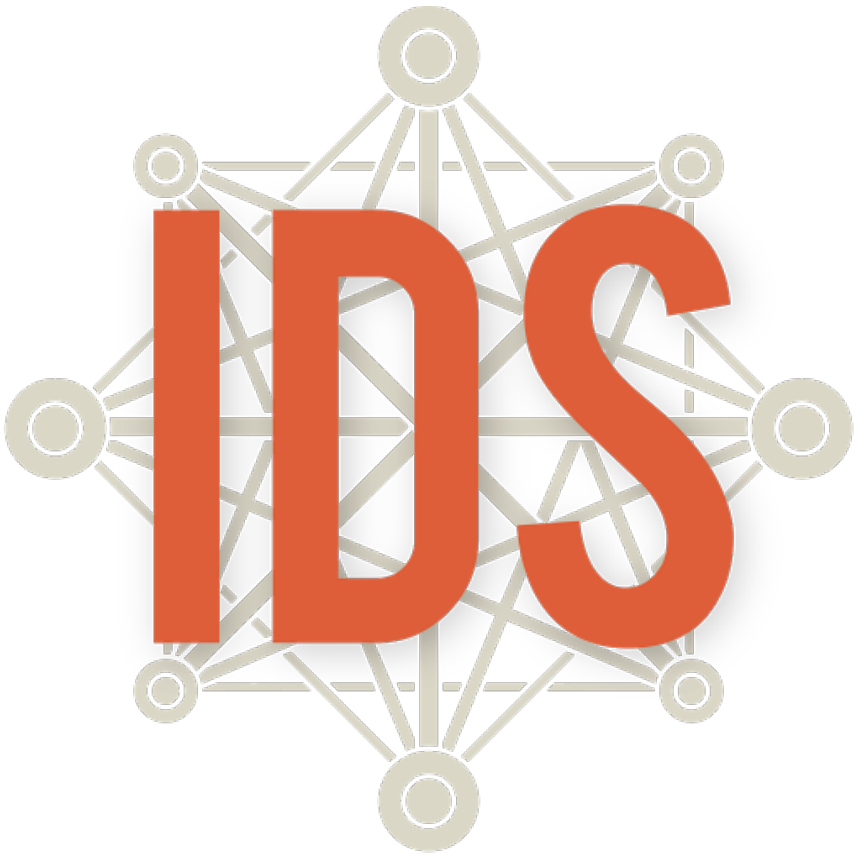


Introduction to Data Science

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://rstudio.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.

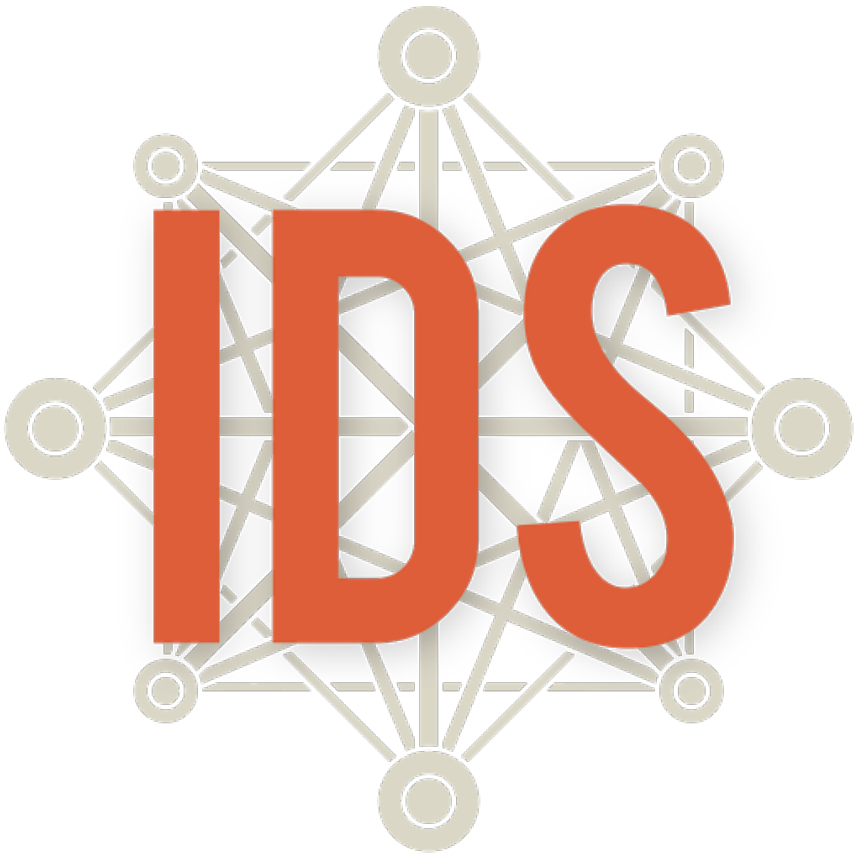
Focus SMPs

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	41
	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	80
	18	Lab 1D: Zooming Through Data		Subsetting	85
	19	Lab 1E: What's the Relationship?		Multi-variable plots	89
	20	Practicum: The Data Cycle & My Food Habits	Food Habits	Data cycle, variability	92
	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	96
	23	Lab 1F: A Diamond In the Rough	Time Use - data	Cleaning names, categories, and strings	101
	24	Lesson 16: Categorical Associations	Time Use - data	Joint relative frequencies in 2-way tables	106
	25	Lesson 17: Interpreting Two-Way Tables	Time Use - data	Marginal & conditional relative frequencies	108
	26+	Lab 1G: What's the FREQ?	Time Use – data	2-way tables, tally	113
	27	Practicum: Teen Depression	Time Use	Statistical questions, interpreting plots	116
	28	Practicum Presentations		Statistical questions, interpreting plots	-
	29-30	Lab 1H: Our Time		Data cycle, synthesis	118
Unit 1 Project (5 Days)	31-35	End of Unit Project and Oral Presentation: Analyzing Data to Evaluate Claims		Data cycle	119

^=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:

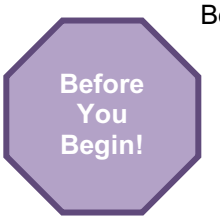
1. Video: *The Target Story* found at: https://www.youtube.com/watch?v=XvSA-6BJkx4&feature=youtu.be
2. Data Science (DS) journal (quad-ruled composition book or similar); MUST be available for every lesson
3. *Data Diary* handout (LMR_1.1_Data Diary)
4. Video: *Terms and Conditions* found at: https://www.youtube.com/watch?v=ZcjtEKNP05c

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 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

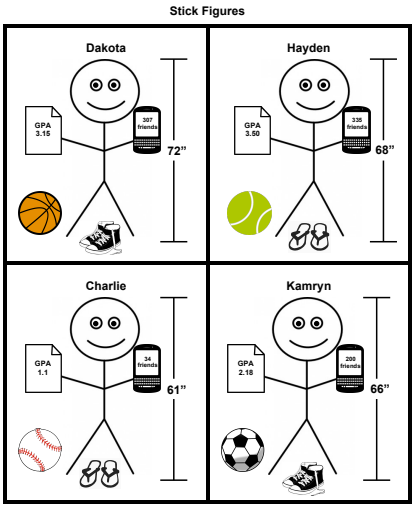
Essential Concepts: Data consist of records of particular characteristics of people or objects. Data can be organized in many different ways, and some ways make it easier than others for achieving particular purposes.

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.



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."

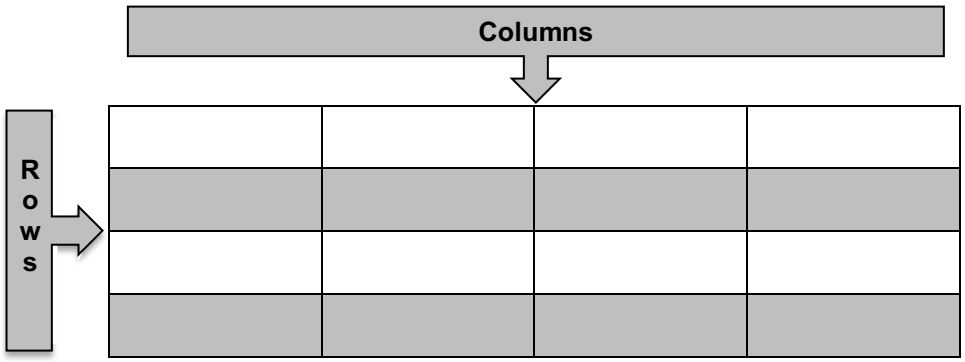
    a. The values from "Shoe" are either "sneakers" or "sandals".

        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

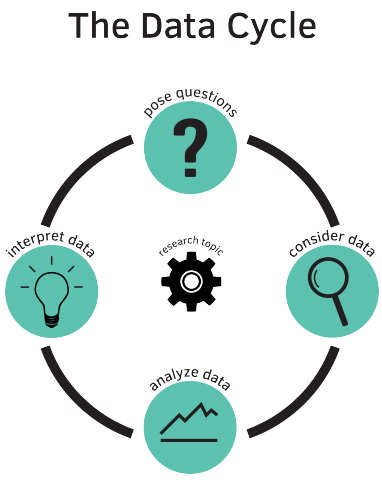
Vocabulary:

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.

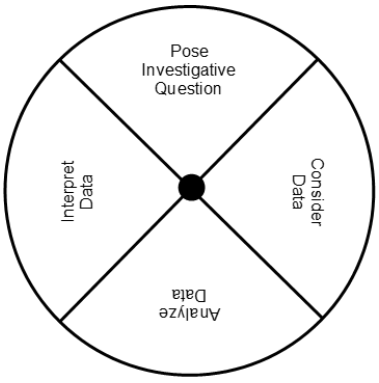
   d. **Interpret Data**: The statistical investigative questions are answered by referring to the tables, graphs, and summaries made in the Data Analysis phase.

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.***

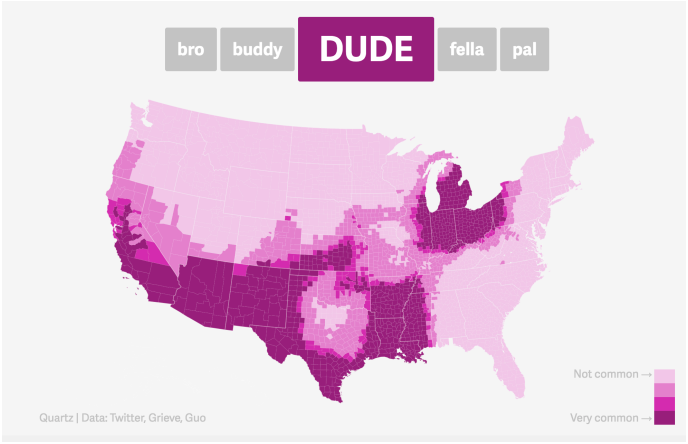
b. What questions or wonderings do we have about the artifact? ***Students should begin developing statistical habits of mind. They should be interrogating the data by asking questions such as: Who is this data about? What was the purpose of collecting the data? What was the survey question asked to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



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: _____

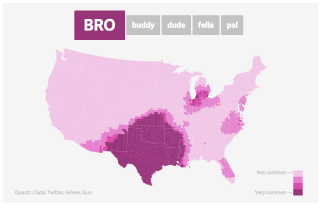
_____ _____

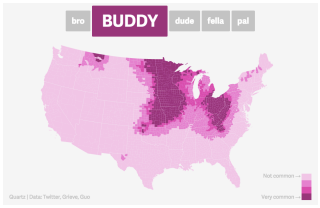
The dude map: How Americans refer to their bros

(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)

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)

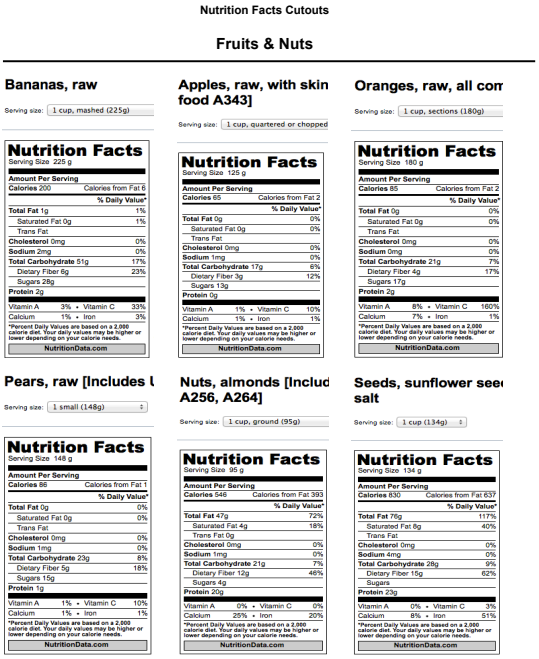
3. *Food Habits Data Collection* handout (LMR_1.8_Food Habits Data Collection)

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.



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

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

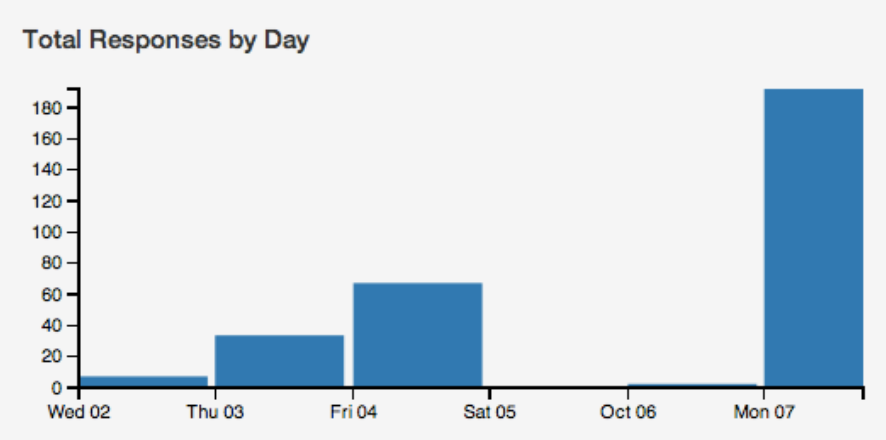
When should you take the survey? If possible, take the survey every time you eat a snack or at the end of the day. Reminders can be set to ensure survey completion.

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:

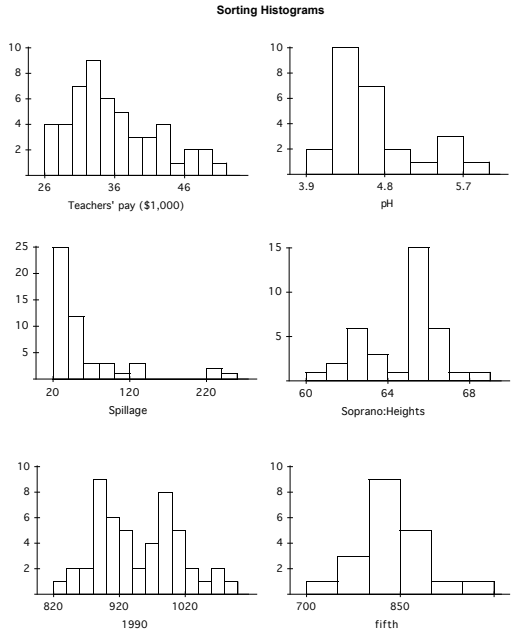
symmetric, left-skewed, right-skewed, unimodal, bimodal

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?

   _____
   _____
   _____

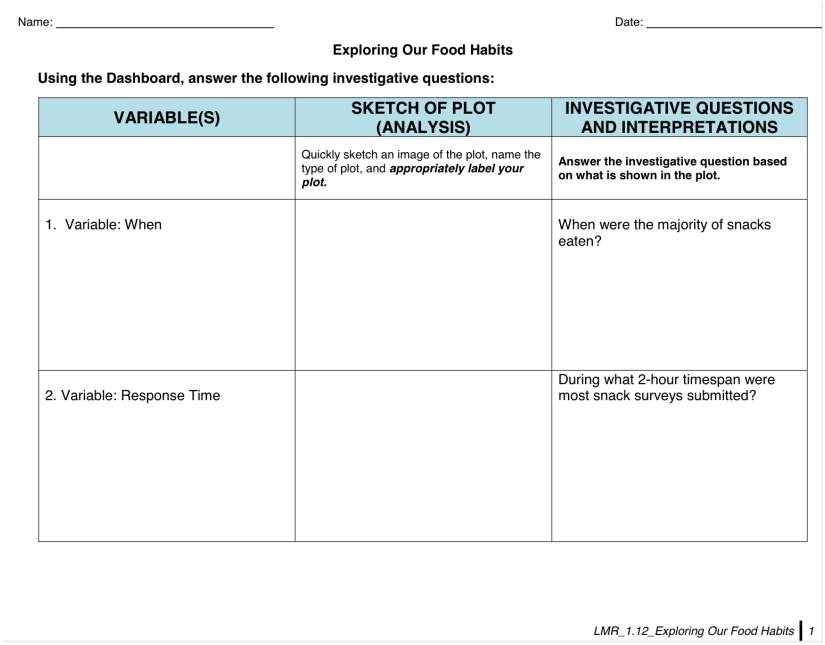
LMR_1.11

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)
```

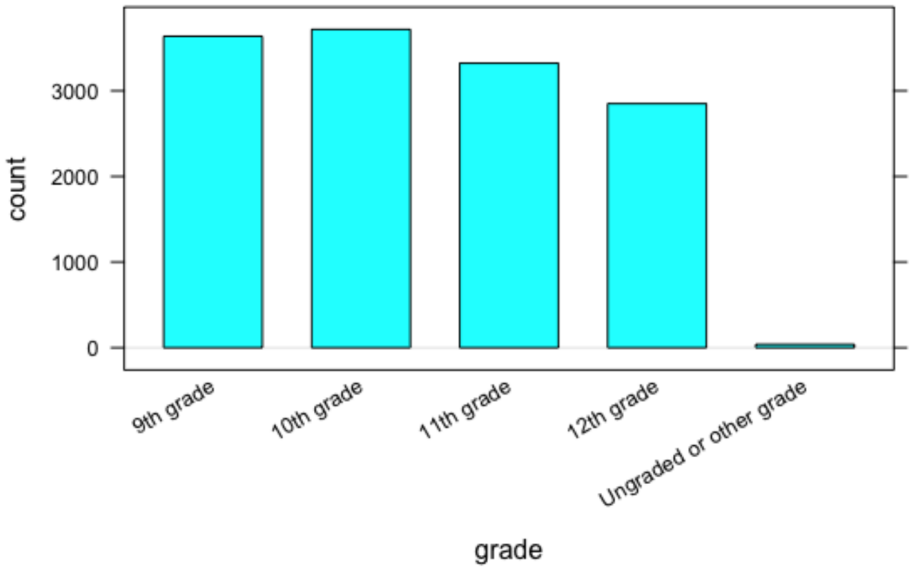
R's most important syntax

- 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 `str`ucture 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.

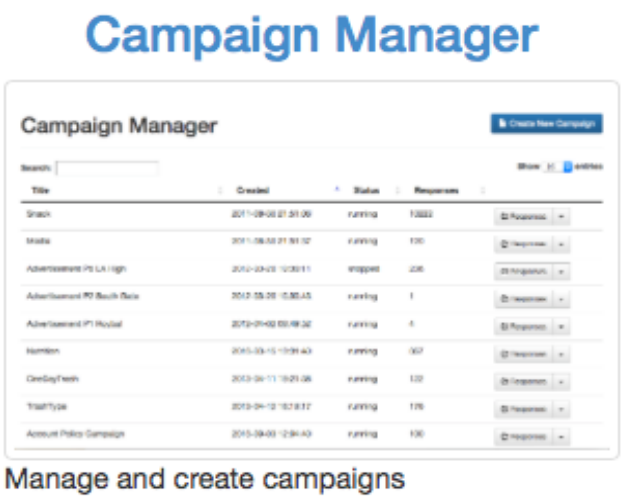
- Throughout the previous labs, we've been using data that was already loaded in RStudio.
  - But what if we want to analyze our own data?
- This lab is all about learning how to load our own participatory sensing data into RStudio

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

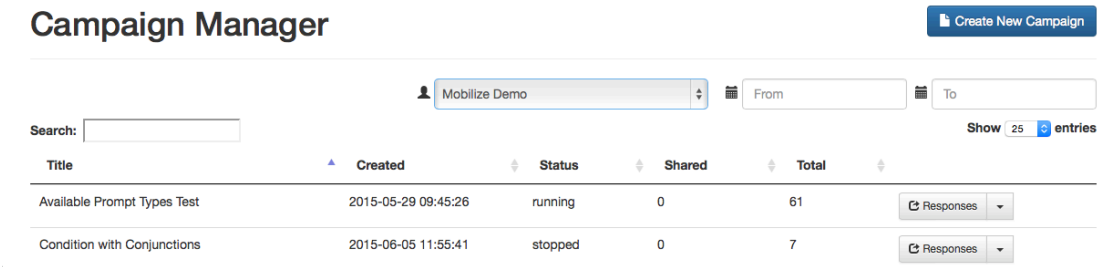
- To begin, go to the *IDS* Tools page.
  - Click on the Campaign Manager
  - Fill in your username and password and click "Sign in."



Manage and create campaigns

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

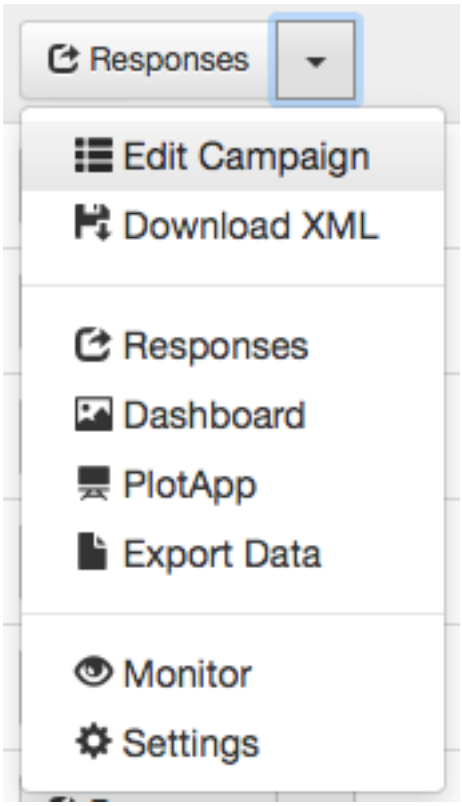
Campaign Manager



- 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.

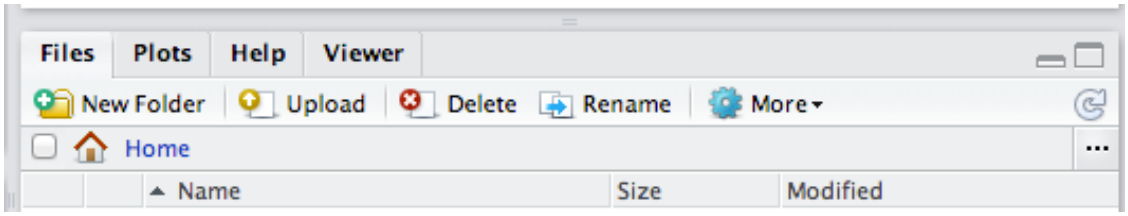


- Look for the option labeled Export Data. Click it.
  - Remember where you save your file!

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

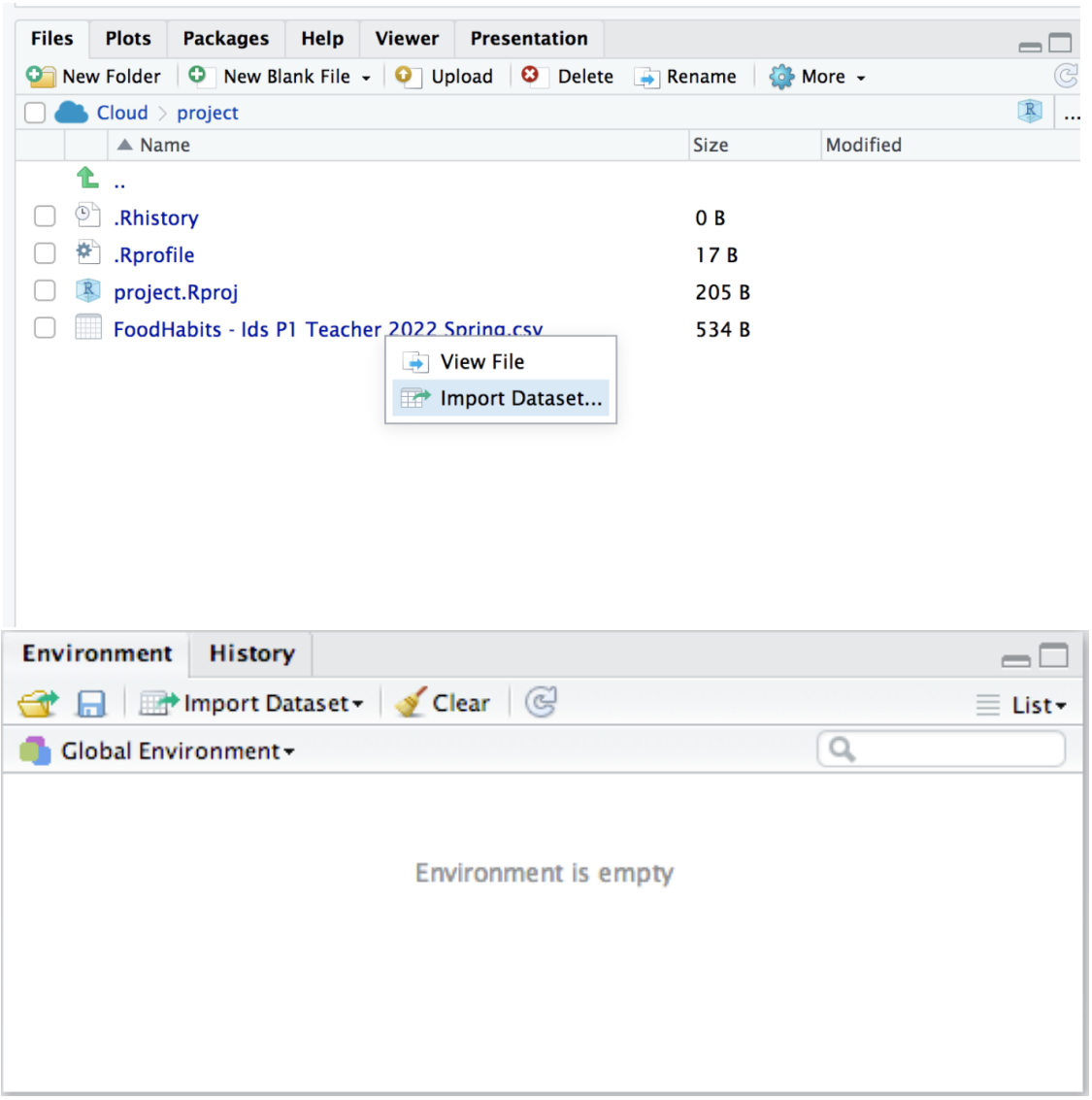
- Look at the four different *panes* in RStudio.
  - Find the *pane* with a Files tab.
  - Click it!
- 

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

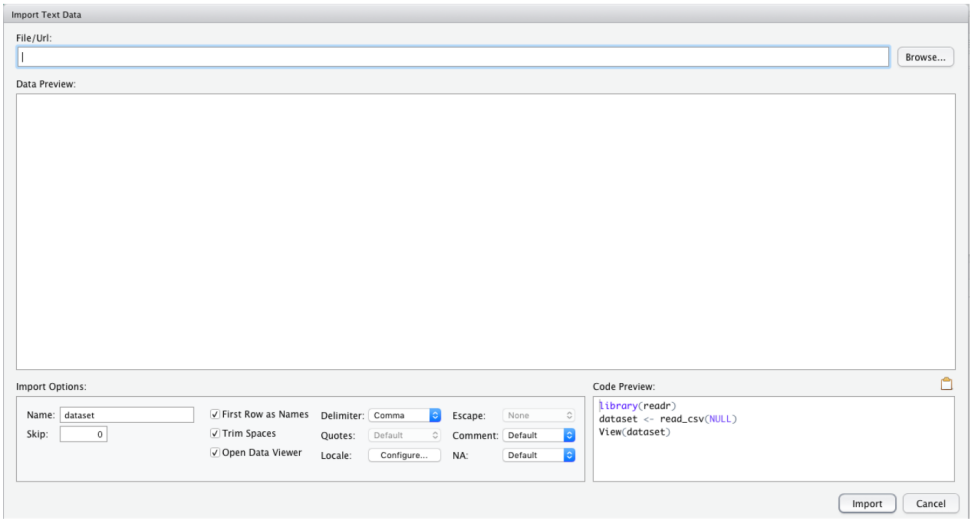
- By Uploading your data into RStudio you've really only given yourself access to it.
  - Don't believe me? Look at the Environment pane ... where's your data?
- To actually use the data we need to Import it into your computer's memory.
- To compute more quickly and efficiently, R will only keep a few 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



- 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



- 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.

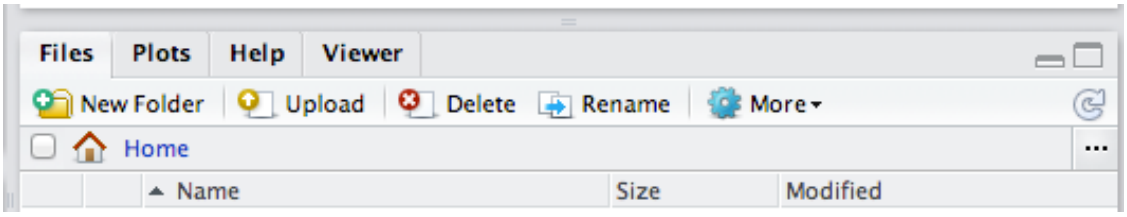
read.csv()

- 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 e*xport*, 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.

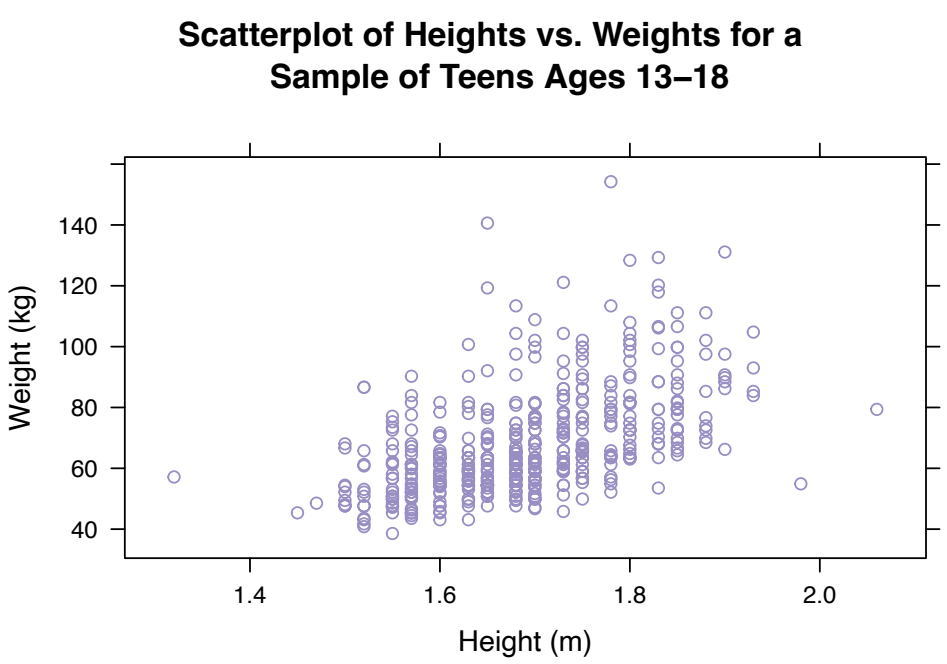
Lesson:

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.

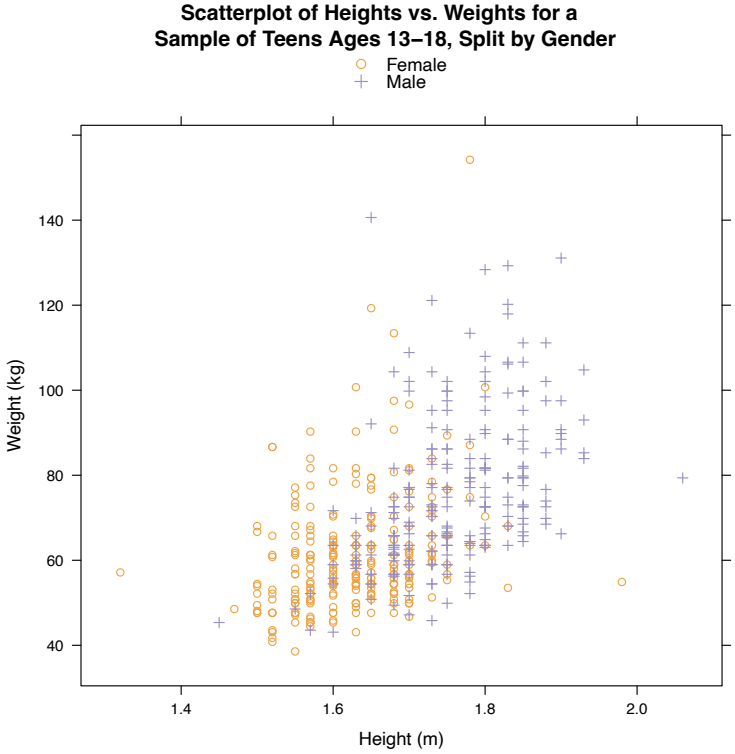
2. Display the following plot to the class (LMR_1.13) so they can see some actual data:



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):

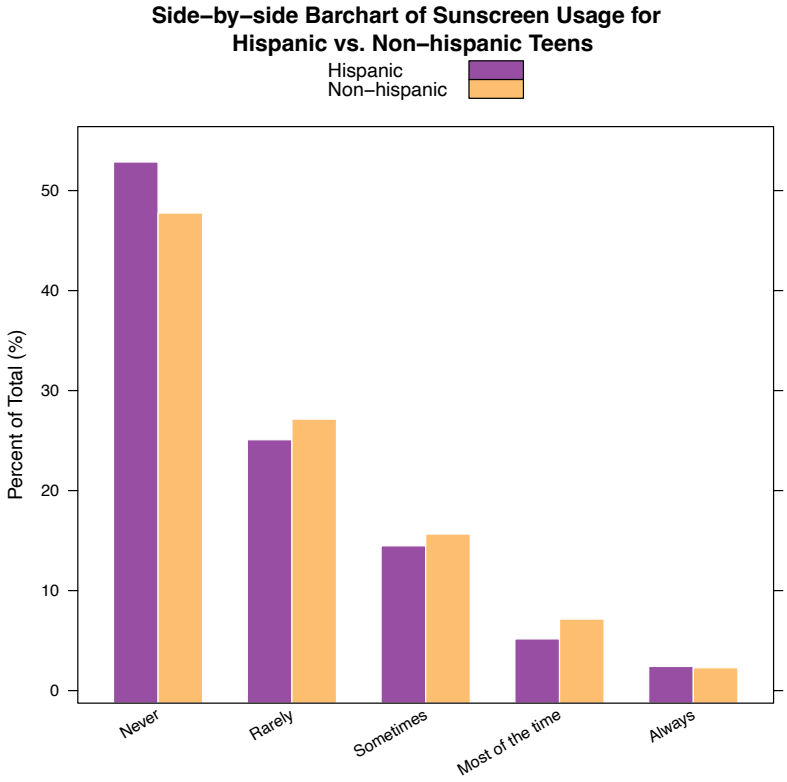


LMR_1.14

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.***

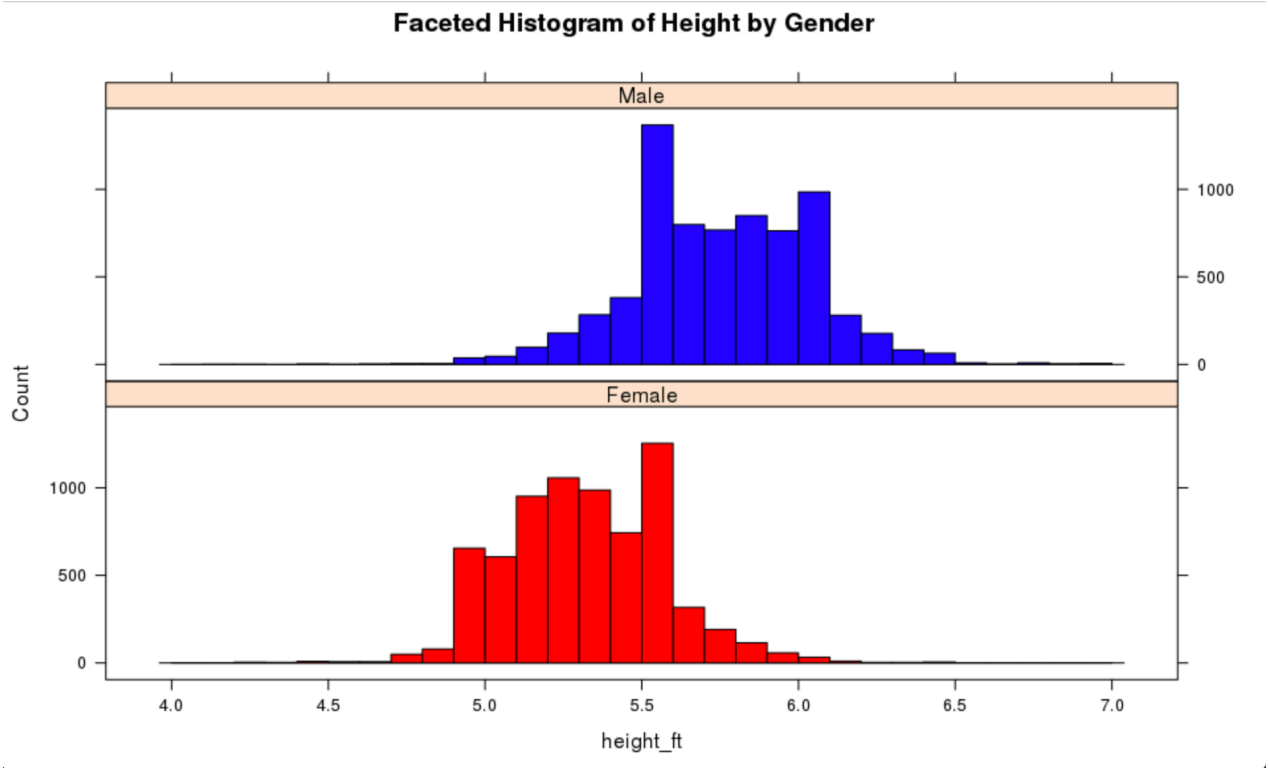
d. What questions can we ask that this graph might answer? ***Who is taller, boys or girls? Who weighs more? Is the association between height and weight the same for boys as it is for girls?***

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

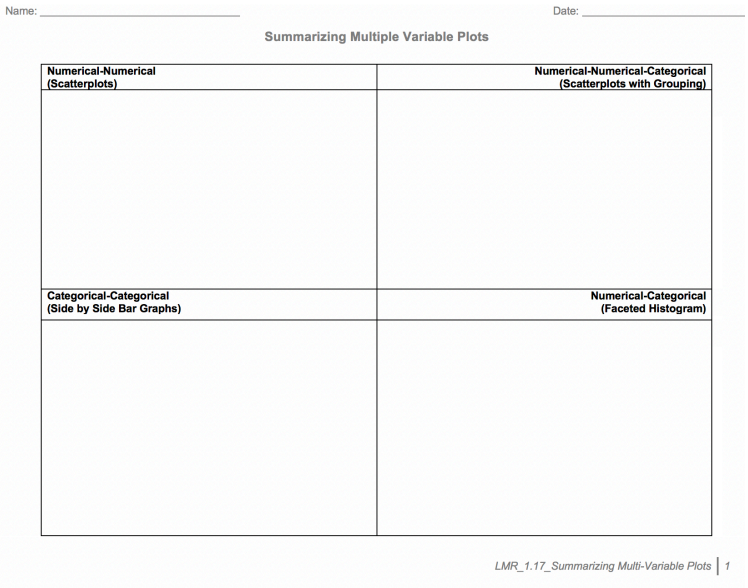
12. Ask students to individually answer the following questions on the handout (LMR_1.15):
    a. What kind of plot is this and how will you remember its features? ***Side-by-side bar chart.***
    b. How many variables are displayed in this plot? Name the variable(s). ***Two variables: whether or not 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"
```

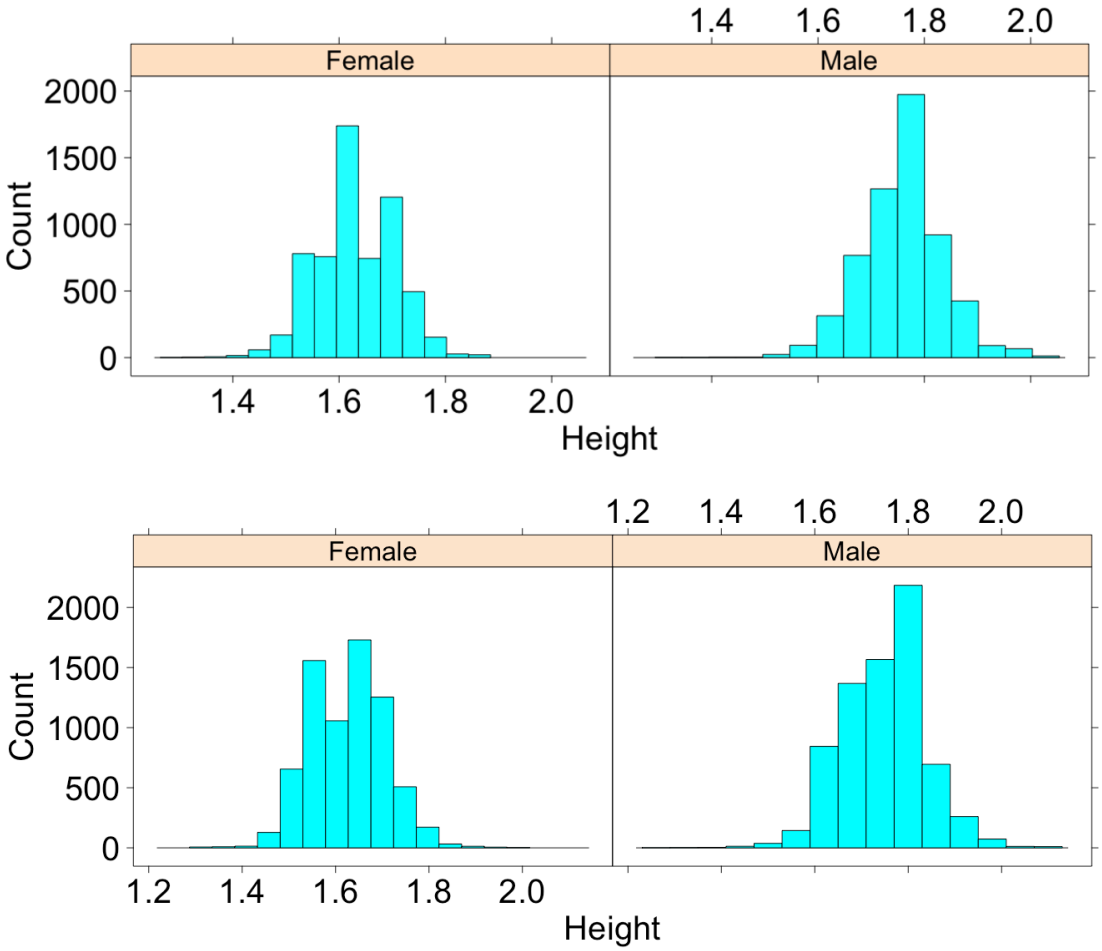
Lab 1E - What's the Relationship?

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?

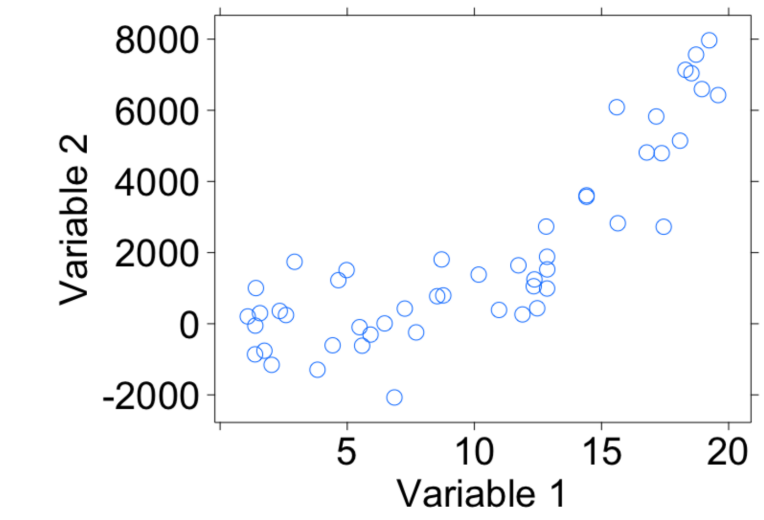


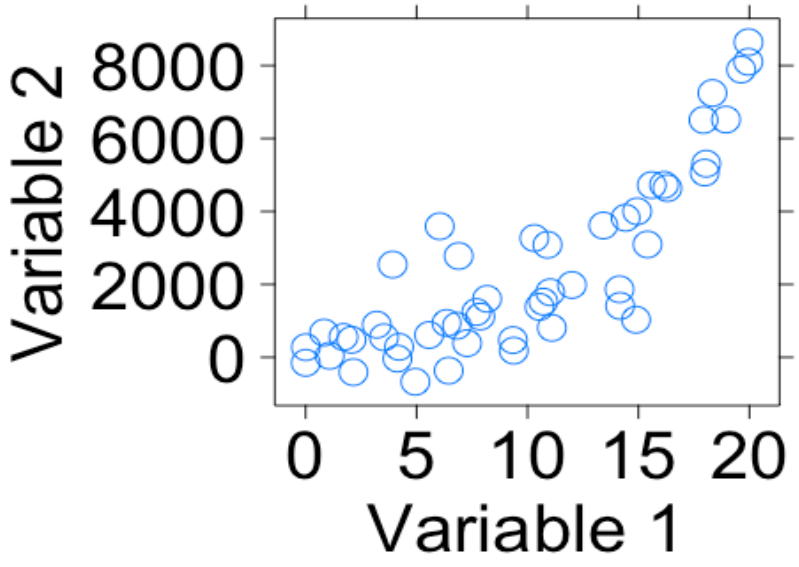
- **How many variables were used to create this plot? Which variables were used and how were they used?**

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 the *New York Times* multimedia graphic titled *How Different Groups Spend Their Time* to spark their curiosity about how they spend their own time. The graphic can be found at http://www.nytimes.com/interactive/2009/07/31/business/20080801-metrics-graphic.html?_r=0.

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 The New York Times multimedia graphic titled *How Different Groups Spend Their Time* 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: *The New York Times' How Different Groups Spend Their Time found at:* http://www.nytimes.com/interactive/2009/07/31/business/20080801-metrics-graphic.html?_r=0
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 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 `str`ucture 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

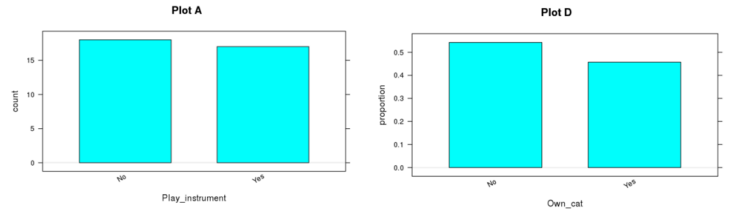
Note: This is a conditional relative frequency by row. We can also calculate conditional relative frequencies by column if we were interested in knowing the difference in cat preference for those who play instruments versus those who don't.

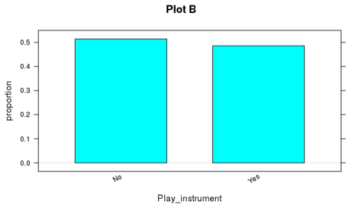
19. Distribute one full set of cards from the *Analyzing Categorical Variables* file (LMR_1.18) to each student team.

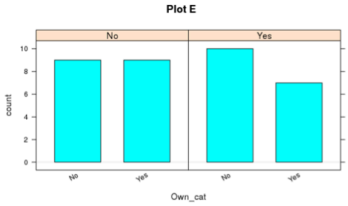
    Advanced preparation required:

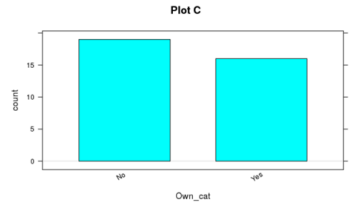
    Print the *Analyzing Categorical Variables* file (LMR_1.18). The handouts can then be cut into a total of 20 cards (12 visuals, 8 numerical summaries). You will need enough sets of the cards for each student team to share one full set. For example, if there are 5 student teams in a class, then 5 copies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

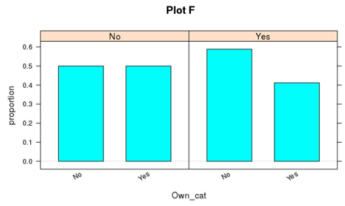
Analyzing Categorical Variables











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?

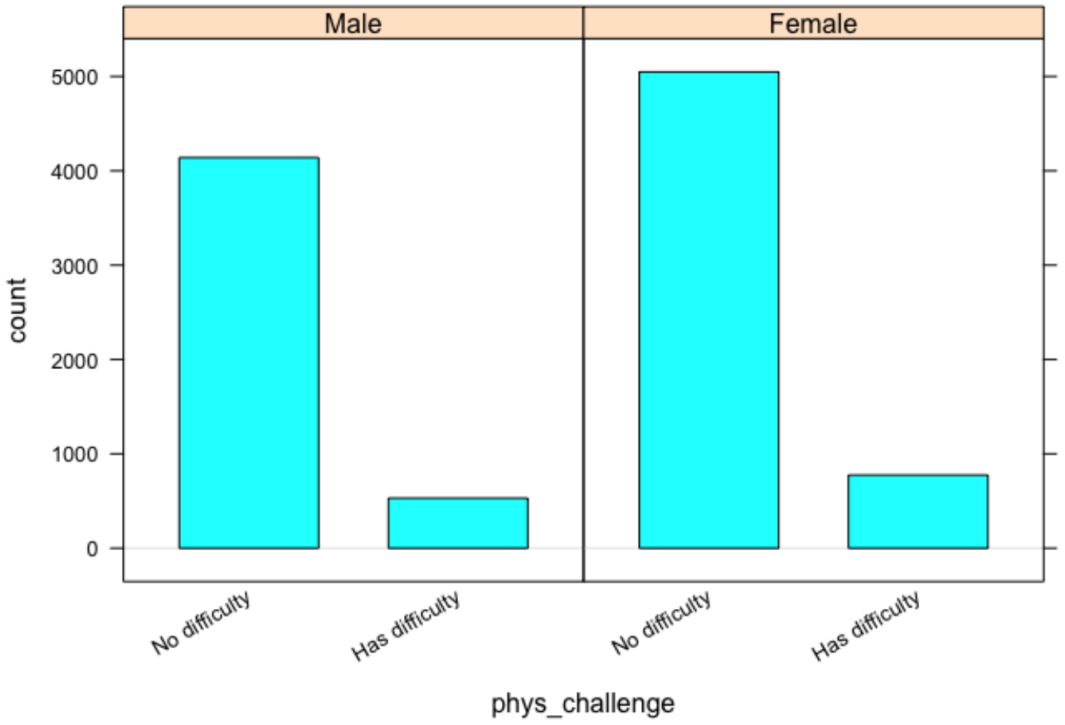
- When it comes to categories, about all you can do is *count* or *tally* how often each category comes up in the data.
- Fill in the blanks below to answer the following: **How many more *females* than *males* are there in our ATU data??**

```
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.