When using the TV-Radio model to predict number of items sold, our predictions will typically be off by 11,609 units.***

b. Which model was the least reliable in predicting the number of items sold? ***Answer: The Newspaper model is the least reliable predictor of number of items sold because it had the most amount of error, on average. When using the Newspaper model to predict number of items sold, our predictions will typically be off by 45,384 units.***

c. What else do you notice about the models? ***Answer: It appears that combining the variables into one model is much better than any of the single-variable models.***

8. Inform the students that, in the next lab, they will find out how to create the line of best fit for models that include many variables.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

Ask students to think of a reason or reasons about why it would not be a good idea to make a scatterplot for models that include more than 3 predictor variables? ***The answer is mainly because humans are limited to seeing things in 3 dimensions. For example, the model that combines all of the variables together is a 4 dimensional model. What does that look like?***

LAB 4E: This Model is Big Enough for All of Us

Complete Lab 4E prior to Practicum.

Lab 4E - This model is big enough for all of us!

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Building better models

- So far, in the labs, we've learned how to make predictions using the *line of best fit*
  - Which we also call *linear models* or *regression models*.
- We've also learned how to measure our model's prediction accuracy by cross-validation.
- In this lab, we'll investigate the following question:

  Will including more variables in our model improve its predictions?

Divide & Conquer

- Start by loading the `movie` data and split it into two sets (See Lab 4C for help). Remember to use `set.seed`.
  - A set named `training` that includes 75% of the data.
  - A set named `testing` that includes the remaining 25%.
- Create a linear model, using the `training` data, that predicts `gross` using `runtime`.
  - Compute the MSE of the model by making predictions for the `testing` data.
- **Do you think that a movie's `runtime` is the only factor that goes into how much a movie will make? What else might affect a movie's `gross`?**

Including more info

- Data scientists often find that including more relevant information in their models leads to better predictions.
  - Fill in the blanks below to predict `gross` using `runtime` and `reviews_num`.

```
lm(____ ~ ____ + ____, data = training)
```

- **Does this new model make more or less accurate predictions? Describe the process you used to arrive at your conclusion.**
- **Write down the code you would use to include a 3rd variable, of your choosing, in your `lm()`.**

Own your own

- **Write down which other variables in the `movie` data you think would help you make better predictions.**
  - **Are there any variables that you think would not improve our predictions?**
- **Create a model for all of the variables you think are relevant.**
  - **Assess whether your model makes more accurate predictions for the `testing` data than the model that included only `runtime` and `reviews_num`**
- **With your neighbors, determine which combination of variables leads to the best predictions for the `testing` data.**

<u>Practicum: Predictions</u>

Objective:
Students will create a linear model to predict the nutritional component that is most closely associated with the amount of sugar contained in a cereal.

Materials:

1. *Predictions Practicum* (LMR_U4_Practicum_Predictions)

Practicum
Predictions

Data about the nutritional components of popular cereal brands has been collected and made available for your team's use. We are interested in determining which other nutritional component is most closely associated with the amount of sugar contained in a cereal.

Your team will use the data to make predictions using linear models and compare the accuracy of your model to the rest of your classmates. Finally, the class will determine which team had the best prediction. Follow the directions below to explore and analyze the data:

1. You will have two data sets: one for training and one for testing. Load both data sets. Write down the code you used.
2. Explore the training data. Make several plots of different variable combinations and fit a linear regression line through them. Select the model that you think best makes the best prediction.
3. For the linear model your team selected:
    a. Describe what the plot shows.
    b. Explain why you selected that particular model.
    c. Compute the mean absolute error of your model using your testing data.
    d. Now make a set of predictions with your testing data. Calculate the mean absolute error for the testing data. Is it better or worse than for the training data, or about the same?
4. Present your team's linear model to the class. Explain why you chose your model and the typical amount of error in its predictions.
5. Give an example of a prediction for one value of x. State that value, give the predicted calories, and describe, based on the testing data, how far off your prediction might actually be.

Lesson 14: Improving Your Model

Objective:

Students will learn to describe associations that are not linear.

Materials:

1. *Describe the Association* handout (LMR_4.14_Describe the Association)

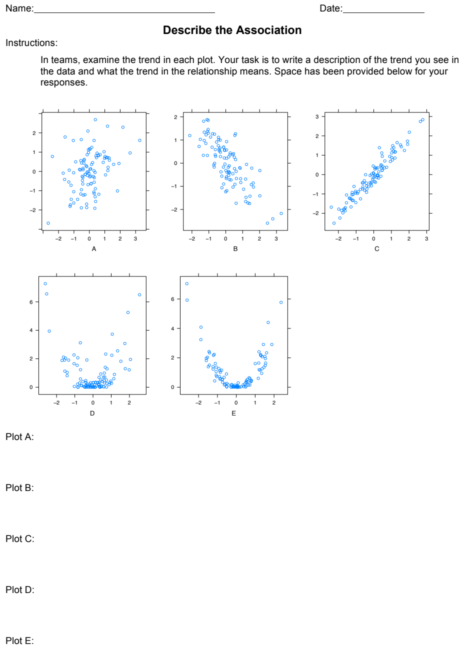
Vocabulary:

non-linear, polynomial trends

Essential Concepts: If a linear model is fit to a non-linear trend, it will not do a good job of predicting. For this reason, we need to identify non-linear trends by looking at a scatterplot or the model needs to match the trend.

Lesson:

1. Remind students that they have been learning a great deal about linear associations. However, there are other types of associations, and today they will learn to describe them.
2. Distribute *Describe the Association* (LMR_4.14). In teams, students will examine the trend of each plot. Their task is to write a description of the trend that they see in the data and what the trend means.



Name:______________________________ Date:_____________

Describe the Association

Instructions:

In teams, examine the trend in each plot. Your task is to write a description of the trend you see in the data and what the trend in the relationship means. Space has been provided below for your responses.

Plot A:

Plot B:

Plot C:

Plot D:

Plot E:

LMR_4.14

3. Allow students time to discuss and record their descriptions for each plot in their DS journals. Walk around the room monitoring student teamwork. Look for descriptions that are interesting to share with the whole class.
4. Select a team to present a description of one plot to the class. Teams will listen to each presentation, compare it to their description of the plot, and as a team they will agree or disagree. If there is disagreement, lead a discussion that guides students to reason toward the correct description.
5. Summarize the discussion for each plot and ask students take notes or revise their descriptions in their DS journals.
6. Repeat steps 4 and 5 for the rest of the plots.

Plot Descriptions for *Describe the Association* (LMR_4.14):

- ***Plot A: There is no trend (perhaps some may see a very, very weak linear trend), so there is no/hardly any association. There is a great deal of scatter in the data. It means that y does not depend on x.***
- ***Plot B: There appears to be a linear trend. The association is negative and appears somewhat strong. It means that as x increases, y decreases.***
- ***Plot C: There is a linear trend. The association is positive and it is very strong. It means that the y-value increases at approximately the same rate for every increase in x value. This is a line.***
- ***Plot D: The trend is non-linear. There seems to be a weak association because there is scatter in the data. Cannot tell if the association is positive or negative. It has the shape of a parabola; therefore, it is quadratic. For smaller x-values, the y-value is decreasing and for larger x values, the y value is increasing.***
- ***Plot E: The trend is non-linear. There seems to be a strong association because there is little scatter in the data. It is also in the shape of a parabola, so it is quadratic.***

7. Using the *Cheat Notes* strategy, ask teams to write notes about how to describe associations.
8. Plots A, B, and C should be familiar to the students by now. However, plots D and E show a different type of trend. Although the trends are non-linear, they can still tell us important information about the y-values based on values of x. Ask:
   - What happens if we were to fit a linear model to these non-linear trends? Would it still make good predictions? ***No. They would not make good predictors***.
9. To examine why they would not make good predictors, draw an approximate linear best-fit line and get students to understand that in some regions, the model would almost always over-predict, and in others would almost always under-predict. We want a model that goes, more or less, through the 'middle' of the points. Ask:
   - How can we get a model that goes, more or less, through the middle of all the data points? ***Answer: We need to change the model.***
10. Trends like the quadratic ones show in plots D and E can be described as **polynomial trends**. Plots that follow quadratic, cubic, quartic, etc. shapes all exhibit polynomial trends. We need to adjust the model. You may show students several choices of equations (quadratic, trinomial, linear) along with their graphs and ask them which might be a good candidate.
11. When investigating the data for trends, the model needs to fit the data.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

Students may finish their *Cheat Notes* for homework, if not completed in class.

LAB 4F: Some Models Have Curves

Complete Lab 4F prior to Lesson 15.

Lab 4F - Some models have curves

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Making models do yoga

- So far, we have only worked with prediction models that fit the *line of best fit* to the data.
- But what happens if the true relationship between the data is nonlinear?
- In this lab, we will learn about prediction models that fit *best fitting curves* to data.
- **Before moving on, load the `movie` data and split it into two sets:**
  - **A set named `training` that includes 75% of the data.**
  - **And a set named `testing` that includes the remaining 25%.**
  - Remember to use `set.seed`.

Problems with lines

- Before learning how to fit curves, let's first fit a linear model for reference.
- **Train a linear model predicting `audience_rating` based on `critics_rating` for the `training` data. Assign this model to `movie_linear`.**
- **Fill in the blanks below to create a scatterplot with `audience_rating` on the y-axis and `critics_rating` on the x-axis using your `testing` data.**

```
xyplot(____ ~ ____, data = ____)
```

- Previously, you used `add_line` to plot the *line of best fit*. An alternative function for plotting the *line of best fit* is `add_curve`, which takes the name of the model as an argument.
- **Run the code below to add the line of best fit for the `training` data to plot.**

```
add_curve(movie_linear)
```

- **Describe, in words, how the line fits the data. Are there any values for `critics_rating` that would make obviously poor prediction?**
  - Hint: how does the linear model perform on very low and very high values of `critics_rating`?
- **Compute the MSE of the model for the `testing` data and write it down for later.**
  - Hint: refer to lab 4B.

Adding flexibility

- You don't need to be a full-fledged Data Scientist to realize that trying to fit a line to curved data is a poor modeling choice.
- If our data is curved, we should try model it with a curve.
- Instead of fitting a line, with equation of the form

$$y = a + bx$$

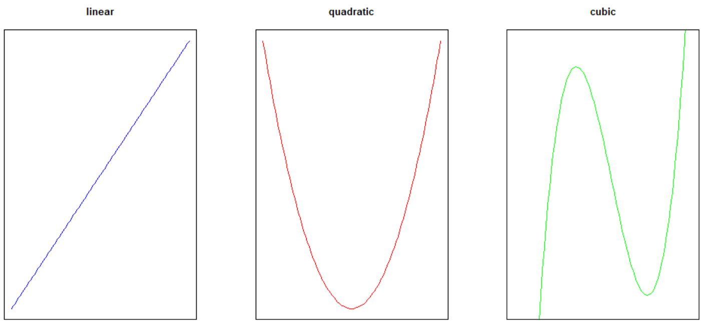
- we might consider fitting a *quadratic curve*, with equation of the form

$$y = ax + bx + cx^2$$

- or even a *cubic curve*, with equation of the form

$$y = a + bx + cx^2 + dx^3$$

- In general, the more coefficients in the model, the more flexible its predictions can be.



Making bend-y models

- To fit a quadratic model in `R`, we can use the `poly()` function.
  - **Fill in the blanks below to train a quadratic model predicting `audience_rating` from `critics_rating`, and assign that model to `movie_quad`.**

```
movie_quad <- lm(____ ~ poly(____, 2), data = training)
```

- **What is the role of the number 2 in the `poly()` function?**

Comparing lines and curves

- **Fill in the blanks to**
  - **create a scatterplot with `audience_rating` on the y-axis and `critics_rating` on the x-axis using your `testing` data, and**
  - **add the *line of best fit* and *best fitting quadratic curve*.**
  - Hint: the `col` argument is added to the `add_curve` functions to help distinguish the two curves.

```
xyplot(____ ~ ____, data = ____)
add_curve(____, col = "blue")
add_curve(____, col = "red")
```

- **Compare how the *line of best fit* and the *quadratic* model fit the data. Which do you think has a lower `test` MSE?**
- **Compute the MSE of the quadratic model for the `test` data and write it down for later.**
- **Use the difference in each model's `test` MSE to describe why one model fits better than the other.**

On your own

- Create a model that predicts `audience_rating` using a cubic curve (polynomial with degree `3`), and assign this model to `movie_cubic`.
- Create a scatterplot with `audience_rating` on the y-axis and `critics_rating` on the x-axis using your `test` data.
- Using the names of the three models you have trained, add the *line of best fit*, *best fitting quadratic curve*, and *best fitting cubic curve* for the `training data` to the plot.
- Based on the plot, which model do you think is the best at predicting the `testing` data?
- Use the difference in testing MSE to verify which model is the best at predicting the `testing data`.

The Growth of Landfills

Instructional Days: 5

Enduring Understandings

Model Eliciting Activities (MEAs) engage students in a complete modeling experience. MEAs are designed to make students' thinking visible and audible by encouraging them to be metacognitive about the process of inventing and testing a model, ask questions as they go through the process, and recognize the iterative nature of modeling.

Engagement

Students will read an excerpt from a CNN article called *Trash City: Inside America's Largest Landfill Site*. This article will set the context of the real-world problem facing many cities—the growth of landfills. The article provides background information as well as baseline data to launch the modeling process.

Learning Objectives

Statistical/Mathematical:

According to the California Common Core State Standards-Mathematics (CCSS-M) Framework:

"Modeling links classroom mathematics and statistics to everyday life, work, and decision-making. Modeling is the process of choosing and using appropriate mathematics and statistics to analyze empirical situations, to understand them better, and to improve decisions. Quantities and their relationships in physical, economic, public policy, social, and everyday situations can be modeled using mathematical and statistical methods. When making mathematical models, technology is valuable for varying assumptions, exploring consequences, and comparing predictions with data.

Modeling is best interpreted not as a collection of isolated topics, but rather in relation to other standards. Making mathematical models is a Standard for Mathematical Practice, and specific modeling standards appear throughout the high school standards indicated by a star symbol (★)."

Every Statistics and Probability standard in the California CCSS-M High School Conceptual Category is considered a modeling standard, indicated by the star symbol; therefore, rather than listing every content standard individually, the modeling activities in this section are designed so that students apply the Statistics and Probability standards learned throughout the curriculum.

Focus Standards for Mathematical Practice:

SMP-4: Model with mathematics.

Data Science:

Students will apply the conceptual understandings learned up to this point in the curriculum.

Applied Computational Thinking using RStudio:

- Previous techniques from the curriculum will be used in order to complete the task.

Real-World Connections:

Engineers, data scientists, and statisticians, to name a few, use modeling in their everyday work. Whether it is for creating a scale model of a bridge or a mathematical model of force impact measures, modeling is an integral part of what they do in the real world.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will use complex sentences to write a letter of recommendation that use data science concepts and skills.
4. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data File:

1. Trash: `data (trash)`

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 15: The Growth of Landfills

Objective:

Students will engage in a modeling activity to learn about reducing the burden of trash landfills.

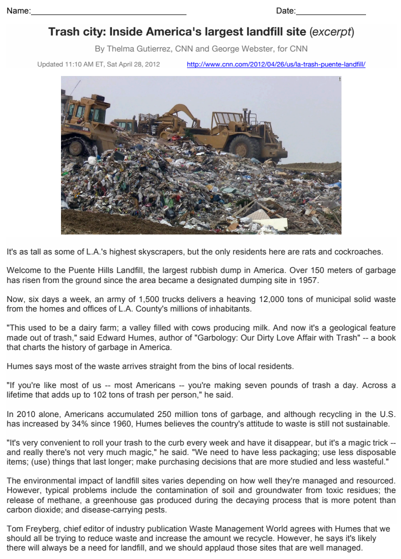
Materials:

1. *Landfill Article* handout (LMR_4.15_Landfill Article)
2. *Landfill Readiness Questions* handout (LMR_4.16_Landfill Readiness Questions)
3. *Landfill Activity* handout (LMR_4.17_Landfill Activity)
4. Computers
5. IDS public dashboard: https://portal.idsucla.org
6. *Trash Data Exploration* handout (LMR_4.18_Trash Data Exploration)

Essential Concepts: Modeling does not always have to produce an equation. Instead, we can create models to answer real-world problems related to our community.

Lesson:

1. Inform students that they will investigate a problem that faces many cities in the United States today: trash. Explain that the next 4 days will be dedicated to completing the investigation and will follow this general structure:
    a. Day 1: Introduce assignment, initial exploration of data, creation of statistical questions.
    b. Day 2: Analysis of data via the IDS public dashboard.
    c. Day 3: Verify analysis via RStudio.
    d. Day 4: Team presentations.
2. Distribute the Landfill Article handout (LMR_4.15_Landfill Article) and explain that the reading is an excerpt from a CNN article titled *Trash City: Inside America's Largest Landfill Site*. The article will set the context for the real-world problem of growing landfills.



Name:______________________________ Date:____________

Trash city: Inside America's largest landfill site (*excerpt*)

By Thelma Gutierrez, CNN and George Webster, for CNN

Updated 11:10 AM ET, Sat April 28, 2012 http://www.cnn.com/2012/04/26/us/la-trash-puente-landfill/

It's as tall as some of L.A.'s highest skyscrapers, but the only residents here are rats and cockroaches.

Welcome to the Puente Hills Landfill, the largest rubbish dump in America. Over 150 meters of garbage has risen from the ground since the area became a designated dumping site in 1957.

Now, six days a week, an army of 1,500 trucks delivers a heaving 12,000 tons of municipal solid waste from the homes and offices of L.A. County's millions of inhabitants.

"This used to be a dairy farm; a valley filled with cows producing milk. And now it's a geological feature made out of trash," said Edward Humes, author of "Garbology: Our Dirty Love Affair with Trash" -- a book that charts the history of garbage in America.

Humes says most of the waste arrives straight from the bins of local residents.

"If you're like most of us -- most Americans -- you're making seven pounds of trash a day. Across a lifetime that adds up to 102 tons of trash per person," he said.

In 2010 alone, Americans accumulated 250 million tons of garbage, and although recycling in the U.S. has increased by 34% since 1960, Humes believes the country's attitude to waste is still not sustainable.

"It's very convenient to roll your trash to the curb every week and have it disappear, but it's a magic trick -- and really there's not very much magic," he said. "We need to have less packaging; use less disposable items; (use) things that last longer; make purchasing decisions that are more studied and less wasteful."

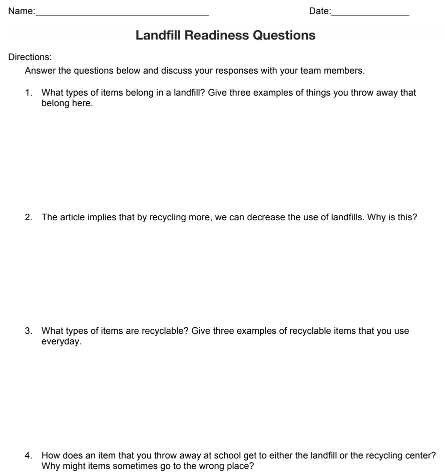
The environmental impact of landfill sites varies depending on how well they're managed and resourced. However, typical problems include the contamination of soil and groundwater from toxic residues; the release of methane, a greenhouse gas produced during the decaying process that is more potent than carbon dioxide; and disease-carrying pests.

Tom Freyberg, chief editor of industry publication Waste Management World agrees with Humes that we should all be trying to reduce waste and increase the amount we recycle. However, he says it's likely there will always be a need for landfill, and we should applaud those sites that are well managed.

LMR_4.15

3. Using the *5 Ws* strategy, ask students to read the article individually and to write down the 5 Ws in their DS journals. The 5 Ws summarize the What, Who, Why, When, and Where of the article.
4. After they have finished reading, students should answer the questions provided on the *Landfill Readiness Questions* handout (LMR_4.16_Landfill Readiness Questions). Then, in teams, students will discuss their insights, questions, and/or reactions to the both the article and the

questions. Follow up the team discussion with a class discussion to gauge what students actually know about trash and recycling.



Name:______________________________ Date:______________

Landfill Readiness Questions

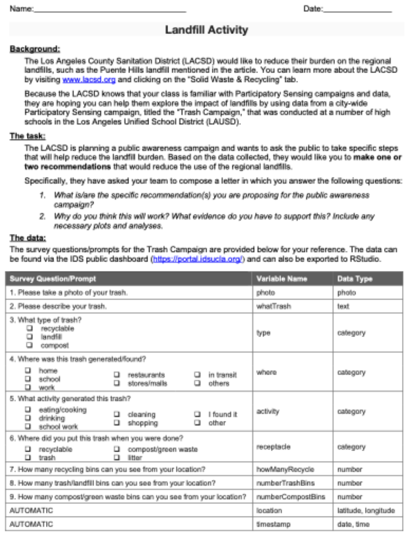
Directions:

Answer the questions below and discuss your responses with your team members.

1. What types of items belong in a landfill? Give three examples of things you throw away that belong here.
2. The article implies that by recycling more, we can decrease the use of landfills. Why is this?
3. What types of items are recyclable? Give three examples of recyclable items that you use everyday.
4. How does an item that you throw away at school get to either the landfill or the recycling center? Why might items sometimes go to the wrong place?

LMR_4.16

5. Next, introduce students to the main task they will be investigating about landfills by distributing the *Landfill Activity* handout (LMR_4.17_Landfill Activity). This handout asks the students to come up with one or two recommendations to help reduce the burden of landfills on the environment. In order to complete the assignment, students will use 2 data analysis tools: the IDS dashboard and RStudio.



Name:______________________________ Date:______________

Landfill Activity

Background:

The Los Angeles County Sanitation District (LACSD) would like to reduce their burden on the regional landfills, such as the Puente Hills landfill mentioned in the article. You can learn more about the LACSD by visiting www.lacsd.org and clicking on the "Solid Waste & Recycling" tab.

Because the LACSD knows that your class is familiar with Participatory Sensing campaigns and data, they are hoping you can help them explore the impact of landfills by using data from a city-wide Participatory Sensing campaign, titled the "Trash Campaign," that was conducted at a number of high schools in the Los Angeles Unified School District (LAUSD).

The task:

The LACSD is planning a public awareness campaign and wants to ask the public to take specific steps that will help reduce the landfill burden. Based on the data collected, they would like you to **make one or two recommendations** that would reduce the use of the regional landfills.

Specifically, they have asked your team to compose a letter in which you answer the following questions:

1. *What is/are the specific recommendation(s) you are proposing for the public awareness campaign?*
2. *Why do you think this will work? What evidence do you have to support this? Include any necessary plots and analyses.*

The data:

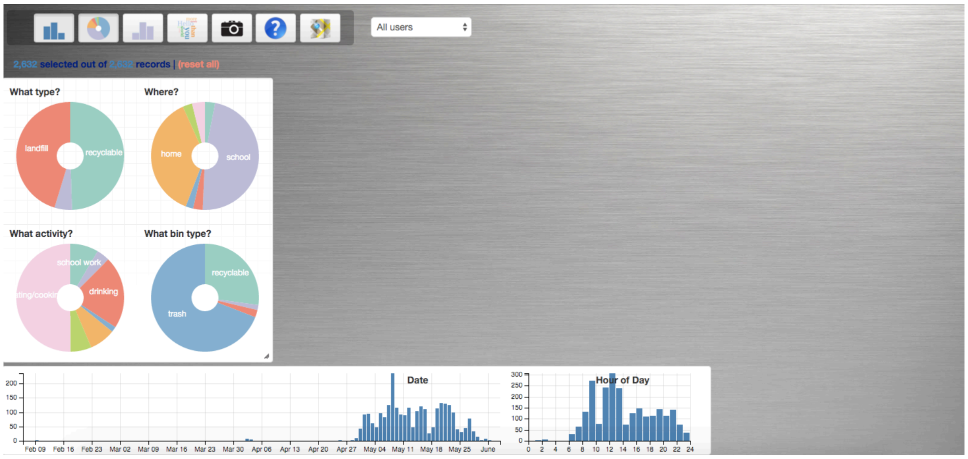
The survey questions/prompts for the Trash Campaign are provided below for your reference. The data can be found via the IDS public dashboard (https://portal.idsucla.org/) and can also be exported to RStudio.

Survey Question/Prompt	Variable Name	Data Type
1. Please take a photo of your trash.	photo	photo
2. Please describe your trash.	whatTrash	text
3. What type of trash? □ recyclable □ landfill □ compost	type	category
4. Where was this trash generated/found? □ home □ restaurants □ in transit □ school □ stores/malls □ others □ work	where	category
5. What activity generated this trash? □ eating/cooking □ cleaning □ I found it □ drinking □ shopping □ other □ school work	activity	category
6. Where did you put this trash when you were done? □ recyclable □ compost/green waste □ trash □ litter	receptacle	category
7. How many recycling bins can you see from your location?	howManyRecycle	number
8. How many trash/landfill bins can you see from your location?	numberTrashBins	number
9. How many compost/green waste bins can you see from your location?	numberCompostBins	number
AUTOMATIC	location	latitude, longitude
AUTOMATIC	timestamp	date, time

LMR_4.17

6. Once all students have read the assignment, use the following questions to check for understanding of what the task is:
   a. What organization is asking for your help? ***The Los Angeles County Sanitation District (LACSD).***
   b. What type of data did the organization collect, and whom did they collect it from? ***Participatory Sensing data via the Trash campaign. The campaign was city-wide and taken by high school students in LAUSD.***
   c. How many recommendations will you present to the organization? ***One or two.***
   d. What does the organization hope to do with your recommendations? ***Create a public awareness campaign to help reduce the burden on landfills.***

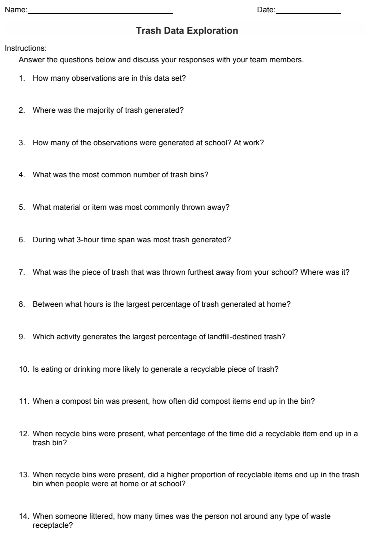
7. At this point, students will begin exploring the data via the IDS public dashboard: https://portal.idsucla.org/
8. They should use the “Trash” campaign data and select “Dashboard” from the “Action” button.
9. The dashboard is a visual tool for exploring and analyzing data. An example screenshot of the Trash campaign in the dashboard is shown below.



All users

2,632 selected out of 2,632 records | (reset all)

10. Students do not need to complete any analyses during today’s lesson. Instead, they should simply “play” with the data and brainstorm possible statistical questions that will help them complete the activity.
11. To assist students’ interaction with the dashboard, distribute the *Trash Dashboard Exploration* handout (LMR_4.18_Trash Data Exploration).



Name:________________________________ Date:_____________

Trash Data Exploration

Instructions:

Answer the questions below and discuss your responses with your team members.

1. How many observations are in this data set?
2. Where was the majority of trash generated?
3. How many of the observations were generated at school? At work?
4. What was the most common number of trash bins?
5. What material or item was most commonly thrown away?
6. During what 3-hour time span was most trash generated?
7. What was the piece of trash that was thrown furthest away from your school? Where was it?
8. Between what hours is the largest percentage of trash generated at home?
9. Which activity generates the largest percentage of landfill-destined trash?
10. Is eating or drinking more likely to generate a recyclable piece of trash?
11. When a compost bin was present, how often did compost items end up in the bin?
12. When recycle bins were present, what percentage of the time did a recyclable item end up in a trash bin?
13. When recycle bins were present, did a higher proportion of recyclable items end up in the trash bin when people were at home or at school?
14. When someone littered, how many times was the person not around any type of waste receptacle?

LMR_4.18

12. Leave 10-15 minutes at the end of class to share out and discuss some of these statistical questions. During this time, the teacher should also check for data understanding.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Lesson 16: Exploring Trash via the Dashboard

Objective:

Students will continue to investigate landfills and perform analyses via the IDS public dashboard.

Materials:

1. Computers
2. IDS public dashboard: https://portal.idsucla.org/

Essential Concepts: Exploring the IDS Dashboard provides a visual approach to data analysis.

Lesson:

1. Today students will continue their data exploration of the Trash campaign via the IDS public dashboard.
2. As a team, students should select statistical questions and provide appropriate plots and summaries from the dashboard to answer those questions.
3. Leave 10-15 minutes at the end of class to share out some of the findings from each student team.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students will brainstorm possible RStudio commands to complement their initial analyses from the dashboard. It is up to the teacher to ask for a minimum number of commands from each student.

Lesson 17: Exploring Trash via RStudio

Objective:

Students will continue to investigate landfills and perform analyses via RStudio.

Materials:

1. Computers
2. RStudio

Essential Concepts: RStudio can be used to verify initial results/findings from data analysis done via the IDS Dashboard.

Lesson:

1. Today students will continue their data analysis of the Trash campaign via RStudio.
2. They should share their answers from the previous lesson's homework assignment in their teams to help them get started with their code. After having shared their initial code, they should spend some time discussing other ideas
3. The Recorder/Reporter should keep a list of the code that the team has agreed to use.
4. By the end of class, students should begin writing their recommendations for reducing the burden on landfills.
5. Inform students that each team will prepare their presentations during the next class period.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next 2 Days

☑ Students will finalize their recommendations for reducing the burden on landfills and have a draft of their letter to send to LACSD and prepare for their team presentations.

Decisions, Decisions

Instructional Days: 3

Enduring Understandings

Decision trees are used to classify observations into similar groupings based on known characteristics. Yes/no questions are asked, then the observations are sorted based on the responses to the questions. After a specified number of iterations, a final group membership is decided. One particular modeling tool we use for decision trees is known as CART (Classification and Regression Trees).

Engagement

Students will participate in the *CART Activity* described in Lesson 12. They will classify football and soccer players into categories based on player characteristics.

Learning Objectives

Statistical/Mathematical:

S-IC 2: Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation.

Data Science:

Understand that classification and regression trees can be used to predict membership in groups.

Applied Computational Thinking using RStudio:

- Create classification and regression trees.

Real-World Connections:

Cardiologists may use a decision tree to diagnose whether people are or are not having a heart attack. Since the late 1870’s, this method has been found to correctly diagnose a heart attack in over 95% of cases compared to correct diagnoses based on individual doctors’ expertise, which ranged between 75 and 90%.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it is used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

Data File or Data Collection Method

Data File:

1. Titanic: `data(titanic)`

Legend for Activity Icons

🎥	💬	📖	☑	📜
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 18: Grow Your Own Decision Tree

Objective:

Students will learn what decision trees look like and how they can be used to classify people or objects into groups. They will engage in an activity to see how making slight changes to the tree can lead to drastic rises or reductions in misclassifications.

Materials:

1. *CART Activity Player Stats* (LMR_4.19_CART Player Stats)
2. *CART Activity Round 1 Questions* (LMR_4.20_CART Round 1)
3. *CART Activity Round 2 Questions* (LMR_4.21_CART Round 2)

Advanced preparation required (see Step 8 below)

Vocabulary:

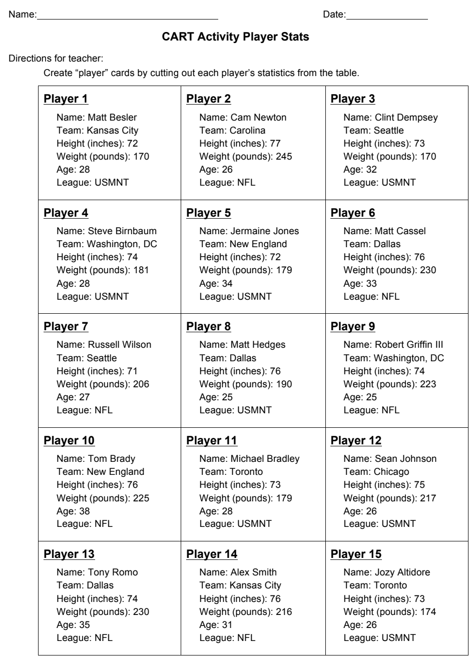
classify, decision tree, Classification and Regression Trees (CART), nodes, misclassifications

> **Essential Concepts**: Some trends are not linear, so the approaches we've done so far won't be helpful. We need to model such trends differently. Decision trees are a non-linear tool for classifying observations into groups when the trend is non-linear.

Lesson:

1. Inform students that, during today's lesson, they will be participating in an activity to try to **classify** professional athletes into one of two groups: (1) soccer players on the US Men's National Team, OR (2) football players in the National Football League (NFL).
2. Remind students that this unit has focused on linear models and making predictions. In the real world, data can be modeled in a variety of ways, many of which are non-linear, and because of this, we can't easily write down a mathematical equation to help us make predictions. However, we can use what we have learned so far to determine whether or not other models can provide a good fit to the data.
3. Introduce the topic of **decision trees** and explain that it is simply a non-linear way to model data.
4. Explain that decision trees are "grown" by using algorithms, or rules, to test many, many different decision trees to find the one that makes the best predictions.
5. A decision tree is basically a series of questions that are asked sequentially. Observations start by answering the first question (at the root of the tree), and then proceed along the different branches based on the answers they give to the questions that follow. At the end, based on all of the questions asked, observations are then classified as one of k classifications.
6. Remind students that algorithms are a series of steps that are repeated a large number of times. For decision trees, this enables us to (1) explore many possible paths, beginning from the same initial point, or (2) find different starting points based on where we ended during the previous iteration.
7. Ask students to recall that they created and worked with *linear models* earlier in the unit. We are continuing our work with models and will learn another method of modeling called **CART**, which stands for **Classification and Regression Trees**. This is another name for decision trees.
8. CART Activity: to get a sense of how decision trees work, the students will see one in action. We are going to try to classify 15 professional athletes into either soccer or football players based on some of their characteristics.

   Note: Advanced preparation required. The cards in each of the LMRs listed above (and displayed below) need to be cut out prior to class time.



Name:________________________________ Date:_______________

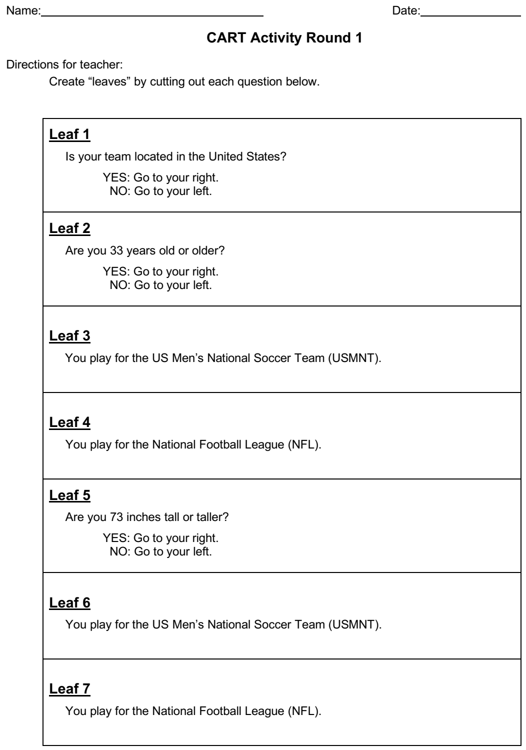
CART Activity Player Stats

Directions for teacher:

Create “player” cards by cutting out each player’s statistics from the table.

Player 1 Name: Matt Besler Team: Kansas City Height (inches): 72 Weight (pounds): 170 Age: 28 League: USMNT	**Player 2** Name: Cam Newton Team: Carolina Height (inches): 77 Weight (pounds): 245 Age: 26 League: NFL	**Player 3** Name: Clint Dempsey Team: Seattle Height (inches): 73 Weight (pounds): 170 Age: 32 League: USMNT
Player 4 Name: Steve Birnbaum Team: Washington, DC Height (inches): 74 Weight (pounds): 181 Age: 28 League: USMNT	**Player 5** Name: Jermaine Jones Team: New England Height (inches): 72 Weight (pounds): 179 Age: 34 League: USMNT	**Player 6** Name: Matt Cassel Team: Dallas Height (inches): 76 Weight (pounds): 230 Age: 33 League: NFL
Player 7 Name: Russell Wilson Team: Seattle Height (inches): 71 Weight (pounds): 206 Age: 27 League: NFL	**Player 8** Name: Matt Hedges Team: Dallas Height (inches): 76 Weight (pounds): 190 Age: 25 League: USMNT	**Player 9** Name: Robert Griffin III Team: Washington, DC Height (inches): 74 Weight (pounds): 223 Age: 25 League: NFL
Player 10 Name: Tom Brady Team: New England Height (inches): 76 Weight (pounds): 225 Age: 38 League: NFL	**Player 11** Name: Michael Bradley Team: Toronto Height (inches): 73 Weight (pounds): 179 Age: 28 League: USMNT	**Player 12** Name: Sean Johnson Team: Chicago Height (inches): 75 Weight (pounds): 217 Age: 26 League: USMNT
Player 13 Name: Tony Romo Team: Dallas Height (inches): 74 Weight (pounds): 230 Age: 35 League: NFL	**Player 14** Name: Alex Smith Team: Kansas City Height (inches): 76 Weight (pounds): 216 Age: 31 League: NFL	**Player 15** Name: Jozy Altidore Team: Toronto Height (inches): 73 Weight (pounds): 174 Age: 26 League: USMNT

LMR_4.19



Name:________________________________ Date:_______________

CART Activity Round 1

Directions for teacher:

Create “leaves” by cutting out each question below.

Leaf 1

Is your team located in the United States?

YES: Go to your right.
NO: Go to your left.

Leaf 2

Are you 33 years old or older?

YES: Go to your right.
NO: Go to your left.

Leaf 3

You play for the US Men’s National Soccer Team (USMNT).

Leaf 4

You play for the National Football League (NFL).

Leaf 5

Are you 73 inches tall or taller?

YES: Go to your right.
NO: Go to your left.

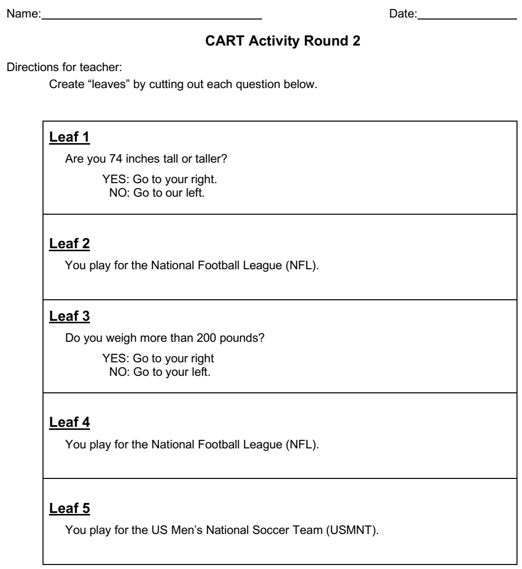
Leaf 6

You play for the US Men’s National Soccer Team (USMNT).

Leaf 7

You play for the National Football League (NFL).

LMR_4.20



Name:______________________________ Date:______________

CART Activity Round 2

Directions for teacher:

Create "leaves" by cutting out each question below.

Leaf 1

Are you 74 inches tall or taller?

YES: Go to your right.
NO: Go to our left.

Leaf 2

You play for the National Football League (NFL).

Leaf 3

Do you weigh more than 200 pounds?

YES: Go to your right
NO: Go to your left.

Leaf 4

You play for the National Football League (NFL).

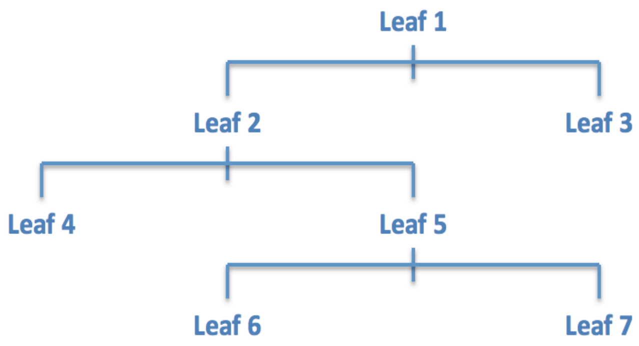
Leaf 5

You play for the US Men's National Soccer Team (USMNT).

LMR_4.21

9. Ask for 15 volunteers and hand each of them a data card from the *CART Activity Player Stats* handout (LMR_4.19). These students will be known as the "players." Each card lists the following variables for 15 different professional athletes:
   a. team location
   b. name
   c. age
   d. height (in inches)
   e. weight (in pounds)
   f. league
10. The "players" will only be allowed to say "yes" or "no" in this activity. No other talking is permitted.
11. Now, ask for 7 additional volunteers to be the ***nodes***, or *leaves*, on the decision tree. Each student will be known as a "leaf."
12. Distribute one question/classification from the *CART Activity Round 1 Questions* (LMR_4.20) to each "leaf."
13. Arrange the 7 "leaves" in the room as depicted by the graphic below:

Round 1 Tree Diagram

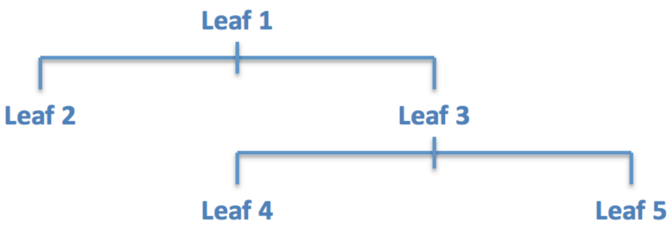


14. Now, each “player,” one at a time, will approach *Leaf 1*, who will ask the “player” the question listed on his/her card. Depending on the player’s answer, *Leaf 1* will direct the “player” to the next “leaf.”
15. The “player” continues through the nodes until a “leaf” declares the “player” to be either (1) a soccer player on the US Men’s National Team, OR (2) a football player in the National Football League (NFL).
16. Allow all the “players” to go through the “leaves” until each one is classified as either a soccer or football player.
17. After each player has been classified, tally the number of correct and incorrect classifications and display a simple table (see example below) on the board.

	Classified Correctly	**Classified Incorrectly**
USMNT Soccer Player		
NFL Football Player		

18. Ask students to calculate the misclassification rate (MCR) which is the proportion of observations who were predicted to be in one category but were actually in another. If all the activity player stats cards were used, the misclassification rate would be 5/15.
19. After proceeding through “Round 1,” ask an additional 5 students to come up as more “leaves,” distribute the cards from the CART Activity Round 2 Questions file (LMR_4.21), and arrange the students like the diagram below:

Round 2 Tree Diagram



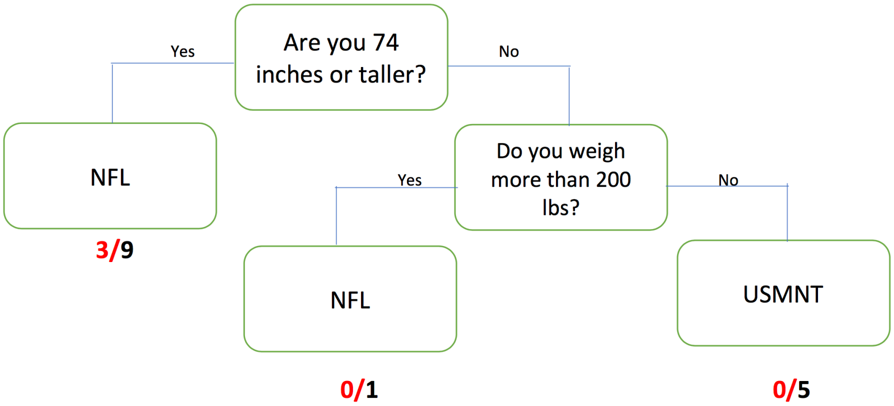
20. Have each “player” go through this new set of “leaves” until they are re-classified by these new rules.
21. Again, tally the number of correct and incorrect classifications and display them on the board and calculate the misclassification rate. If all the activity player stats cards were used, the misclassification rate would be 3/15.
22. Once the activity has been completed, ask students the following questions:
    a. How do decision trees classify objects/people as being a member of a group? ***By asking a series of questions, one at a time, and sending the participant down a particular path until he/she is classified.***
    b. Did we do as well, worse, or better in Round 2 compared to Round 1 at correctly guessing which sport the “players” participate in? Explain. ***Answers will vary according to results of the activity***.
    c. How can we figure out what questions to ask and in what order to minimize the number of **misclassifications**? (This one might not be obvious. The point is for the students to wrestle with how they might think it can be done.)

23. Also have the students discuss the following questions:

    a. How is a decision tree/CART similar to or different than a linear model?

    b. Can we really call a decision tree a model? Why or why not?

24. In lab 4G you will use RStudio to create tree models that will make good predictions without needing a lot of branches. RStudio can also calculate the misclassification rate. However, you might find the visual a little confusing to interpret, so we will use the Round 2 classification tree to see what the output from RStudio might look like.



25. Project the image above and explain to the students that if all 15 of the activity player stats cards were used, then RStudio would give ratios for each of the leaves where a classification was made. The denominator tells us how many observations ended up in that leaf and the numerator tells us the number of misclassifications. Ask students:

    a. What does the output 3/9 represent? ***Answer: The 9 tells us that nine players were classified as NFL players based solely on the fact that they were taller than 74 inches. The 3 represents soccer players that were misclassified so 6 football players were classified correctly.***

    b. How would you calculate the misclassification rate (MCR)? ***Answer: You would add all of the numerators which represent the misclassifications and divide by the total number of observations which you could obtain by adding all the denominators. (3+0+0)/(9+1+5)=3/15.***

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework

Students will record their responses to the following discussion questions about CARTs and submit them the following day:

a. How is a decision tree/CART similar to or different than a linear model?
b. Can we really call a decision tree a model? Why or why not?

Lesson 19: Data Scientists or Doctors?

Objective:

Students will create their own decision trees based on training data (i.e., the data from the previous day's lessons), and then see how well their decision tree works on new test data.

Materials:

1. *Decision Tree for Heart Attack Risk* graphic (LMR_4.22_CART Heart Attacks)
2. *Make Your Own Decision Tree* handout (LMR_4.23_Your Own Decision Tree)

Vocabulary:

training data, testing data

> **Essential Concepts**: We can determine the usefulness of decision trees by comparing the number of misclassifications in each.

Lesson:

1. Ask students the following question:

   If a close friend or family member were having chest pains, would you want to take that person to a doctor or to a data scientist?

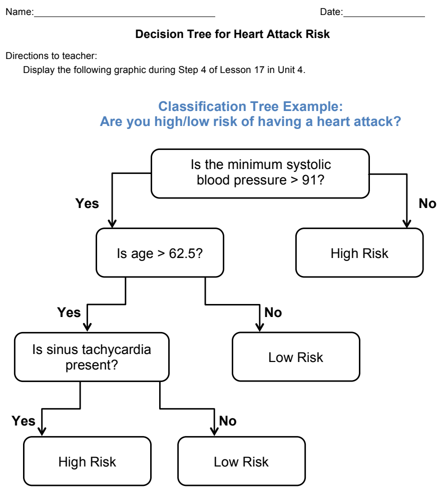
2. Give the students some time to think about the question and have a few of them share out their responses with the class.

   Note: It's likely that most students will choose to bring their loved one to a doctor.

3. As it turns out, back in the late 1970s, a cardiologist (and early data scientist) named Lee Goldman developed a decision tree based on millions of patient observations. The decision tree was made to diagnose whether people were or were not having a heart attack. Interestingly, the results of the decision tree compared to how actual doctor diagnoses are shown below:

   a. Correct diagnoses using the decision tree were above 95%.
   b. Correct diagnoses based on individual doctors' expertise? Anywhere between 75-90%.

4. Display the graphic from the *Decision Tree for Heart Attack Risk* file (LMR_4.22_CART Heart Attacks) and explain that this is one example of what the decision tree might have looked like. **Note:** This is NOT the actual tree Goldman developed.



Name:______________________________ Date:______________

Decision Tree for Heart Attack Risk

Directions to teacher:

Display the following graphic during Step 4 of Lesson 17 in Unit 4.

Classification Tree Example:
Are you high/low risk of having a heart attack?
Is the minimum systolic blood pressure > 91?
Yes
No
Is age > 62.5?
High Risk
Yes
No
Is sinus tachycardia present?
Low Risk
Yes
No
High Risk
Low Risk

LMR_4.22

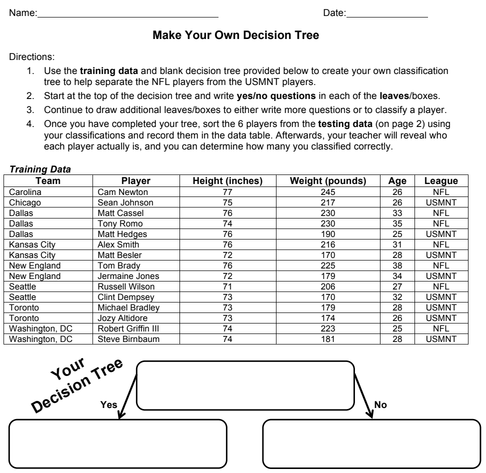
5. Using a *Pair-Share*, ask students to discuss the following questions using the graphic above, as well as what they learned during the previous lesson's activity.

    a. What are decision trees?
    b. How do they work at classifying data into groups?

6. Then display the following data (the same data from the player cards used in the previous lesson):

Team	Player	Height (inches)	Weight (pounds)	Age	League
Carolina	Cam Newton	77	245	26	NFL
Chicago	Sean Johnson	75	217	26	USMNT
Dallas	Matt Cassel	76	230	33	NFL
Dallas	Tony Romo	74	230	35	NFL
Dallas	Matt Hedges	76	190	25	USMNT
Kansas City	Alex Smith	76	216	31	NFL
Kansas City	Matt Besler	72	170	28	USMNT
New England	Tom Brady	76	225	38	NFL
New England	Jermaine Jones	72	179	34	USMNT
Seattle	Russell Wilson	71	206	27	NFL
Seattle	Clint Dempsey	73	170	32	USMNT
Toronto	Michael Bradley	73	179	28	USMNT
Toronto	Jozy Altidore	73	174	26	USMNT
Washington, D.C.	Robert Griffin III	74	223	25	NFL
Washington, D.C.	Steve Birnbaum	74	181	28	USMNT

7. Distribute the *Make Your Own Decision Tree* handout (LMR_4.23_Your Own Decision Tree) and give students time to come up with their own decision trees based on the **training data** they are given. Students may work in pairs or teams. They should follow the directions on page 1 of the handout and come up with a series of possible yes/no questions that they could ask to classify each player into his correct league (the NFL or the USMNT).



Name:______________________________ Date:_______________

Make Your Own Decision Tree

Directions:

1. Use the **training data** and blank decision tree provided below to create your own classification tree to help separate the NFL players from the USMNT players.
2. Start at the top of the decision tree and write **yes/no questions** in each of the **leaves**/boxes.
3. Continue to draw additional leaves/boxes to either write more questions or to classify a player.
4. Once you have completed your tree, sort the 6 players from the **testing data** (on page 2) using your classifications and record them in the data table. Afterwards, your teacher will reveal who each player actually is, and you can determine how many you classified correctly.

Training Data

Team	Player	Height (inches)	Weight (pounds)	Age	League
Carolina	Cam Newton	77	245	26	NFL
Chicago	Sean Johnson	75	217	26	USMNT
Dallas	Matt Cassel	76	230	33	NFL
Dallas	Tony Romo	74	230	35	NFL
Dallas	Matt Hedges	76	190	25	USMNT
Kansas City	Alex Smith	76	216	31	NFL
Kansas City	Matt Besler	72	170	28	USMNT
New England	Tom Brady	76	225	38	NFL
New England	Jermaine Jones	72	179	34	USMNT
Seattle	Russell Wilson	71	206	27	NFL
Seattle	Clint Dempsey	73	170	32	USMNT
Toronto	Michael Bradley	73	179	28	USMNT
Toronto	Jozy Altidore	73	174	26	USMNT
Washington, DC	Robert Griffin III	74	223	25	NFL
Washington, DC	Steve Birnbaum	74	181	28	USMNT

LMR_4.23

8. Once the students have finished creating their decision trees, ask the following questions:

    a. Will you be able to classify other players from a new data set correctly using this particular decision tree?
    b. Is this decision tree too specific to the training data?

9. Inform the students that they should now use the **testing data** on page 2 of the handout to try to classify 5 *mystery players* into one of the two leagues. They should record the classification that their tree outputs in the data table on page 2.

10. Let the students compare their decision trees and league assignments with one another. Hopefully, there will be a bit of variety in terms of the trees and the classifications.
11. Next, show students the correct league classifications for the 5 mystery players. The mystery player names are also included in this table.

Team	Player	Height (inches)	Weight (pounds)	Age	League
Toronto	Michael Bradley	74	175	28	USMNT
New York	Eli Manning	76	218	34	NFL
New Orleans	Drew Brees	72	209	36	NFL
Washington, DC	Perry Kitchen	72	160	23	USMNT
New England	Lee Nguyen	68	150	29	USMNT

12. By a show of hands, ask:
    a. How many students misclassified all of the players in the testing data?
    b. How many misclassified 4 of the 5 players?
    c. How many misclassified 3 of the 5 players?
    d. How many misclassified 2 of the 5 players?
    e. Did anyone correctly classify ALL 5 mystery players? If so, ask those students to share their decision trees with the rest of the class.
13. Inform students that, when faced with much more data, creating classification trees becomes much harder to make by hand. It is so difficult, in fact, that data scientists rely on software to grow their trees for them. Students will learn how to create decision trees in RStudio during the next lab.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

☑ Write a paragraph describing the role testing data and training data play in creating a classification tree.

LAB 4G: Growing Trees

Complete Lab 4G prior to Lesson 20.

Lab 4G - Growing trees

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Trees vs. Lines

- So far in the labs, we've learned how we can fit linear models to our data and use them to make predictions.
- In this lab, we'll learn how to make predictions by growing trees.
  - Instead of creating a line, we split our data into branches based on a series of *yes* or *no* questions.
  - The branches help sort our data into *leaves* which can then be used to make predictions.
- Start, by loading the `titanic` data.

Our first tree

- Use the `tree()` function to create a *classification* tree that predicts whether a person `survived` the Titanic based on their `gender`.
  - A *classification* tree tries to predict which category a categorical variable would belong to based on other variables.
  - The syntax for `tree` is similar to that of the `lm()` function.
  - Assign this model the name `tree1`.
- **Why can't we just use a *linear model* to predict whether a passenger on the Titanic `survived` or not based on their `gender`?**

Viewing trees

- To actually look at and interpret our `tree1`, place the model into the `treeplot` function.
  - **Write down the labels of the two *branches*.**
  - **Write down the labels of the two *leaves*.**
- Answer the following, based on the `treeplot`:
  - **Which `gender` does the model predict will survive?**
  - **Where does the plot tell you the number of people that get sorted into each leaf? How do you know?**
  - **Where does the plot tell you the number of people that have been sorted *incorrectly* in each leaf?**

Leafier trees

- Similar to how you included multiple variables for a linear model, create a `tree` that predicts whether a person `survived` based on their `gender`, `age`, `class`, and where they `embarked`.
  - Call this model `tree2`.
- Create a `treeplot` for this model and answer the following question:
  - **Mrs. Cumings was a 38 year old female with a 1st class ticket from Cherbourg. Does the model predict that she survived?**
  - **Which variable ended up not being used by `tree`?**

Tree complexity

- By default, the `tree()` function will fit a *tree model* that will make good predictions without needing lots of branches.
- We can increase the complexity of our trees by changing the complexity parameter, `cp`, which equals `0.01` by default.
- We can also change the minimum number of observations needed in a leaf before we split it into a new branch using `minsplit`, which equals `20` by default.
- Using the same variables that you used in `tree2`, create a model named `tree3` but include `cp = 0.005` and `minsplit = 10` as arguments.
  - **How is `tree3` different from `tree2`?**

Misclassification rate

- Similar to how we use the *mean squared error* to describe how well our model predicts numerical variables, we use the *misclassification rate* to describe how well our model predicts categorical variables.
  - The *misclassification rate* (MCR) is the number of people who were predicted to be in one category but were actually in another.
  - Fill in the blanks to create a function to calculate the MCR

```
calc_mcr <- function(actual, predicted) {
  sum(____ != ____) / length(____)
}
```

Predictions and Cross-validation

- Just like with *linear models*, we can use cross-validation to measure our *classification trees* prediction accuracy.
  - Use the `data` function to load the `titanic_test` data.
  - Fill in the blanks below to predict whether people in the `titanic_test` data survived or not using `tree1`.

```
titanic_test <- mutate(____, prediction = predict(____, newdata = ____, t
ype = "class"))
```

- Then run the following to calculate the MCR

```
summarize(titanic_test, mcr = calc_mcr(survived, prediction))
```

On your own

- **In your own words, explain what the *misclassification rate* is.**
- **Which model (`tree1`, `tree2` or `tree3`) had the lowest misclassification rate for the `titanic_test` data?**
- Create a 4th model using the same variables used in `tree2`. This time though, change the *complexity parameter* to `0.0001`. Then answer the following
  - **Does creating a more complex *classification tree* always lead to better predictions? Why not?**
- A *regression tree* is a tree model that predicts a numerical variable. Create a *regression tree* model to predict the Titanic's passenger's ages and calculate the MSE.
  - Plots of regression trees are often too complex to plot.

Ties that Bind

Instructional Days: 3

Enduring Understandings

Clustering is another way to classify data into groups. We classify observations based on numerical characteristics and their similarities. We use k-means to determine the mean value for each group of k clusters by randomly assigning an initial value for the mean and then moving the mean based on its proximity to the points.

Networks classify people into groupings based on who knows whom. Nodes are formed when a relationship between two people is present.

Engagement

Students will participate in the *Find the Clusters Activity* described in Lesson 14. They will determine which points in a plot should be grouped as football players and which points should be grouped as swimmers.

Learning Objectives

Statistical/Mathematical:

S-IC 2: Decide if a specified model is consistent with results from a given data-generating process, e.g., using simulation.

Data Science:

Understand what RStudio is doing when using the k-means function to find clusters in a group of data and when creating networks in order to learn how to classify data into groups.

Applied Computational Thinking using RStudio:

- Use the k-means function to find clusters in a group of data.
- Plot the data with the cluster assignments based on the k-means function.

Real-World Connections:

Network analysis is used by many private and public entities such as the National Security Agency when they want to find terrorist networks to have maximum impact on communications. The k-means algorithm is a technique for grouping entities according to the similarity of their attributes. For example, dividing countries into similar groups using k-means to make fair comparisons is applicable.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will use complex sentences to write informative short reports that use data science concepts and skills.

Legend for Activity Icons

🎥	💬	📖	☑	✍
Video clip	Discussion	Articles/Reading	Assessments	Class Scribes

Lesson 20: Where Do I Belong?

Objective:

Students will learn what clustering is and how to classify groups of people into clusters based on unknown similarities.

Materials:

1. *Find the Clusters* handout (LMR_4.24_Find the Clusters)

Vocabulary:

clustering, cluster, k-means

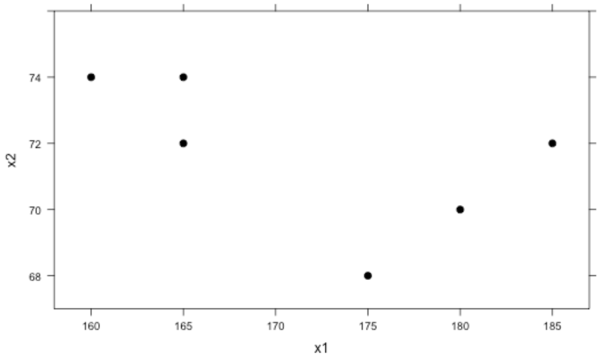
Essential Concepts: We can identify groups, or "clusters," in data based on a few characteristics. For example, it is easy to classify a classroom into males and females, but what if you only knew each person's arm span? How well could you classify their genders now?

Lesson:

1. Inform the students that they will continue to explore different types of models, and today they will be focusing on **clustering**. Clustering is the process of grouping a set of objects (or people) together in such a way that people in the same group (called a **cluster**) are more similar to each other than to those in other groups.
2. Have the students recall that, in the previous lessons, they used decision trees and CART to classify people into different groups based on whether or not a person had a specific characteristic (e.g., whether or not a professional athlete's team is based in the US).
3. But, sometimes we don't know what these specific characteristics are. We are simply given numerical variables and asked to find similarities. This is where clustering comes in – similar people will congregate towards each other, and we want to see if we can identify their groupings.
4. We will look at a very basic example first. Suppose the following 6 observations are given:

Obs	X_1	X_2
1	160	74
2	165	72
3	165	74
4	175	68
5	180	70
6	185	72

5. Plot the X_1 and X_2 points on a scatterplot either on the board or on poster paper (X_1 can be on the horizontal axis and X_2 can be on the vertical axis). The graph should look like the one below:

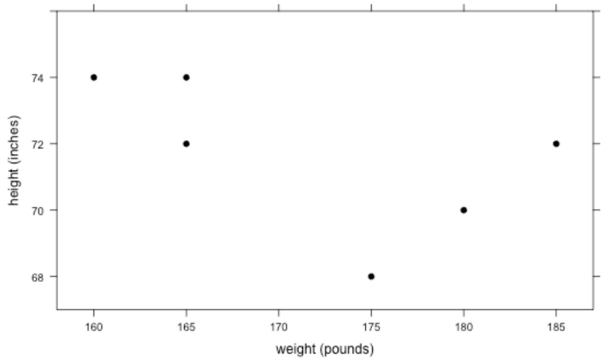


6. Ask students if they think there are any clusters, or groups, that stand out to them. It is likely that they will say there are 2 clusters in the graph: the top left corner 3 points, and the bottom right 3 points.

7. Now pose the following scenario that further describes the data:

    a. A doctor provides yearly physicals to the men's football and men's swimming teams at a local high school.

    b. He has collected data over the past few years on each player's weight (in pounds) and height (in inches). He informs us that weight was coded as the variable X_1, and height was coded as the variable X_2. You can re-label the scatterplot with this new information.



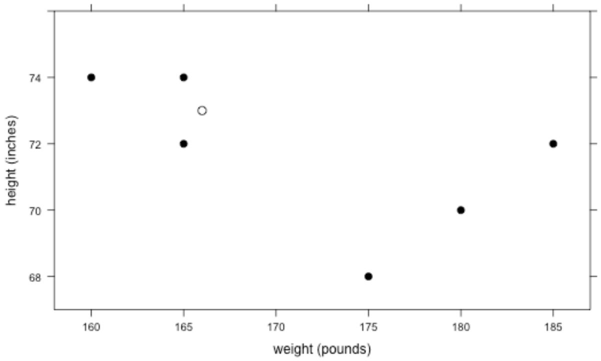
    c. Unfortunately, the doctor never recorded what sport each person played.

8. Using the information about height and weight, ask the students to decide:

    a. Which group of points most likely represents players from the swimming team? ***The points in the upper left corner are probably swimmers because swimmers are usually tall (and have large arm spans) and thin.***

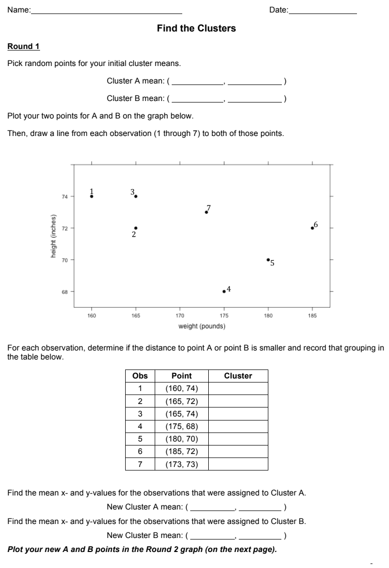
    b. Which group of points most likely represents players from the football team? ***The points in the bottom right corner are probably football players because they tend to be heavier and more muscular.***

9. Now suppose a new player comes into the doctor's office for a physical. His weight and height are recorded as 166 pounds and 73 inches, respectively, but the doctor forgets to ask what sport he plays. Plot this point on the graph and ask students to determine which sport they think this student plays. ***This student is most likely a swimmer because he is tall and thin, and his point is near the swimming cluster.***



10. That was an easy one! But what if a player comes in and has the following measurements: weight = 173 pounds, height = 73 inches?

11. Distribute the *Find the Clusters* handout (LMR_4.24) and tell the students that the new point has been added to the "Round 1" graph.



Name:______________________________ Date:______________

Find the Clusters

Round 1

Pick random points for your initial cluster means.

Cluster A mean: (___________, ___________)

Cluster B mean: (___________, ___________)

Plot your two points for A and B on the graph below.

Then, draw a line from each observation (1 through 7) to both of those points.

For each observation, determine if the distance to point A or point B is smaller and record that grouping in the table below.

Obs	Point	Cluster
1	(160, 74)	
2	(165, 72)	
3	(165, 74)	
4	(175, 68)	
5	(180, 70)	
6	(185, 72)	
7	(173, 73)	

Find the mean x- and y-values for the observations that were assigned to Cluster A.

New Cluster A mean: (__________, __________)

Find the mean x- and y-values for the observations that were assigned to Cluster B.

New Cluster B mean: (__________, __________)

Plot your new A and B points in the Round 2 graph (on the next page).

LMR_4.24

12. Ask students:

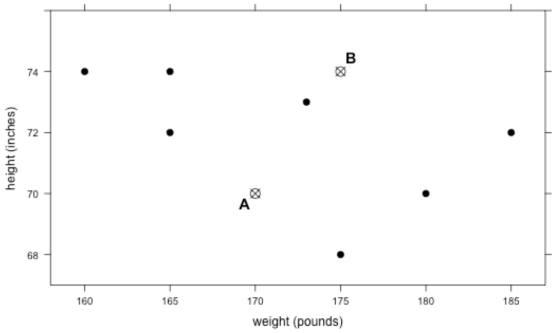
    a. On which team do you think this person plays? ***It is much more difficult to tell now because it looks like it is right in between the two clusters.***

13. In order to determine group placement, we can use an algorithm called **k-means clustering**. With this method, we select k clusters that we want to identify. Since we know we only have 2 types of athletes, football players and swimmers, we will be finding k = 2 clusters.

14. To introduce the students to this idea, circle the 3 points in the upper left corner (the ones that are likely the swimmers) and have students find the "mean point." This means that they should find the mean x-value and the mean y-value of the 3 points. They can then plot this new point and use it as the mean of this particular group, or cluster.

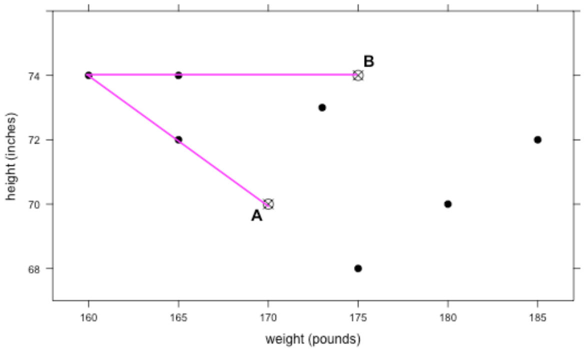
15. The goal of this algorithm is to keep recalculating means as the possible clusters change. To begin, we will randomly pick 2 arbitrary points on the plot (we can call them A and B) to be our starting means for each cluster. There is no incorrect way to pick the starting means, but the further away the means are from the actual points, the longer it will take the algorithm to complete. If you would like to use the point found in Step 14 and label it as "A," that is completely fine. You can simply pick just one other random point and label it as "B."

16. For now, we will start with the following two points as guesses for the means of each group: A: (170, 70) and B: (175, 74). In the "Round 1" plot on the *Find the Clusters* handout, each student should plot and label these two points.

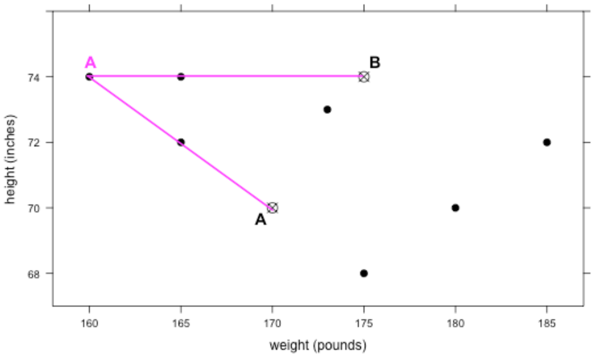


17. Inform the students that they will be drawing lines from each original point to both means. Then, they will decide if the point is closer to mean A or mean B and label the point with that letter.

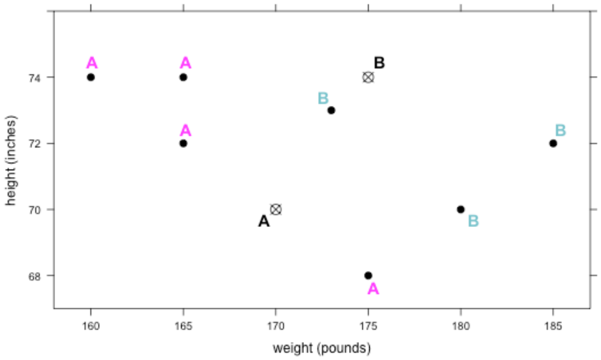
Lines have been draw from the top left point to the means in the plot below as a guide. You can draw this on the board as a reference for the students as well.



18. Since the line to point A is smaller, we would classify that point as being in cluster A (as shown below).



19. The students should draw similar lines for every point on the graph and make a decision as to which cluster each belongs in. They can simply eyeball it. Even if they guess incorrectly, the algorithm should be able to find the correct groups after some time. The correct classifications for Round 1 are as follows:



20. Once the class has agreed on the first round's cluster classifications, they should compute new values for the k-means (A and B). For mean A, they simply need to find the mean x-value for the 4 points and the mean y-value for the 4 points. The new means for A and B have been calculated below. The students should be calculating these on their own and recording their new means on the handout.

x-value for A = (160 + 165 + 165 + 175)/4 = 166.25
y-value for A = (74 + 72 + 74 + 68)/4 = 72

x-value for B = (173 + 180 + 185)/3 = 179.3
y-value for B = (73 + 70 + 72)/3 = 71.67

new A = (166.25, 72)
new B = (179.3, 71.67)

21. Have the students continue working through the handout until the cluster membership remains the same between 2 consecutive rounds. This means that, from one iteration to the next, the points in each cluster do not change.

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Homework & Next Day

☑ Write a paragraph that describes k-means clustering in your own words.

LAB 4H: Finding Clusters

Complete Lab 4H prior to Lesson 21.

Lab 4H - Finding clusters

Directions: Follow along with the slides and answer the questions in **bold** font in your journal.

Clustering data

- We've seen previously that data scientists have methods to predict values of specific variables.
  - We used *regression* to predict numerical values and *classification* to predict categories.
- *Clustering* is similar to classification in that we want to group people into categories. But there's one important difference:
  - In *clustering*, we don't know how many groups to use because we're not predicting the value of a known variable!
- In this lab, we'll learn how to use the k-means clustering algorithm to group our data into clusters.

The k-means algorithm

- The k-means algorithm works by splitting our data into *k* different clusters.
  - The number of clusters, the value of *k*, is chosen by the data scientist.
- The algorithm works *only* for numerical variables and *only* when we have no missing data.
- To start, use the `data` function to load the `futbol` data set.
  - This data contains 23 players from the US Men's National Soccer team (USMNT) and 22 quarterbacks from the National Football League (NFL).
- Create a scatterplot of the players `ht_inches` and `wt_lbs` and color each dot based on the `league` they play for.

Running k-means

- After plotting the player's heights and weights, we can see that there are two clusters, or different types, of players:
  - Players in the NFL tend to be taller and weigh more than the shorter and lighter USMNT players.
- Fill in the blanks below to use k-means to cluster the same height and weight data into two groups:

```
kclusters(____~____, data = futbol, k = ____)
```

- Use this code and the mutate function to add the values from `kclusters` to the `futbol` data. Call the variable `clusters`.

k-means vs. ground-truth

- In comparing our football and soccer players, we *know* for certain which league each player plays in.
  - We call this knowledge *ground-truth*.
- Knowing the *ground-truth* for this example is helpful to illustrate how k-means works, but in reality, data-scientists would run k-means not knowing the *ground-truth*.
- **Compare the clusters chosen by k-means to the ground-truth. How successful was k-means at recovering the `league` information?**

On your own

- Load your class' `timeuse` data (remember to run `timeuse_format` so each row represents the mean time each student spent participating in the various activities).
- Create a scatterplot of `homework` and `videogames` variables.
  - Based on this graph, identify and remove any outliers by using the `subset` function.
- Use `kclusters` with `k=2` for `homework` and `videogames`.
  - **Describe how the groups differ from each other in terms of how long each group spends playing `videogames` and doing `homework.`**

Lesson 21: Our Class Network

Objective:

Students will participate in an activity to map out their own network based on acquaintances between two people.

Materials:

1. *Friend Network Graphic* (LMR_4.25_Friend Network Graphic)
2. Index cards
3. *Network Code* file (LMR_4.26_Network Code R Script)

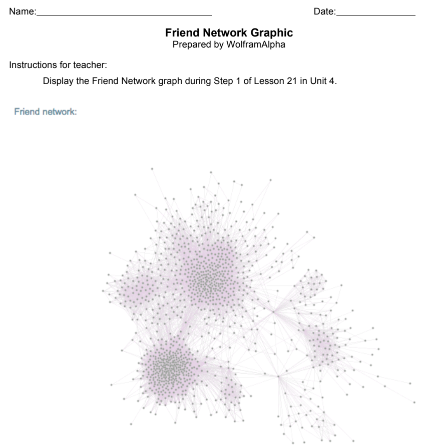
Vocabulary:

network

Essential Concepts: Networks are made when observations are interconnected. In a social setting, we can examine how different people are connected by finding relationships between other people in a network.

Lesson:

1. Display the *Friend Network Graphic* (LMR_4.25), which shows a WolframAlpha visualization of someone's Facebook friends. Inform the students that this type of model is called a **network**, which is simply a group of people or things that are interconnected in some way.



Name:______________________________ Date:______________

Friend Network Graphic

Prepared by WolframAlpha

Instructions for teacher:

Display the Friend Network graph during Step 1 of Lesson 21 in Unit 4.

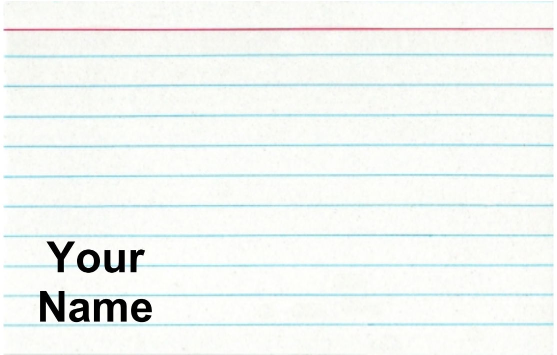
Friend network:

LMR_4.25

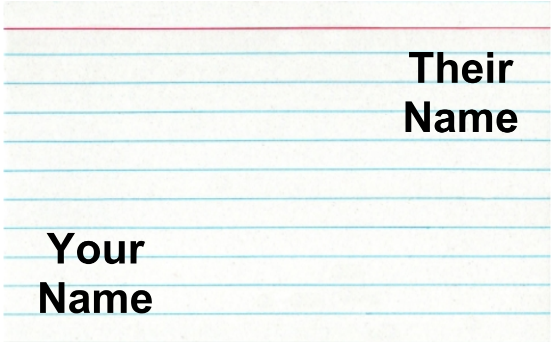
2. Ask the following questions about the graphic:

   a. What does each dot represent? ***Each dot represents one person.***
   b. What does each line represent? ***Each line represents a friendship between two people.***
   c. How are all the people in this graphic connected to each other? ***They are all friends with the person whose Facebook this is.***
   d. Why are some areas denser than others? ***A lot of people in the darker spots know each other, so there are more connections/friendships.***
   e. Why are some people not in groups at all (the dots at the edges of the graphic)? ***The main person does not have any friends in common with this person.***
   f. What might some of the groupings (the denser spots) represent? ***Answers will vary. Some examples include high school friends, college friends, graduate school friends, family members, or people who participate in similar hobbies.***

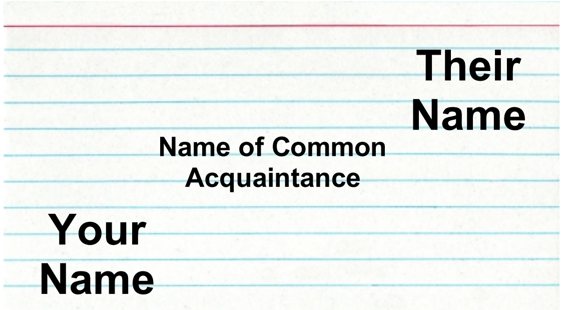
3. Ask the students what other types of social networks, other than Facebook, they belong to? Responses will most likely include TikTok, Twitter, Instagram, Snapchat, LinkedIn, Google+, etc.
4. Next, inform the students that networks can be as big or as small as we want. We can even determine our own class's social network and create visualizations from it!
5. Network Activity:
    a. Distribute index cards to students. Each student will need enough cards to make a connection with every other person in the class. For example, if there are 20 students in a class, then each student needs 19 cards.
    b. On EVERY index card, the student should write his/her first AND last name in the lower left-hand corner (see image below).



    c. Next, each student will walk around the classroom and put another student's first AND last name in the top right-hand corner of an index card (see image below).

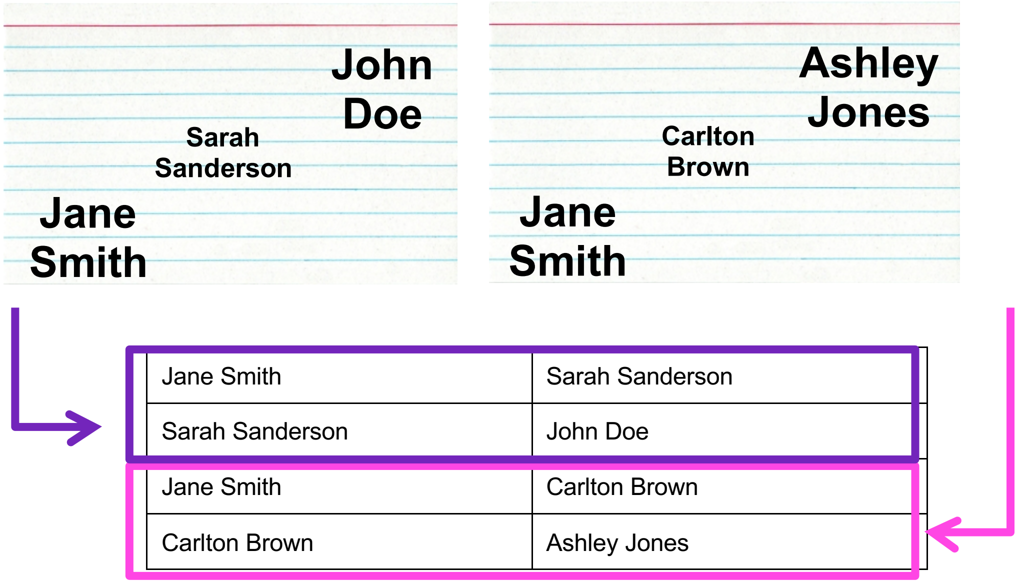


    d. In the center of the index card, the students should write the name of the *closest* 3rd person that they BOTH know (see image below). The person can be someone in the class, someone outside of the class, or someone who doesn't even attend the same school.



    e. Once all of the students have completed their cards, they will turn them in to the teacher so the teacher can create a visualization of the network.
    f. This will probably take an entire class period to complete, which is fine because the graphics can be created and shown the next day.

6. At this point, the teacher will need to manually input the data from the index cards into a spreadsheet. ***It is recommended that the spreadsheet be saved as a .csv file.*** Two sample index cards are included, along with how you would input the data.



Jane Smith	Sarah Sanderson
Sarah Sanderson	John Doe
Jane Smith	Carlton Brown
Carlton Brown	Ashley Jones

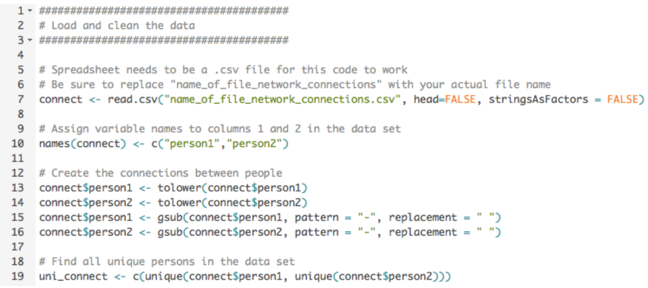
<u>Note:</u> The first index card corresponds to rows 1 and 2 in the spreadsheet (the purple box). The second index card corresponds to rows 3 and 4 in the spreadsheet (the pink box). So, each card will take up two rows in the spreadsheet.

<u>Note:</u> It is probably best to input the data after class and present the visualization during the next day.

7. Once all data has been input into a spreadsheet, use the code provided in the *Network Code* file (LMR_4.26) to produce graphs for the class's social network.

<u>Note:</u> The R Script file can be opened and viewed in the "source" pane of RStudio. There are 2 places where the code needs to be edited by the teacher:

   a. Be sure to change the file name when reading in the .csv file in Line 7 of the code.
   b. Read the comments in Lines 91-96 to help find the 5 most popular people in the class's network. This may require some edits to Lines 97 and 108



```
 1 ########################################
 2 # Load and clean the data
 3 ########################################
 4
 5 # Spreadsheet needs to be a .csv file for this code to work
 6 # Be sure to replace "name_of_file_network_connections" with your actual file name
 7 connect <- read.csv("name_of_file_network_connections.csv", head=FALSE, stringsAsFactors = FALSE)
 8
 9 # Assign variable names to columns 1 and 2 in the data set
10 names(connect) <- c("person1","person2")
11
12 # Create the connections between people
13 connect$person1 <- tolower(connect$person1)
14 connect$person2 <- tolower(connect$person2)
15 connect$person1 <- gsub(connect$person1, pattern = "-", replacement = " ")
16 connect$person2 <- gsub(connect$person2, pattern = "-", replacement = " ")
17
18 # Find all unique persons in the data set
19 uni_connect <- c(unique(connect$person1, unique(connect$person2)))
```

LMR_4.26

Class Scribes:

One team of students will give a brief talk to discuss what they think the 3 most important topics of the day were.

Next Day

Students will end their water usage campaign data collection after today's lesson. Starting the next day, they will analyze their data as part of the End of Unit 4 project.

End of Unit Design Project and Oral Presentation: Water Usage

Objective:

Students will apply their learning of the third and fourth units of the curriculum by completing an end of unit design project.

Materials:

1. Computers
2. *IDS Unit 4 – Project and Oral Presentation* (LMR_U4_Design Project)

End of Unit 4 Project and Oral Presentation: Water Usage

At the beginning of this unit, you explored a 2010 data set from the Los Angeles Department of Water and Power (DWP). You also created a Participatory Sensing campaign to investigate water usage around your community.

For this assignment, you will use both data sets to apply what you have learned in unit 4 and to answer the research question from the beginning of the unit:

How can we help city officials use Participatory Sensing to find out how water is being used around your neighborhood?

Your assignment is as follows:

1. You and a partner will predict water usage for the month of June using a subset of the dwp_2010 data set, which is called dwp_student.
    - Load the dwp_student data set.
    - Using this data, create two data sets: training and testing. Name these data sets `student_train and student_test`.
    - Create the best prediction model that you can based on your training data. Remember to `set.seed(123)` when creating your own training and testing data.
    - You're building this model with data from July 2010 to May 2011. You will use your model to predict water usage for June 2011.
    - After you settle on a specific model, submit your model (code) to your teacher. If you created any new variables, submit the code you used to create them as well.
    - ***What do the variables included in your prediction model say about how Angelenos use water?***
    - You will evaluate the prediction accuracy based on a separate set of data. Your teacher will give you another data set. Use this data set to evaluate your prediction. The pair with the smallest prediction error based on mean squared error (MSE), is the winner.
2. Using your Participatory Sensing data, explain how water is being used in your neighborhood. Make sure you use evidence from your PS data analysis. Be sure to answer the research question and your statistical questions.

Create a 5-minute presentation comprising of 4 to 5 slides that explains your model, the predicted value for June 2011 water consumption, and the findings using your campaign data. Be sure to include a detailed explanation of how you and your partner decided to create your prediction model, and how it performed on the test data set your teacher provided in your presentation. Each person must participate in the presentation. In addition to the presentation, submit a 2-4 page, double-spaced summary of your analysis including plots/graphs.

Note to teacher about the testing data set: The data set you will provide for students to test their prediction models is called **`dwp_teacher`**. It is recommended that you provide the data set's name upon students' submission of the code for their prediction models.