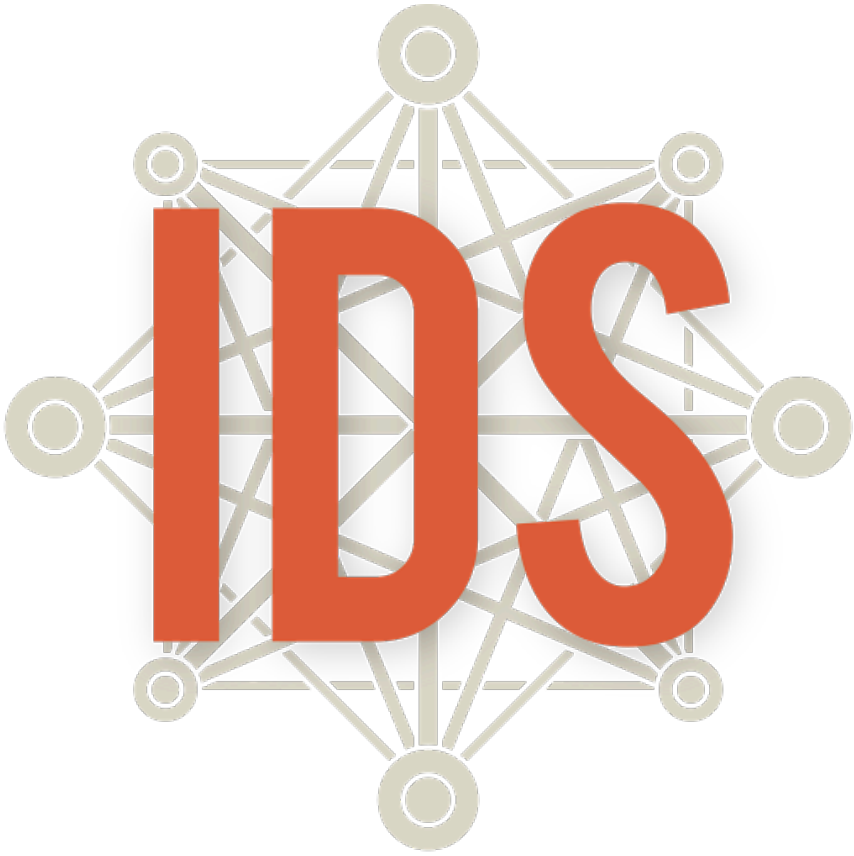


Introduction to Data Science

Unit 4

Introduction to Data Science
Daily Overview: Unit 4

Theme	Day	Lessons and Labs	Campaign	Topics	Page
Predictions and Models (15 days)	1	Lesson 1: Water Usage		Data cycle, official data sets	7
	2	Lesson 2: Exploring Water Usage		Exploratory data analysis, campaign creation	10
	3	Lesson 3: Evaluating and Implementing a Water Campaign	Water Campaign—data	Statistical questions, evaluate & mock implement campaign	12
	4^	Lesson 4: Refining the Water Campaign	Water Campaign—data	Revise and edit campaign, data collection	14
	5	Lesson 5: Statistical Predictions Using One Variable	Water Campaign—data	One-variable predictions using a rule	16
	6	Lesson 6: Statistical Predictions by Applying the Rule	Water Campaign—data	Predictions applying mean square deviation, mean absolute error	19
	7	Lesson 7: Statistical Predictions Using Two Variables	Water Campaign—data	Two-variable statistical predictions, scatterplots	23
	8	*LAB 4A: If the Line Fits…*	Water Campaign—data	Estimate line of best fit	26
	9	*LAB 4B: What's the Score?*	Water Campaign—data	Comparing predictions to real data	28
	10	Lesson 8: What's the Trend?	Water Campaign—data	Trend, associations, linear model	30
	11	Lesson 9: Spaghetti Line	Water Campaign—data	Estimate line of best fit, single linear regression	34
	12	*LAB 4C: Cross-Validation*	Water Campaign—data	Use training and testing data for predictions	37
	13	Lesson 10: Predicting Values	Water Campaign—data	Predictions based on linear models	40
	14	Lesson 11: How Strong Is It?	Water Campaign—data	Correlation coefficient, strength of trend	43
	15	*LAB 4D: Interpreting Correlations*	Water Campaign—data	Use correlation coefficient to determine best model	45
Piecing it Together (6 days)	16	Lesson 12: More Variables to Make Better Predictions	Water Campaign—data	Multiple linear regression	49
	17	Lesson 13: Combination of Variables	Water Campaign—data	Multiple linear regression	52
	18	*LAB 4E: This Model Is Big Enough for All of Us*	Water Campaign—data	Multiple linear regression	55
	19	Practicum: Predictions	Water Campaign—data	Linear regression	56
	20	Lesson 14: Improving Your Model	Water Campaign—data	Non-linear regression	57
	21	*LAB 4F: Some Models Have Curves*	Water Campaign—data	Non-linear regression	59
The Growth of Landfills (5 days)	22	Lesson 15: The Growth of Landfills	Water Campaign—data	Modeling to answer real-world problems	63
	23	Lesson 16: Exploring Trash via the Dashboard	Water Campaign—data	Analyze data to improve models	66
	24	Lesson 17: Exploring Trash via RStudio	Water Campaign—data	Analyze data to improve models	67
	25	Prepare Team Presentations	Water Campaign—data	Modeling with statistics	-
	26	Present Team Recommendations	Water Campaign—data	Modeling with statistics	-
Decisions, Decisions! (3 days)	27	Lesson 18: Grow Your Own Classification Tree	Water Campaign—data	Multiple predictors, classifying into groups, decision trees	69
	28	Lesson 19: Data Scientists or Doctors?	Water Campaign—data	Decision trees based on training and testing data	74
	29	*LAB 4G: Growing Trees*	Water Campaign—data	Decision trees to classify observations	77
Ties that Bind (3 days)	30	Lesson 20: Where Do I Belong?	Water Campaign—data	Clustering, k-means	80
	31	*LAB 4H: Finding Clusters*	Water Campaign—data	Clustering, k-means	85
	32+	Lesson 21: Our Class Network	Water Campaign—data	Clustering, networks	87
End of Unit Project (7 days)	33-40	End of Unit 3 and 4 Design Project and Oral Presentations: Water Usage	Water Campaign	Synthesis of above	90

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

IDS Unit 4: Essential Concepts

Lesson 1: Water Usage

Data can be used to make predictions. Official data sets rely on censuses or random samples and can be used to make generalizations. On the other hand, data from Participatory Sensing campaigns are not random and rely on the sensors, in our case, humans, to be gathered and limits the ability to generalize.

Lesson 2: Exploring Water Usage

Exploring different data sets can give us insight about the same processes. Information from an official data set compared with a Participatory Sensing data set can yield more information than one data set alone. Research questions provide an overall direction to make comparisons between data sets.

Lesson 3: Evaluating and Implementing a Water Campaign

Statistical questions guide a Participatory Sensing campaign so that we can learn about a community or ourselves. These campaigns should be evaluated before implementing to make sure they are reasonable and ethically sound.

Lesson 4: Learning About Our Water Campaign

Statistical questions guide a Participatory Sensing campaign so that we can learn about a community or ourselves. These campaigns should be tried before implementing to make sure they are collecting the data they are meant to collect and refined accordingly.

Lesson 5: Statistical Predictions using One Variable

Anyone can make a prediction. But statisticians measure the success of their predictions. This lesson encourages the classroom to consider different measures of success.

Lesson 6: Statistical Predictions by Applying the Rule

If we use the squared residuals rule, then the mean of our current data is the best prediction of future values. If we use the mean absolute error rule, then the median of the current data is the best prediction of future values.

Lesson 7: Statistical Predictions Using Two Variables

When predicting values of a variable *y*, and *if y* is associated with *x*, then we can get improved predictions by using our knowledge about *x*. Basically, we "subset" the data for a given value of *x*, and use the mean *y* for those subset values. If the resulting means follow a trend, we can model this trend to generalize to as-yet unseen values of *x*.

Lesson 8: What's the Trend?

Associations are important because they help us make better predictions; the stronger the trend, the better the prediction we can make. "Better" in this case means that our mean squared residuals can be made smaller.

Lesson 9: Spaghetti Line

We can often use a straight line to summarize a trend. "Eye balling" a straight line to a scatterplot is one way to do this.

Lesson 10: Predicting Values

The regression line can be used to make good predictions about values of *y* for any given value of *x*. This works for exactly the same reason the mean works well for one variable: the predictions will make your score on the mean squared residuals as small as possible.

Lesson 11: How Strong Is It?

A high absolute value for correlation means a strong linear trend. A value close to 0 means a weak linear trend.

Lesson 12: More Variables to Make Better Predictions

We can use scatterplots to assess which variables might lead to strong predictive models. Sometimes using several predictors in one model can produce stronger models.

Lesson 13: Combination of Variables

If multiple predictors are associated with the response variable, a better predictive model will be produced, as measured by the mean absolute error.

Lesson 14: Improving your Model

If a linear model is fit to a non-linear trend, it will not do a good job of predicting. For this reason, we need to identify non-linear trends by looking at a scatterplot or the model needs to match the trend.

Lesson 15: The Growth of Landfills

Modeling does not always have to produce an equation. Instead, we can create models to answer real-world problems related to our community.

Lesson 16: Exploring Trash via the Dashboard

Exploring the IDS Dashboard provides a visual approach to data analysis.

Lesson 17: Exploring Trash via RStudio
RStudio can be used to verify initial results/findings from data analysis done via the IDS Dashboard.

Lesson 18: Grow Your Own Classification Tree
Many data sets have multiple predictors and are very non-linear. We can still use this data, but need to model it differently, such as in a decision tree. Decision trees are a useful tool for classifying observations into groups.

Lesson 19: Data Scientists or Doctors?
We can determine the usefulness of decision trees by comparing the number of misclassifications in each.

Lesson 20: Where Do I Belong?
We can identify groups, or "clusters," in data based on a few characteristics. For example, it is easy to classify a classroom into males and females, but what if you only knew each student's arm span? How well could you classify their genders now?

Lesson 21: Our Class Network
Networks are made when observations are interconnected. In a social setting, we can examine how different people are connected by finding relationships between other people in a network.

Predictions and Models

Instructional Days: 16

Enduring Understandings

The regression line is a prediction machine. We give it an x-value, it gives us a predicted y-value. The regression line summarizes the trend in the data, but there may still remain variability in the dependent variable that is not explained by the independent variable. Although the regression line provides optimal predictions when the association is linear, other models are needed for when it is not linear.

Engagement

Students will analyze a map from the Medical Daily website. The map and its article called *How Twitter Can Predict Heart Disease: Negative Tweets Associated With Stress, Higher Risk Of Disease*, shows a side-by-side comparison of CDC heart attack deaths data and Twitter's predicted data. They will engage in a discussion comparing and contrasting the visualization. The map can be found at:

http://www.medicaldaily.com/how-twitter-can-predict-heart-disease-negative-tweets-associated-stress-higher-risk-318830

Learning Objectives

Statistical/Mathematical:

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

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

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

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

S-IC 6: Evaluate reports based on data. *

*This standard is woven throughout the course. It is a recurring standard for every unit.

Focus Standards for Mathematical Practice for All of Unit 4:

SMP-2: Reason abstractly and quantitatively.

SMP-4: Model with mathematics.

SMP-7: Look for and make use of structure.

Data Science:

Judge whether or not the linear model is appropriate. Learn to interpret a correlation coefficient in a linear model and interpret slope and intercept. Evaluate the strength of a linear association. Evaluate the potential error in a linear model.

Applied Computational Thinking using RStudio:

- Use linear regression models to predict response values based on sets of predictors.
- Fit a regression line to data and predict outcomes.
- Compute the correlation coefficient of a linear model.
- Create a Participatory Sensing campaign using a campaign Authoring Tool.

Real-World Connections:

Many studies are published in which predictions are made, and media reports often cite data that make predictions. They involve one or more explanatory variable and a response variable, such as income vs. education, weight vs. exercise, and cost of insurance vs. age. Understanding linear regression helps evaluate these studies and reports.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will use complex sentences to write informative short reports that use data science concepts and skills.
4. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data File:

1. LA DWP (`dwp_2010`)
2. Movies (`movie`)

Data Collection:

Students will collect data for their water usage campaign.

Legend for Activity Icons


Video clip


Discussion


Articles/Reading


Assessments


Class Scribes

Lesson 1: Water Usage

Objective:

Students will compare and contrast an official data set versus a Participatory Sensing data set. They will begin to analyze an official data set from 2010 provided by the Los Angeles Department of Water and Power (DWP) to help them understand how water was used in the Los Angeles area in the recent past, before the drought.

Materials:

1. *Video: California Drought Crisis Reaches Worst Level as It Spreads North* http://www.nbcnews.com/storyline/california-drought/california-drought-crisis-reaches-worst-level-it-spreads-north-n169516
2. *Webpage: Twitter vs. Heart Disease* Webpage (Found at: http://www.medicaldaily.com/how-twitter-can-predict-heart-disease-negative-tweets-associated-stress-higher-risk-318830)
3. Class Created Campaign Information (from Unit 3, Lessons 17-19)

Vocabulary:

census

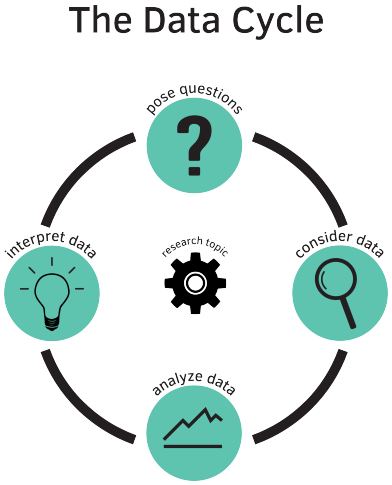
Essential Concepts: Data can be used to make predictions. Official data sets rely on censuses or random samples and can be used to make generalizations. On the other hand, data from Participatory Sensing campaigns are not random and rely on the sensors, in our case, humans, to be gathered and limits the ability to generalize.

Lesson:

1. Ask students to recall that statistics are used to make predictions about population parameters.
2. Project the map found on the Medical Daily website. Inform students that Twitter data was compared to CDC heart disease deaths data on side-by-side maps. Using a *Think, Pair, Share* ask students to discuss:
    a. What is the source of the data on each map? ***A: Tweets as predicted by Twitter and heart attacks as listed on a death certificate and recorded by the CDC***.
    b. What do the colors on the map mean? ***A: On the spectrum from green to red, green means fewer deaths by heart attack and red means a greater number of deaths by heart attacks***.
    c. How are the maps the same? How are they different?
    d. How reliable are the methods used to report these data? ***A: In the case of the CDC data, we have verifiability (death certificates). On the other hand, the Twitter data predicts based on a person's word.***
    e. How scalable are the methods and can they be generalized? ***A: Official data sets are usually censuses or random samples; they address things at a high level. Participatory Sensing or, in this case the Twitter data, is not random, but addresses things at a personal or local level; however, because it is not a census nor a random sample, it is difficult to be precise about uncertainty or ability to generalize***.
3. Quickly share student responses to the discussion. Then, inform them that this unit focuses on data to make predictions.
4. Set the context for the next three lessons. Inform students that they will be delving into the topic of water usage. In California, water usage is extremely important, given that the state has been in an exceptional drought in the second decade of the 21st century.
5. Using the K-L-W strategy in their DS journals, give students a couple of minutes to write what they *Know* about droughts. Then, students will write what they *Learned* about the California drought as they watch the brief NB News video clip titled *California Drought Crisis Reaches Worst Level as It Spreads North*. Finally, they will write 2-3 questions about what they *Want to*

know/learn about droughts. Video clip is found at: http://www.nbcnews.com/storyline/california-drought/california-drought-crisis-reaches-worst-level-it-spreads-north-n169516.

6. Do a quick *Whip Around* to share some of the students' responses to the *K-L-W*.
7. Inform students that they will be learning about water usage in their own neighborhoods. The LADWP data provides information at a high level about some (most) neighborhoods in L.A. Students will investigate how they can learn more using participatory sensing. In statistics, the Data Cycle provides a process by which we can learn about or investigate a particular topic of interest. To review the components of the Data Cycle, give students two minutes to *Quick Sketch* each phase of the cycle in their DS Journals.
8. After students have had an opportunity to do their sketches, display the Data Cycle graphic below to review it:



9. Next, review their class created campaign from Unit 3. Using a Pair-Share strategy, ask students to discuss when a Participatory Sensing campaign should be used rather than a survey. ***A: Answers will vary. Research questions that include variation across time or across locations are good candidates for Participatory Sensing campaigns; therefore, a trigger is necessary in order to record observations at multiple time points and locations. If a question needs to be answered only once, then a survey is a better method.***
10. Remind students that in the last unit, they created one campaign for the entire class. In this unit, each student team will be creating and implementing a campaign on the topic of water usage.
11. Before they start creating their campaign, they are going to explore an official data set provided by the Los Angeles Department of Water and Power (DWP) to learn more about water usage in the Los Angeles area. Some students may receive services from the DWP.
12. Explain how the data were collected:
    - The data you will see reflects the water usage in Los Angeles in the fiscal year that began in 2010 (July 2010-June 2011). At this time, L.A. was entering a drought, but water conservation efforts had not yet begun.
    - The DWP supplies water to businesses and addresses within its boundaries. It records the amount of water delivered to each address each month. For privacy purposes, it doesn't report how much water a single address uses.
    - Instead, it combines these into neighborhoods. These neighborhoods are defined by the U.S. Census and called Census Blocks. A census block is usually one, sometimes two, square blocks.
    - The DWP reports separate water usage figures for businesses, government structures (such as schools), and residences. For privacy purposes, this data set eliminated any Census Block that had fewer than 15 addresses.

- Water use is reported in Hundreds of Cubic Feet (HCF) per month (one HCF is about 748 gallons). Display the picture below, which shows a truck that holds about 6 HCF, so students can get a sense of the amount of water as reported.

13. Load and display the DWP data set in RStudio using the command **`data(dwp_2010)`**. Then, expand the spreadsheet ( ) and explain what each variable in the data set means (they may want to record these in their DS journals):
    a. census = census block
    b. sector_type = category of facility
    c. longitude, latitude = GPS coordinates for center of census block location
    d. census_pop = population of census block (number of water users)
    e. total = total number of HFCs used by sector type per block in 2010
    f. july through june = number of HFCs used by sector type per month
    g. count = number of facilities per census block for that sector type
14. Next, load an interactive map of the DWP 2010 data by visiting: https://labs.idsucla.org/extras/animations/watermap/watermap.html
15. Lead a discussion about what is on the page. Ask:
    a. What do the colors and percentages on the legend mean?
    b. What trends do you see?
16. Then click on a marker (circle) to show the popup. Ask:
    a. What information is the popup displaying?
    b. How is the popup displaying the information?
17. Then, click on the Size by census_pop circle under the legend. Ask:
    a. What do you notice about the markers?
    b. What is the size of each marker telling us?
    c. What else do you see?
18. Now that they know what the variables mean, ask student teams to generate two statistical questions about the data. Below are three examples of possible statistical questions:
    a. What month uses the most water?
    b. Typically, how much water do residences consume during that month?
    c. Does this change if you factor in the number of people living in that census block?
19. If time permits, conduct a share-out of the teams' statistical questions.

Class Scribes:

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

Lesson 2: Exploring Water Usage

Objective:

Students will engage in exploratory data analysis with a Los Angeles Department of Water and Power (DWP) data set and begin the creation of a water usage Participatory Sensing campaign to observe patterns of water use in their neighborhoods.

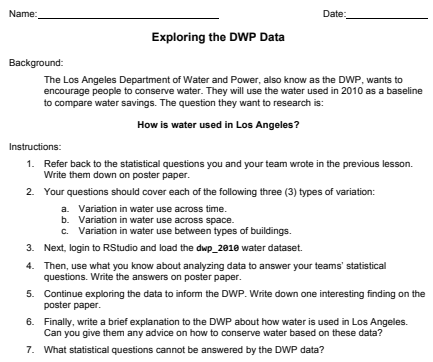
Materials:

1. *Exploring the DWP Data* (LMR_4.1_Exploring DWP Data)
2. *Water Campaign* (LMR_4.2_Water Campaign)
3. Poster paper
4. Markers
5. Class Created Campaign Information (from Unit 3, Lessons 17-19)

Essential Concepts: Exploring different datasets can give us insight about the same processes. Information from an official dataset compared with a participatory sensing dataset can yield more information than one dataset alone. Research questions provide an overall direction to make comparisons between datasets.

Lesson:

1. Display the DWP data using RStudio. In pairs, ask students to recall what each of the variables mean.
2. Next, ask student teams to refer back to the statistical questions they generated in the previous lesson - they will need it for the data exploration.
3. Distribute the *Exploring the DWP Data* handout (LMR_4.1). In their teams, allow students about 20-30 minutes to explore the DWP data set and complete the handout.



Name:_____ Date:_____

Exploring the DWP Data

Background:

The Los Angeles Department of Water and Power, also know as the DWP, wants to encourage people to conserve water. They will use the water used in 2010 as a baseline to compare water savings. The question they want to research is:

How is water used in Los Angeles?

Instructions:

1. Refer back to the statistical questions you and your team wrote in the previous lesson. Write them down on poster paper.
2. Your questions should cover each of the following three (3) types of variation:
   a. Variation in water use across time.
   b. Variation in water use across space.
   c. Variation in water use between types of buildings.
3. Next, login to RStudio and load the `dwp_2010` water dataset.
4. Then, use what you know about analyzing data to answer your teams' statistical questions. Write the answers on poster paper.
5. Continue exploring the data to inform the DWP. Write down one interesting finding on the poster paper.
6. Finally, write a brief explanation to the DWP about how water is used in Los Angeles. Can you give them any advice on how to conserve water based on these data?
7. What statistical questions cannot be answered by the DWP data?

LMR_4.1

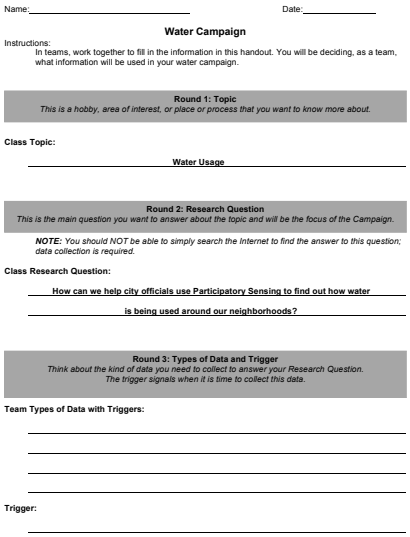
4. After students have had time to explore the DWP data, conduct a whole class discussion based on the following (answers will vary based on student teams' data exploration):
   1. What were some interesting findings?
   2. Which number statistics provided you information to help answer the research question?
   3. Which plots gave you some insights into Los Angeles' water usage?
   4. Based on your findings and by citing evidence from your analysis, what would you say about how water is being used in Los Angeles and who is using it?
5. To prepare for the creation of the Participatory Sensing campaign, ask students to discuss the following in their teams:
   a. How do you think water usage has changed since 2010?

b. What sectors do you think have changed the most?

6. Now that students have an idea about water usage in Los Angeles based on the DWP 2010 data exploration, inform them that they will create a water usage campaign. The research question fro this campaign is:

How can we save water in our neighborhood?

7. Quickly review their class campaign from Unit 3; placing emphasis on the trigger and at how the data they decided to collect answers the research question.
8. Distribute the *Water Campaign* handout (LMR_4.2). Ask students to notice that Rounds 1 and 2 are completed. Their task is to design the rest of the campaign by completing the remaining rounds.



Name:_____ Date:_____

Water Campaign

Instructions:
In teams, work together to fill in the information in this handout. You will be deciding, as a team, what information will be used in your water campaign.

Round 1: Topic
This is a hobby, area of interest, or place or process that you want to know more about.

Class Topic:

Water Usage

Round 2: Research Question
This is the main question you want to answer about the topic and will be the focus of the Campaign.

***NOTE:** You should NOT be able to simply search the Internet to find the answer to this question; data collection is required.*

Class Research Question:

How can we help city officials use Participatory Sensing to find out how water is being used around our neighborhoods?

Round 3: Types of Data and Trigger
Think about the kind of data you need to collect to answer your Research Question. The trigger signals when it is time to collect this data.

Team Types of Data with Triggers:

Trigger:

LMR_4.2

9. Round 3: Allow student teams a reasonable amount of time to engage in a *Brainstorm,* in which they will discuss what kind of data needs to be collected in order to answer this research question and when is the best time to trigger the data collection/completion of the survey. Before they begin, ask students to keep the following question in mind: Which of these data will give us information that addresses our research question?
10. Facilitate the student teams' *Brainstorm* session by circulating around the room to check for understanding. If teams need help with deciding which data they should collect, you may ask them to ponder the following:
    a. What are some water sources?
    b. What do we use water for?
    c. Where might we see water as we walk around our neighborhoods?
    d. What would you consider wasted water?
    e. What are some uses of water that cannot be avoided?
11. Ask students to record the information from each round on poster paper - in this lesson, Rounds 1-3.
12. Inform students that they will be completing Rounds 4 and 5 during the next lesson.

Class Scribes:

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

Lesson 3: Evaluating and Implementing a Water Campaign

Objective:

Students will complete the design of their water usage Participatory Sensing campaign, create the campaign using the campaign authoring tool, and implement a mock campaign to evaluate the feasibility of the campaign.

Materials:

1. *Water Campaign* (LMR_4.2_Water Campaign) from previous lesson
2. Posters from previous lesson
3. Markers
4. *Campaign Authoring Instructions* handout (LMR_4.3_Campaign Authoring)

Essential Concepts: Statistical questions guide a participatory sensing campaign so that we can learn about a community or ourselves. These campaigns should be evaluated before implementing to make sure they are reasonable and ethically sound.

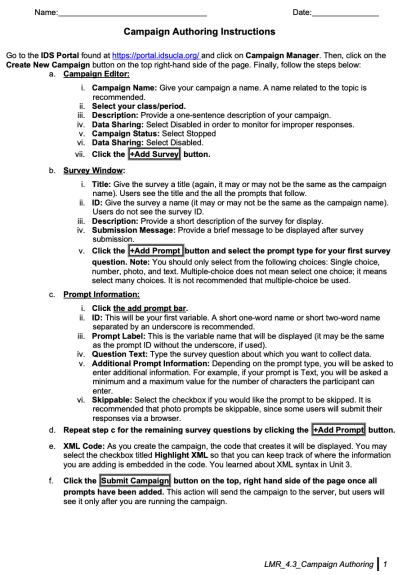
Lesson:

1. Student teams will continue designing their water usage Participatory Sensing campaign. Allow them a couple of minutes to review the information on their posters before moving on to round 4.
2. Round 4: Now that the teams have decided on a trigger and the type of data needed, they will discuss and create survey questions/prompts to ask when the trigger is set. The questions should consider all of the possible data they might collect at this trigger event.
3. Once teams have created their survey questions/prompts, they will evaluate each survey question. For each question they should consider:
    a. What type of survey question/prompt will this be (e.g. single choice, text, photo, numerical, discrete numerical, categorical, location)?
    b. How does this question/prompt help address the research question?
    c. Does the question/prompt need to be reworded? (Is it clear what is being asked for? Do they know how to answer it?) One way to do this is to pair teams and take turns asking each other prompts. The team that is being asked may explain what information they think the question is asking for.
4. If survey questions need to be rewritten, students will decide as a team on the changes.
5. Once finalized, they will record the survey question/prompt that goes along with each data variable on their *Water Campaign* handout (LMR_4.2), being cognizant of question bias.
6. Round 5: In teams, students will now generate three statistical questions that they might answer with the data they will collect and to guide their campaign. They need to make sure that their statistical questions are interesting and relevant to the water usage topic. They will record these statistical questions on their posters. Remind students that they will also have data about the date, time, and place of data collection.
7. Confirm that the questions are statistical and that they can be answered with the data the students propose to collect by circulating around the room to check on each team. Each team will decide on no more than 3 statistical questions to guide their campaign.
8. Now that they have all the pieces of the campaign, teams will evaluate whether their campaign is reasonable and ethically sound. Each team will hold a discussion on the following questions:
    a. Are answers to your survey questions likely to *vary* when the trigger occurs? (If not, you'll get bored entering the same data again and again)
    b. Can the team carry out the campaign?
    c. Do triggers occur so rarely that you'll have very little data? Do they occur so often that you'll get frustrated entering too much data?

d. Ethics: Would sharing these data with strangers or friends be embarrassing or undermine someone's privacy?
e. Can you change your trigger or survey questions to improve your evaluation?
f. Will you be able to gather enough relevant data from your survey questions to be able to answer your statistical questions?

9. Students have collaboratively created their Water Usage Participatory Sensing campaign. They will now use the Campaign Authoring tool to create a campaign like the ones they see on their smart devices or the computer.

10. Distribute the *Campaign Authoring Instructions* handout (LMR_4.3). Each team will select a member to type the information required to create their campaign. Then, they will follow the instructions on the handout.



Name:_____ Date:_____

Campaign Authoring Instructions

Go to the **IDS Portal** found at https://portal.idsucla.org/ and click on **Campaign Manager**. Then, click on the **Create New Campaign** button on the top right-hand side of the page. Finally, follow the steps below:

a. **Campaign Editor:**
   i. **Campaign Name:** Give your campaign a name. A name related to the topic is recommended.
   ii. **Select your class/period.**
   iii. **Description:** Provide a one-sentence description of your campaign.
   iv. **Data Sharing:** Select Disabled in order to monitor for improper responses.
   v. **Campaign Status:** Select Stopped
   vi. **Data Sharing:** Select Disabled.
   vii. **Click the +Add Survey button.**

b. **Survey Window:**
   i. **Title:** Give the survey a title (again, it may or may not be the same as the campaign name). Users see the title and the all the prompts that follow.
   ii. **ID:** Give the survey a name (it may or may not be the same as the campaign name). Users do not see the survey ID.
   iii. **Description:** Provide a short description of the survey for display.
   iv. **Submission Message:** Provide a brief message to be displayed after survey submission.
   v. **Click the +Add Prompt button and select the prompt type for your first survey question. Note:** You should only select from the following choices: Single choice, number, photo, and text. Multiple-choice does not mean select one choice; it means select many choices. It is not recommended that multiple-choice be used.

c. **Prompt Information:**
   i. **Click the add prompt bar.**
   ii. **ID:** This will be your first variable. A short one-word name or short two-word name separated by an underscore is recommended.
   iii. **Prompt Label:** This is the variable name that will be displayed (it may be the same as the prompt ID without the underscore, if used).
   iv. **Question Text:** Type the survey question about which you want to collect data.
   v. **Additional Prompt Information:** Depending on the prompt type, you will be asked to enter additional information. For example, if your prompt is Text, you will be asked a minimum and a maximum value for the number of characters the participant can enter.
   vi. **Skippable:** Select the checkbox if you would like the prompt to be skipped. It is recommended that photo prompts be skippable, since some users will submit their responses via a browser.

d. **Repeat step c for the remaining survey questions by clicking the +Add Prompt button.**

e. **XML Code:** As you create the campaign, the code that creates it will be displayed. You may select the checkbox titled **Highlight XML** so that you can keep track of where the information you are adding is embedded in the code. You learned about XML syntax in Unit 3.

f. **Click the Submit Campaign button on the top, right hand side of the page once all prompts have been added.** This action will send the campaign to the server, but users will see it only after you are running the campaign.

LMR_4.3_Campaign Authoring | 1

LMR_4.3

11. To name their campaign, a naming convention is suggested. Otherwise, you will have multiple campaigns with the same name. For example, teams may include their team name or number in order to easily identify their campaigns.

12. Once their campaign is authored, students will save their work and make edits after they mock implement the campaign for a few days.

13. They will collect data during the mock implementation of their campaign using the information they recorded on the *Water Campaign* handout (LMR_4.2) and record the answers to the survey/prompts on paper. They will make observations about how well their campaign worked and what improvements or changes need to be made.

14. Round 6 will be completed once students have mock-implemented their campaigns.

Class Scribes:

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

Homework

Students will collect data by mock implementing their Water Usage Participatory Sensing campaign.

Lesson 4: Refining the Water Campaign

Objective:

Students will revise their water usage Participatory Sensing campaign according to the finding from the mock implementation of their campaign to refine it. Student teams will then share their final campaigns with the rest of the class.

Materials:

1. Posters from previous lesson
2. Markers

Essential Concepts: Statistical questions guide a participatory sensing campaign so that we can learn about a community or ourselves. These campaigns should be tried before implementing to make sure they are collecting the data they are meant to collect and refined accordingly.

Lesson:

1. Student teams will come together to discuss their findings regarding the mock implementation of their campaign. Allow them time to share their findings with their team members.
2. Next, ask student teams to discuss the revisions they need to make according to their findings.
3. Once they have made a decision on the revisions, they will reflect these changes on their posters.
4. Now they will edit their campaigns. The Recorder/Reporter will type for his/her team and will login by going to the IDS Portal and clicking on Campaign Manager.
5. Then, ask students to find their team's campaign. If the campaign is not visible, they may do a **Search** by typing in their campaign name.
6. Once the campaign is found, the Recorder/Reporter will click on the drop down menu to the right of the campaign information and select **Edit Campaign**.
7. On the **Campaign Editor**, students may scroll down to see their prompts. Students may do the following actions to the prompts:
    - **Edit**: Click on the prompt's name to expand and make changes as needed. To close the prompt, they may click on the **X** that appears on the expanded prompt.
    - **Delete:** Click on the **X** that appears on the prompt's name (non-expanded prompt).
    - **Add:** Scroll down to the **+Add Prompt** button.
8. Once finished making the edits/revisions to the campaign, the Reporter/Recorder will change the **Campaign Status** to **Running** (green).
9. To run the campaign and begin collecting data via the mobile app or web browser, the Reporter/Recorder will click on **Update Campaign** on the top right hand side of the Campaign Editor.
10. Ask teams to refresh their campaigns on their smartphones or the web browser to verify that their campaign appears as one of the choices.
11. Now that they are finished with their campaigns, student teams will share out their campaigns with the rest of the class by engaging in a Gallery Walk of the posters that show their work.
12. Encourage teams to ask questions or make comments as they visit each poster.
13. Before moving to the next round (from poster to poster), ask the teams if they have questions or need clarification or would like to make a comment to the team that created the poster. Repeat until teams have visited all the posters.

Class Scribes:

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

Homework

Students may begin collecting data by implementing their Water Usage Participatory Sensing campaign. They will have a longer period of time to collect these data—about a month—to ensure they collect a sufficient amount of data since only the members of their team will be collecting data for their campaign and not the entire class.

Lesson 5: Statistical Predictions in One Variable

Objective:

Students will devise a rule to determine how to choose a winner when predicting the typical height of all students in a large high school and measure the success of their prediction. They will consider different measures of success.

Materials:

1. *Heights of Students at a Large High School* handout (LMR_4.4_HS Student Heights)

Vocabulary:

rule

Essential Concepts: Anyone can make a prediction. But statisticians measure the success of their predictions. This lesson encourages the classroom to consider different measures of success.

Lesson:

1. Inform the class that for this lesson, our class will help judge a contest held at a particular high school. This school held a contest in which they selected students at random from a classroom and reported their height.
2. The information in Steps 3 – 7 is included in the *Heights of Students at a Large High School* handout (LMR_4.4).

Name: _____ Date: _____

Heights of Students at a Large High School

Background:

Our class will help judge a contest held at a particular high school. This school held a contest in which they selected students at random from a classroom and reported their heights.

They gave all of the students the data for the first 20 selected students (but you will not see these data!).

Each student was asked to predict the heights of the last 10 students. Here is the catch: **students were allowed to give only ONE number that had to be used to predict all 10 heights.**

Three student teams made predictions about the height of the last 10 students. The judges of this contest want you to tell them how they should determine the winner.

Instructions:

1. Your team's job is to determine the winning team, the team that came in second place, and the team that came in third place. Your team must come up with two things:
   a. You must support your choice of a winner by using a **rule** for calculating a total score for each team. The rule must be applied to each team's guess to determine their placement and your team must be able to explain how your rule helped select the winner.
   b. You must write instructions to the judges that explain how to use your rule to select a winner. For example, do they choose the team with the largest score? The smallest?
2. Each team's predictions are provided here, along with related plots (see page 2) for your reference.
3. Answer the questions that follow each of the plots.

Team Predictions:

Team A: 69 inches
Team B: 70 inches
Team C: 66 inches

Table of Prediction and Actual Outcomes:

Prediction of Team A	Prediction of Team B	Prediction of Team C	Plot A Heights	Plot B Heights
69	70	66	63	74
69	70	66	69	72
69	70	66	65.5	67.5
69	70	66	63.6	67
69	70	66	69	70
69	70	66	74	73
69	70	66	66	73
69	70	66	70	63
69	70	66	80	67
69	70	66	68	71.5

LMR_4.4

3. They gave all of the students the data for the first 20 selected students, but you will not see these data! Each student was asked to predict the heights of the last 10 students. Here is the catch: **students were allowed to give only ONE number that had to be used to predict all 10 heights.** .
4. Three student teams made predictions about the height of the last 10 students. The judges of this contest want you to tell them how they should determine the winner.
5. Your team's job is to determine the winning team, the team that came in second place, and the team that came in third place. Your team must come up with two things:
   a. You must support your choice of a winner by using a **rule** for calculating a total score for each team. The rule must be applied to each team's guess to determine their placement and your team must be able to explain how your rule helped select the winner.

b. You must write instructions to the judges that explain how to use your rule to select a winner. For example, do they choose the team with the largest score? The smallest?

6. Here are the predictions of the three teams:

   Team A: 69 inches
   Team B: 70 inches
   Team C: 66 inches

7. Display Plot A, found on page 2 of the *Heights of Students at a High School* handout (LMR_4.4). This dotplot displays the heights shown in the Actual Outcome column of the table. Inform students that this dotplot and table is provided to help them come up with a method to determine a winner. The table is to visualize the predicted heights side-by-side with the randomly selected heights.

   Notes to teacher:

   a. Students may have to be reminded that negative values with large absolute value are larger than positive values with small absolute values (e.g., 10 is larger than 3).

   b. Let students struggle for a little bit. A prompt to get them started: Look at the difference between a team's prediction and the actual outcomes (e.g., for the first height, Team A predicted 69, actual outcome was 63, so 69-63=6). They might also need to be nudged towards the *sum* of these differences – they need to produce a single score, not 10 separate scores.

   c. Here are some rules you can "feed" to the class to move them along. Ask them (a) Describe this rule in words. (b) Is it better to get a high score or a low score or some other score? (c) Which teams win for each? (Note, some of these rules produce ties).

      i. Rule 1: sum(heights-predicted.value == 0) ***words: the number of exactly correct predictions***

      ii. Rule 2: sum(heights-predicted.value) ***words: the sum of the differences between predicted value and the actual heights***

      iii. Rule 3: sum(abs(heights-estimate)) ***words: the sum of the absolute values of the deviations***

      iv. Rule 4: sum((heights-estimate)^2) ***words: the sum of the squared deviations***

      Note: It is unlikely that students will think of the last two. That's okay, because we will introduce them in a future lesson, but you might want to present one (or both) to see what they think about these rules.

8. Allow student teams time to discuss and complete the task for Plot A.

9. Do not share their responses to Plot A. Instead, display the following questions:

   a. What if we had a different set of 10 randomly selected students and plotted their heights?

   b. Would the same team win?

10. Allow teams to discuss the questions, then share a couple of responses to the questions in the previous step.

11. Display Plot B, found on page 2 of the *Heights of Students at a High School* handout (LMR_4.4), then have them find the winner using this new sample. Is it the same as they chose before?

    Note: We do NOT know the value of the true population mean/typical value. This is what we are really trying to predict.

12. Teams will take turns to share their work as follows:

    a. Which team did you select as the winner using Plot A?
    b. Explain the method, or **rule**, your team used to declare the winner.

c. Which team did you select as the winner using Plot B? Is the winner the same?
d. Did you use the same rule to select a winner or did it change? If it changed, explain.

13. During the share out, students will take notes about the other teams' rules in their DS journals.
14. Teams may continue to share at the start of the next lesson, if they run out of time.

Class Scribes:

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

Lesson 6: Statistical Predictions Applying the Rule

Objective:

Students will apply the rule statisticians use to determine the best method for predicting heights for students at a high school.

Materials:

1. Each team's rule for determining a winner (from previous lesson)
2. *Prediction Games* handout (LMR_4.5_Prediction Games)

Vocabulary:

mean squared deviation, mean absolute error

Essential Concepts: If we use the squared residuals rule, then the mean of our current data is the best prediction of future values. If we use the mean absolute error rule, then the median of the current data is the best prediction of future values

Lesson:

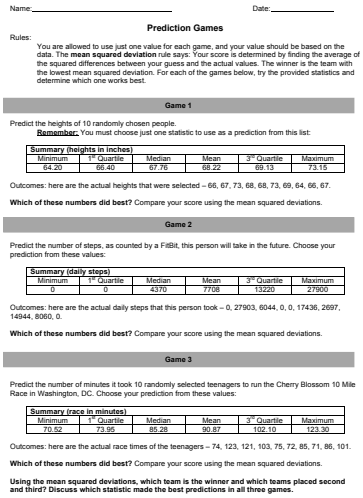
1. Ask students to recall that in the previous lesson, each student team created a rule to determine a winner. Which team's rules worked well for determining a winner?
2. Remind them that in their DS Journals, they took notes about each team's rule as they presented. This time, they will be switching roles – instead of creating a rule to judge the given predictions, they will be given a rule and it's their job to find the best procedure to win the contest.
3. Inform students that the question we are trying to answer is:

   How can we create a general rule that will always select the BEST guess to win no matter what 10 data points we are given?

4. Explain to students that data scientists use the "mean squared deviation" rule (also called the "mean squared error" or "mean squared residual" rule or "residual sums of squares" rule, the latter term being the most common). A "*deviation*" is the difference between our prediction and the actual outcome (as in MAD) and is sometimes called a "residual."

   Note to Teacher: Basically, students are being asked to determine which of these predicted values is "closest" to the data. One issue that comes up is dealing with positive and negative differences.

5. Distribute the *Prediction Games* handout (LMR_4.5).



Name:_____ Date:_____

Prediction Games

Rules:

You are allowed to use just one value for each game, and your value should be based on the data. The **mean squared deviation** rule says: Your score is determined by finding the average of the squared differences between your guess and the actual values. The winner is the team with the lowest mean squared deviation. For each of the games below, try the provided statistics and determine which one works best.

Game 1

Predict the heights of 10 randomly chosen people.

Remember: You must choose just one statistic to use as a prediction from this list:

Summary (heights in inches)					
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
64.20	66.40	67.76	68.22	69.13	73.15

Outcomes: here are the actual heights that were selected – 66, 67, 73, 68, 68, 73, 69, 64, 66, 67.

Which of these numbers did best? Compare your score using the mean squared deviations.

Game 2

Predict the number of steps, as counted by a FitBit, this person will take in the future. Choose your prediction from these values:

Summary (daily steps)					
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0	0	4370	7708	13220	27900

Outcomes: here are the actual daily steps that this person took – 0, 27903, 6044, 0, 0, 17436, 2697, 14944, 8060, 0.

Which of these numbers did best? Compare your score using the mean squared deviations.

Game 3

Predict the number of minutes it took 10 randomly selected teenagers to run the Cherry Blossom 10 Mile Race in Washington, DC. Choose your prediction from these values:

Summary (race in minutes)					
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
70.52	73.95	85.28	90.87	102.10	123.30

Outcomes: here are the actual race times of the teenagers – 74, 123, 121, 103, 75, 72, 85, 71, 86, 101.

Which of these numbers did best? Compare your score using the mean squared deviations.

Using the mean squared deviations, which team is the winner and which teams placed second and third? Discuss which statistic made the best predictions in all three games.

LMR_4.5

6. Explain the rules of the game as follows:

   You are allowed to use just one value for each game, and your value should be based on the data. The **mean squared deviation** rule says: Your score is determined by finding the average of the squared differences between your guess and the actual values. The winner is the team with the lowest mean squared deviation. For each of the games below, try the provided statistics and determine which one works best.

$$MSD = \frac{\sum_{i=1}^{n}(x_i - \hat{x})^2}{n}$$

   a. Game 1: Predict the heights of 10 randomly chosen people.
   Remember: You must choose just one statistic to use as a prediction from this list:

	Summary (heights in inches)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	64.20	66.40	67.76	68.22	69.13	73.15
MSD	**22.9**	**10.58**	**7.8056**	**7.7044**	**8.7509**	**33.1925**

	Summary (heights in inches)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	64.20	66.40	67.76	68.22	69.13	73.15
MAE	**3.94**	**2.34**	**2.1**	**2.188**	**2.578**	**5.05**

   Outcomes: here are the actual heights that were selected – 66, 67, 73, 68, 68, 73, 69, 64, 66, 67. Which of these numbers did best? Compare your score using the mean squared deviations.

   For example, using the minimum and outcomes above, gives you a mean squared deviation of:

$$\frac{\sum(66-64.20)^2+(67-64.20)^2+\ldots+(67-64.20)^2}{10} = \frac{229}{10} = 22.9 \text{ square in.}$$

   Note to teacher: The value of the mean squared deviation will always be in square units. In order to convert back to the original units, simply take the square root of the mean squared deviation.

   Interpretation: When using the minimum height to make predictions about all heights, our predictions will typically be off by $\sqrt{22.9} = 4.79$ ***inches.***

   b. Game 2: Predict the number of steps, as counted by a FitBit, this person will take in the future. Choose your prediction from these values:

	Summary (daily steps)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	0	0	4370	7708	13220	27900
MSD	**141,468,199**	**141,468,199**	**93,193,683**	**82,048,768**	**112,426,503**	**489,749,479**

	Summary (daily steps)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	0	0	4370	7708	13220	27900
MAE	**7708.4**	**7708.4**	**7169**	**7501.8**	**9636.2**	**20192.2**

Outcomes: here are the actual daily steps that this person took – 0, 27903, 6044, 0, 0, 17436, 2697, 14944, 8060, 0. Which of these numbers did best? Compare your score using the mean squared deviations.

Note to teacher: For Game 2, you might consider allowing students to utilize RStudio to calculate the mean squared deviation. The example below can be used to calculate the mean squared deviation for predicting daily steps using the minimum. Before revealing the codes, elicit a class discussion about how RStudio can be used to calculate the MSD.

Step 1: Create a vector of the given daily steps

```
> steps<-c(0,27903,6044,0,0,17436,2697,14944,8060,0)
```

Step 2: Store the squared deviations

```
> sqr_dev<-((steps-0)^2)
```

Step 3: Find the mean of the squared deviations

```
> mean(sqr_dev)
```

The code can be shortened to two steps if you apply a composition of the last two functions

```
> mean((steps-0)^2)
```

c. Game 3: Predict the number of minutes it took 10 randomly selected teenagers to run the Cherry Blossom 10 Mile Race in Washington, D.C.

	Summary (race in minutes)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	70.52	73.95	85.28	90.87	102.10	123.30
MSD	**777.0264**	**647.6125**	**387.3624**	**353.5429**	**474.49**	**1390.33**

	Summary (race in minutes)					
	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
	70.52	73.95	85.28	90.87	102.10	123.30
MAE	**20.58**	**18.13**	**15.7**	**16.674**	**19.14**	**32.2**

Outcomes: here are the actual race times of the teenagers – 74, 123, 121, 103, 75, 72, 85, 71, 86, 101. Which of these numbers did best? Compare your score using the mean squared deviations.

7. Using the mean squared deviations, which statistic is the winner and which statistics placed second and third? Discuss which statistic made the best predictions in all three games.

   Note to teacher: Explain that the mean worked best for all three contests. Data scientists (and mathematicians) can prove that the mean will **always** work best (except in a few weird cases from time to time). So if you want to predict the future, the mean is the best single guess you can make.

8. Ask: What if another data science class has a best rule that is different from ours?

9. Another agreed upon method that data scientists and statisticians often use is the **mean absolute error**. It's unlikely that students will figure this out on their own. The reasons why we do it in statistics date back to the 18th century, so it won't make a lot of sense; but it's what statisticians do. The mean absolute error is expressed as (where $\hat{x}$ stands for the predicted value):

$$MAE = \frac{\sum_{i=1}^{n} |x_i - \hat{x}|}{n}$$

10. Explain that each team will now use the statisticians' method for declaring a winner. Display the mean absolute error formula and discuss what each symbol means.
11. Using our previous examples, recalculate your predictions using the MAE.
12. Using the mean absolute error, which statistic is the winner and which statistics placed second and third?

Answers:

	Summary (heights in inches)					
	Minimum	***1st Quartile***	***Median***	***Mean***	***3rd Quartile***	***Maximum***
	64.20	***66.40***	***67.76***	***68.22***	***69.13***	***73.15***
MAE	***3.94***	***2.34***	***2.1***	***2.188***	***2.578***	***5.05***

	Summary (daily steps)					
	Minimum	***1st Quartile***	***Median***	***Mean***	***3rd Quartile***	***Maximum***
	0	***0***	***4370***	***7708***	***13220***	***27900***
MAE	***7708.4***	***7708.4***	***7169***	***7501.8***	***9636.2***	***20192.2***

	Summary (race in minutes)					
	Minimum	***1st Quartile***	***Median***	***Mean***	***3rd Quartile***	***Maximum***
	70.52	***73.95***	***85.28***	***90.87***	***102.10***	***123.30***
MAE	***20.58***	***18.13***	***15.7***	***16.674***	***19.14***	***32.2***

Note to teacher: Explain that in this instance, the median is the "winner." This means that the way you play the game depends on the rules of the game. If we used squared deviations, play with the mean. If we use the mean absolute error (MAE), play with the median.

Class Scribes:

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

Lesson 7: Statistical Predictions Using Two Variables

Objective:
Students will learn how to predict height using arm span data - and vice versa - visually on a scatterplot.

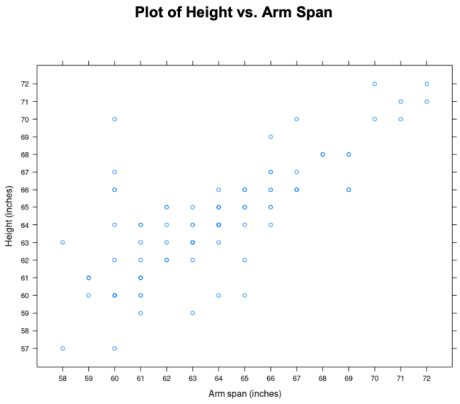
Materials:

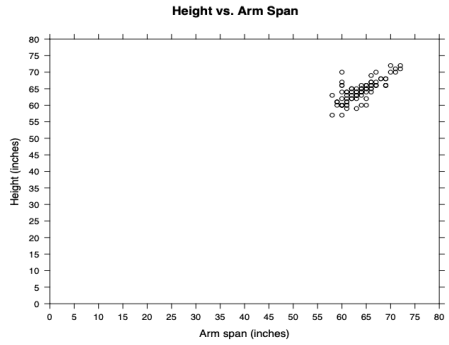
1. *Arm span vs. Height Scatterplot* (LMR_4.6_Arm Span vs Height)
   Note: This handout will be referenced in subsequent lessons.
2. Assorted color markers (dry erase or overhead)—See step 3 of lesson.
3. Overhead or LCD projector

Essential Concepts: When predicting values of a variable *y* - and if *y* is associated with *x* - then we can get improved predictions by using our knowledge about *x*. We essentially "subset" the data for a given value of *x* and use the mean *y* for those subset values. If the resulting means follow a trend, we can model this trend to generalize to as-yet unseen values of *x*.

Lesson:

1. Remind students that in the previous lessons they were working with height data to predict the typical height of all the students at a large high school, implementing a method used by statisticians to help them make good predictions.
2. In addition to the height data, it turns out that each student's arm span data was also collected and recorded.
3. Display the *Arm Span vs. Height Scatterplot* (LMR_4.6) on a white board or overhead projector (you will write on the board or the transparency later in the lesson—see step 9).





LMR_4.6

4. Distribute the *Arm Span vs. Height* handout (LMR_4.6). Students will refer to this handout again later in a subsequent lesson.

5. In teams, ask students to analyze the plot and discuss the following questions:

   - What kind of plot is this? ***Scatterplot.***
   - How many variables are displayed in this plot? ***Two variables.***
   - Which variable is shown on the x-axis? On the y-axis? ***Arm span is shown on the x-axis and height is shown on the y-axis.***
   - What is this plot showing? ***It is showing the relationship between a person's height and the person's corresponding arm span measurement.***
   - How can I find out the height of the person whose arm span measures 68 inches? ***Find 68 on the x-axis. Then find the data point located at 68. Place finger on the data point and track its location on the y-axis. The height is also 68 inches.***

6. Using Talk Moves, conduct a class discussion of the questions in step 5.
7. Remind students that we've learned that the mean is the best way of predicting heights. The mean heights of these people is 64 inches.
8. Ask the students: Do you think we can do better? Is 64 a good prediction for someone whose arm span is 72"? What about 60"? How can you come up with a rule for determining the best predicted height *if you know the person's arm span*?

   Note to teacher: Lead students to realize that they can do this by "subsetting" the data for the fixed *x* value. For example, if arm span is 60, they should consider only the heights of people whose arm span is 60 and find the mean.

9. In teams, ask students to approximate the mean height for people whose arm span is 60, 64, 68, and 72.

   Note: Because the plot does not clearly show duplicate ordered pairs, an approximation is sufficient at this point. You may have students use RStudio to calculate the mean height for the specific armspans. Refer to the OPTIONAL section at the end of this lesson.

10. Then plot these points on the graph. We'll use this later – the points should be roughly along a straight line. ***These arm spans have a range of height values associated with them. Students may take a mean of the heights, but answers may vary.***
11. Ask students if they see any patterns or rules they can use from this to help with predictions. Because there were multiple height values associated with each arm span length, you will likely get multiple answers from students. The goal now is to come up with a rule that suggests a plausible height value for anyone with a particular arm span.
12. A sentence starter to guide students: If a person has a bigger arm span, then we should predict____ [a bigger height]. If time permits, you might push them to be more precise. Let's take someone who has a 60 inch arm span. You predicted a height of ____. How much should we increase our prediction for people with a 62 inch arm span? Can you do this without subsetting the data and re-calculating?
13. Conceptually, students are wrestling with the notion of the slope of the regression line but there's no need to point this out just yet. Important: The equation of the line of best fit will be revealed in lesson 10.

OPTIONAL FOR ITEM 9 If you want to obtain the exact mean height for each arm span value in step 9, copy the code below and run it in an RScript.

```
xyplot(height~armspan, data = arm_span,
       scales = list(x = list(at = seq(58, 72, 1)),
       y = list(at = seq(52, 72, 1))),
       xlab = "Arm span (inches)", ylab = "Height (inches)")

armspan_60 <- filter(arm_span, armspan==60)
mean(~height, data = armspan_60)
#62.66667
```

```r
armspan_64 <- filter(arm_span, armspan==64)
mean(~height, data = armspan_64)
#64

armspan_68 <- filter(arm_span, armspan==68)
mean(~height, data = armspan_68)
#68

armspan_72 <- filter(arm_span, armspan==72)
mean(~height, data = armspan_72)
#71.5

#Base R Code
#syntax to create a scatterplot using base R
plot(arm_span$height, arm_span$armspan)

#Points function in base R is more user friendly
points(60, 62.66667, col = "red", cex = 2)
points(64, 64, col = "red", cex = 2)
points(68, 68, col = "red", cex = 2)
points(72, 71.5, col = "red", cex = 2)
```

Class Scribes:

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

Homework & Next 2 Days

LAB 4A: If the Line Fits…

LAB 4B: What's the Score?

Complete Labs 4A and 4B prior to Lesson 8.

Lab 4A - If the line fits ...

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

How to make predictions

- Anyone can make predictions.
  - Data scientists use data to inform their predictions by using the information learned from the sample to make predictions for the whole population.
- In this lab, we'll learn how to make predictions by finding the *line of best-fit*.
  - You will also learn how to use the information from one variable to make predictions about another variable.

Predicting heights

- Use the `data()` function to load the `arm_span` data.
- This data comes from a sample of 90 people in the Los Angeles area.
  - The measurements of `height` and `armspan` are in inches.
  - A person's `armspan` is the maximum distance between their fingertips when they spread their arms out wide.
- Make a plot of the `height` variable.
  - **If you had to predict the height of someone in the LA area, what single height would you choose and why?**
  - **Would you describe this as a *good* guess? What might you try to improve your predictions?**

Predicting heights knowing arm spans

- Create two subsets of our `arm_span` data:
  - One for `armspan >= 61 & armspan <= 63`.
  - A second for `armspan >= 64 & armspan <= 66`.
- Create a histogram for the `height` of people in each subset. Answer the following based on the data:
  - **What `height` would you predict if you knew a person had an `armspan` around 62 inches?**
  - **What `height` would you predict if you knew a person had an armspan around 65 inches?**
  - **Does knowing someone's `armspan` help you predict their height. Why or why not?**

Fitting lines

- Notice that there is a trend that people with a larger `armspan` also tend to have a larger mean `height`.
  - One way of describing this sort of trend is with a line.
- Data scientists often *fit* lines to their data to make predictions.
  - What we mean by *fit* is to come up with a line that's close to as many of the data points as possible.
- Create a scatterplot for `height` and `armspan`. Then run the following code. Draw a line by clicking twice on the *Plot* pane.

```
add_line()
```

Predicting with lines

- Draw a line that you think is a good *fit* and write down its equation. Using this equation:
  - **Predict how tall a person with a 62 and a person with a 65 inch `armspan` would be.**
- Using a line to make predictions also lets us make predictions for `armspans` that aren't in our data.
  - **How tall would you predict a person with a 63.5 inch `armspan` to be?**
- **Compare your answers with a neighbor's. Did both of you come up with the same equation for a line? If not, can you tell which line fits the data best?**

Regression lines

- If you were to go around your class, each student would have created a different line that they feel *fit* the data best.
  - Which is a problem because everyone's line will make slightly different predictions.
- To avoid this variation in predictions, data scientists will use *regression lines*.
  - This line connects the mean `height` of people with similar `arm_spans`.
  - Fill in the blanks below to create the a *regression line* using an `lm`, or *linear model*:

```
lm(____ ~ ____, data = arm_span)
```

Predicting with regression lines

- Use the output of the code from the previous slide to write down the equation of the *regression line* in the form

  `y = a + bx`.
- Add this line to a scatterplot by filling in the blanks below:

```
add_line(intercept = ____, slope = ____)
```

- Predict the height of a person with a 63.5 inch `armspan` and compare it with a neighbor. Ensure you both arrive at the same predicted value.
- **Measure your `armspan` and use the regression line to predict your height. How close was the prediction?**

Lab 4B - What's the score?

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

Previously

- In the previous lab, we learned we could make predictions about one variable by utilizing the information of another.
- In this lab, we will learn how to measure the accuracy of our predictions.
  - This in turn will let us evaluate how well a model performs at making predictions.
  - We'll also use this information later to compare different models to find which model makes the best predictions.

Predictions using a line

- Load the `arm_span` data again.
  - Create an `xyplot` with `height` on the y-axis and `armspan` on the x-axis.
  - Type `add_line()` to run the `add_line` function; you'll be prompted to click twice in the plot window to create a line that you think fits the data well.
- Fill in the blanks below to create a function that will make predictions of people's `heights` based on their `armspan`:

```
predict_height <- function(armspan) {
  ____ * armspan + ____
}
```

Make your predictions

- Fill in the blanks to include your predictions in the `arm_span` data.

```
____ <- mutate(____, predicted_height = ____(____))
```

- Now that we've made our predictions, we'll need to figure out a way to decide how accurate our predictions are.
  - We'll want to compare our *predicted heights* to the *actual heights*.
  - At the end, we'll want to come up with a single number summary that describes our model's accuracy.

Sums of differences

- A residual is the difference between the actual and predicted value of a quantity of interest.
- Fill in the blanks below to create a function which calculates the sum of differences:
- **What do residuals measure?**

```
____ <- mutate(____, residual = ____(____))
```

- **What do residuals measure?**
- One method we might consider to measure our model's accuracy is to sum the residuals.
- Fill in the blanks below to calculate our accuracy summary.

```
summarize(____, sum(____))
```

- Hint: Like `mutate`, the firs argument of `summarize` is a dataframe, and the second argument is the action to perform on a column of the dataframe. Whereas the output of `mutate` is a column, the output of `summarize` is (usually) a single number summary.

Checking our work

- **Describe and interpret, in words, what the output of your accuracy summary means.**
  - **Compare your accuracy summary with a neighbor's. Whose line was more accurate and why?**
- **Write down why adding positive and negative errors together is problematic for assessing prediction accuracy.**
  - **Why does calculating the squared values for the differences solve this problem?**
- The *mean squared error* (MSE) is calculated by squaring all of the residuals, and then taking the mean of the squared residuals.
- Fill in the blanks below to calculate the MSE of your line.

```
summarize(____, mean((____)^2)
```

On your own

- Create a *regression line* as you did in the previous lab, for `height` and `armspan`.
  - We also refer to *regression lines* as *linear models*.
  - Assign this model the name `best_fit`.
- Making predictions with models `R` is familiar with is simpler than with lines, or models, we come up with ourselves.
  - Fill in the blanks to make predictions using `best_fit`:

```
____ <- mutate(____, predicted_height = predict(____))
```

- Hint: the `predict` function takes a linear model as input, and outputs the predictions of that model.
- Calculate the MSE for these new predicted values.

The magic of lm()

- The `lm()` function creates the *line of best fit* equation by finding the line that minimizes the *mean squared error*. Meaning, it's the *best fitting line possible*.
  - Compare the MSE value you calculated using the line you fitted with `add_line()` to the same value you calculated using the `lm` function.
  - Ask your neighbors if any of their lines beat the `lm` line in terms of the MSE. Were any of them successful?
- To see how the `lm` line fits your data, create a scatterplot and then run:

```
add_line(intercept = ____, slope = ____)
```

Lesson 8: What's the Trend?

Objective:
Students will understand that the regression line is a model for a linear association (trend). They will learn to identify the direction and strength of trends.

Materials:

1. *What's the Trend?* handout (LMR_4.7_What's the Trend)
   Note: This handout will be referenced and used in subsequent lessons.
2. *Strength of Association* handout (LMR_4.8_Strength of Association)

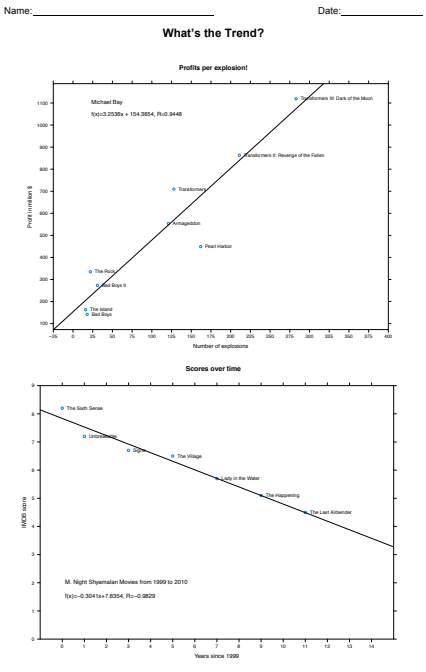
Vocabulary:

trend, positive association, negative association, no association, shape, linear, model, strength of association

Essential Concepts: Associations are important because they help us make better predictions; the stronger the trend, the better the prediction we can make. "Better" in this case means that our mean squared residuals can be made smaller.

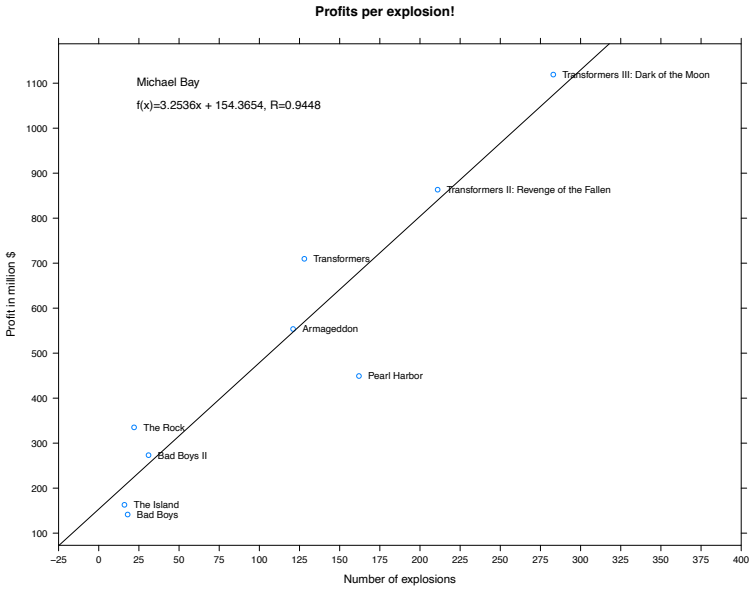
Lesson:

1. Distribute *What's the Trend?* (LMR_4.7). Students will analyze the two scatterplots on the handout. The *Profits per Explosion* plot shows the relationship between the number of explosions in Michael Bay's movies and the profit earned by each movie. The *Scores Over Time* plot shows the relationship between M. Night Shyamalan movies made since *The Sixth Sense* was released in 1999 and their Internet Movie Database (IMBD) scores.

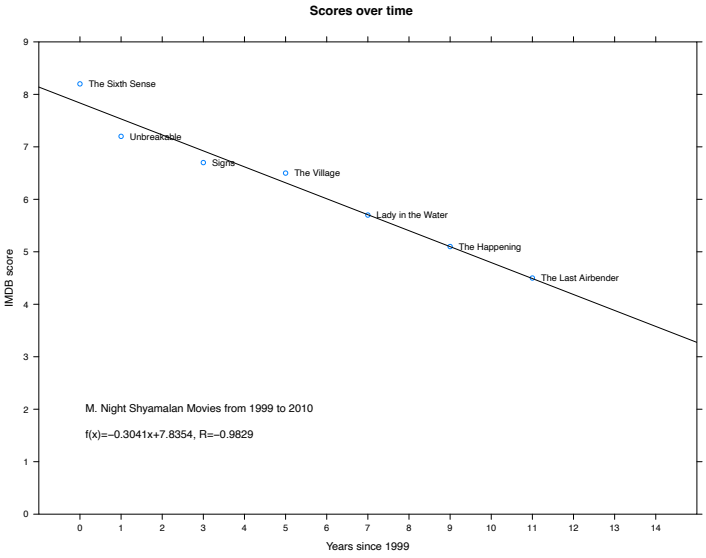


LMR_4.7

2. In teams, students will discuss and record their responses to the following questions for each plot:



a. What kind of plot is this? ***Scatterplot.***
b. What do the numbers on the x-axis represent? What do the numbers on the y-axis represent? ***The x-axis shows number of explosions and y-axis shows profit in millions of dollars.***
c. What is this plot telling us? ***Answers will vary. One example could be that if there are more explosions in a movie, then the movie will earn a greater profit.***



d. What kind of plot is this? ***Scatterplot.***
e. What do the numbers on the x-axis represent? What do the numbers on the y-axis represent? ***The x-axis shows the number of years since 1999 and the y-axis shows the movie's IMDB score.***
f. What is this plot telling us? ***Answers will vary. One example could be that as M. Night Shyamalan has produced more movies, their IMDB ratings have gone down.***

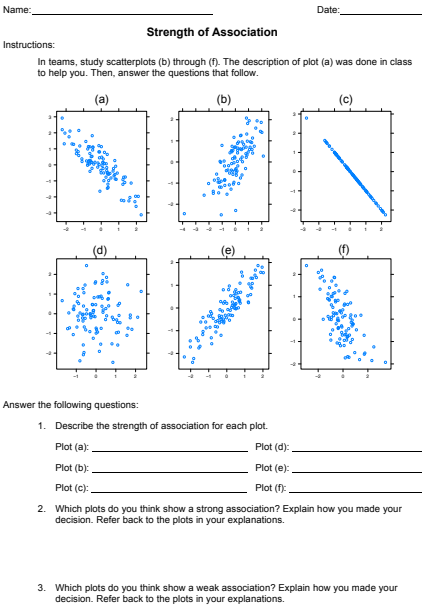
3. Allow students time to discuss and record their answers to the questions.
4. Display both plots, if possible (students may also refer to the plots in their own handout). Discuss the following questions with the whole class:

    - What is happening in each plot? What seems to be the trend? ***Guide students to understand that the Profits per Explosion plot shows an increasing trend, while the***

*Scores Over Time plot shows a decreasing trend. An increasing **trend** is called a **positive association** and a decreasing **trend** is called a **negative association**.*

- What does it mean to have an increasing trend and a positive association? *In Profits per Explosion, it means that as the number of explosions increase, the movie profits also increase.*
- What does it mean to have a decreasing trend and a negative association? *In Scores Over Time, it means that as the years after 1999 pass, the movie IMBD ratings decrease.*

5. Quickwrite: What if we had a plot with **no association**? Ask students to sketch what they think a scatterplot that shows no association looks like. *A correct sketch will show a scatterplot with data points that show no positive or negative association; no trend or pattern. There would be no association or a very weak one. The data would be scattered.*
6. Select a couple of sketches to share with the whole class. Discuss why the sketches show no association.
7. Ask students to discuss their thoughts about why a line was drawn through the points of the two plots and why there are equations for each plot.
8. Conduct a share out of their observations. Guide students to the understanding that the **shape** of both plots is **linear**. The line represents a *model* for the relationship between two variables. The equation shown in the plots above represents the line through the points. It provides a description of the data and the relationship between the variables.
9. Distribute *Strength of Association* (LMR_4.8_Strength of Association). In teams, students will examine the scatterplots (b) through (e). Their task is to discuss the **strength of the association** for each plot. They will determine which plots they think show strong associations and which ones they believe show weak associations. They must explain how they made their decision. Reasons must reference the plots.
10. As an example, demonstrate how to describe plot (a) in the *Strength of Association* handout. *Possible description: Plot (a) shows a negative association, or decreasing trend. The association appears to be fairly strong because the points are relatively close together, forming a moderate linear pattern*.



Name:_____ Date:_____

Strength of Association

Instructions:

In teams, study scatterplots (b) through (f). The description of plot (a) was done in class to help you. Then, answer the questions that follow.

Answer the following questions:

1. Describe the strength of association for each plot.

   Plot (a): _____ Plot (d): _____

   Plot (b): _____ Plot (e): _____

   Plot (c): _____ Plot (f): _____

2. Which plots do you think show a strong association? Explain how you made your decision. Refer back to the plots in your explanations.

3. Which plots do you think show a weak association? Explain how you made your decision. Refer back to the plots in your explanations.

LMR_4.8

11. Once all teams have completed the handout, assign one plot to each team for a share out. If two teams have the same plot, one team will share its explanation first and the second team can agree, disagree, or add to the first team's description.
12. Guide students to understand that a strong association has points closer to each other and a weak association has points more scattered.
13. If students run out of time, they will complete the remainder of the activity for homework.

Class Scribes:

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

Homework

Complete the *Strength of Association* handout (LMR_4.8_Strength of Association).

Lesson 9: The Spaghetti Line

Objective:

Students will estimate the line of best fit for a height and arm span data set using a strand of spaghetti as a modeling tool.

Materials:

1. *The Spaghetti Line* (LMR_4.9_The Spaghetti Line)
   Note: Advance preparation is required. Cut out plots prior to beginning the lesson.
2. *What's the Trend?* (LMR_4.7_What's the Trend) from lesson 8
3. *Arm Span vs. Height Scatterplot* (LMR_4.6_Arm Span vs Height) from Lesson 7
4. 1 lb. of Uncooked Spaghetti
5. Grid Paper
6. Tape or Glue
7. Poster paper

Vocabulary:

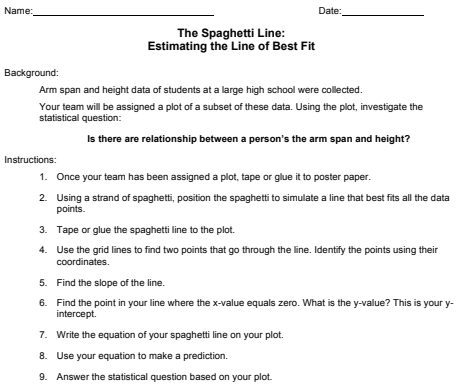
line of best fit, regression line

Essential Concepts: We can often use a straight line to summarize a trend. "Eye balling" a straight line to a scatterplot is one way to do this.

Lesson:

1. If necessary, begin by sharing out the descriptions for the plots in the *Strength of Association* (LMR_4.8_Strength of Association) handout from the previous lesson.
2. Inform students that in this lesson, they will estimate the equation of the **line of best fit** for a height and arm span data set.
3. Refer students back to the plots in the *What's the Trend?* handout (LMR_4.7_What's the Trend). The line in each of the plots is known as the **line of best fit**, or the **regression line**. This is a trend line that best represents or models the data in each scatterplot. Ask students:
   a. Why do you think this line is called "best fit"? ***Some possible answers are that it is a line that is closest to all data points or that it "fits" evenly among the data points. This is a good time to refer back to the discussion about height versus arm span in lesson 7.***
4. Distribute *The Spaghetti Line* (LMR_4.9_The Spaghetti Line) to each student and a couple of spaghetti strands per team. Students will estimate the line of best fit as outlined in the handout. Team solutions should be recorded on poster paper. They will glue their assigned plot on the poster and record their responses to the questions on the poster paper.

   Note to teacher: If necessary, review how to find the slope of a line using two points and how to write an equation using the slope and y-intercept.



Name:_____ Date:_____

The Spaghetti Line:
Estimating the Line of Best Fit

Background:

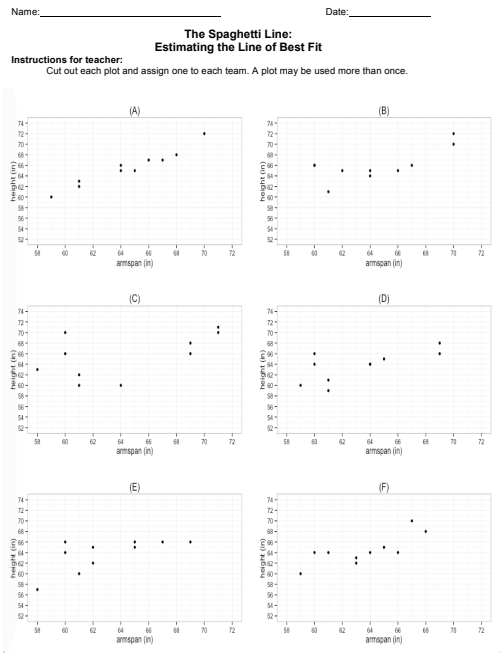
Arm span and height data of students at a large high school were collected.

Your team will be assigned a plot of a subset of these data. Using the plot, investigate the statistical question:

Is there are relationship between a person's the arm span and height?

Instructions:

1. Once your team has been assigned a plot, tape or glue it to poster paper.
2. Using a strand of spaghetti, position the spaghetti to simulate a line that best fits all the data points.
3. Tape or glue the spaghetti line to the plot.
4. Use the grid lines to find two points that go through the line. Identify the points using their coordinates.
5. Find the slope of the line.
6. Find the point in your line where the x-value equals zero. What is the y-value? This is your y-intercept.
7. Write the equation of your spaghetti line on your plot.
8. Use your equation to make a prediction.
9. Answer the statistical question based on your plot.



Name:_____ Date:_____

The Spaghetti Line:
Estimating the Line of Best Fit

Instructions for teacher:
Cut out each plot and assign one to each team. A plot may be used more than once.

LMR_4.9

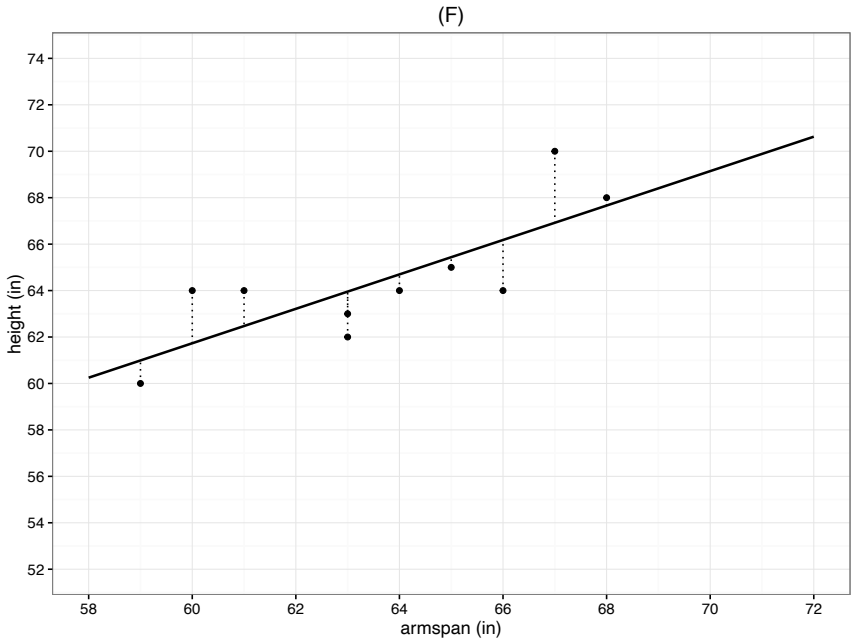
5. Ask teams to post their work around the room. Conduct a *Gallery Walk* so that teams can see each other's work.

6. Lead a discussion about the teams' lines. Ask: Which team has the best line? Why?

   Note to teacher: Push the students a bit by adding an obviously bad line to the graph and asking why their line is better than this one. Push them to come to an understanding that the "best" line comes close to the *most* points.

7. Inform students that data scientists have a way of finding the best line. They choose the line so that the mean squared distances between the points and the line is as small as possible. Discuss with students:
   a. What methods have we used so far? ***We've used Mean Squared Deviations and Mean Absolute Error (Lesson 6).***
   b. How did we use these methods? ***It was best to use Mean Squared Deviations when we are looking at mean and Mean Absolute Error when we are looking at median.***
   c. Which method do you think data scientists use most often? ***Data scientists often use MAE.***

8. [See graphic below] If time permits, ask students to calculate the distances and squares of two different lines so that they can understand what it means. This is the 2D version of the game they played in Lesson 6.



9. Inform students that they will see the equation of the arm span vs. height data in lesson 10.

Class Scribes:

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

Homework & Next Day

Students will use a straight edge to draw a line of best fit for the scatter plot in the *Arm Span vs. Height* handout (LMR_4.6_Arm Span vs. Height) from lesson 7. They will use their knowledge of slope and y-intercept to determine the equation for the line of best fit that they drew.

LAB 4C: Cross-Validation

Complete Lab 4C prior to Lesson 10.

Lab 4C - Cross-Validation

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

Predictions

- In the previous lab, we learned how to:
  - Create a linear model predicting `height` from the `arm_span` data (4A).
  - See how well our model predicts `height` on the `arm_span` data by computing mean squared error (MSE) (4B).
- In this lab, we will see how our model predicts heights of people *we haven't yet measured*.
- To do this, we will use a method called *cross-validation*.
- Cross-validation consists of three steps:
  - Step 1: Split the data into *training* and *test* sets.
  - Step 2: Create a model using the *training* set.
  - Step 3: Use this model to make predictions on the *test* set.

Step 1: train-test split

- Waiting for new observations can take a long time. The U.S. takes a census of its population once every 10 years, for example.
- Instead of waiting for new observations, data scientists will take their current data and divide it into two distinct sets.
- Split the `arm_span` data into `training` and `testing` data sets using the following steps.
- First, fill in the blanks below to randomly select which rows of `arm_span` will go into the `training` set.

```
set.seed(123)
train_rows <- sample(1:____, size = 85)
```

- Second, use the `slice` function to create two dataframes: one called `train` consisting of the `train_rows`, and another called `test` consisting of the remaining rows of `arm_span`.

```
train <- slice(arm_span, ____)
test <- slice(____, - ____)
```

- **Explain these lines of code and describe the `train` and `test` data sets.**

Aside: set.seed()

- When we split data, we're randomly separating our observations into *training* and *testing* sets.
  - It's important to notice that no single observation will be placed in both sets.
- Because we're splitting the data sets randomly, our models can also vary slightly, person-to-person.
  - This is why it's important to use `set.seed`.
- By using `set.seed`, we're able to reproduce the random splitting so that each person's model outputs the same results.
  Whenever you split data into training and testing, always use `set.seed` first.

Aside: train-test ratio

- When splitting data into *training* and *testing* sets, we need to have enough observations in our data so that we can build a good model.
  - This is why we kept 85 observations in our `training` data.
- As data sets grow larger, we can use a larger proportion of the data to *test* with.

Step 2: train the model

- Step 2 is to create a linear model relating `height` and `armspan` using the `training` data.
- Fit a line of best fit model to our `training` data and assign it the name `best_train`.
- Recall that the slope and intercept of our linear model are chosen to minimize MSE.
- Since the MSE being minimized is from the training data, we can call it *training MSE*.

Step 3: test the model

- Step 3 is to use the model we built on the `training` data to make predictions on the `test` data.
- Note that we are NOT recomputing the slope and intercept to fit the test data best. We use the same slope and intercept that were computed in step 2.
- Because we're using the *line of best fit*, we can use the `predict()` function we introduced in the last lab to make predictions.
  - Fill in the blanks below to add predicted heights to our `test` data:

```
test <- mutate(test, ____ = predict(best_train, newdata = ____))
```

- Hint: the `predict` function without the argument `newdata` will output predictions on the `training` data. To output predictions on the `test` data, supply the `test` data to the `newdata` argument.
- **Calculate the MSE in the same way as you did in the previous lab (test MSE is simply MSE of the predictions on the test data).**

Recap

- Another way to describe the three steps is
- Step 1: Split the data into `training` and `test` sets.
- Step 2: Choose a slope and intercept that minimize training MSE.
- Step 3: Using the same slope and intercept from step 2, make predictions on the `test` set, and use those predictions to compute test MSE.
- This begs the question, why do we care about test MSE?

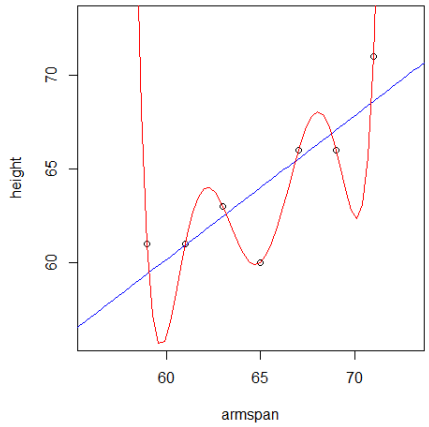
Why cross-validate?

- Why go to all this trouble to compute test MSE when we could just compute MSE on the original dataset?
- When we compute MSE on the original dataset, we are measuring the ability of a model to make predictions on *the current batch of data*.
- Relying on a single dataset can lead to models that are so specific to the current batch of data that they're unable to make good predictions for future observations.
  - This phenomenon is known as *overfitting*.
- By splitting the data into a training and test set, we are *hiding a proportion of the data* from the model. This emulates future observations, which are unseen.

- Test MSE estimates the ability of a model to make predictions of *future observations*.

Example of overfitting

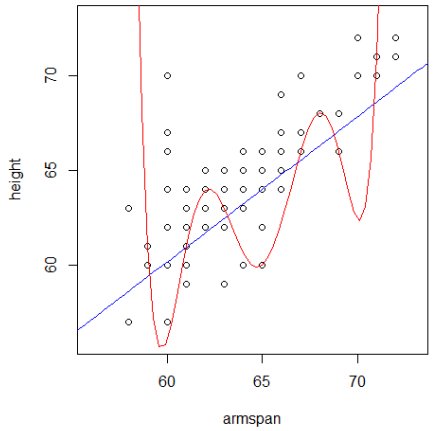
- The following example motivates cross-validation by illustrating the dangers of overfitting.
- We randomly select 7 points from the `arm_span` dataset and fit two models: a linear model, and a polynomial model.
  - You will learn how to fit a polynomial model in lab 4F.
- Below is a plot of these 7 `training` points, and two cures representing the value of height each model would predict given a value of armspan.



- **Which model does a better job of predicting the 7 `training` points?**
- **Which model do you think will do a better job of predicting the rest of the data?**

Example of overfitting, continued

- Below is a plot of the rest of the `arm_span` dataset, along with the predictions each model would make.



- **Which model does a better job of generalizing to the rest of the `arm_span` dataset?**

Lesson 10: Predicting Values

Objective:

Students will learn how to make predictions based on linear models.

Materials:

1. *What's the Trend?* handout (LMR_4.7_What's the Trend) from lesson 8
2. *Predicting Values* handout (LMR_4.10_Predicting Values)

Vocabulary:

observed value, predicted value

Essential Concepts: The regression line can be used to make good predictions about values of *y* for any given value of *x*. This works for exactly the same reason the mean works well for one variable: the predictions will make your score on the mean squared residuals as small as possible.

Lesson:

☑ 1. Entrance ticket: How do you find the equation of a line using two points?

Note to Teacher: Students may share their responses with a partner or you may choose to use this ticket as an assessment.

2. Reveal the equation of the line of best fit for the Arm Span vs. Height data and ask students to check their equations from the homework assignment:

$$\widehat{\text{height}} = 0.7328(\text{armspan}) + 17.4957$$

Note: Any time a *hat* is on top of a variable, this means we are making "predicted values" of that variable.

3. Whose equation came closest to the equation of the regression line? Ask the student whose equation came closest to share how he/she came up with the equation.

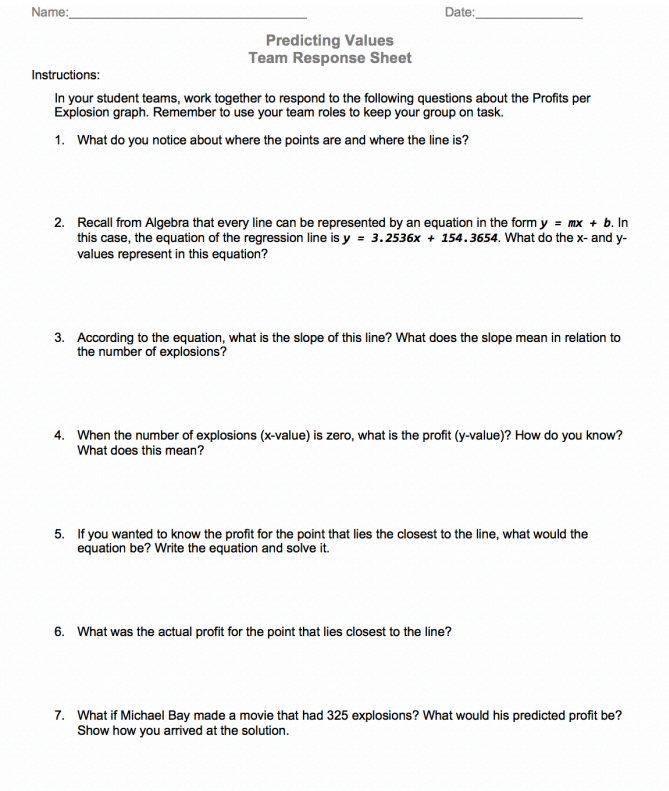
4. Inform students that the equation of the line is a rule that predicts the height based on a second variable, in this case, arm span.

🗨 5. Team discussion question:

Using the equation of the line of best fit provided, how can we predict the height of a student whose arm span is 67 inches?

6. Remind students that lines of best fit are also known as regression lines and they are models that can be used to make predictions. Today, they will explore more about this line.

🗨 7. Ask student teams to refer back to *What's the Trend?* Handout (LMR_4.7). They should discuss the following questions and record their responses on the *Predicting Values* handout (LMR_4.10):



Name:_____ Date:_____

Predicting Values
Team Response Sheet

Instructions:

In your student teams, work together to respond to the following questions about the Profits per Explosion graph. Remember to use your team roles to keep your group on task.

1. What do you notice about where the points are and where the line is?

2. Recall from Algebra that every line can be represented by an equation in the form $y = mx + b$. In this case, the equation of the regression line is $y = 3.2536x + 154.3654$. What do the x- and y-values represent in this equation?

3. According to the equation, what is the slope of this line? What does the slope mean in relation to the number of explosions?

4. When the number of explosions (x-value) is zero, what is the profit (y-value)? How do you know? What does this mean?

5. If you wanted to know the profit for the point that lies the closest to the line, what would the equation be? Write the equation and solve it.

6. What was the actual profit for the point that lies closest to the line?

7. What if Michael Bay made a movie that had 325 explosions? What would his predicted profit be? Show how you arrived at the solution.

LMR_4.10

a. What do you notice about where the points are and where the line is? ***Some points are near the line, others are further away, and one point is exactly on the line. Data points are observed values and points on the line are predicted values.***

b. Recall from Algebra that every line can be represented by an equation in the form $y = mx + b$. In this case, the equation of the regression line is $y = 3.2536x + 154.3654$. What do the x- and y-values represent in this equation? ***The x-values represent the number of explosions and the y-values represent the predicted profit.***

c. According to the equation, what is the slope of this line? What does the slope mean in relation to the number of explosions? ***The slope is 3.2536. It is the rate of change between the number of explosions and the profit. It means that for every explosion increase the profit increases by 3.2536 dollars.***

d. When the number of explosions (x-value) is zero, what is the profit (y-value)? How do you know? What does this mean? ***The profit is 154.3654 million dollars. Students may use the equation to show that they substituted zero for x, so the y-intercept is the profit. It means that if Michael Bay were to make a movie with NO explosions, this would be his projected profit.***

e. If you wanted to know the profit for the point that lies the closest to the line, what would the equation be? Write the equation and solve it. ***Profit=3.2536(211)+154.3654. Profit=840.875 or 840,875,000 million dollars.***

f. What was the actual profit for the point that lies closest to the line? ***The actual profit was 836,303,693 million dollars.***

g. What if Michael Bay made a movie that had 325 explosions? What would his predicted profit be? Show how you arrived at the solution. ***By substituting 325 in the value of x in the equation, predicted profit will be $1,211,785,400 or $ 1,211.7854, or by finding the point on the line or both.***

8. Assign one question to each team for a share out. If two teams have the same question, one team will share its explanation first and the second team can agree, disagree, or add to the first team's explanation.

   Note: If students ask/wonder about the meaning of the R^2, inform them that it is related to R, also known as the correlation coefficient. They will learn about R (not R^2)in lesson 11.

Class Scribes:

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

Homework

Students will answer the following questions about the *Scores Over Time* plot (LMR_4.7):

- What do you notice about where the points are and where the line is?
- What do the y-and x-values represent in this equation?
- According to the equation, what is the slope of this line? What does the slope mean?
- When the x-value is zero, what is the y-value? How do you know? What does this mean?
- What would the predicted value of the score be if M. Night Shyamalan released a movie in 2015? How do you know?

Lesson 11: How Strong Is It?

Objective:

Students will learn that the correlation coefficient is a value that measures the strength in linear associations only.

Materials:

1. *Correlation Coefficient* handout (LMR_4.11_Correlation Coefficient)
   Note: Advance preparation required. This handout is the resource for the plot cutouts. DO NOT distribute as-is to students.

Vocabulary:

correlation coefficient

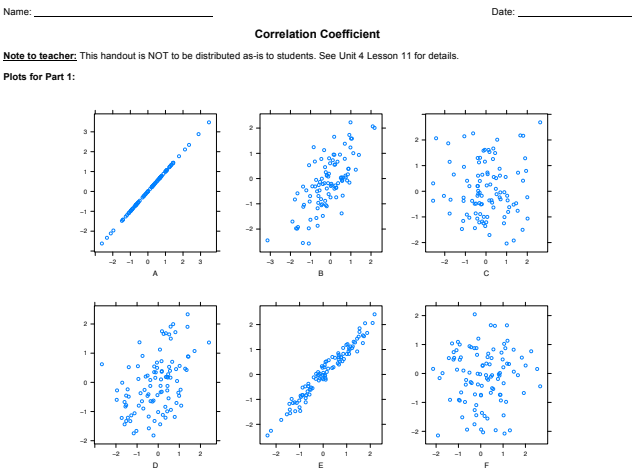
> **Essential Concept**: A high absolute value for correlation means a strong linear trend. A value close to 0 means a weak linear trend.

Lesson:

1. Inform students that, so far, they have been labeling associations as strong, very strong, or weak. A number called the **correlation coefficient** measures strength of association. The correlation coefficient only applies to linear relationships, which must be checked visually with a scatterplot. Later we will learn how to calculate this number using RStudio.

   Note to teacher: Advance preparation is needed for this lesson. Each team needs one envelope with cutouts of plots A-F in LMR_4.11 (Part A). Make envelopes according to the number of teams in the class. This process will be repeated for LMR_4.11 (Part B).

2. Distribute the envelopes to the teams. Students will examine the strength of association in each plot. Their task is to assign the correlation coefficient that corresponds to each plot and to explain why they assigned that correlation coefficient to that particular plot. The only piece of information they will receive is that a correlation coefficient equal to 1 has the strongest linear association and a correlation coefficient equal to 0 has the weakest association.



Name: _____ Date: _____

Correlation Coefficient

Note to teacher: This handout is NOT to be distributed as-is to students. See Unit 4 Lesson 11 for details.

Plots for Part 1:

A B C D E F

LMR_4.11

3. Assign each team one plot. If there are more teams than plots, these teams will be assigned a plot in the next round. Each team will share the correlation coefficient they assigned to their plot and the explanation that goes with it.

4. Using the *Voting Cards* strategy (see Instructional Strategies), the rest of the teams will show whether they approve, disapprove, or are uncertain about the teams' assignment and/or explanation. Repeat for each plot. The correlation coefficients for each plot are:
    - ***Plot A: r = 1.00***
    - ***Plot B: r = 0.72***
    - ***Plot C: r = 0.19***
    - ***Plot D: r = 0.48***
    - ***Plot E: r = 0.98***
    - ***Plot F: r = 0.00***
5. The last set of plots showed positive associations. Now students will assign the correlation coefficients for plots G-L for LMR_4.11 (Part 2).
6. Distribute the envelopes to the teams. Students will examine the strength of association in each plot. Their task is to assign the correlation coefficient that corresponds to each plot and to explain why they assigned that correlation coefficient to that particular plot. The only piece of information they will receive is that a correlation coefficient equal to -1 has the strongest linear association and a correlation coefficient equal to 0 has the weakest association.
7. Teams previously not assigned a plot are now assigned one. Each team will share the correlation coefficient they assigned to their plot and the explanation that goes with it.

8. Using the Voting Cards strategy, the rest of the teams will show whether they approve, disapprove, or are uncertain about the teams' assignment and/or explanation. Lead a class discussion whenever there is disapproval or uncertainty. Repeat for each plot. The correlation coefficients for each plot are:
    - ***Plot G: r = -1.00***
    - ***Plot H: r = 0.72***
    - ***Plot I: r = -0.19***
    - ***Plot J: r = -0.48***
    - ***Plot K: r = 0.98***
    - ***Plot L: r = 0.00***
9. Journal Entry: What is a correlation coefficient, what does it do, and what does it tell us about a scatterplot?

Homework & Next Day

Students will complete journal entry for homework if not completed in class.

LAB 4D: Interpreting Correlations

Complete Lab 4D prior to Lesson 12.

Lab 4D - Interpreting correlations

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

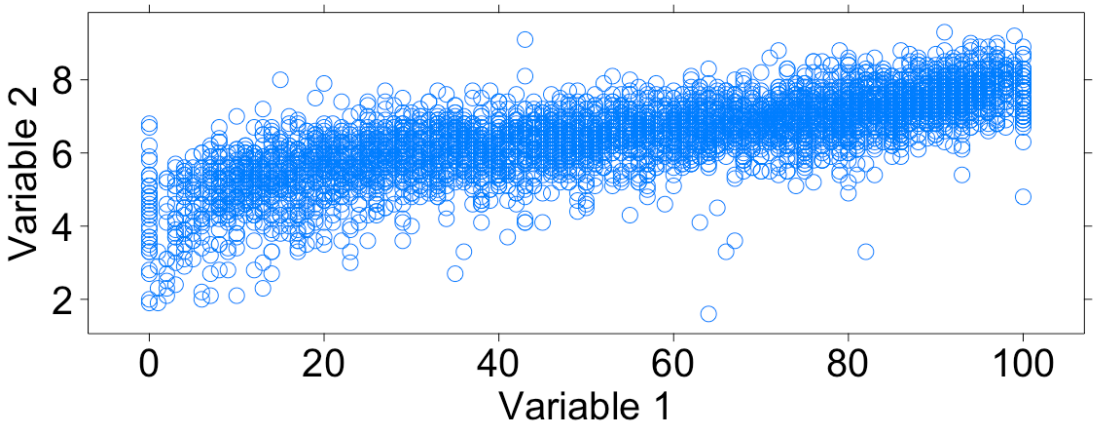
Some background...

- So far, we've learned about measuring the success of a model based on how close its predictions come to the actual observations.
- The *correlation coefficient* is a tool that gives us a fairly good idea of how these predictions will turn out without having to make predictions on future observations.
- For this lab, we will be using the `movie` data set to investigate the following questions:

  Which variables are better predictors of a movie's `audience_rating` when the predictions are made using a line of best fit?

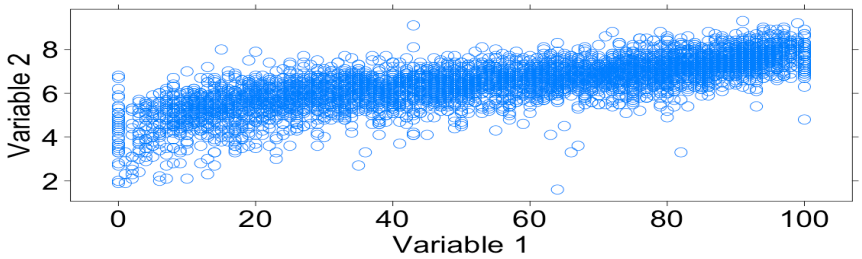
Correlation coefficients

- The *correlation coefficient* describes the *strength* and *direction* of the linear trend.
- It's only useful when the trend is linear and both variables are numeric.



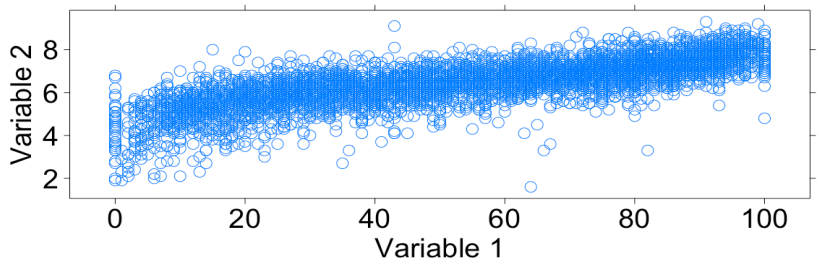
- **Are these variables linearly related? Why or why not?**

Correlation review I



- Correlation coefficients with values close to 1 are very strong with a positive slope. Values close to -1 means the correlation is very strong with a negative slope.
  - **Does this plot have a positive or negative correlation?**

Correlation review II



- **Recall that if there is no linear relation between two numerical variables, the correlation coefficient is close to 0. What do you guess the correlation coefficient will be for these two variables?**

The movie data

- Load the `movie` data using the `data` command.
- The data comes from a variety of sources like *IMDB* and *Rotten Tomatoes*.
  - The `critics_rating` contains values between 0 and 100, 100 being the best.
  - The `audience_rating` contains values that range between 0 and 10, 10 being the best.
  - `n_critics` and `n_audience` describe the number of reviews used for the ratings.
  - `gross` and `budget` describes the amount of money the film made and took to make.

Calculating Correlation Coefficients!

- We can use the `cor()` function to find the particular correlation coefficient of the variables from the previous plot, which happen to be `audience_rating` and `critics_rating`.
  - But note, the `cor()` function removes any observations which contains an `NA` value in either variable.
  - **Calculate the correlation coefficient for these variables using the `cor` function. The inputs to the functions work just like the inputs of the `xyplot` function.**

Now answer the following.

- **What was the value of the correlation coefficient you calculated?**
- **How does this actual value compare with the one you estimated previously?**
- **Does this indicate a strong, weak, or moderate association? Why?**
- **How would the scatter plot need to change in order for the correlation to be stronger?**
- **How would it need to change in order for the correlation to be weaker?**

Correlation and Predictions

- Find the two variables that look to have the strongest correlation with `critics_rating`.
  - Compute the correlation coefficients for `critics_rating` and each of the two variables.
  - **Use the correlation coefficient to determine which variable has a stronger linear relationship with `critics_rating`.**
- Fit two `lm` models to predict `critics_rating` with each variable and compute the MSE for each.

    - **Use the MSE to determine which variable is a better predictor of `critics_rating`.**
- **How are the correlation coefficient and the MSE related?**

On your own

- Select two different numerical variables from the `movie` data.
- Plot the variables using the `xyplot()` function.
    - **Would calculating a correlation coefficient for the two variables be appropriate? Justify your answer.**
    - **Predict what value you think the correlation coefficient will be. Compare this value to the actual value. Finally, interpret what the actual correlation coefficient means.**
- Work with your classmates to determine which two variables have the strongest correlation coefficient.
- **Why do you think these variables are so strongly related? Is using the correlation coefficient to describe the relationship appropriate and why/why not?**

Piecing it Together

Instructional Days: 5

Enduring Understandings

Real-life phenomena are often complex. Data scientists use multiple regression models to create simple equations to help explain and predict these phenomena. Data scientists can also use polynomial transformations to add flexibility to rigid linear models.

Engagement

Students will read the article titled *How Long Can a Spinoff Like Better Call Saul Last?* that will set the context for students to begin thinking about more than one explanatory variable to make better predictions. The article can be found at:
http://fivethirtyeight.com/features/how-long-can-a-spinoff-like-better-call-saul-last/

Learning Objectives

Statistical/Mathematical:

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

a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. *Use given functions or choose a function suggested by the context. Emphasize linear models*.

Data Science:

Understand that multiple regression can be a better tool for predicting that simple linear regression and know when it is appropriate to use multiple regression versus simple linear regression. Understand when linear models are not appropriate based on the shape of the scatterplot.

Applied Computational Thinking using RStudio:

- Use multiple linear regression models with other predictor variables
- Fit regression lines to data and predict outcomes.
- Create non-linear models to look for relationships.
- Fit polynomials functions to data.

Real-World Connections:

Economists and marketing firms use multiple regression to predict changes in the market and adjust strategies to fit the demands of changes in the marketplace.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data Set:

1. NFL data set
2. USMNT data set

Data File:

3. Movies: `data(movie)`

Legend for Activity Icons


Video clip


Discussion


Articles/Reading


Assessments


Class Scribes

Lesson 12: More Variables to Make Better Predictions

Objective:

Students will see that information from different variables can be used together to create linear models that make more accurate predictions.

Materials:

1. *Advertising Plots Part 1* handout (LMR_4.12_Advertising Plots 1)
2. *Advertising Plots Part 2* handout (LMR_4.13_Advertising Plots 2)
3. *Article: How Long Can a Spinoff Like 'Better Call Saul' Last?* http://fivethirtyeight.com/features/how-long-can-a-spinoff-like-better-call-saul-last/

Vocabulary:

market

Essential Concepts: We can use scatterplots to assess which variables might lead to strong predictive models. Sometimes using several predictors in one model can produce stronger models.

Lesson:

1. Remind students that models are used to make predictions. Ask a volunteer to think of a TV show that had a "spinoff" and to name both of the shows. Ask if he/she knows whether or not the original was more or less successful than the spinoff. Then, ask the class: Is there a way to predict spinoff success?

2. Next, using the *Talking to the Text* instructional strategy, ask students to read the article titled: *How Long Can a Spinoff Like Better Call Saul Last?*

   Note: If this is the first time using this strategy with your students, make sure you model/explain it before they begin reading it. See Instructional Strategies in Teacher Resources for a description.

3. After reading the article, ask students to discuss three *Talking to the Text* responses with a partner. You may set a time limit for each student to share with his/her partner.

4. Then, in teams, students will answer the following questions pertaining to the article:

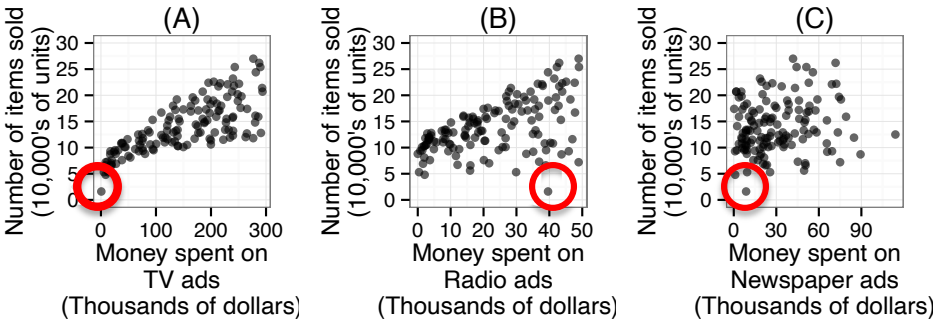
   a. What is the article trying to predict?
   b. How many variables are used?
   c. What other variables might affect a spinoff?
   d. The dotted line in the plot is not a regression line. How would you draw a regression line to make predictions?
   e. What other information would you like to know to predict a spinoff's success?

5. Allow students time to discuss and record their answers. Then conduct a share out of their responses to the discussion questions.

6. Discuss the following questions with the class:

   a. What effect does advertising have on retail sales?
   b. Where do stores advertise (What mediums do they use)? Does each method of advertisement reach the same people?
   c. Does each method of advertisement have a similar effect? Or are some methods more effective than others?

7. Distribute the 3 plots from the *Advertising Plots Part 1* handout (LMR_4.12) and inform the students about the data using the details below:



LMR_4.12

(Plots are presented separately in the LMR)

a. These 3 plots show the number of items sold by a retailer (in 200 different markets) and the amount of money the company spent on *TV*, *Radio* and *Newspaper* advertisements.

b. The data has 200 observations, one for each different market. A **market** is simply a location where an item is sold. For example, Los Angeles and San Francisco are two different markets.

c. Each observation has 4 variables: (1) The number of items sold (in 10's of thousands of units), (2) the money spent on TV ads (in thousands of dollars), (3) the money spent on radio ads (in thousands of dollars), and (4) the money spent on newspaper ads (in thousands of dollars).

d. The data were collected using an observational study.

8. To illustrate a-d above, ask students to refer to plot A (TV ads) and circle the market in which this retailer sold the least number of items (see circles in plots above). Ask: How many items did this market sell? ***About 20,000 items. The actual number of items sold was 1.6 (in 10,000's of units) which is 16,000 items***. How much money did this retailer spend on TV ads in this market? ***This retailer spent zero dollars on TV ads***. ***The actual amount the retailer spent on TV ads was 0.7 thousands of dollars, which is $700.***

9. Students should then refer to plot B (Radio ads), find the same market (the one in which the retailer sold about 20,000 items) and circle it. Ask: How much money did the retailer spend on Radio ads in the same market? ***About 40 thousand dollars***. ***The actual amount spent on Radio ads was 39.6 thousands of dollars, which is $39,600***.

10. Finally, ask students to refer to plot C (Newspaper ads), find the same market (the one in which the retailer sold about 20,000 items), and circle it. Ask: How much money did the retailer spend on Newspaper ads in the same market? ***About 10 thousand dollars***. ***About The actual amount spent on Newspaper ads is 8.7 thousands of dollars, which is $8,700***

TV	Radio	Newspaper	Sales
0.7	39.6	8.7	1.6

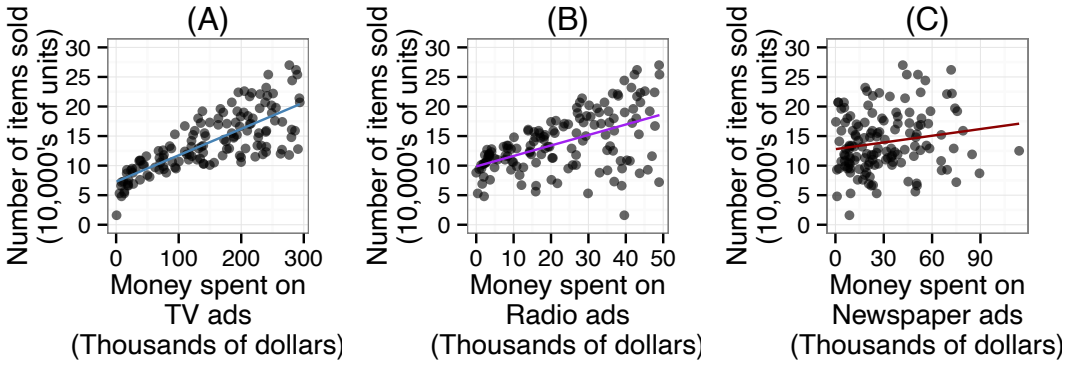
.

11. Based on the above plots, use a Pair-Share to discuss the following:

a. Describe the relationship between advertisements and the number of items sold.

b. Which type of advertisement is the most strongly correlated with the number of units sold? How can you tell?

12. Distribute the *Advertising Plots Part 2* handout (LMR_4.13_Advertising Plots 2), which contains plots A-C, but now include the line of best fit.



LMR_4.13

(Plots are presented separately in the LMR)

13. Ask students to recall from Lesson 6 that a method statisticians use to figure out which predicted values is closest to the actual data is the mean absolute error (MAE).

    Note to teacher: In the labs, students will use the mean squared error (MSE) - also learned in Lesson 6 - which calculates the regression line. In the lessons, we discuss the issue more generally using the mean absolute error (MAE).

14. In teams, ask students to discuss the following:

    a. How would you use the mean absolute error to determine which plot would make the most accurate predictions? ***Answers will vary, but you would expect to hear something like: "the prediction line that has the least amount of distance to all the points on the plot would make the most accurate prediction because the predicted values will be closer to the actual data."***

15. Next, have students select a statement they think is best (a or b), then write a justification for their selection based on what they learned in this lesson. This may be completed as homework.

    a. Combining multiple variables (e.g., TV and Newspaper ads, TV and Radio ads, TV, Radio, and Newspaper ads, etc.) into one model will lead to worse predictions because the variables that make poor predictions will contaminate those that make good predictions.

    b. Combining multiple variables (e.g., TV and Newspaper ads, TV and Radio ads, TV, Radio, and Newspaper ads, etc.) into one model will lead to better predictions because the model can use more information to make predictions.

16. Inform students that RStudio has the capability of creating models that combine multiple variables to make predictions about another variable. For example, it can make a model to predict number of items sold using both money spent on TV and money spent on Newspaper ads. Students will learn more about it during the next lesson.

Class Scribes:

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

Homework

Students may continue writing their justifications for the selected statement in item 15 if they were unable to finish.

Lesson 13: Combination of Variables

Objective:

Students will learn that we can make better predictions by including more variables. Then they will wrestle with how the information should be combined.

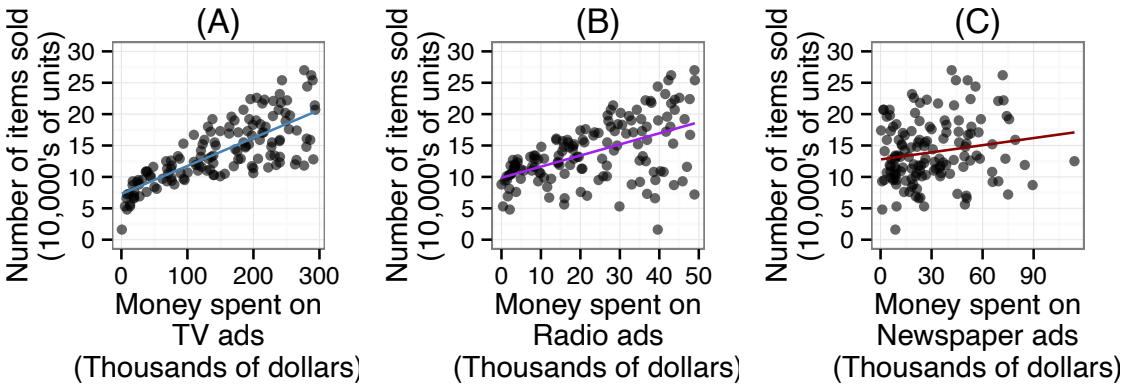
Materials:

1. *Advertising Plots Part 2* handout (LMR_4.13_AdvertisingPlots2) from Lesson 12

Essential Concepts: If multiple predictors are associated with the response variable, a better predictive model will be produced, as measured by the mean absolute error.

Lesson:

1. Display the plots and statements from the previous day:



LMR_4.13
(Plots are presented separately in the LMR)

   a. Combining multiple variables (e.g., money spent on TV and Newspaper ads, TV and Radio ads, TV, Radio, and Newspaper ads, etc.) into one model will lead to worse predictions because the variables that make poor predictions will contaminate those that make good predictions.
   b. Combining multiple variables (e.g., TV and Newspaper ads, TV and Radio ads, TV, Radio, and Newspaper ads, etc.) into one model will lead to better predictions because the model can use more information to make predictions.

2. Ask the students to share out their opinions in an Active Debate (see Unit 2 Lesson 6 as an example).
3. Next, inform teams that they will have 2 minutes to come up with as many combinations of ads (variables) as they can think of (e.g., TV + Newspaper ads, TV+ Radio ads, TV + Radio + Newspaper ads, etc.)
4. After 2 minutes, list all the different combinations by conducting a Whip Around and eliciting a combination from each team.
5. By a show of hands, ask students to select which combination or single model will be the best predictor for the number of items sold by the retailer.

6. Then inform students that we will determine which of the statements is true by comparing the mean absolute error (MAE) of single models (like the ones we showed in the previous lesson) vs. combined models. But first, use the line of best fit for the combined variables:

$$\widehat{sales} = 0.045449(tv) + 0.186570(radio) - 0.004952(newspaper) + 3.029878$$

Note: The function that produced the line of best fit using RStudio was
`lm(Sales ~ TV + Radio + Newspaper, data= retail)`

   a. Use this equation to predict the amount of sales for the same market they circled in the previous lesson. ***Students' calculation should yield the predicted value in (b), below***.

   Note: Remind students that they need to substitute the values as they appear in the x-axis of the plots without converting to thousands of dollars. For example, the circled market spent about 10 thousand dollars on newspaper ads, so students should substitute 10 instead of the expanded value in the equation.

TV	Radio	Newspaper	Sales
0.7	39.6	8.7	1.6

   b. Does the predicted value (10.407) seem like a plausible number of sales? Why? ***It is not a plausible number of sales because the prediction is too high. The prediction says the retailer will sell about 104,070 units, when the actual sales were about 16,000 units. Although the model did not make a very good prediction for this market, it is not surprising because as LMR_4.13 displays, that market did not fit the overall pattern in any of the scatterplots.***

7. Reveal that RStudio calculated the mean absolute error for different combinations plus the single models, and the results are displayed on the table below. This means that, for example, when using the TV model to predict number of items sold, our predictions will typically be off by about 2.337808 (in 10,000s) of units or 23,378 units. Then ask students:

Model	Mean Absolute Error
TV	2.337808
Radio	3.565113
Newspaper	4.538444
TV-Radio	1.160937
TV-Newspaper	2.344971
Radio-Newspaper	2.93832
TV-Radio-Newspaper	1.161068

a. Which model is the best predictor of number of items sold? ***Answer: The TV-Radio model is the best predictor of number of items sold because it had the least amount of error, on average. When using the TV-Radio model to predict number of items sold, our predictions will typically be off by 11,609 units.***

b. Which model was the least reliable in predicting the number of items sold? ***Answer: The Newspaper model is the least reliable predictor of number of items sold because it had the most amount of error, on average. When using the Newspaper model to predict number of items sold, our predictions will typically be off by 45,384 units.***

c. What else do you notice about the models? ***Answer: It appears that combining the variables into one model is much better than any of the single-variable models.***

8. Inform the students that, in the next lab, they will find out how to create the line of best fit for models that include many variables.

Class Scribes:

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

Homework & Next Day

Ask students to think of a reason or reasons about why it would not be a good idea to make a scatterplot for models that include more than 3 predictor variables? ***The answer is mainly because humans are limited to seeing things in 3 dimensions. For example, the model that combines all of the variables together is a 4 dimensional model. What does that look like?***

LAB 4E: This Model is Big Enough for All of Us

Complete Lab 4E prior to Practicum.

Lab 4E - This model is big enough for all of us!

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

Building better models

- So far, in the labs, we've learned how to make predictions using the *line of best fit*
  - Which we also call *linear models* or *regression models*.
- We've also learned how to measure our model's prediction accuracy by cross-validation.
- In this lab, we'll investigate the following question:

  Will including more variables in our model improve its predictions?

Divide & Conquer

- Start by loading the `movie` data and split it into two sets (See Lab 4C for help). Remember to use `set.seed`.
  - A set named `training` that includes 75% of the data.
  - A set named `testing` that includes the remaining 25%.
- Create a linear model, using the `training` data, that predicts `gross` using `runtime`.
  - Compute the MSE of the model by making predictions for the `testing` data.
- **Do you think that a movie's `runtime` is the only factor that goes into how much a movie will make? What else might affect a movie's `gross`?**

Including more info

- Data scientists often find that including more relevant information in their models leads to better predictions.
  - Fill in the blanks below to predict `gross` using `runtime` and `reviews_num`.

```
lm(____ ~ ____ + ____, data = training)
```

- **Does this new model make more or less accurate predictions? Describe the process you used to arrive at your conclusion.**
- **Write down the code you would use to include a 3rd variable, of your choosing, in your `lm()`.**

Own your own

- **Write down which other variables in the `movie` data you think would help you make better predictions.**
  - **Are there any variables that you think would not improve our predictions?**
- **Create a model for all of the variables you think are relevant.**
  - **Assess whether your model makes more accurate predictions for the `testing` data than the model that included only `runtime` and `reviews_num`**
- **With your neighbors, determine which combination of variables leads to the best predictions for the `testing` data.**

Practicum: Predictions

Objective:
Students will create a linear model to predict the nutritional component that is most closely associated with the amount of sugar contained in a cereal.

Materials:

1. *Predictions Practicum* (LMR_U4_Practicum_Predictions)

Practicum
Predictions

Data about the nutritional components of popular cereal brands has been collected and made available for your team's use. We are interested in determining which other nutritional component is most closely associated with the amount of sugar contained in a cereal.

Your team will use the data to make predictions using linear models and compare the accuracy of your model to the rest of your classmates. Finally, the class will determine which team had the best prediction. Follow the directions below to explore and analyze the data:

1. You will have two data sets: one for training and one for testing. Load both data sets. Write down the code you used.
2. Explore the training data. Make several plots of different variable combinations and fit a linear regression line through them. Select the model that you think best makes the best prediction.
3. For the linear model your team selected:
    a. Describe what the plot shows.
    b. Explain why you selected that particular model.
    c. Compute the mean absolute error of your model using your testing data.
    d. Now make a set of predictions with your testing data. Calculate the mean absolute error for the testing data. Is it better or worse than for the training data, or about the same?
4. Present your team's linear model to the class. Explain why you chose your model and the typical amount of error in its predictions.
5. Give an example of a prediction for one value of x. State that value, give the predicted calories, and describe, based on the testing data, how far off your prediction might actually be.

Lesson 14: Improving Your Model

Objective:

Students will learn to describe associations that are not linear.

Materials:

1. *Describe the Association* handout (LMR_4.14_Describe the Association)

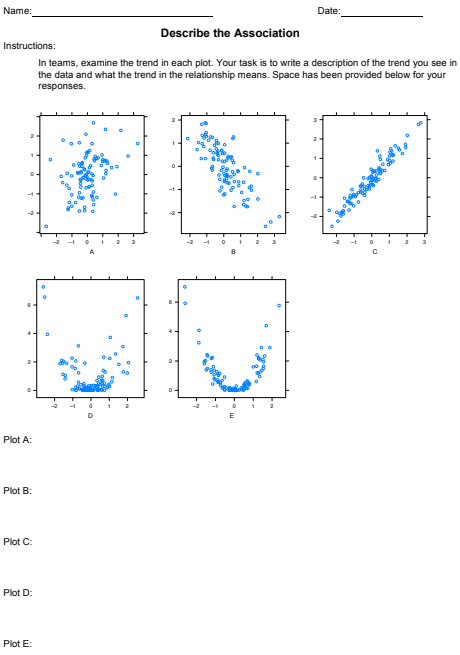
Vocabulary:

non-linear, polynomial trends

Essential Concepts: If a linear model is fit to a non-linear trend, it will not do a good job of predicting. For this reason, we need to identify non-linear trends by looking at a scatterplot or the model needs to match the trend.

Lesson:

1. Remind students that they have been learning a great deal about linear associations. However, there are other types of associations, and today they will learn to describe them.
2. Distribute *Describe the Association* (LMR_4.14). In teams, students will examine the trend of each plot. Their task is to write a description of the trend that they see in the data and what the trend means.



Name:_____ Date:_____

Describe the Association

Instructions:

In teams, examine the trend in each plot. Your task is to write a description of the trend you see in the data and what the trend in the relationship means. Space has been provided below for your responses.

A B C D E

Plot A:

Plot B:

Plot C:

Plot D:

Plot E:

LMR_4.14

3. Allow students time to discuss and record their descriptions for each plot in their DS journals. Walk around the room monitoring student teamwork. Look for descriptions that are interesting to share with the whole class.
4. Select a team to present a description of one plot to the class. Teams will listen to each presentation, compare it to their description of the plot, and as a team they will agree or disagree. If there is disagreement, lead a discussion that guides students to reason toward the correct description.
5. Summarize the discussion for each plot and ask students take notes or revise their descriptions in their DS journals.

6. Repeat steps 4 and 5 for the rest of the plots.

   Plot Descriptions for *Describe the Association* (LMR_4.14):

   - ***Plot A: There is no trend (perhaps some may see a very, very weak linear trend), so there is no/hardly any association. There is a great deal of scatter in the data. It means that y does not depend on x.***
   - ***Plot B: There appears to be a linear trend. The association is negative and appears somewhat strong. It means that as x increases, y decreases.***
   - ***Plot C: There is a linear trend. The association is positive and it is very strong. It means that the y-value increases at approximately the same rate for every increase in x value. This is a line.***
   - ***Plot D: The trend is non-linear. There seems to be a weak association because there is scatter in the data. Cannot tell if the association is positive or negative. It has the shape of a parabola; therefore, it is quadratic. For smaller x-values, the y-value is decreasing and for larger x values, the y value is increasing.***
   - ***Plot E: The trend is non-linear. There seems to be a strong association because there is little scatter in the data. It is also in the shape of a parabola, so it is quadratic.***

7. Using the *Cheat Notes* strategy, ask teams to write notes about how to describe associations.
8. Plots A, B, and C should be familiar to the students by now. However, plots D and E show a different type of trend. Although the trends are non-linear, they can still tell us important information about the y-values based on values of x. Ask:
   - What happens if we were to fit a linear model to these non-linear trends? Would it still make good predictions? ***No. They would not make good predictors***.
9. To examine why they would not make good predictors, draw an approximate linear best-fit line and get students to understand that in some regions, the model would almost always over-predict, and in others would almost always under-predict. We want a model that goes, more or less, through the 'middle' of the points. Ask:
   - How can we get a model that goes, more or less, through the middle of all the data points? ***Answer: We need to change the model.***
10. Trends like the quadratic ones show in plots D and E can be described as **polynomial trends**. Plots that follow quadratic, cubic, quartic, etc. shapes all exhibit polynomial trends. We need to adjust the model. You may show students several choices of equations (quadratic, trinomial, linear) along with their graphs and ask them which might be a good candidate.
11. When investigating the data for trends, the model needs to fit the data.

Class Scribes:

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

Homework & Next Day

Students may finish their *Cheat Notes* for homework, if not completed in class.

LAB 4F: Some Models Have Curves

Complete Lab 4F prior to Lesson 15.

Lab 4F - Some models have curves

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

Making models do yoga

- So far, we have only worked with prediction models that fit the *line of best fit* to the data.
- But what happens if the true relationship between the data is nonlinear?
- In this lab, we will learn about prediction models that fit *best fitting curves* to data.
- **Before moving on, load the `movie` data and split it into two sets:**
  - **A set named `training` that includes 75% of the data.**
  - **And a set named `testing` that includes the remaining 25%.**
  - Remember to use `set.seed`.

Problems with lines

- Before learning how to fit curves, let's first fit a linear model for reference.
- **Train a linear model predicting `audience_rating` based on `critics_rating` for the `training` data. Assign this model to `movie_linear`.**
- **Fill in the blanks below to create a scatterplot with `audience_rating` on the y-axis and `critics_rating` on the x-axis using your `testing` data.**

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

- Previously, you used `add_line` to plot the *line of best fit*. An alternative function for plotting the *line of best fit* is `add_curve`, which takes the name of the model as an argument.
- **Run the code below to add the line of best fit for the `training` data to plot.**

```
add_curve(movie_linear)
```

- **Describe, in words, how the line fits the data. Are there any values for `critics_rating` that would make obviously poor prediction?**
  - Hint: how does the linear model perform on very low and very high values of `critics_rating`?
- **Compute the MSE of the model for the `testing` data and write it down for later.**
  - Hint: refer to lab 4B.

Adding flexibility

- You don't need to be a full-fledged Data Scientist to realize that trying to fit a line to curved data is a poor modeling choice.
- If our data is curved, we should try model it with a curve.
- Instead of fitting a line, with equation of the form

$$y = a + bx$$

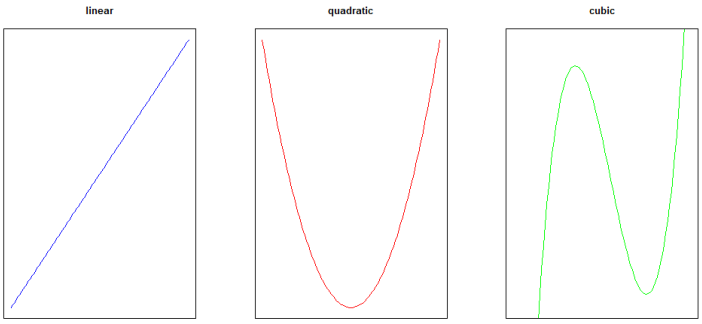
- we might consider fitting a *quadratic curve*, with equation of the form

$$y = ax + bx + cx^2$$

- or even a *cubic curve*, with equation of the form

$$y = a + bx + cx^2 + dx^3$$

- In general, the more coefficients in the model, the more flexible its predictions can be.



Making bend-y models

- To fit a quadratic model in `R`, we can use the `poly()` function.
  - **Fill in the blanks below to train a quadratic model predicting `audience_rating` from `critics_rating`, and assign that model to `movie_quad`.**

```
movie_quad <- lm(____ ~ poly(____, 2), data = training)
```

- **What is the role of the number 2 in the `poly()` function?**

Comparing lines and curves

- **Fill in the blanks to**
  - **create a scatterplot with `audience_rating` on the y-axis and `critics_rating` on the x-axis using your `testing` data, and**
  - **add the *line of best fit* and *best fitting quadratic curve*.**
  - Hint: the `col` argument is added to the `add_curve` functions to help distinguish the two curves.

```
xyplot(____ ~ ____, data = ____)
add_curve(____, col = "blue")
add_curve(____, col = "red")
```

- **Compare how the *line of best fit* and the *quadratic* model fit the data. Which do you think has a lower `test` MSE?**
- **Compute the MSE of the quadratic model for the `test` data and write it down for later.**
- **Use the difference in each model's `test` MSE to describe why one model fits better than the other.**

On your own

- Create a model that predicts `audience_rating` using a cubic curve (polynomial with degree `3`), and assign this model to `movie_cubic`.
- Create a scatterplot with `audience_rating` on the y-axis and `critics_rating` on the x-axis using your `test` data.
- Using the names of the three models you have trained, add the *line of best fit*, *best fitting quadratic curve*, and *best fitting cubic curve* for the `training data` to the plot.
- Based on the plot, which model do you think is the best at predicting the `testing` data?
- Use the difference in testing MSE to verify which model is the best at predicting the `testing data`.

The Growth of Landfills

Instructional Days: 5

Enduring Understandings

Model Eliciting Activities (MEAs) engage students in a complete modeling experience. MEAs are designed to make students' thinking visible and audible by encouraging them to be metacognitive about the process of inventing and testing a model, ask questions as they go through the process, and recognize the iterative nature of modeling.

Engagement

Students will read an excerpt from a CNN article called *Trash City: Inside America's Largest Landfill Site*. This article will set the context of the real-world problem facing many cities—the growth of landfills. The article provides background information as well as baseline data to launch the modeling process.

Learning Objectives

Statistical/Mathematical:

According to the California Common Core State Standards-Mathematics (CCSS-M) Framework:

"Modeling links classroom mathematics and statistics to everyday life, work, and decision-making. Modeling is the process of choosing and using appropriate mathematics and statistics to analyze empirical situations, to understand them better, and to improve decisions. Quantities and their relationships in physical, economic, public policy, social, and everyday situations can be modeled using mathematical and statistical methods. When making mathematical models, technology is valuable for varying assumptions, exploring consequences, and comparing predictions with data.

Modeling is best interpreted not as a collection of isolated topics, but rather in relation to other standards. Making mathematical models is a Standard for Mathematical Practice, and specific modeling standards appear throughout the high school standards indicated by a star symbol (★)."

Every Statistics and Probability standard in the California CCSS-M High School Conceptual Category is considered a modeling standard, indicated by the star symbol; therefore, rather than listing every content standard individually, the modeling activities in this section are designed so that students apply the Statistics and Probability standards learned throughout the curriculum.

Focus Standards for Mathematical Practice:

SMP-4: Model with mathematics.

Data Science:

Students will apply the conceptual understandings learned up to this point in the curriculum.

Applied Computational Thinking using RStudio:

- Previous techniques from the curriculum will be used in order to complete the task.

Real-World Connections:

Engineers, data scientists, and statisticians, to name a few, use modeling in their everyday work. Whether it is for creating a scale model of a bridge or a mathematical model of force impact measures, modeling is an integral part of what they do in the real world.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will use complex sentences to write a letter of recommendation that use data science concepts and skills.
4. Students will read informative texts to evaluate claims based on data.

Data File or Data Collection Method

Data File:

1. Trash: `data (trash)`

Legend for Activity Icons

 Video clip

 Discussion

 Articles/Reading

 Assessments

 Class Scribes

Lesson 15: The Growth of Landfills

Objective:

Students will engage in a modeling activity to learn about reducing the burden of trash landfills.

Materials:

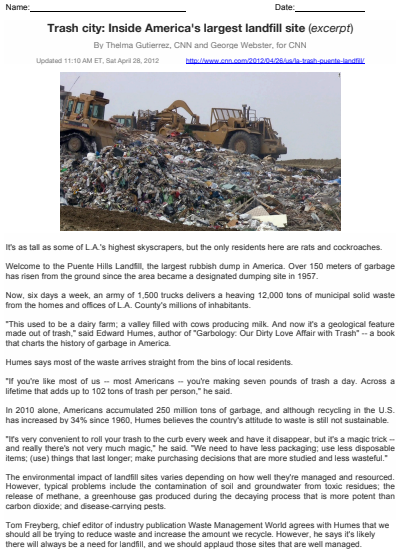
1. *Landfill Article* handout (LMR_4.15_Landfill Article)
2. *Landfill Readiness Questions* handout (LMR_4.16_Landfill Readiness Questions)
3. *Landfill Activity* handout (LMR_4.17_Landfill Activity)
4. Computers
5. IDS public dashboard: https://portal.idsucla.org
6. *Trash Data Exploration* handout (LMR_4.18_Trash Data Exploration)

Essential Concepts: Modeling does not always have to produce an equation. Instead, we can create models to answer real-world problems related to our community.

Lesson:

1. Inform students that they will investigate a problem that faces many cities in the United States today: trash. Explain that the next 4 days will be dedicated to completing the investigation and will follow this general structure:
    a. Day 1: Introduce assignment, initial exploration of data, creation of statistical questions.
    b. Day 2: Analysis of data via the IDS public dashboard.
    c. Day 3: Verify analysis via RStudio.
    d. Day 4: Team presentations.

2. Distribute the Landfill Article handout (LMR_4.15_Landfill Article) and explain that the reading is an excerpt from a CNN article titled *Trash City: Inside America's Largest Landfill Site*. The article will set the context for the real-world problem of growing landfills.



Name:_____ Date:_____

Trash city: Inside America's largest landfill site (*excerpt*)

By Thelma Gutierrez, CNN and George Webster, for CNN

Updated 11:10 AM ET, Sat April 28, 2012 http://www.cnn.com/2012/04/26/us/la-trash-puente-landfill/

It's as tall as some of L.A.'s highest skyscrapers, but the only residents here are rats and cockroaches.

Welcome to the Puente Hills Landfill, the largest rubbish dump in America. Over 150 meters of garbage has risen from the ground since the area became a designated dumping site in 1957.

Now, six days a week, an army of 1,500 trucks delivers a heaving 12,000 tons of municipal solid waste from the homes and offices of L.A. County's millions of inhabitants.

"This used to be a dairy farm; a valley filled with cows producing milk. And now it's a geological feature made out of trash," said Edward Humes, author of "Garbology: Our Dirty Love Affair with Trash" -- a book that charts the history of garbage in America.

Humes says most of the waste arrives straight from the bins of local residents.

"If you're like most of us -- most Americans -- you're making seven pounds of trash a day. Across a lifetime that adds up to 102 tons of trash per person," he said.

In 2010 alone, Americans accumulated 250 million tons of garbage, and although recycling in the U.S. has increased by 34% since 1960, Humes believes the country's attitude to waste is still not sustainable.

"It's very convenient to roll your trash to the curb every week and have it disappear, but it's a magic trick -- and really there's not very much magic," he said. "We need to have less packaging; use less disposable items; (use) things that last longer; make purchasing decisions that are more studied and less wasteful."

The environmental impact of landfill sites varies depending on how well they're managed and resourced. However, typical problems include the contamination of soil and groundwater from toxic residues; the release of methane, a greenhouse gas produced during the decaying process that is more potent than carbon dioxide; and disease-carrying pests.

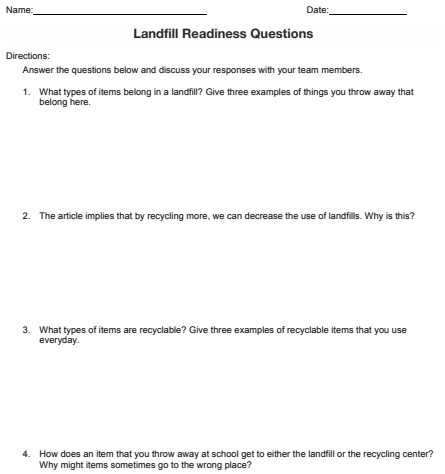
Tom Freyberg, chief editor of industry publication Waste Management World agrees with Humes that we should all be trying to reduce waste and increase the amount we recycle. However, he says it's likely there will always be a need for landfill, and we should applaud those sites that are well managed.

LMR_4.15

3. Using the *5 Ws* strategy, ask students to read the article individually and to write down the 5 Ws in their DS journals. The 5 Ws summarize the What, Who, Why, When, and Where of the article.

4. After they have finished reading, students should answer the questions provided on the *Landfill Readiness Questions* handout (LMR_4.16_Landfill Readiness Questions). Then, in teams,

students will discuss their insights, questions, and/or reactions to the both the article and the questions. Follow up the team discussion with a class discussion to gauge what students actually know about trash and recycling.



Name:_____ Date:_____

Landfill Readiness Questions

Directions:

Answer the questions below and discuss your responses with your team members.

1. What types of items belong in a landfill? Give three examples of things you throw away that belong here.

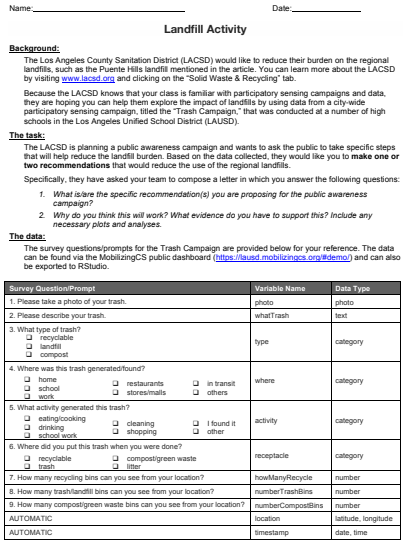
2. The article implies that by recycling more, we can decrease the use of landfills. Why is this?

3. What types of items are recyclable? Give three examples of recyclable items that you use everyday.

4. How does an item that you throw away at school get to either the landfill or the recycling center? Why might items sometimes go to the wrong place?

LMR_4.16

5. Next, introduce students to the main task they will be investigating about landfills by distributing the *Landfill Activity* handout (LMR_4.17_Landfill Activity). This handout asks the students to come up with one or two recommendations to help reduce the burden of landfills on the environment. In order to complete the assignment, students will use 2 data analysis tools: the IDS dashboard and RStudio.



Name:_____ Date:_____

Landfill Activity

Background:

The Los Angeles County Sanitation District (LACSD) would like to reduce their burden on the regional landfills, such as the Puente Hills landfill mentioned in the article. You can learn more about the LACSD by visiting www.lacsd.org and clicking on the "Solid Waste & Recycling" tab.

Because the LACSD knows that your class is familiar with participatory sensing campaigns and data, they are hoping you can help them explore the impact of landfills by using data from a city-wide participatory sensing campaign, titled the "Trash Campaign," that was conducted at a number of high schools in the Los Angeles Unified School District (LAUSD).

The task:

The LACSD is planning a public awareness campaign and wants to ask the public to take specific steps that will help reduce the landfill burden. Based on the data collected, they would like you to **make one or two recommendations** that would reduce the use of the regional landfills.

Specifically, they have asked your team to compose a letter in which you answer the following questions:

1. *What is/are the specific recommendation(s) you are proposing for the public awareness campaign?*
2. *Why do you think this will work? What evidence do you have to support this? Include any necessary plots and analyses.*

The data:

The survey questions/prompts for the Trash Campaign are provided below for your reference. The data can be found via the MobilizingCS public dashboard (https://lausd.mobilizingcs.org/#demo/) and can also be exported to RStudio.

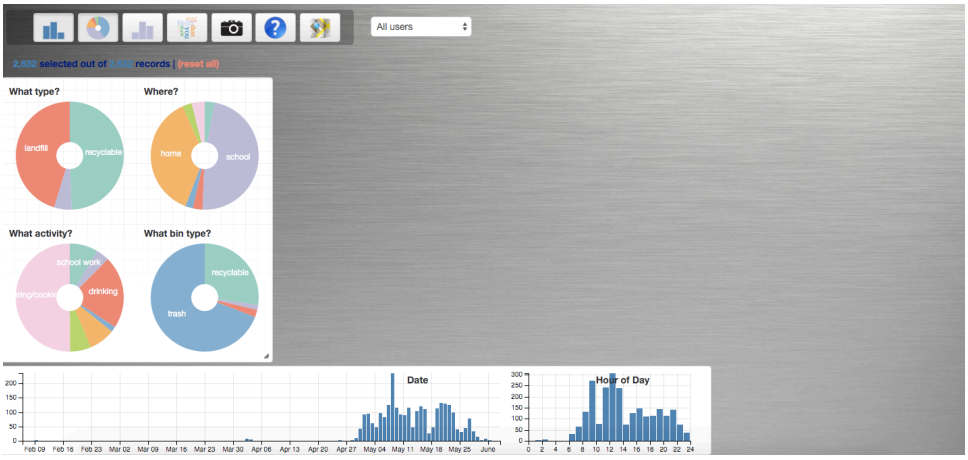
Survey Question/Prompt	Variable Name	Data Type
1. Please take a photo of your trash.	photo	photo
2. Please describe your trash.	whatTrash	text
3. What type of trash? ❑ recyclable ❑ landfill ❑ compost	type	category
4. Where was this trash generated/found? ❑ home ❑ restaurants ❑ in transit ❑ school ❑ stores/malls ❑ others ❑ work	where	category
5. What activity generated this trash? ❑ eating/cooking ❑ cleaning ❑ I found it ❑ drinking ❑ shopping ❑ other ❑ school work	activity	category
6. Where did you put this trash when you were done? ❑ recyclable ❑ compost/green waste ❑ trash ❑ litter	receptacle	category
7. How many recycling bins can you see from your location?	howManyRecycle	number
8. How many trash/landfill bins can you see from your location?	numberTrashBins	number
9. How many compost/green waste bins can you see from your location?	numberCompostBins	number
AUTOMATIC	location	latitude, longitude
AUTOMATIC	timestamp	date, time

LMR_4.17

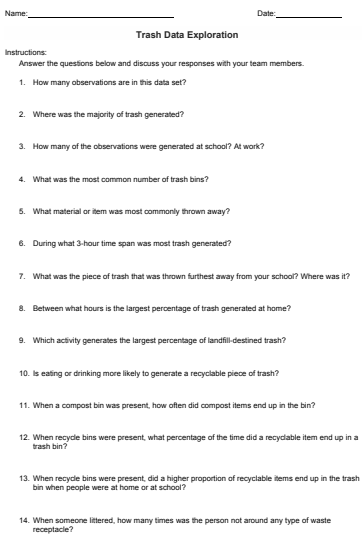
6. Once all students have read the assignment, use the following questions to check for understanding of what the task is:
    a. What organization is asking for your help? ***The Los Angeles County Sanitation District (LACSD).***
    b. What type of data did the organization collect, and whom did they collect it from? ***Participatory Sensing data via the Trash campaign. The campaign was city-wide and taken by high school students in LAUSD.***
    c. How many recommendations will you present to the organization? ***One or two.***

d. What does the organization hope to do with your recommendations? ***Create a public awareness campaign to help reduce the burden on landfills.***

7. At this point, students will begin exploring the data via the IDS public dashboard: https://portal.idsucla.org/
8. They should use the "Trash" campaign data and select "Dashboard" from the "Action" button.
9. The dashboard is a visual tool for exploring and analyzing data. An example screenshot of the Trash campaign in the dashboard is shown below.



10. Students do not need to complete any analyses during today's lesson. Instead, they should simply "play" with the data and brainstorm possible statistical questions that will help them complete the activity.
11. To assist students' interaction with the dashboard, distribute the *Trash Dashboard Exploration* handout (LMR_4.18_Trash Data Exploration).



Name:_____ Date:_____

Trash Data Exploration

Instructions:
Answer the questions below and discuss your responses with your team members.

1. How many observations are in this data set?
2. Where was the majority of trash generated?
3. How many of the observations were generated at school? At work?
4. What was the most common number of trash bins?
5. What material or item was most commonly thrown away?
6. During what 3-hour time span was most trash generated?
7. What was the piece of trash that was thrown furthest away from your school? Where was it?
8. Between what hours is the largest percentage of trash generated at home?
9. Which activity generates the largest percentage of landfill-destined trash?
10. Is eating or drinking more likely to generate a recyclable piece of trash?
11. When a compost bin was present, how often did compost items end up in the bin?
12. When recycle bins were present, what percentage of the time did a recyclable item end up in a trash bin?
13. When recycle bins were present, did a higher proportion of recyclable items end up in the trash bin when people were at home or at school?
14. When someone littered, how many times was the person not around any type of waste receptacle?

LMR_4.18

12. Leave 10-15 minutes at the end of class to share out and discuss some of these statistical questions. During this time, the teacher should also check for data understanding.

Class Scribes:

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

Lesson 16: Exploring Trash via the Dashboard

Objective:

Students will continue to investigate landfills and perform analyses via the IDS public dashboard.

Materials:

1. Computers
2. IDS public dashboard: https://portal.idsucla.org/

Essential Concepts: Exploring the IDS Dashboard provides a visual approach to data analysis.

Lesson:

1. Today students will continue their data exploration of the Trash campaign via the IDS public dashboard.
2. As a team, students should select statistical questions and provide appropriate plots and summaries from the dashboard to answer those questions.
3. Leave 10-15 minutes at the end of class to share out some of the findings from each student team.

Class Scribes:

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

Homework

Students will brainstorm possible RStudio commands to complement their initial analyses from the dashboard. It is up to the teacher to ask for a minimum number of commands from each student.

Lesson 17: Exploring Trash via RStudio

Objective:

Students will continue to investigate landfills and perform analyses via RStudio.

Materials:

1. Computers
2. RStudio

Essential Concepts: RStudio can be used to verify initial results/findings from data analysis done via the IDS Dashboard.

Lesson:

1. Today students will continue their data analysis of the Trash campaign via RStudio.
2. They should share their answers from the previous lesson's homework assignment in their teams to help them get started with their code. After having shared their initial code, they should spend some time discussing other ideas
3. The Recorder/Reporter should keep a list of the code that the team has agreed to use.
4. By the end of class, students should begin writing their recommendations for reducing the burden on landfills.
5. Inform students that each team will prepare their presentations during the next class period.

Class Scribes:

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

Next 2 Days

Students will finalize their recommendations for reducing the burden on landfills and have a draft of their letter to send to LACSD and prepare for their team presentations.

Decisions, Decisions

Instructional Days: 3

Enduring Understandings

Decision trees are used to classify observations into similar groupings based on known characteristics. Yes/no questions are asked, then the observations are sorted based on the responses to the questions. After a specified number of iterations, a final group membership is decided. One particular modeling tool we use for decision trees is known as CART (Classification and Regression Trees).

Engagement

Students will participate in the *CART Activity* described in Lesson 12. They will classify football and soccer players into categories based on player characteristics.

Learning Objectives

Statistical/Mathematical:

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

Data Science:

Understand that classification and regression trees can be used to predict membership in groups.

Applied Computational Thinking using RStudio:

- Create classification and regression trees.

Real-World Connections:

Cardiologists may use a decision tree to diagnose whether people are or are not having a heart attack. Since the late 1870's, this method has been found to correctly diagnose a heart attack in over 95% of cases compared to correct diagnoses based on individual doctors' expertise, which ranged between 75 and 90%.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it is used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.

Data File or Data Collection Method

Data File:

1. Titanic: `data(titanic)`

Legend for Activity Icons


Video clip


Discussion


Articles/Reading


Assessments


Class Scribes

Lesson 18: Grow Your Own Decision Tree

Objective:

Students will learn what decision trees look like and how they can be used to classify people or objects into groups. They will engage in an activity to see how making slight changes to the tree can lead to drastic rises or reductions in misclassifications.

Materials:

1. *CART Activity Player Stats* (LMR_4.19_CART Player Stats)
2. *CART Activity Round 1 Questions* (LMR_4.20_CART Round 1)
3. *CART Activity Round 2 Questions* (LMR_4.21_CART Round 2)
   Advanced preparation required (see Step 8 below)

Vocabulary:

classify, decision tree, Classification and Regression Trees (CART), nodes, misclassifications

Essential Concepts: Some trends are not linear, so the approaches we've done so far won't be helpful. We need to model such trends differently. Decision trees are a non-linear tool for classifying observations into groups when the trend is non-linear.

Lesson:

1. Inform students that, during today's lesson, they will be participating in an activity to try to **classify** professional athletes into one of two groups: (1) soccer players on the US Men's National Team, OR (2) football players in the National Football League (NFL).
2. Remind students that this unit has focused on linear models and making predictions. In the real world, data can be modeled in a variety of ways, many of which are non-linear, and because of this, we can't easily write down a mathematical equation to help us make predictions. However, we can use what we have learned so far to determine whether or not other models can provide a good fit to the data.
3. Introduce the topic of **decision trees** and explain that it is simply a non-linear way to model data.
4. Explain that decision trees are "grown" by using algorithms, or rules, to test many, many different decision trees to find the one that makes the best predictions.
5. A decision tree is basically a series of questions that are asked sequentially. Observations start by answering the first question (at the root of the tree), and then proceed along the different branches based on the answers they give to the questions that follow. At the end, based on all of the questions asked, observations are then classified as one of k classifications.
6. Remind students that algorithms are a series of steps that are repeated a large number of times. For decision trees, this enables us to (1) explore many possible paths, beginning from the same initial point, or (2) find different starting points based on where we ended during the previous iteration.
7. Ask students to recall that they created and worked with *linear models* earlier in the unit. We are continuing our work with models and will learn another method of modeling called **CART**, which stands for **Classification and Regression Trees**. This is another name for decision trees.
8. CART Activity: to get a sense of how decision trees work, the students will see one in action. We are going to try to classify 15 professional athletes into either soccer or football players based on some of their characteristics.

   Note: Advanced preparation required. The cards in each of the LMRs listed above (and displayed below) need to be cut out prior to class time.

Name:_____ Date:_____

CART Activity Player Stats

Directions for teacher:

Create "player" cards by cutting out each player's statistics from the table.

Player 1 Name: Matt Besler Team: Kansas City Height (inches): 72 Weight (pounds): 170 Age: 28 League: USMNT	**Player 2** Name: Cam Newton Team: Carolina Height (inches): 77 Weight (pounds): 245 Age: 26 League: NFL	**Player 3** Name: Clint Dempsey Team: Seattle Height (inches): 73 Weight (pounds): 170 Age: 32 League: USMNT
Player 4 Name: Steve Birnbaum Team: Washington, DC Height (inches): 74 Weight (pounds): 181 Age: 28 League: USMNT	**Player 5** Name: Jermaine Jones Team: New England Height (inches): 72 Weight (pounds): 179 Age: 34 League: USMNT	**Player 6** Name: Matt Cassel Team: Dallas Height (inches): 76 Weight (pounds): 230 Age: 33 League: NFL
Player 7 Name: Russell Wilson Team: Seattle Height (inches): 71 Weight (pounds): 206 Age: 27 League: NFL	**Player 8** Name: Matt Hedges Team: Dallas Height (inches): 76 Weight (pounds): 190 Age: 25 League: USMNT	**Player 9** Name: Robert Griffin III Team: Washington, DC Height (inches): 74 Weight (pounds): 223 Age: 25 League: NFL
Player 10 Name: Tom Brady Team: New England Height (inches): 76 Weight (pounds): 225 Age: 38 League: NFL	**Player 11** Name: Michael Bradley Team: Toronto Height (inches): 73 Weight (pounds): 179 Age: 28 League: USMNT	**Player 12** Name: Sean Johnson Team: Chicago Height (inches): 75 Weight (pounds): 217 Age: 26 League: USMNT
Player 13 Name: Tony Romo Team: Dallas Height (inches): 74 Weight (pounds): 230 Age: 35 League: NFL	**Player 14** Name: Alex Smith Team: Kansas City Height (inches): 76 Weight (pounds): 216 Age: 31 League: NFL	**Player 15** Name: Jozy Altidore Team: Toronto Height (inches): 73 Weight (pounds): 174 Age: 26 League: USMNT

LMR_4.19

Name:_____ Date:_____

CART Activity Round 1

Directions for teacher:

Create "leaves" by cutting out each question below.

Leaf 1 Is your team located in the United States? YES: Go to your right. NO: Go to your left.
Leaf 2 Are you 33 years old or older? YES: Go to your right. NO: Go to your left.
Leaf 3 You play for the US Men's National Soccer Team (USMNT).
Leaf 4 You play for the National Football League (NFL).
Leaf 5 Are you 73 inches tall or taller? YES: Go to your right. NO: Go to your left.
Leaf 6 You play for the US Men's National Soccer Team (USMNT).
Leaf 7 You play for the National Football League (NFL).

LMR_4.20

Name:_____ Date:_____

CART Activity Round 2

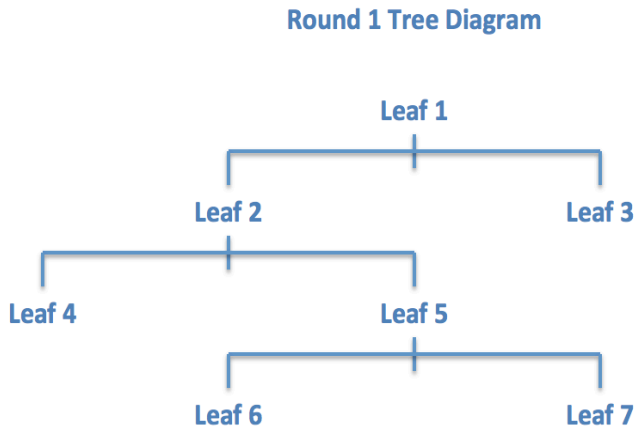
Directions for teacher:

Create "leaves" by cutting out each question below.

Leaf 1 Are you 74 inches tall or taller? YES: Go to your right. NO: Go to our left.
Leaf 2 You play for the National Football League (NFL).
Leaf 3 Do you weigh more than 200 pounds? YES: Go to your right NO: Go to your left.
Leaf 4 You play for the National Football League (NFL).
Leaf 5 You play for the US Men's National Soccer Team (USMNT).

LMR_4.21

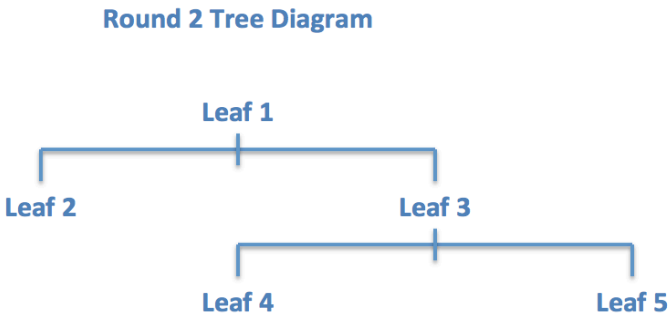
9. Ask for 15 volunteers and hand each of them a data card from the *CART Activity Player Stats* handout (LMR_4.19). These students will be known as the "players." Each card lists the following variables for 15 different professional athletes:
    a. team location
    b. name
    c. age
    d. height (in inches)
    e. weight (in pounds)
    f. league
10. The "players" will only be allowed to say "yes" or "no" in this activity. No other talking is permitted.
11. Now, ask for 7 additional volunteers to be the ***nodes***, or *leaves*, on the decision tree. Each student will be known as a "leaf."
12. Distribute one question/classification from the *CART Activity Round 1 Questions* (LMR_4.20) to each "leaf."
13. Arrange the 7 "leaves" in the room as depicted by the graphic below:



14. Now, each "player," one at a time, will approach *Leaf 1*, who will ask the "player" the question listed on his/her card. Depending on the player's answer, *Leaf 1* will direct the "player" to the next "leaf."
15. The "player" continues through the nodes until a "leaf" declares the "player" to be either (1) a soccer player on the US Men's National Team, OR (2) a football player in the National Football League (NFL).
16. Allow all the "players" to go through the "leaves" until each one is classified as either a soccer or football player.
17. After each player has been classified, tally the number of correct and incorrect classifications and display a simple table (see example below) on the board.

	Classified Correctly	**Classified Incorrectly**
USMNT Soccer Player		
NFL Football Player		

18. Ask students to calculate the misclassification rate (MCR) which is the proportion of observations who were predicted to be in one category but were actually in another. If all the activity player stats cards were used, the misclassification rate would be 5/15.
19. After proceeding through "Round 1," ask an additional 5 students to come up as more "leaves," distribute the cards from the CART Activity Round 2 Questions file (LMR_4.21), and arrange the students like the diagram below:

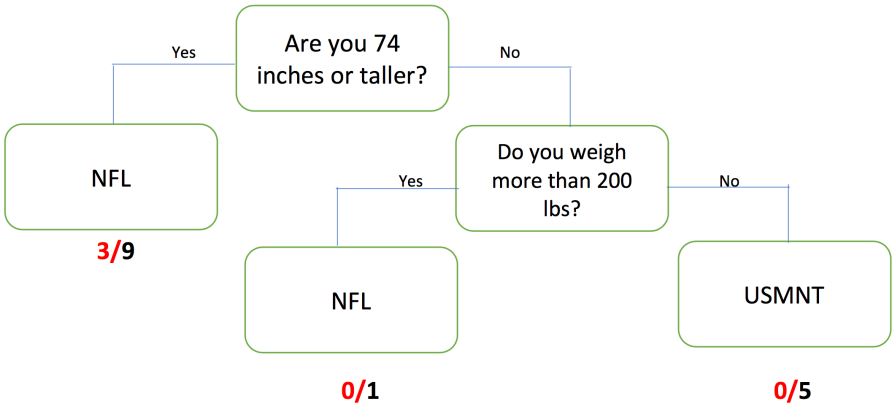


20. Have each "player" go through this new set of "leaves" until they are re-classified by these new rules.
21. Again, tally the number of correct and incorrect classifications and display them on the board and calculate the misclassification rate. If all the activity player stats cards were used, the misclassification rate would be 3/15.
22. Once the activity has been completed, ask students the following questions:
    a. How do decision trees classify objects/people as being a member of a group? ***By asking a series of questions, one at a time, and sending the participant down a particular path until he/she is classified.***
    b. Did we do as well, worse, or better in Round 2 compared to Round 1 at correctly guessing which sport the "players" participate in? Explain. ***Answers will vary according to results of the activity.***
    c. How can we figure out what questions to ask and in what order to minimize the number of **misclassifications**? (This one might not be obvious. The point is for the students to wrestle with how they might think it can be done.)

23. Also have the students discuss the following questions:

    a. How is a decision tree/CART similar to or different than a linear model?
    b. Can we really call a decision tree a model? Why or why not?

24. In lab 4G you will use RStudio to create tree models that will make good predictions without needing a lot of branches. RStudio can also calculate the misclassification rate. However, you might find the visual a little confusing to interpret, so we will use the Round 2 classification tree to see what the output from RStudio might look like.



25. Project the image above and explain to the students that if all 15 of the activity player stats cards were used, then RStudio would give ratios for each of the leaves where a classification was made. The denominator tells us how many observations ended up in that leaf and the numerator tells us the number of misclassifications. Ask students:

    a. What does the output 3/9 represent? ***Answer: The 9 tells us that nine players were classified as NFL players based solely on the fact that they were taller than 74 inches. The 3 represents soccer players that were misclassified so 6 football players were classified correctly.***

    b. How would you calculate the misclassification rate (MCR)? ***Answer: You would add all of the numerators which represent the misclassifications and divide by the total number of observations which you could obtain by adding all the denominators. (3+0+0)/(9+1+5)=3/15.***

Class Scribes:

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

Homework

Students will record their responses to the following discussion questions about CARTs and submit them the following day:

a. How is a decision tree/CART similar to or different than a linear model?
b. Can we really call a decision tree a model? Why or why not?

Lesson 19: Data Scientists or Doctors?

Objective:

Students will create their own decision trees based on training data (i.e., the data from the previous day's lessons), and then see how well their decision tree works on new test data.

Materials:

1. *Decision Tree for Heart Attack Risk* graphic (LMR_4.22_CART Heart Attacks)
2. *Make Your Own Decision Tree* handout (LMR_4.23_Your Own Decision Tree)

Vocabulary:

training data, testing data

Essential Concepts: We can determine the usefulness of decision trees by comparing the number of misclassifications in each.

Lesson:

1. Ask students the following question:

   If a close friend or family member were having chest pains, would you want to take that person to a doctor or to a data scientist?

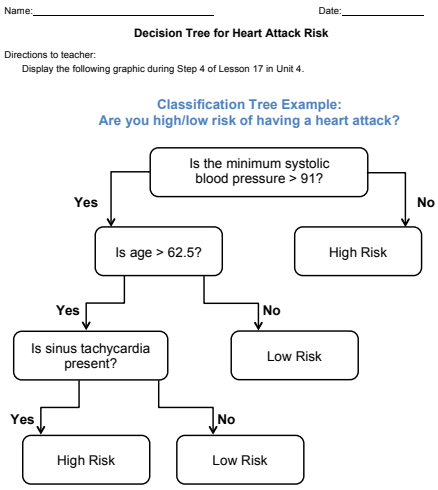
2. Give the students some time to think about the question and have a few of them share out their responses with the class.

   Note: It's likely that most students will choose to bring their loved one to a doctor.

3. As it turns out, back in the late 1970s, a cardiologist (and early data scientist) named Lee Goldman developed a decision tree based on millions of patient observations. The decision tree was made to diagnose whether people were or were not having a heart attack. Interestingly, the results of the decision tree compared to how actual doctor diagnoses are shown below:

   a. Correct diagnoses using the decision tree were above 95%.
   b. Correct diagnoses based on individual doctors' expertise? Anywhere between 75-90%.

4. Display the graphic from the *Decision Tree for Heart Attack Risk* file (LMR_4.22_CART Heart Attacks) and explain that this is one example of what the decision tree might have looked like. **Note:** This is NOT the actual tree Goldman developed.



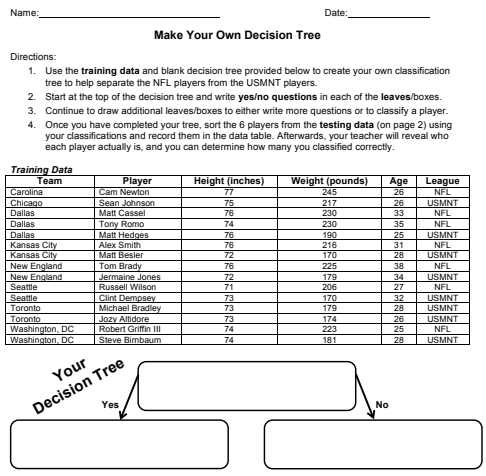
5. Using a *Pair-Share*, ask students to discuss the following questions using the graphic above, as well as what they learned during the previous lesson's activity.

    a. What are decision trees?
    b. How do they work at classifying data into groups?

6. Then display the following data (the same data from the player cards used in the previous lesson):

Team	Player	Height (inches)	Weight (pounds)	Age	League
Carolina	Cam Newton	77	245	26	NFL
Chicago	Sean Johnson	75	217	26	USMNT
Dallas	Matt Cassel	76	230	33	NFL
Dallas	Tony Romo	74	230	35	NFL
Dallas	Matt Hedges	76	190	25	USMNT
Kansas City	Alex Smith	76	216	31	NFL
Kansas City	Matt Besler	72	170	28	USMNT
New England	Tom Brady	76	225	38	NFL
New England	Jermaine Jones	72	179	34	USMNT
Seattle	Russell Wilson	71	206	27	NFL
Seattle	Clint Dempsey	73	170	32	USMNT
Toronto	Michael Bradley	73	179	28	USMNT
Toronto	Jozy Altidore	73	174	26	USMNT
Washington, D.C.	Robert Griffin III	74	223	25	NFL
Washington, D.C.	Steve Birnbaum	74	181	28	USMNT

7. Distribute the *Make Your Own Decision Tree* handout (LMR_4.23_Your Own Decision Tree) and give students time to come up with their own decision trees based on the **training data** they are given. Students may work in pairs or teams. They should follow the directions on page 1 of the handout and come up with a series of possible yes/no questions that they could ask to classify each player into his correct league (the NFL or the USMNT).



Name:_____ Date:_____

Make Your Own Decision Tree

Directions:

1. Use the **training data** and blank decision tree provided below to create your own classification tree to help separate the NFL players from the USMNT players.
2. Start at the top of the decision tree and write **yes/no questions** in each of the **leaves**/boxes.
3. Continue to draw additional leaves/boxes to either write more questions or to classify a player.
4. Once you have completed your tree, sort the 6 players from the **testing data** (on page 2) using your classifications and record them in the data table. Afterwards, your teacher will reveal who each player actually is, and you can determine how many you classified correctly.

Training Data

Team	Player	Height (inches)	Weight (pounds)	Age	League
Carolina	Cam Newton	77	245	26	NFL
Chicago	Sean Johnson	75	217	26	USMNT
Dallas	Matt Cassel	76	230	33	NFL
Dallas	Tony Romo	74	230	35	NFL
Dallas	Matt Hedges	76	190	25	USMNT
Kansas City	Alex Smith	76	216	31	NFL
Kansas City	Matt Besler	72	170	28	USMNT
New England	Tom Brady	76	225	38	NFL
New England	Jermaine Jones	72	179	34	USMNT
Seattle	Russell Wilson	71	206	27	NFL
Seattle	Clint Dempsey	73	170	32	USMNT
Toronto	Michael Bradley	73	179	28	USMNT
Toronto	Jozy Altidore	73	174	26	USMNT
Washington, DC	Robert Griffin III	74	223	25	NFL
Washington, DC	Steve Birnbaum	74	181	28	USMNT

LMR_4.23

8. Once the students have finished creating their decision trees, ask the following questions:

    a. Will you be able to classify other players from a new data set correctly using this particular decision tree?
    b. Is this decision tree too specific to the training data?

9. Inform the students that they should now use the **testing data** on page 2 of the handout to try to classify 5 *mystery players* into one of the two leagues. They should record the classification that their tree outputs in the data table on page 2.

10. Let the students compare their decision trees and league assignments with one another. Hopefully, there will be a bit of variety in terms of the trees and the classifications.
11. Next, show students the correct league classifications for the 5 mystery players. The mystery player names are also included in this table.

Team	Player	Height (inches)	Weight (pounds)	Age	League
Toronto	Michael Bradley	74	175	28	USMNT
New York	Eli Manning	76	218	34	NFL
New Orleans	Drew Brees	72	209	36	NFL
Washington, DC	Perry Kitchen	72	160	23	USMNT
New England	Lee Nguyen	68	150	29	USMNT

12. By a show of hands, ask:
    a. How many students misclassified all of the players in the testing data?
    b. How many misclassified 4 of the 5 players?
    c. How many misclassified 3 of the 5 players?
    d. How many misclassified 2 of the 5 players?
    e. Did anyone correctly classify ALL 5 mystery players? If so, ask those students to share their decision trees with the rest of the class.
13. Inform students that, when faced with much more data, creating classification trees becomes much harder to make by hand. It is so difficult, in fact, that data scientists rely on software to grow their trees for them. Students will learn how to create decision trees in RStudio during the next lab.

Class Scribes:

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

Homework & Next Day

Write a paragraph describing the role testing data and training data play in creating a classification tree.

LAB 4G: Growing Trees

Complete Lab 4G prior to Lesson 20.

Lab 4G - Growing trees

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

Trees vs. Lines

- So far in the labs, we've learned how we can fit linear models to our data and use them to make predictions.
- In this lab, we'll learn how to make predictions by growing trees.
  - Instead of creating a line, we split our data into branches based on a series of *yes* or *no* questions.
  - The branches help sort our data into *leaves* which can then be used to make predictions.
- Start, by loading the `titanic` data.

Our first tree

- Use the `tree()` function to create a *classification* tree that predicts whether a person `survived` the Titanic based on their `gender`.
  - A *classification* tree tries to predict which category a categorical variable would belong to based on other variables.
  - The syntax for `tree` is similar to that of the `lm()` function.
  - Assign this model the name `tree1`.
- **Why can't we just use a *linear model* to predict whether a passenger on the Titanic `survived` or not based on their `gender`?**

Viewing trees

- To actually look at and interpret our `tree1`, place the model into the `treeplot` function.
  - **Write down the labels of the two *branches*.**
  - **Write down the labels of the two *leaves*.**
- Answer the following, based on the `treeplot`:
  - **Which `gender` does the model predict will survive?**
  - **Where does the plot tell you the number of people that get sorted into each leaf? How do you know?**
  - **Where does the plot tell you the number of people that have been sorted *incorrectly* in each leaf?**

Leafier trees

- Similar to how you included multiple variables for a linear model, create a `tree` that predicts whether a person `survived` based on their `gender`, `age`, `class`, and where they `embarked`.
  - Call this model `tree2`.
- Create a `treeplot` for this model and answer the following question:
  - **Mrs. Cumings was a 38 year old female with a 1st class ticket from Cherbourg. Does the model predict that she survived?**
  - **Which variable ended up not being used by `tree`?**

Tree complexity

- By default, the `tree()` function will fit a *tree model* that will make good predictions without needing lots of branches.
- We can increase the complexity of our trees by changing the complexity parameter, `cp`, which equals `0.01` by default.
- We can also change the minimum number of observations needed in a leaf before we split it into a new branch using `minsplit`, which equals `20` by default.
- Using the same variables that you used in `tree2`, create a model named `tree3` but include `cp = 0.005` and `minsplit = 10` as arguments.
  - **How is `tree3` different from `tree2`?**

Misclassification rate

- Similar to how we use the *mean squared error* to describe how well our model predicts numerical variables, we use the *misclassification rate* to describe how well our model predicts categorical variables.
  - The *misclassification rate* (MCR) is the number of people who were predicted to be in one category but were actually in another.
  - Fill in the blanks to create a function to calculate the MCR

```
calc_mcr <- function(actual, predicted) {
  sum(____ != ____) / length(____)
}
```

Predictions and Cross-validation

- Just like with *linear models*, we can use cross-validation to measure our *classification trees* prediction accuracy.
  - Use the `data` function to load the `titanic_test` data.
  - Fill in the blanks below to predict whether people in the `titanic_test` data survived or not using `tree1`.

```
titanic_test <- mutate(____, prediction = predict(____, newdata = ____, t
ype = "class"))
```

- Then run the following to calculate the MCR

```
summarize(titanic_test, mcr = calc_mcr(survived, prediction))
```

On your own

- **In your own words, explain what the *misclassification rate* is.**
- **Which model (`tree1`, `tree2` or `tree3`) had the lowest misclassification rate for the `titanic_test` data?**
- Create a 4th model using the same variables used in `tree2`. This time though, change the *complexity parameter* to `0.0001`. Then answer the following
  - **Does creating a more complex *classification tree* always lead to better predictions? Why not?**
- A *regression tree* is a tree model that predicts a numerical variable. Create a *regression tree* model to predict the Titanic's passenger's ages and calculate the MSE.
  - Plots of regression trees are often too complex to plot.

Ties that Bind

Instructional Days: 3

Enduring Understandings

Clustering is another way to classify data into groups. We classify observations based on numerical characteristics and their similarities. We use k-means to determine the mean value for each group of k clusters by randomly assigning an initial value for the mean and then moving the mean based on its proximity to the points.

Networks classify people into groupings based on who knows whom. Nodes are formed when a relationship between two people is present.

Engagement

Students will participate in the *Find the Clusters Activity* described in Lesson 14. They will determine which points in a plot should be grouped as football players and which points should be grouped as swimmers.

Learning Objectives

Statistical/Mathematical:

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

Data Science:

Understand what RStudio is doing when using the k-means function to find clusters in a group of data and when creating networks in order to learn how to classify data into groups.

Applied Computational Thinking using RStudio:

- Use the k-means function to find clusters in a group of data.
- Plot the data with the cluster assignments based on the k-means function.

Real-World Connections:

Network analysis is used by many private and public entities such as the National Security Agency when they want to find terrorist networks to have maximum impact on communications. The k-means algorithm is a technique for grouping entities according to the similarity of their attributes. For example, dividing countries into similar groups using k-means to make fair comparisons is applicable.

Language Objectives

1. Students will use complex sentences to construct summary statements about their understanding of data, how it is collected, how it used, and how to work with it.
2. Students will engage in partner and whole group discussions and presentations to express their understanding of data science concepts.
3. Students will use complex sentences to write informative short reports that use data science concepts and skills.

Legend for Activity Icons


Video clip


Discussion


Articles/Reading


Assessments


Class Scribes

Lesson 20: Where Do I Belong?

Objective:

Students will learn what clustering is and how to classify groups of people into clusters based on unknown similarities.

Materials:

1. *Find the Clusters* handout (LMR_4.24_Find the Clusters)

Vocabulary:

clustering, cluster, k-means

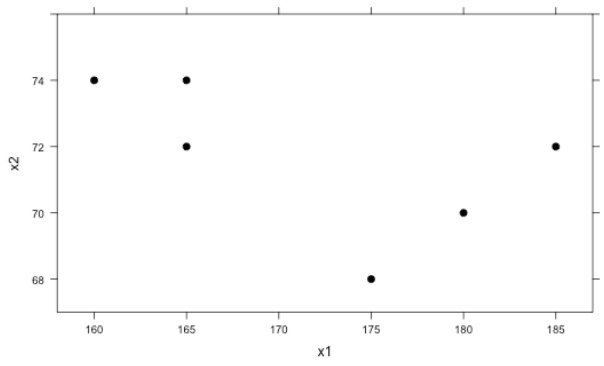
Essential Concepts: We can identify groups, or "clusters," in data based on a few characteristics. For example, it is easy to classify a classroom into males and females, but what if you only knew each person's arm span? How well could you classify their genders now?

Lesson:

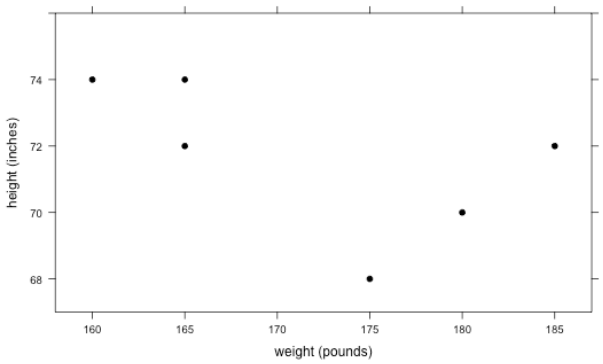
1. Inform the students that they will continue to explore different types of models, and today they will be focusing on **clustering**. Clustering is the process of grouping a set of objects (or people) together in such a way that people in the same group (called a **cluster**) are more similar to each other than to those in other groups.
2. Have the students recall that, in the previous lessons, they used decision trees and CART to classify people into different groups based on whether or not a person had a specific characteristic (e.g., whether or not a professional athlete's team is based in the US).
3. But, sometimes we don't know what these specific characteristics are. We are simply given numerical variables and asked to find similarities. This is where clustering comes in – similar people will congregate towards each other, and we want to see if we can identify their groupings.
4. We will look at a very basic example first. Suppose the following 6 observations are given:

Obs	X_1	X_2
1	160	74
2	165	72
3	165	74
4	175	68
5	180	70
6	185	72

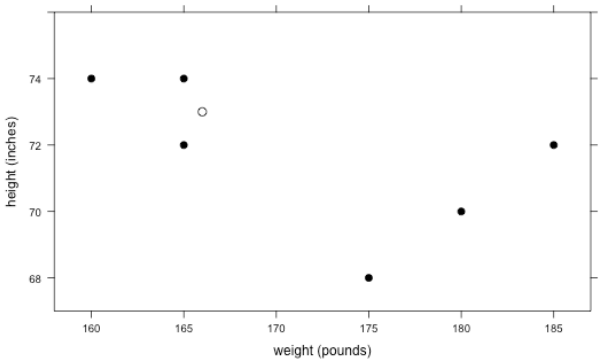
5. Plot the X_1 and X_2 points on a scatterplot either on the board or on poster paper (X_1 can be on the horizontal axis and X_2 can be on the vertical axis). The graph should look like the one below:



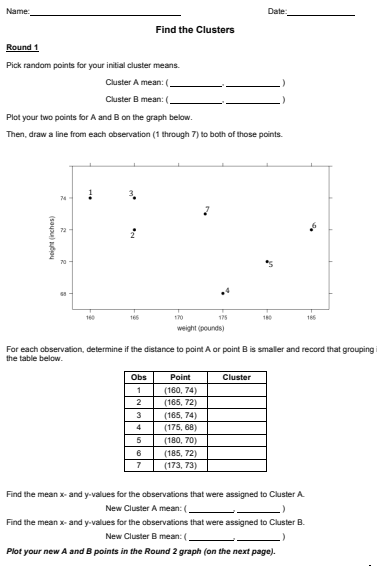
6. Ask students if they think there are any clusters, or groups, that stand out to them. It is likely that they will say there are 2 clusters in the graph: the top left corner 3 points, and the bottom right 3 points.
7. Now pose the following scenario that further describes the data:
    a. A doctor provides yearly physicals to the men's football and men's swimming teams at a local high school.
    b. He has collected data over the past few years on each player's weight (in pounds) and height (in inches). He informs us that weight was coded as the variable X_1, and height was coded as the variable X_2. You can re-label the scatterplot with this new information.



    c. Unfortunately, the doctor never recorded what sport each person played.
8. Using the information about height and weight, ask the students to decide:
    a. Which group of points most likely represents players from the swimming team? ***The points in the upper left corner are probably swimmers because swimmers are usually tall (and have large arm spans) and thin.***
    b. Which group of points most likely represents players from the football team? ***The points in the bottom right corner are probably football players because they tend to be heavier and more muscular.***
9. Now suppose a new player comes into the doctor's office for a physical. His weight and height are recorded as 166 pounds and 73 inches, respectively, but the doctor forgets to ask what sport he plays. Plot this point on the graph and ask students to determine which sport they think this student plays. ***This student is most likely a swimmer because he is tall and thin, and his point is near the swimming cluster.***



10. That was an easy one! But what if a player comes in and has the following measurements: weight = 173 pounds, height = 73 inches?
11. Distribute the *Find the Clusters* handout (LMR_4.24) and tell the students that the new point has been added to the "Round 1" graph.



Name:_____ Date:_____

Find the Clusters

Round 1

Pick random points for your initial cluster means.

Cluster A mean: (_____ , _____)

Cluster B mean: (_____ , _____)

Plot your two points for A and B on the graph below.

Then, draw a line from each observation (1 through 7) to both of those points.

For each observation, determine if the distance to point A or point B is smaller and record that grouping in the table below.

Obs	Point	Cluster
1	(160, 74)	
2	(165, 72)	
3	(165, 74)	
4	(175, 68)	
5	(180, 70)	
6	(185, 72)	
7	(173, 73)	

Find the mean x- and y-values for the observations that were assigned to Cluster A.

New Cluster A mean: (_____ , _____)

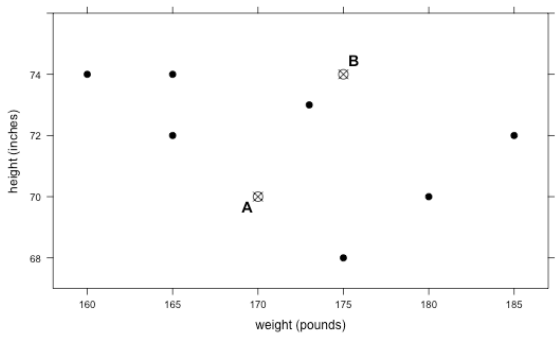
Find the mean x- and y-values for the observations that were assigned to Cluster B.

New Cluster B mean: (_____ , _____)

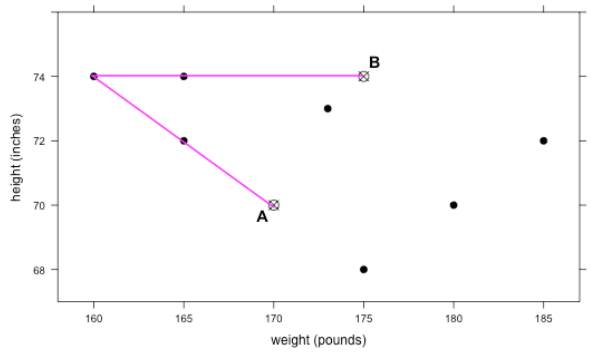
Plot your new A and B points in the Round 2 graph (on the next page).

LMR_4.24

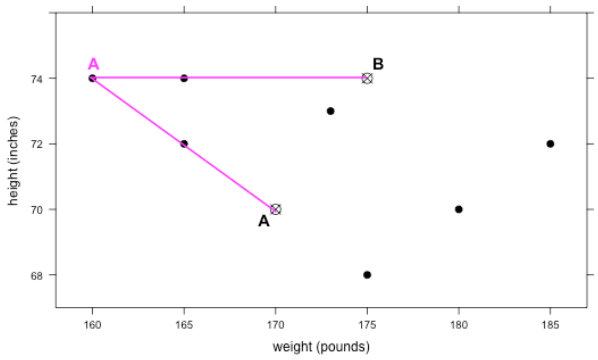
12. Ask students:
    a. On which team do you think this person plays? ***It is much more difficult to tell now because it looks like it is right in between the two clusters.***
13. In order to determine group placement, we can use an algorithm called **k-means clustering**. With this method, we select k clusters that we want to identify. Since we know we only have 2 types of athletes, football players and swimmers, we will be finding k = 2 clusters.
14. To introduce the students to this idea, circle the 3 points in the upper left corner (the ones that are likely the swimmers) and have students find the “mean point.” This means that they should find the mean x-value and the mean y-value of the 3 points. They can then plot this new point and use it as the mean of this particular group, or cluster.
15. The goal of this algorithm is to keep recalculating means as the possible clusters change. To begin, we will randomly pick 2 arbitrary points on the plot (we can call them A and B) to be our starting means for each cluster. There is no incorrect way to pick the starting means, but the further away the means are from the actual points, the longer it will take the algorithm to complete. If you would like to use the point found in Step 14 and label it as “A,” that is completely fine. You can simply pick just one other random point and label it as “B.”
16. For now, we will start with the following two points as guesses for the means of each group: A: (170, 70) and B: (175, 74). In the “Round 1” plot on the *Find the Clusters* handout, each student should plot and label these two points.



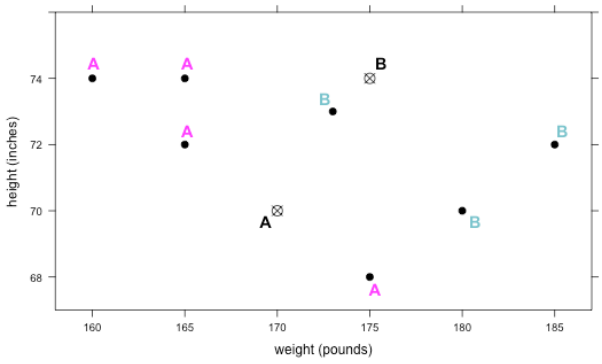
17. Inform the students that they will be drawing lines from each original point to both means. Then, they will decide if the point is closer to mean A or mean B and label the point with that letter. Lines have been draw from the top left point to the means in the plot below as a guide. You can draw this on the board as a reference for the students as well.



18. Since the line to point A is smaller, we would classify that point as being in cluster A (as shown below).



19. The students should draw similar lines for every point on the graph and make a decision as to which cluster each belongs in. They can simply eyeball it. Even if they guess incorrectly, the algorithm should be able to find the correct groups after some time. The correct classifications for Round 1 are as follows:



20. Once the class has agreed on the first round's cluster classifications, they should compute new values for the k-means (A and B). For mean A, they simply need to find the mean x-value for the 4 points and the mean y-value for the 4 points. The new means for A and B have been calculated below. The students should be calculating these on their own and recording their new means on the handout.

x-value for A = (160 + 165 + 165 + 175)/4 = 166.25
y-value for A = (74 + 72 + 74 + 68)/4 = 72

x-value for B = (173 + 180 + 185)/3 = 179.3
y-value for B = (73 + 70 + 72)/3 = 71.67

new A = (166.25, 72)
new B = (179.3, 71.67)

21. Have the students continue working through the handout until the cluster membership remains the same between 2 consecutive rounds. This means that, from one iteration to the next, the points in each cluster do not change.

Class Scribes:

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

Homework & Next Day

Write a paragraph that describes k-means clustering in your own words.

LAB 4H: Finding Clusters

Complete Lab 4H prior to Lesson 21.

Lab 4H - Finding clusters

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

Clustering data

- We've seen previously that data scientists have methods to predict values of specific variables.
  - We used *regression* to predict numerical values and *classification* to predict categories.
- *Clustering* is similar to classification in that we want to group people into categories. But there's one important difference:
  - In *clustering*, we don't know how many groups to use because we're not predicting the value of a known variable!
- In this lab, we'll learn how to use the k-means clustering algorithm to group our data into clusters.

The k-means algorithm

- The k-means algorithm works by splitting our data into *k* different clusters.
  - The number of clusters, the value of *k*, is chosen by the data scientist.
- The algorithm works *only* for numerical variables and *only* when we have no missing data.
- To start, use the `data` function to load the `futbol` data set.
  - This data contains 23 players from the US Men's National Soccer team (USMNT) and 22 quarterbacks from the National Football League (NFL).
- Create a scatterplot of the players `ht_inches` and `wt_lbs` and color each dot based on the `league` they play for.

Running k-means

- After plotting the player's heights and weights, we can see that there are two clusters, or different types, of players:
  - Players in the NFL tend to be taller and weigh more than the shorter and lighter USMNT players.
- Fill in the blanks below to use k-means to cluster the same height and weight data into two groups:

```
kclusters(____~____, data = futbol, k = ____)
```

- Use this code and the `mutate` function to add the values from `kclusters` to the `futbol` data. Call the variable `clusters`.

k-means vs. ground-truth

- In comparing our football and soccer players, we *know* for certain which league each player plays in.
  - We call this knowledge *ground-truth*.
- Knowing the *ground-truth* for this example is helpful to illustrate how k-means works, but in reality, data-scientists would run k-means not knowing the *ground-truth*.
- **Compare the clusters chosen by k-means to the ground-truth. How successful was k-means at recovering the `league` information?**

On your own

- Load your class' `timeuse` data (remember to run `timeuse_format` so each row represents the mean time each student spent participating in the various activities).
- Create a scatterplot of `homework` and `videogames` variables.
  - Based on this graph, identify and remove any outliers by using the `subset` function.
- Use `kclusters` with `k=2` for `homework` and `videogames`.
  - **Describe how the groups differ from each other in terms of how long each group spends playing `videogames` and doing `homework`.**

Lesson 21: Our Class Network

Objective:

Students will participate in an activity to map out their own network based on acquaintances between two people.

Materials:

1. *Friend Network Graphic* (LMR_4.25_Friend Network Graphic)
2. Index cards
3. *Network Code* file (LMR_4.26_Network Code R Script)

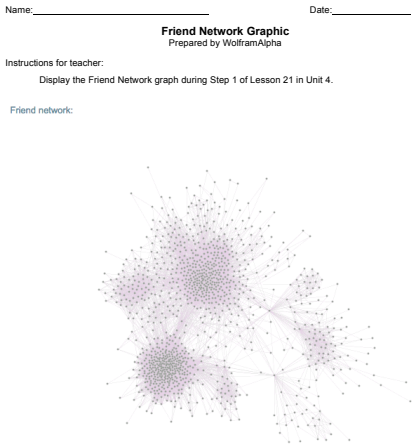
Vocabulary:

network

Essential Concepts: Networks are made when observations are interconnected. In a social setting, we can examine how different people are connected by finding relationships between other people in a network.

Lesson:

1. Display the *Friend Network Graphic* (LMR_4.25), which shows a WolframAlpha visualization of someone's Facebook friends. Inform the students that this type of model is called a **network**, which is simply a group of people or things that are interconnected in some way.


Name:_____ Date:_____

Friend Network Graphic
Prepared by WolframAlpha

Instructions for teacher:

Display the Friend Network graph during Step 1 of Lesson 21 in Unit 4.

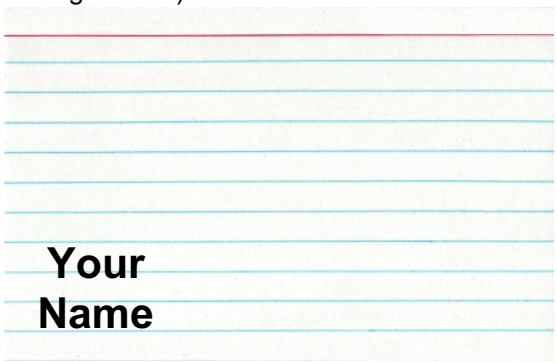
Friend network:

LMR_4.25

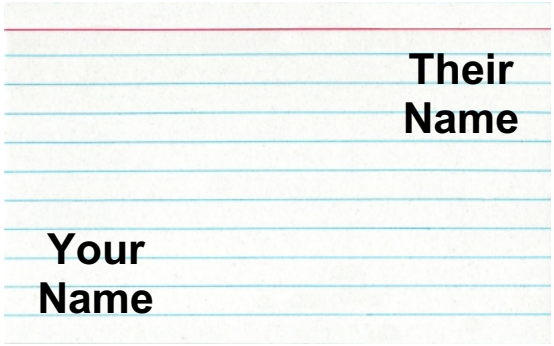
2. Ask the following questions about the graphic:

   a. What does each dot represent? ***Each dot represents one person.***
   b. What does each line represent? ***Each line represents a friendship between two people.***
   c. How are all the people in this graphic connected to each other? ***They are all friends with the person whose Facebook this is.***
   d. Why are some areas denser than others? ***A lot of people in the darker spots know each other, so there are more connections/friendships.***
   e. Why are some people not in groups at all (the dots at the edges of the graphic)? ***The main person does not have any friends in common with this person.***
   f. What might some of the groupings (the denser spots) represent? ***Answers will vary. Some examples include high school friends, college friends, graduate school friends, family members, or people who participate in similar hobbies.***

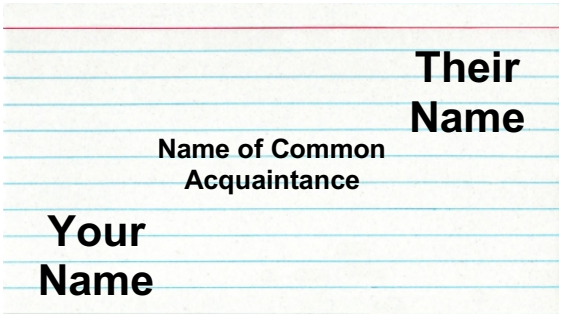
3. Ask the students what other types of social networks, other than Facebook, they belong to? Responses will most likely include TikTok, Twitter, Instagram, Snapchat, LinkedIn, Google+, etc.
4. Next, inform the students that networks can be as big or as small as we want. We can even determine our own class's social network and create visualizations from it!
5. Network Activity:
    a. Distribute index cards to students. Each student will need enough cards to make a connection with every other person in the class. For example, if there are 20 students in a class, then each student needs 19 cards.
    b. On EVERY index card, the student should write his/her first AND last name in the lower left-hand corner (see image below).



    c. Next, each student will walk around the classroom and put another student's first AND last name in the top right-hand corner of an index card (see image below).

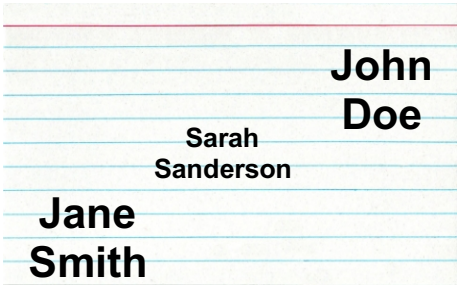


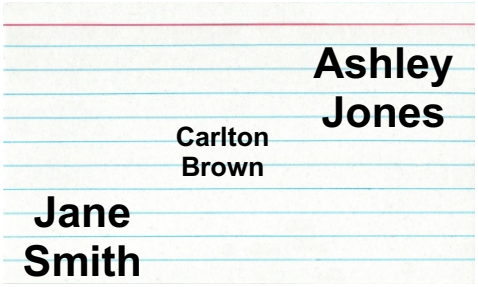
    d. In the center of the index card, the students should write the name of the *closest* 3rd person that they BOTH know (see image below). The person can be someone in the class, someone outside of the class, or someone who doesn't even attend the same school.



    e. Once all of the students have completed their cards, they will turn them in to the teacher so the teacher can create a visualization of the network.
    f. This will probably take an entire class period to complete, which is fine because the graphics can be created and shown the next day.

6. At this point, the teacher will need to manually input the data from the index cards into a spreadsheet. ***It is recommended that the spreadsheet be saved as a .csv file.*** Two sample index cards are included, along with how you would input the data.





Jane Smith	Sarah Sanderson
Sarah Sanderson	John Doe
Jane Smith	Carlton Brown
Carlton Brown	Ashley Jones

Note: The first index card corresponds to rows 1 and 2 in the spreadsheet (the purple box). The second index card corresponds to rows 3 and 4 in the spreadsheet (the pink box). So, each card will take up two rows in the spreadsheet.

Note: It is probably best to input the data after class and present the visualization during the next day.

7. Once all data has been input into a spreadsheet, use the code provided in the *Network Code* file (LMR_4.26) to produce graphs for the class's social network.

   Note: The R Script file can be opened and viewed in the "source" pane of RStudio. There are 2 places where the code needs to be edited by the teacher:

   a. Be sure to change the file name when reading in the .csv file in Line 7 of the code.
   b. Read the comments in Lines 91-96 to help find the 5 most popular people in the class's network. This may require some edits to Lines 97 and 108

```
1  ######################################
2  # Load and clean the data
3  ######################################
4
5  # Spreadsheet needs to be a .csv file for this code to work
6  # Be sure to replace "name_of_file_network_connections" with your actual file name
7  connect <- read.csv("name_of_file_network_connections.csv", head=FALSE, stringsAsFactors = FALSE)
8
9  # Assign variable names to columns 1 and 2 in the data set
10 names(connect) <- c("person1","person2")
11
12 # Create the connections between people
13 connect$person1 <- tolower(connect$person1)
14 connect$person2 <- tolower(connect$person2)
15 connect$person1 <- gsub(connect$person1, pattern = "-", replacement = " ")
16 connect$person2 <- gsub(connect$person2, pattern = "-", replacement = " ")
17
18 # Find all unique persons in the data set
19 uni_connect <- c(unique(connect$person1, unique(connect$person2)))
```

LMR_4.26

Class Scribes:

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

Next Day

Students will end their water usage campaign data collection after today's lesson. Starting the next day, they will analyze their data as part of the End of Unit 4 project.

End of Unit Design Project and Oral Presentation: Water Usage

Objective:

Students will apply their learning of the third and fourth units of the curriculum by completing an end of unit design project.

Materials:

1. Computers
2. *IDS Unit 4 – Project and Oral Presentation* (LMR_U4_Design Project)

End of Unit 4 Project and Oral Presentation: Water Usage

At the beginning of this unit, you explored a 2010 data set from the Los Angeles Department of Water and Power (DWP). You also created a Participatory Sensing campaign to investigate water usage around your community.

For this assignment, you will use both data sets to apply what you have learned in unit 4 and to answer the research question from the beginning of the unit:

How can we help city officials use Participatory Sensing to find out how water is being used around your neighborhood?

Your assignment is as follows:

1. You and a partner will predict water usage for the month of June using a subset of the dwp_2010 data set, which is called dwp_student.
    - Load the dwp_student data set.
    - Using this data, create two data sets: training and testing. Name these data sets `student_train and student_test`.
    - Create the best prediction model that you can based on your training data. Remember to `set.seed(123)` when creating your own training and testing data.
    - You're building this model with data from July 2010 to May 2011. You will use your model to predict water usage for June 2011.
    - After you settle on a specific model, submit your model (code) to your teacher. If you created any new variables, submit the code you used to create them as well.
    - ***What do the variables included in your prediction model say about how Angelenos use water?***
    - You will evaluate the prediction accuracy based on a separate set of data. Your teacher will give you another data set. Use this data set to evaluate your prediction. The pair with the smallest prediction error based on mean squared error (MSE), is the winner.

2. Using your Participatory Sensing data, explain how water is being used in your neighborhood. Make sure you use evidence from your PS data analysis. Be sure to answer the research question and your statistical questions.

Create a 5-minute presentation comprising of 4 to 5 slides that explains your model, the predicted value for June 2011 water consumption, and the findings using your campaign data. Be sure to include a detailed explanation of how you and your partner decided to create your prediction model, and how it performed on the test data set your teacher provided in your presentation. Each person must participate in the presentation. In addition to the presentation, submit a 2-4 page, double-spaced summary of your analysis including plots/graphs.

Note to teacher about the testing data set: The data set you will provide for students to test their prediction models is called **`dwp_teacher`**. It is recommended that you provide the data set's name upon students' submission of the code for their prediction models.