

# Introduction to Data Science

## Unit 4

# Introduction to Data Science

## Daily Overview: Unit 4

Theme	Day	Lessons and Labs	Campaign	Topics	Page
Campaigns and Community (5 days)	1	Lesson 1: Trash		Modeling to answer real world problems, official datasets	6
	2	Lesson 2: Drought		Exploratory data analysis, campaign creation	10
	3	Lesson 3: Community Connection		Community topic research, campaign creation	12
	4	Lesson 4: Evaluate and Implement the Campaign		Statistical questions, evaluate & mock implement campaign	15
	5^	Lesson 5: Refine and Create the Campaign	Team Campaign—data	Revise and edit campaign, data collection	17
Predictions and Models (13 days)	6	Lesson 6: Statistical Predictions Using One Variable	Team Campaign—data	One-variable predictions using a rule	20
	7	Lesson 7: Statistical Predictions by Applying the Rule	Team Campaign—data	Predictions applying mean square deviation, mean absolute error	22
	8	Lesson 8: Statistical Predictions Using Two Variables	Team Campaign—data	Two-variable statistical predictions, scatterplots	26
	9	Lesson 9: Spaghetti Line	Team Campaign—data	Estimate line of best fit, single linear regression	29
	10	<i>LAB 4A: If the Line Fits...</i>	Team Campaign—data	Estimate line of best fit	31
	11	Lesson 10: What's the Best Line?	Team Campaign—data	Predictions based on linear models	33
	12	<i>LAB 4B: What's the Score?</i>	Team Campaign—data	Comparing predictions to real data	36
	13	<i>LAB 4C: Cross-Validation</i>	Team Campaign—data	Use training and test data for predictions	38
	14	Lesson 11: What's the Trend?	Team Campaign—data	Trend, associations, linear model	41
	15	Lesson 12: How Strong Is It?	Team Campaign—data	Correlation coefficient, strength of trend	45
	16	<i>LAB 4D: Interpreting Correlations</i>	Team Campaign—data	Use correlation coefficient to determine best model	48
	17	Lesson 13: Improving Your Model	Team Campaign—data	Non-linear regression	51
	18	<i>LAB 4E: Some Models Have Curves</i>	Team Campaign—data	Non-linear regression	53
Piecing it Together (4 days)	19	Lesson 14: More Variables to Make Better Predictions	Team Campaign—data	Multiple linear regression	56
	20	Lesson 15: Combination of Variables	Team Campaign—data	Multiple linear regression	59
	21	<i>LAB 4F: This Model Is Big Enough for All of Us</i>	Team Campaign—data	Multiple linear regression	62
	22	Practicum: Predictions	Team Campaign—data	Linear regression	63
Decisions, Decisions! (3 days)	23	Lesson 16: Football or Futbol?	Team Campaign—data	Multiple predictors, classifying into groups, decision trees	65
	24	Lesson 17: Grow Your Own Decision Tree	Team Campaign—data	Decision trees based on training and test data	71
	25	<i>LAB 4G: Growing Trees</i>	Team Campaign—data	Decision trees to classify observations	75
Ties that Bind (3 days)	26	Lesson 18: Where Do I Belong?	Team Campaign—data	Clustering, k-means	79
	27	<i>LAB 4H: Finding Clusters</i>	Team Campaign—data	Clustering, k-means	85
	28+	Lesson 19: Our Class Network	Team Campaign—data	Clustering, networks	87
End of Unit Project (7 days)	29-36	End of Unit 4 Modeling Activity Project	Team Campaign	Synthesis of above	90

^=Data collection window begins.

+=Data collection window ends.

## **IDS Unit 4: Essential Concepts**

### **Lesson 1: Trash**

Exploring different datasets can give us insight about the same processes. Data from our Participatory Sensing campaigns rely on human sensors and limit the ability to generalize to the greater population.

### **Lesson 2: Drought**

Data can be used to make predictions. Official datasets rely on censuses or random samples and can be used to make generalizations.

### **Lesson 3: Community Connection**

Data collected through Participatory Sensing campaigns will be used to create models that answer real-world problems related to our community.

### **Lesson 4: Evaluate and Implement the Campaign**

Statistical investigative 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 5: Refine and Create the Campaign**

Statistical investigative 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 6: 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 7: Statistical Predictions by Applying the Rule**

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

### **Lesson 8: Statistical Predictions Using Two Variables**

When predicting values of a variable  $y$ , and if  $y$  is linearly associated with  $x$ , then we can get improved predictions by using our knowledge about  $x$ . For every value of  $x$ , find the mean of the  $y$  values for that value of  $x$ . If the resulting mean follows a trend, we can model this trend to generalize to unseen values of  $x$ .

### **Lesson 9: Spaghetti Line**

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

### **Lesson 10: What's the Best Line?**

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 errors as small as possible.

### **Lesson 11: 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 errors can be made smaller.

### **Lesson 12: 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 13: 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 14: 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 15: 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 16: Football or Futbol?**

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 17: Grow Your Own Decision Tree**

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

### **Lesson 18: 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 group of people into football players and swimmers, but what if you only knew each person's arm span? How well could you classify them into football players and swimmers now?

### **Lesson 19: 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.

# Campaigns and Community

Instructional Days: 5

## Enduring Understandings

Modeling Activities are posed as open-ended problems that are designed to challenge students to build models that are used to solve complex, real-world problems. They may be used to engage students in statistical reasoning and provide a means to better understand students' thinking.

## Engagement

Students will watch a video called *Fighting Pollution Through Data*. This video will set the context of the real-world problem facing many cities around the world — trash. The video provides background information as well as a baseline topic to launch ideas for the modeling process. The video can be found at: <https://www.youtube.com/watch?v=xOYAIXjHveA>

## 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."

### 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:

- Create a Participatory Sensing campaign using a campaign Authoring Tool

### 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 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 1: Trash**

### **Objective:**

Students will learn about reducing the burden of trash landfills.

### **Materials:**

1. *Video: Fighting Pollution Through Data*  
<https://www.youtube.com/watch?v=xOYAIxjHveA>
2. *Landfill Readiness Questions* handout (LMR\_4.1\_Landfill Readiness Questions)
3. *Trash Campaign Exploration* handout (LMR\_4.2\_Trash Campaign Exploration)
4. *Trash Campaign Creation* handout (LMR\_4.3\_Trash Campaign Creation)

**Essential Concepts:** Exploring different datasets can give us insight about the same processes. Data from our Participatory Sensing campaigns rely on human sensors and limit the ability to generalize to the greater population.

### **Lesson:**

1. Inform students that they will investigate a problem that faces many cities around the world today: trash.
2. Using the 5 Ws strategy, ask students to write down the 5 Ws in their DS journals as they watch a video about trash. The 5 Ws summarize the What, Who, Why, When, and Where of the resource.
3. Play the *Fighting Plastic Pollution Through Data* video, found at:  
<https://www.youtube.com/watch?v=xOYAIxjHveA>



**Note:** While the video is just over 13 minutes long, students should be able to answer the 5 Ws from the content of the first 10 minutes.



4. After they have finished watching the video, engage in a class discussion around the following question and discuss their insights, questions, and/or reactions to the video:
  - a. What types of data did they collect in the video? **Answer: Photos, location, brand names, weight (kg), waste categories (plastic bags, straws, toothpaste, etc.), and number of packaging.**
  - b. How are they useful in fighting plastic pollution? **Answer: Brand accountability, community involvement, cleaning rivers/dumping sites, and finding sources of pollution.**
5. Now we'll take inventory of our own understanding of landfills and how trash travels there. Distribute the *Landfill Readiness Questions* handout (LMR\_4.1). Allow students private think-time before having them discuss in their teams.

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. What types of items are recyclable? Give three examples of recyclable items that you use every day.

3. 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\_4.1

- Let students know that they will be exploring data from a trash Participatory Sensing campaign, titled the “Trash Campaign,” that was conducted at a number of high schools in the Los Angeles Unified School District (LAUSD).
- Concerned students in LAUSD engaged in a model eliciting activity and created a trash participatory sensing campaign to investigate possible trash issues in their communities. Based on the data collected, they made recommendations to the Los Angeles County Sanitation District (LACSD) that would help reduce the use of the regional landfills.
- Distribute the *Trash Campaign Exploration* handout (LMR\_4.2) to assist in students' interaction with the IDS public dashboard.

Name: \_\_\_\_\_ Date: \_\_\_\_\_

### Trash Campaign Exploration

**Instructions:**  
The survey questions/prompts for the Trash Campaign, a Participatory Sensing Campaign that was conducted at a number of high schools in the Los Angeles Unified School District (LAUSD), are provided below for your reference. Use the IDS public dashboard, (<https://portal.idsucla.org/>), to provide the questions on the next page.

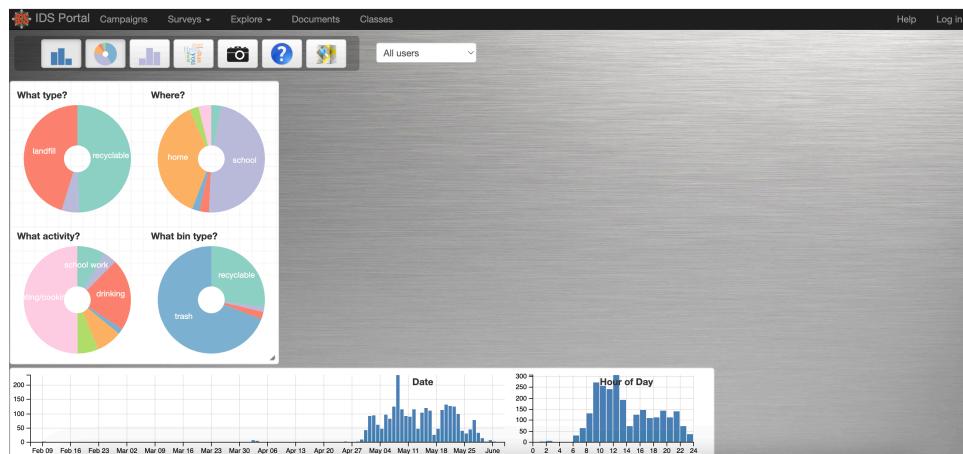
Survey Question/Prompt	Variable Name	Data Type
1. Please take a photo of your trash.	photo	photo
2. Please describe your trash.	trash_name	text
3. What type of trash? <input type="checkbox"/> recyclable <input type="checkbox"/> landfill <input type="checkbox"/> compost	trash_type	category
4. Where was this trash generated/found? <input type="checkbox"/> home <input type="checkbox"/> school <input type="checkbox"/> work <input type="checkbox"/> restaurants <input type="checkbox"/> stores/malls <input type="checkbox"/> in transit <input type="checkbox"/> others	where	category
5. What activity generated this trash? <input type="checkbox"/> eating/cooking <input type="checkbox"/> drinking <input type="checkbox"/> school work <input type="checkbox"/> cleaning <input type="checkbox"/> shopping <input type="checkbox"/> I found it <input type="checkbox"/> other	activity	category
6. Where did you put this trash when you were done? <input type="checkbox"/> recyclable <input type="checkbox"/> trash <input type="checkbox"/> compost/green waste <input type="checkbox"/> litter	disposal_bin	category
7. How many recycling bins can you see from your location?	recycle_bins	number

LMR\_4.2

- Navigate students to the IDS public dashboard: <https://portal.idsucla.org/>

- They should use the “Trash” campaign data and select “Dashboard” from the “Action” button.

11. The dashboard is a visual tool for exploring and analyzing data. An example screenshot of the Trash campaign in the dashboard is shown below.



12. Students should “play” with the data and think about characteristics of the campaign. Answers to the questions in the *Trash Campaign Exploration* handout (LMR\_4.2) are provided here for reference.
- a. How many observations are in this dataset? **Answer: 2,631.**
  - b. Where was the majority of trash generated? **Answer: School (1,254).**
  - c. How many observations were generated at school? **At work? Answer: School has 1,254 and Work has 59.**
  - d. What material or item was most commonly thrown away? **Answer: Recyclable (1,302) or Paper (477).**
  - e. Between what hours is the largest percentage of trash generated at home? **Answer: Between 2100 and 220, which is 9pm to 10pm (106).**
  - f. Which activity generates the largest percentage of landfill-destined trash? **Answer: Eating/cooking with 69.02% (822/1,191)**
  - g. Is eating or drinking more likely to generate a recyclable piece of trash? **Answer: Drinking, because it resulted in 480 recyclables versus Eating resulted in 399 recyclables, out of 1,302 total.**
  - h. When recycle bins are present, did a higher proportion of recyclable items end up in the trash bin when people were at home or at school? **Answer: School (134/225) had a higher proportion of recyclable items end up in the trash than Home (76/225).**
  - i. When someone littered, how many times was the person not around any type of waste receptacle? **Answer: 15 out of 58.**
13. Take time at the end of class to share out and discuss the components of the Trash Campaign. If needed, you may use the *Trash Campaign Creation* handout (LMR\_4.3) as an additional resource to help with the deconstruction of the Trash Campaign.

Name: \_\_\_\_\_ Date: \_\_\_\_\_

Trash Campaign Creation

Instructions:  
As a class, work together to fill in the information in this handout. You will be deconstructing the information that was used for the Trash Campaign.

**Round 1: Topic**  
*What do you think was the area of interest that students wanted to know more about with the Trash Campaign?*

**Topic:**  
\_\_\_\_\_

**Round 2: Research Question**  
*What do you think was the main question that students wanted to answer about the topic?  
This would have been the focus of the Trash Campaign.*

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

**Research Question:**  
\_\_\_\_\_

LMR\_4.3

**Class Scribes:**



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

## Lesson 2: Drought

### **Objective:**

Students will learn about the cause of California droughts and its effect on other states.

### **Materials:**

1. Article: *California, 'America's garden', is drying out*  
<https://yaleclimateconnections.org/2021/06/california-americas-garden-is-drying-out/>
2. U.S. drought data interactive map  
<https://www.drought.gov/historical-information?dataset=0&selectedDateUSDM=20120710>
3. *Team Campaign Creation* handout (LMR\_4.4\_Team Campaign Creation)

**Essential Concepts:** Data can be used to make predictions. Official datasets rely on censuses or random samples and can be used to make generalizations.

### **Lesson:**

1. Begin the lesson with the quote: "The consequences of drought in California are felt well outside the state's borders. California is effectively America's garden – it produces two-thirds of all fruits and nuts grown in the U.S."
2. Using the K-L-W strategy in their DS journals, give students a couple of minutes to write what they Know about droughts.
-  3. Then, students will write what they Learned about the California drought as they read the article titled *California, 'America's garden', is drying up*. The article is found at:  
<https://yaleclimateconnections.org/2021/06/california-americas-garden-is-drying-out/>
4. Finally, they will write 2-3 questions about what they Want to know/learn about droughts.
5. Do a quick Whip Around to share some of the students' responses to the K-L-W.
6. Next, load an interactive map of U.S. drought data by visiting:  
<https://www.drought.gov/historical-information?dataset=0&selectedDateUSDM=20120710>
-  7. Lead a discussion about what is on the page. Ask:
  - a. What do the colors and percentages on the legend mean? **Answer: The color is the drought type (Abnormally Dry, Moderate Drought, etc.) and the percentage is the percentage of area of the U.S. that is that type of drought.**
  - b. What do the colors, percentages, and years on the graph mean? **Answer: The colors correspond to the drought type (as mentioned in the previous question), the percentages correspond to the percentage of area of the U.S. that is that type of drought (as mentioned in the previous question), and the years signify the dates for which the data is available.**
  - c. What information are the map and graph displaying? **Answer: The map tells us the areas of drought and the graph tells us percentages of drought over time.**
  - d. Which date is selected by default and what percent of the U.S. is in Exceptional Drought that day? **Answer: July 10, 2012, and 0.62% is in Exceptional Drought.**

**Note:** If you used a different link other than was listed here, you may have a different default date. In order to see the types of droughts and their percentages, hover your cursor over the graph to line up with the date.
-  8. Then click on a new date to update the map. Ask:
  - a. What date is displayed now? **Answers will vary.**
  - b. What information is the map displaying? **Answers will depend on the chosen date above.**
  - c. Which states are affected by drought? **Answers will depend on the chosen date above.**

**Note:** You can click on a state to display its name.



9. Then click on the Combine States option, choose a state, and click Combine. Ask:
  - a. What information is the map displaying now? *Answers will vary but students should see the chosen state as the new map.*
  - b. What else do you see? *Answers will depend on the chosen state above.*
  - c. What are some wonderings you have about the data? *Answers will vary.*

**Note:** There are multiple options here. You can choose to display multiple states by adding another state (or states) and choosing Combine. You can also designate a specific time period in the Time Series Options.

10. Now that they know the overall layout of the interactive map and its options, ask student teams to complete the *Team Campaign Creation* handout (LMR\_4.4).

**Note:** The interactive map has a download data feature that allows customization/filtering of U.S. drought data. Keep this in mind as an option for students who may be interested in this topic.

Name: \_\_\_\_\_ Date: \_\_\_\_\_

Team Campaign Creation

Instructions:  
As a team, work together to fill in the information in this handout. You will be deciding, as a team, what information will be used for your participatory sensing campaign.

Article that addresses community concern or focus:  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

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

Team Ideas of Topics:  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

Team Decided Topic:  
\_\_\_\_\_

LMR\_4.4

### 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: Community Connection**

#### **Objective:**

Students will research and review articles from accredited sources to guide the design of their community focused Participatory Sensing campaign.

#### **Materials:**

1. Resource: *Web Literacy for Student Fact-Checkers*  
<https://pressbooks.pub/webliteracy/>
2. Video: *Online Verification Skills - Investigate the Source* found at:  
[https://www.youtube.com/watch\\_popup?v=hB6qjlxKltA](https://www.youtube.com/watch_popup?v=hB6qjlxKltA)
3. Video: *Online Verification Skills - Verifying Images and Videos* found at:  
[https://www.youtube.com/watch\\_popup?v=7eKG9RuqUE4](https://www.youtube.com/watch_popup?v=7eKG9RuqUE4)
4. Video: *Online Verification Skills - Find the Original Source* found at:  
[https://www.youtube.com/watch\\_popup?v=tRZ-N3OvvUs](https://www.youtube.com/watch_popup?v=tRZ-N3OvvUs)
5. *Team Campaign Creation* handout (LMR\_4.4\_Team Campaign Creation)  
**Note:** This handout was used in Lesson 2 but a new copy should be used for this lesson.
6. Article: *Great Pacific Garbage Patch: The World's Biggest Landfill in the Pacific Ocean*  
<https://science.howstuffworks.com/environmental/earth/oceanography/great-pacific-garbage-patch.htm>
7. Data file: *US Landfills*  
<https://www.epa.gov/lmop/landfill-technical-data>
8. Article: *3 reasons why California's drought isn't really over, despite the rain*  
<https://www.npr.org/2023/03/23/1165378214/3-reasons-why-californias-drought-isnt-really-over-despite-all-the-rain/>
9. Data file: *US Drought*  
<https://www.drought.gov/historical-information?dataset=0&selectedDateUSDM=20120710>
10. Article: *The number of homeless people in America grew in 2023 as high cost of living took a toll*  
<https://www.usatoday.com/story/news/nation/2023/12/15/homelessness-in-america-grew-2023/71926354007/>
11. Data file: *Annual Homeless Assessment Report (AHAR)*:  
<https://www.huduser.gov/portal/datasets/ahar.html>
12. Article: *COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide*  
<https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide>
13. Data file: *National Health Interview Survey (NHIS)*:  
<https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm>

**Essential Concepts:** Data collected through Participatory Sensing campaigns will be used to create models that answer real-world problems related to our community.

#### **Lesson:**

1. In this lesson, students will decide on a topic and research question for their Team Participatory Sensing campaign. They will review articles and choose at least one to set the context for the real-world problem that they will be addressing with their campaign.
2. It is important to check the reliability of your sources so that you are not spreading misinformation or basing your research on untrustworthy or inaccurate information.
3. An option for checking the credibility of sources is the SIFT method (the “SIFT” method is adapted from <https://pressbooks.pub/webliteracy/> by Michael A. Caulfield):
  - a. **Stop** – First, ask yourself if you know and/or trust the source of the page. Don’t read it or share it until you know what it is. Second, if you have started working through the moves and find yourself in a “click cycle”, STOP and remind yourself what your goal is – don’t lead yourself astray down a rabbit hole of facts/links.



- b. **Investigate the Source** – Take time to figure out where it is from before reading it. If it's from an unfamiliar source, do a Google or Wikipedia search. Here's a video to help with investigating a source:  
[https://www.youtube.com/watch\\_popup?v=hB6qjIxKltA](https://www.youtube.com/watch_popup?v=hB6qjIxKltA)
- c. **Find Trusted Coverage** – This is where we determine if a claim is reliable. Do a Google News search for relevant stories and use known fact-checking sites like snopes.com, factcheck.org, or politifact.com. If there are images in your source, you can do a reverse image search via Google to check its validity or if it's used elsewhere. Here's a helpful video on how to do a reverse image search:  
[https://www.youtube.com/watch\\_popup?v=7eKG9RuqUE4](https://www.youtube.com/watch_popup?v=7eKG9RuqUE4)
- d. **Trace claims, quotes, and media back to the original context** – A lot of what we see on the web is a re-reporting or commentary of a story. When you see phrases like "As reported by..." or "According to...", it can be an indication of this. Here's a video to help navigate finding an original source:  
[https://www.youtube.com/watch\\_popup?v=tRZ-N3OvvUs](https://www.youtube.com/watch_popup?v=tRZ-N3OvvUs)

4. Now that you know how to check the credibility of sources, we will review characteristics of a Participatory Sensing campaign.
5. Refer back to their class created campaign from Unit 3 to review. Using a *Pair-Share* strategy, ask students to discuss when a Participatory Sensing campaign should be used rather than a survey. *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.*
6. 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 a topic that addresses a community concern or interest.
7. Distribute the *Team Campaign Creation* handout (LMR\_4.4). Use the remainder of the class for students to find and review articles that will be the basis for their team participatory sensing campaigns. They will decide on a topic today and continue to create their campaigns in the next class period.

Name: \_\_\_\_\_ Date: \_\_\_\_\_

Team Campaign Creation

Instructions:  
As a team, work together to fill in the information in this handout. You will be deciding, as a team, what information will be used for your participatory sensing campaign.

Article that addresses community concern or focus:  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

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

Team Ideas of Topics:  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

Team Decided Topic:  
\_\_\_\_\_

LMR\_4.4



8. Facilitate the student teams' brainstorm session by circulating around the room to check for understanding. If teams need help with finding an article and choosing a topic, you may recommend one of the following:
  - a. Article: *Great Pacific Garbage Patch: The World's Biggest Landfill in the Pacific Ocean*  
<https://science.howstuffworks.com/environmental/earth/oceanography/great-pacific-garbage-patch.htm>
    - i. Data file: *US Landfills*  
<https://www.epa.gov/lmop/landfill-technical-data>

**Note:** This article and supporting data align with the *trash* topic introduced in lesson 1.

- b. Article: *3 reasons why California's drought isn't really over, despite the rain*  
<https://www.npr.org/2023/03/23/1165378214/3-reasons-why-californias-drought-isnt-really-over-despite-all-the-rain/>
    - i. Data file: *US Drought*  
<https://www.drought.gov/historical-information?dataset=0&selectedDateUSDM=20120710>

**Note:** This article and supporting data align with the *drought* topic introduced in lesson 2.
  - c. Article: *The number of homeless people in America grew in 2023 as high cost of living took a toll*  
<https://www.usatoday.com/story/news/nation/2023/12/15/homelessness-in-america-grew-2023/71926354007/>
    - i. Data file: *Annual Homeless Assessment Report (AHAR)*:  
<https://www.huduser.gov/portal/datasets/ahar.html>

**Note:** This article and supporting data align with the topic of homelessness.
  - d. Article: *COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide*  
<https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide>
    - i. Data file: *National Health Interview Survey (NHIS)*:  
<https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm>

**Note:** This article and supporting data align with the topic of mental health.
9. Take time near the end of class to have student teams share out their chosen topics.

#### 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 come up with at least 2 possible research questions for their Participatory Sensing campaign. They will come to a consensus about their research question with their team in the next class period.

## **Lesson 4: Evaluate and Implement the Campaign**

### **Objective:**

Students will complete the design of their community focused Participatory Sensing campaign and implement a mock campaign to evaluate the feasibility of the campaign.

### **Materials:**

1. *Team Campaign Creation* handout (LMR\_4.4\_Team Campaign Creation) from previous lesson

**Essential Concepts:** Statistical investigative 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 team Participatory Sensing campaign. Allow them some time to review their possible research questions with their team and to decide on a team research question before moving on to round 3.
2. **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?
3. Once teams have decided on their types of data and trigger, they will create survey questions/prompts to collect this type of data with this trigger.
4. **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.
5. 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. number, text, photo, time, single choice)?
  - 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.
6. If survey questions need to be rewritten, students will decide as a team on the changes.
7. Once finalized, they'll record the survey question/prompt that goes along with each data variable on their *Team Campaign Creation* handout (LMR\_4.4), being cognizant of question bias.

Name: \_\_\_\_\_ Date: \_\_\_\_\_

#### Team Campaign Creation

Instructions:  
As a team, work together to fill in the information in this handout. You will be deciding, as a team, what information will be used for your participatory sensing campaign.

Article that addresses community concern or focus:  
\_\_\_\_\_  
\_\_\_\_\_

#### Round 1: Topic

*This is a hobby, area of interest, or place or process that you want to know more about.*

Team Ideas of Topics:  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

Team Decided Topic:  
\_\_\_\_\_

-  8. **Round 5:** In teams, students will now generate three statistical investigative questions that they might answer with the data they will collect and to guide their campaign. They need to make sure that their statistical investigative questions are interesting and relevant to their chosen topic. Remind students that they will also have data about the date, time, and place of data collection.
9. 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 investigative questions to guide their campaign.
-  10. 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 investigative questions?
11. During their discussion about whether their campaign is reasonable and ethically sound, if teams discover that they need to make changes, they can make adjustments at this time.
12. Students have collaboratively created their Team Participatory Sensing campaign.

**Class Scribes:**



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

**Homework**

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

## **Lesson 5: Refine and Create the Campaign**

### **Objective:**

Students will revise their community focused Participatory Sensing campaign according to the finding from the mock implementation of their campaign to refine it. Student teams will then create their campaigns using the Campaign Authoring tool.

### **Materials:**

1. *Campaign Authoring* handout (LMR\_4.5\_Campaign Authoring)

**Essential Concepts:** Statistical investigative 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

### **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. Now they will use the Campaign Authoring tool to create a campaign like the ones they see on their smart devices or the computer.
- 4. Distribute the *Campaign Authoring* handout (LMR\_4.5). 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**

Follow the instructions below to create your campaign. If you would like a video tutorial to accompany the written instructions, it can be found here: <https://youtu.be/PzWmCHQohnI>

Go to the IDS Home Page found at <https://portal.idsucia.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 Info:**

- 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. **Campaign Type:** Select Research.
- v. **Editable Survey:** Set to Disabled in order to monitor for improper responses.
- vi. **Editable Responses:** 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 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 (optional – may be the same as the Description in Campaign Info).
- iv. **Submission Message:** Provide a brief message to be displayed after survey submission. Note: it is helpful to include a reminder to click the green button to submit the survey.
- v. Click the **[+Add Prompt]** button and select the prompt type for your first survey question.

c. **Prompt Information:**

LMR\_4.5

**Note:** 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.

5. Ask teams to refresh their campaigns on their smartphones or their web browser to verify that their campaign appears as one of the choices.
6. Now that they are finished with their campaigns, student teams will use the remaining time to plan the expectations for their End of Unit 4 Modeling Activity Project.

### **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 Team Participatory Sensing campaign. They will have until the end of Unit 4 to collect data. Since only the members of their team will be involved in gathering data for their campaign, this allows them time to ensure they have a sufficient amount of data.

# Predictions and Models

Instructional Days: 13

## 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 explore and make predictions with a dataset consisting of arm span and height values from a group of Los Angeles high school students. The Arm Span vs. Height data allows for a real-world connection while learning about linear models and predictions. They will engage in multiple discussions as they build their understanding of linear models, refine how they make their predictions, and test the accuracy of those predictions.

## 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.

### 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 Mean Squared Error and Mean Absolute Error.
2. Students will engage in partner and whole group discussions to express their understanding of linear regression and how to measure its accuracy.
3. Students will use mathematical vocabulary to explain orally and in writing the attributes of various scatterplots.
4. Students will make connections, in writing, between predictions using different types of models (i.e., linear, quadratic, cubic).

## Data File or Data Collection Method

### Data File:

1. Arm Spans vs. Heights: `data(arm_span)`
2. Movies: `data(movie)`

### Data Collection:

Students will collect data for their Team Participatory Sensing campaign.

## Legend for Activity Icons



Video clip



Discussion



Articles/Reading



Assessments



Class Scribes

## **Lesson 6: 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\_4.6\_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\_4.6).

Name: \_\_\_\_\_ Date: \_\_\_\_\_

#### **Heights of Students at a Large High School**

##### **Background:**

Our class will help judge a contest held at a particular high school to see who can make the best predictions.

Height data for 40 randomly selected students were provided to three teams.

Using this data, each team was asked to predict the heights of a random sample of 10 students. Here is the catch: teams were allowed to give only ONE number that had to be used to predict all 10 heights.

As the judges of this contest, you will determine the winner.

##### **Instructions:**

1. Your job is to determine the winning team. You 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.
  - b. The rule must be applied to each team's prediction, and you must be able to explain how your rule helped select the winner. For example, do you choose the team with the largest score? The smallest?
2. Each team's predictions are provided here for your reference.
3. Answer the questions that follow.

##### **Team Predictions:**

Team A: 67.9 inches

Team B: 68.1 inches

Team C: 70.9 inches

## **LMR\_4.6**

3. Our class will help judge a contest held at a particular high school to see who can make the best predictions. Height data for 40 randomly selected students were provided to three teams. Using this data, each team was asked to predict the heights of a random sample of 10 students. Here is the catch: teams were allowed to give only ONE number that had to be used to predict all 10 heights. As the judges of this contest, you will determine the winner.
4. Your job is to determine the winning team. You 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.
  - b. The rule must be applied to each team's prediction, and you must be able to explain how your rule helped select the winner. For example, do they choose the team with the largest score? The smallest?
5. Here are the predictions of the three teams:

Team A: 67.9 inches

Team B: 68.1 inches

Team C: 70.9 inches

6. Display Dataset A, found on page 1 of the *Heights of Students at a High School* handout (LMR\_4.6).

**Notes to teacher:**

- a. Students may have to be reminded that negative values with large absolute values are larger than positive values with small absolute values (e.g.,  $|-10|$  is larger than  $|3|$  because 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 67.9, actual outcome was 70.1, so  $67.9 - 70.1 = -2.2 = |-2.2| = 2.2$ ). 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:  $\text{sum}(\text{heights}-\text{predicted.value} == 0)$   
*Translated into words: the number of exactly correct predictions*
  - ii. Rule 2:  $\text{sum}(\text{heights}-\text{predicted.value})$   
*Translated into words: the sum of the differences between predicted value and the actual heights*
  - iii. Rule 3:  $\text{sum}(\text{abs}(\text{heights}-\text{estimate}))$   
*Translated into words: the sum of the absolute values of the deviations/errors*
  - iv. Rule 4:  $\text{sum}((\text{heights}-\text{estimate})^2)$   
*Translated into words: the sum of the squared deviations/errors*

**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.



7. Allow student teams time to discuss and complete the task for Dataset A.
  8. Do not share their responses to Dataset A. Instead, display the following questions:
    - a. What if we had a different set of 10 randomly selected students?
    - b. Would the same team win?
  9. Allow teams to discuss the questions, then share a couple of responses to the questions in the previous step.
  10. Display Dataset B, found on page 2 of the *Heights of Students at a High School* handout (LMR\_4.6), 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.
11. Teams will take turns to share their work as follows:
    - a. Which team did you select as the winner using Dataset A?
    - b. Explain the method, or **rule**, your team used to declare the winner.
    - c. Which team did you select as the winner using Dataset 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.
  12. During the share out, students will take notes about the other teams' rules in their DS journals.
  13. 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 7: 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. *Heights of Students at a Large High School* handout (LMR\_4.6\_HS Student Heights)
3. *A Tale of Two Rules* handout (LMR\_4.7\_A Tale of Two Rules)
4. *Prediction Games* handout (LMR\_4.8\_Prediction Games)

### **Vocabulary:**

training data, test data, mean squared error, mean absolute error, residual

**Essential Concepts:** If we use the mean squared errors rule, then the mean of our current data is the best prediction of future values. If we use the mean absolute errors 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 is their job to find the best procedure to win the contest.
3. Have students refer back to the *Heights of Students at a Large High School* handout (LMR\_4.6) from the previous lesson.
4. Recall that the student teams were provided with height data on 40 selected students to come up with their predictions for future observations. This is a common practice with statisticians and data scientists. The first dataset of 40 students is called the **training data** where we train a model to make predictions. Then we use the **test data** (Dataset A and Dataset B) to test those predictions. Using the training data, the teams used different statistics for their predictions:
  - a. Team A used the mean.
  - b. Team B used the median.
  - c. Team C used the third quartile.
5. In the previous lesson, you created your own rules to determine the winner. Today, you will learn rules that statisticians and data scientists use. The first is called the **mean squared error** rule.  
**Note to teacher:** acknowledge any groups who came up with MSE or MAE on their own in the previous lesson.
6. An "error" is the difference between our prediction and the actual outcome and is sometimes called a "residual". The mean squared error is also called:
  - a. Mean squared deviation
  - b. Mean squared residual
  - c. Residual sum of squares

The formula looks like this, where  $\hat{x}$  stands for the predicted value:

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

Using this formula, the teams' scores are determined by finding the average of the squared differences between their predictions and the actual values. The winner is the team with the lowest mean squared error.

7. Let's use R to do the heavy lifting for us. Demonstrate how to find the mean squared error by typing the following commands in the console (throughout the process, show what is happening in the Environment and the dataframe):

```
#First, let's create a vector of the heights in our first dataset:  
height <- c(70.1, 61, 70.1, 68.1, 63, 66.1, 61, 70.1, 72.8, 70.9)  
  
#Next, convert this vector into a dataframe:  
datasetA <- data.frame(height)  
  
#Now we find the residuals using one of the statistics.  
#For this example, we'll use the first quartile from the training data (65 inches):  
datasetA <- mutate(datasetA, residual = height-65)  
  
#Next, we will square each residual:  
datasetA <- mutate(datasetA, sq_res = residual^2)  
  
#Finally, we use the mean function to:  
#sum up the squared residuals and divide by 10 to find our mean squared deviation:  
mean(~sq_res, data = datasetA)
```

This process gives us the mean squared error of 22.05.

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

Interpretation: When using the Q<sub>1</sub> height to make predictions about all heights, our predictions will typically be off by  $\sqrt{22.05} = 4.6957$  inches.

Here is the vector of heights for Dataset B:

```
heightB <- c(70.1, 72, 68.9, 61.8, 70.9, 59.8, 72, 65, 66.1, 68.9)
```

8. Distribute the *A Tale of Two Rules* handout (LMR\_4.7).

Name: \_\_\_\_\_ Date: \_\_\_\_\_

### A Tale of Two Rules

Background: How consistent were our rules for the contest in determining a winner?

Statisticians and Data Scientists use rules, like mean squared error, for determining the accuracy of their predictions.

- The mean squared error rule says: the score is determined by finding the average of the squared differences between the prediction and the actual values.

Instructions:

For each of the test datasets below, calculate the mean squared error and determine which prediction is best.

#### Dataset A

Here are the heights from Dataset A – 70.1, 61, 70.1, 68.1, 63, 66.1, 61, 70.1, 72.8, 70.9

Summary of 40 Students (heights in inches)					
	Team A	Team B	Team C	3 <sup>rd</sup> Quartile	Maximum
Minimum	65	67.9	68.1	70.9	76
MSE	22.05				

Which team/statistic made the best prediction using MSE?

#### Dataset B

Here are the heights from Dataset B – 70.1, 72, 68.9, 61.8, 70.9, 59.8, 72, 65, 66.1, 68.9

LMR\_4.7

9. Let's see how well our teams' predictions did on the heights of the test data. Students will work in teams to answer: Using the mean squared errors, which statistic is the winner? Discuss which statistic/team made the best predictions. **See answers below:**

### Dataset A

Here are the heights from Dataset A – 70.1, 61, 70.1, 68.1, 63, 66.1, 61, 70.1, 72.8, 70.9

Summary of 40 Students (heights in inches)					
	Team A	Team B	Team C		
Minimum	1 <sup>st</sup> Quartile	Mean	Median	3 <sup>rd</sup> Quartile	Maximum
59.1	65	67.9	68.1	70.9	76
MSE	84.236	22.05	17.004	17.276	29.484
					92.01

### Dataset B

Here are the heights from Dataset B – 70.1, 72, 68.9, 61.8, 70.9, 59.8, 72, 65, 66.1, 68.9

Summary of 40 Students (heights in inches)					
	Team A	Team B	Team C		
Minimum	1 <sup>st</sup> Quartile	Mean	Median	3 <sup>rd</sup> Quartile	Maximum
59.1	65	67.9	68.1	70.9	76
MSE	87.673	22.273	16.393	16.573	27.493
					87.673

**Note to teacher:** Explain that the mean/Team A was the winner of this contest. 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.

10. Ask: What if another data science class has a best rule that is different from ours?
11. 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 can be proven mathematically but it's beyond the scope of this course. The mean absolute error is expressed as (where  $\hat{x}$  stands for the predicted value):

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

12. 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.

Here is the script for MAE:

```
#Find the absolute value of the residuals using one of the statistics
#for this example, we'll use the first quartile from the training data (65 inches):
datasetA <- mutate(datasetA, residual = abs(heightA-65))

#Finally, we use the mean function to:
#sum up the residuals and divide by 10 to find our mean absolute error:
mean(~residual, data = datasetA)
```

This process gives us the mean absolute error of 4.32.

13. Using our previous examples, recalculate your predictions using the MAE.
14. Using the mean absolute error, which statistic/team is the winner? **See answers below:**

### Dataset A

Here are the heights from Dataset A – 70.1, 61, 70.1, 68.1, 63, 66.1, 61, 70.1, 72.8, 70.9

Summary of 40 Students (heights in inches)					
	Team A	Team B	Team C		
Minimum	1 <sup>st</sup> Quartile	Mean	Median	3 <sup>rd</sup> Quartile	Maximum
59.1	65	67.9	68.1	70.9	76
MAE	8.22	4.32	3.52	3.48	3.96
					8.68

### Dataset B

Here are the heights from Dataset B – 70.1, 72, 68.9, 61.8, 70.9, 59.8, 72, 65, 66.1, 68.9

Summary of 40 Students (heights in inches)					
	Team A	Team B	Team C		
Minimum	1 <sup>st</sup> Quartile	Mean	Median	3 <sup>rd</sup> Quartile	Maximum
59.1	65	67.9	68.1	70.9	76
MAE	8.45	4.23	3.43	3.39	3.79
					8.45

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

15. Optional practice: Students can practice finding the mean squared error and mean absolute error using the mean and median with the *Prediction Games* handout (LMR\_4.8). The LMR includes the five number summary, if they were curious how the MSE and MAE for other statistics compare.

Name: \_\_\_\_\_ Date: \_\_\_\_\_

### Prediction Games

Instructions:

For each of the games below, calculate the statistics and determine which one works best.

- The **mean squared error** rule says: the score is determined by finding the average of the squared differences between the prediction and the actual values.
- The **mean absolute error** rule says: the score is determined by finding the average of the absolute value of the differences between the prediction and the actual values.

#### Game 1

Given the following statistics from randomly selected height training data, determine the best predictor for the following test data.

Here are the heights from the test data – 66, 67, 73, 68, 68, 73, 69, 64, 66, 67

Training Data Summary (Heights in Inches)					
	Minimum	1 <sup>st</sup> Quartile	Median	Mean	3 <sup>rd</sup> Quartile
MSE	64.20	66.40	67.76	68.22	69.13
MAE					73.15

Which statistic did best with MSE?

Which statistic did best with MAE?

LMR\_4.8

### Class Scribes:



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

## Lesson 8: Statistical Predictions Using Two Variables

### **Objective:**

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

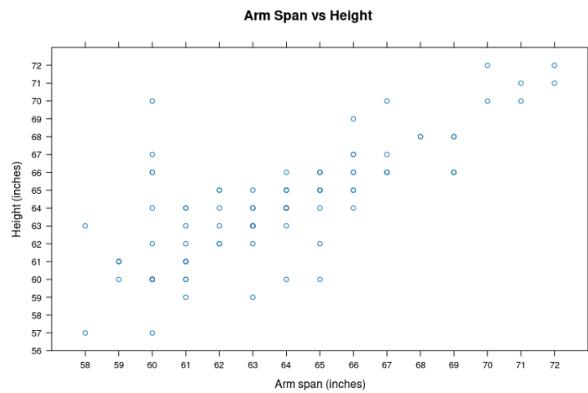
### **Materials:**

1. *Arm span vs. Height* Scatterplot (LMR\_4.9\_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 linearly associated with  $x$ , then we can get improved predictions by using our knowledge about  $x$ . For every value of  $x$ , find the mean of the  $y$  values for that value of  $x$ . If the resulting mean follows a trend, we can model this trend to generalize to unseen values of  $x$ .

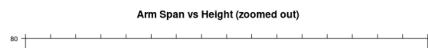
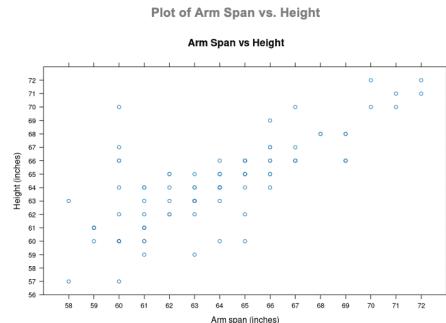
### **Lesson:**

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 from LMR\_4.9 on a white board or overhead projector (you will write on the board or the transparency later in the lesson — see step 9).



4. Distribute the *Arm Span vs. Height* handout (LMR\_4.9). Students will refer to this handout again later in a subsequent lesson.

Name: \_\_\_\_\_ Date: \_\_\_\_\_



LMR\_4.9

5. In teams, ask students to analyze the plot and discuss the following questions:
  - What kind of plot is this? **Answer: Scatterplot.**
  - How many variables are displayed in this plot? **Answer: Two variables.**
  - Which variable is shown on the x-axis? On the y-axis? **Answer: Arm span is shown on the x-axis and height is shown on the y-axis.**
  - What is this plot showing? **Answer: 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? **Answer: 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 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 arm spans. Refer to the OPTIONAL section at the end of this lesson.
  10. Then plot these points on the graph. We will use this later – the points should be roughly along a straight line.
- Note:** 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 9.

**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.

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

```

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

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

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

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

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

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

```

### Class Scribes:



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

## Lesson 9: The Spaghetti Line

### **Objective:**

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

### **Materials:**

1. *The Spaghetti Line* (LMR\_4.10\_The Spaghetti Line)

**Note:** Advance preparation is required. Cut out plots prior to beginning the lesson.

2. 1 lb. of Uncooked Spaghetti
3. Grid Paper
4. Tape or Glue
5. Poster paper

### **Vocabulary:**

line of best fit

**Essential Concepts:** We can often use a straight line to summarize a trend. “Eyeballing” a straight line to a scatterplot is one way to do this.

### **Lesson:**

**Note:** Lab 4A may be done in the same class period as this lesson.

1. Inform students that in this lesson, they will estimate the equation of the **line of best fit** for a height and arm span dataset.
2. Distribute *The Spaghetti Line* handout (LMR\_4.10) 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. Students may need extra help with finding the correct y-intercept because they cannot rely on the y-axis values in this case since the plots are not starting at the origin, (0, 0).

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 investigative question:

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

**Instructions:**

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

LMR\_4.10

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



4. 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 and this line is called the **line of best fit**.

**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**

**LAB 4A: If the Line Fits...**

Complete Lab 4A prior to Lesson 10.

## Lab 4A - If the line fits ...

Directions: Follow along with the slides, completing the questions in **blue** on your computer, and answering the questions in **red** 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` and `armspan <= 63`.
  - A second for `armspan >= 64` and `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.**  
`add_line()`

- **On the Plot pane, click two data points to draw a line through.**
- NOTE: If your line does not appear or it appears but is above the points you selected, zoom out on your browser (typically 50% if you have a Mac, 80% on Windows). Or if your line appears below the points you selected, zoom in on your browser. Then run the `add_line()` function again and click on two points. Zoom out (or in) until your line appears through the points you selected.

### 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-inch `armspan` 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. Did both of you come up with the same equation for a line? If not, can you tell which line fits the data best?**

## Lesson 10: What's the Best Line?

### **Objective:**

Students will understand that the mean squared error (MSE) is a way to assess the fit of a linear model. The MSE measures the total squared distances between all the data values from the line of best fit and divides it by the number of observations in the dataset.

### **Materials:**

1. *Arm Span vs. Height* Scatterplot (LMR\_4.9\_Arm Span vs Height) from lesson 8
2. *Testing Line of Best Fit* handout (LMR\_4.11\_Testing Line of Best Fit)

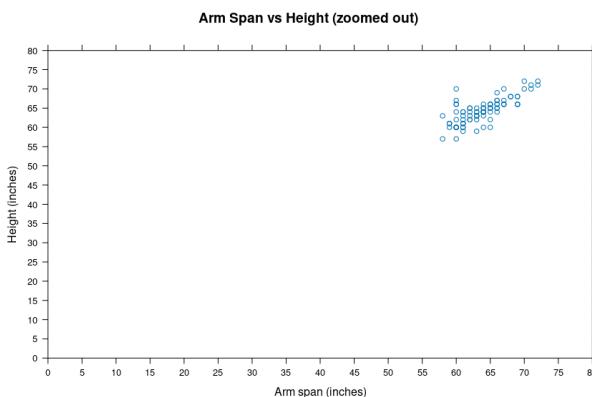
### **Vocabulary:**

regression line, 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 errors as small as possible.

### **Lesson:**

1. Ask student teams to refer back to the *Arm Span vs. Height* handout (LMR\_4.9) but this time have them look at the zoomed out scatterplot.



2. Using their understanding of a line of best fit from The Spaghetti Line lesson and Lab 4A, have them draw (or use strands of spaghetti) what they believe to be the line of best fit for the data.  
**Note:** They can use their equations from Lab 4A as a guide but note that it will be difficult to plot decimals on this scatterplot.
3. Ask students: How does this line compare to the lines from the team posters in The Spaghetti Line lesson? *Answers will vary but students may notice that the y-intercepts may be similar or that the overall slope appears similar (they are not writing an equation for the line in step 2, but they may notice where their line intercepts the y-axis and/or the steepness of the line, i.e. slope).*
4. Reveal the equation of the line of best fit for the Arm Span vs. Height data and ask students to compare it to their equations from Lab 4A:

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

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

5. 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.

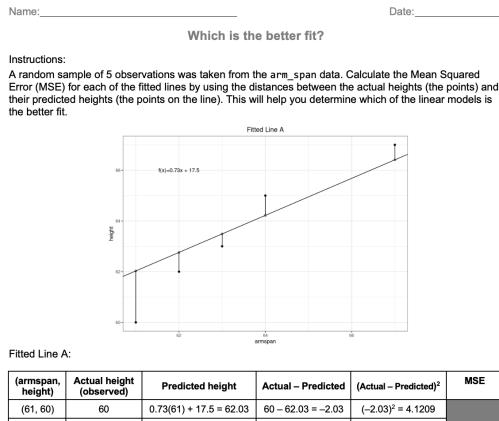
- Inform students that the equation of the line is a rule that predicts the height based on a second variable, in this case, arm span. Data points are **observed values** and points on the line are **predicted values**.



- 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?**

- What was the actual height for someone with an arm span of 67 inches? *Answer: There are three points on our Arm Span vs Height scatterplot at an arm span of 67 inches; 66-inch height, 67-inch height, and 70-inch height.*
- How close was our predicted height? *Answer: Our line of best fit predicted that someone with an arm span of 67 inches has a height of 66.5933 inches, which rounded to the nearest whole number is 67 inches. This is pretty close as it matches one of the possible heights for someone with an arm span of 67 inches in our data.*
- Remind students that lines of best fit are also known as **regression lines** and they are models that can be used to make predictions.
- 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:
  - What methods have we used so far? *Answer: We've used Mean Squared Error and Mean Absolute Error (Lesson 7).*
  - When is it appropriate to use each method? *Answer: It is best to use Mean Squared Error when we were looking at mean and Mean Absolute Error when we were looking at median.*
- Distribute the *Testing Line of Best Fit* handout (LMR\_4.11). Students will calculate MSE by using the distances between the actual heights (the points) and their predicted heights (the points on the line) of two different lines. They do this so that they can understand what those distances mean – that together they form our “error” that help us determine the best fitting line.



LMR\_4.11

- Discuss with students:

- What did you have to do to your MSE value to make it useable for interpretation? *Answer: We had to take the square root of our MSE value in order to convert it back to inches.*
- Which linear model was the better fit? How do you know? *Answers will vary but this is where students should compare the MSE values – a smaller MSE indicates a smaller error, and therefore a better fit.*

**Note:** Students may ask for an easier and/or faster way to calculate MSE. They will be using RStudio to calculate MSE in 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**

**LAB 4B: What's the Score?**

**Lab 4C: Cross-Validation**

Complete Labs 4B and 4C prior to Lesson 11.

## Lab 4B - What's the score?

Directions: Follow along with the slides, completing the questions in **blue** on your computer, and answering the questions in **red** 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 `xypplot` 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 add a column of residuals to `arm_span`:  
`____ <- 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 first 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.
- Describe and interpret, in words, what the output of your accuracy summary means.
- Write down why adding positive and negative errors together is problematic for assessing prediction accuracy.

## Mean squared error

- When adding residuals, the positive errors in our predictions (underestimates) are cancelled out by negative errors (overestimates) which lead to the impression that our model is making better predictions than it actually is.
- To solve this problem we calculate the squared values of the errors because squared values are always positive.
- 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))`
- Compare your MSE with a neighbor. Whose line was more accurate and why?**

## 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 use *regression lines*.
  - We also refer to *regression lines* as *linear models*.
  - This line connects the mean `height` of people with similar `armspans`.
  - Fill in the blanks below to create a *regression line* using `lm`, which stands for *linear model*:**

```
best_fit <- lm(____ ~ ____, data = arm_span)
```

## Plotting regression lines

- Type `best_fit` into the console to see the slope and intercept of the regression line.**
- Add this line to a scatterplot by filling in the blanks below.**  
`add_line(intercept =____, slope = ____)`

## Predicting with regression lines

- 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.

## 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.
- Calculate the MSE for the values predicted using the regression line.**
- Compare the MSE value you calculated using the line you fitted with `add_line()` to the linear model obtained with `lm()`. Which linear model performed better?**
- Ask your neighbors if any of their lines beat the `lm` line in terms of the MSE. Were any of them successful?**

## Lab 4C - Cross-Validation

Directions: Follow along with the slides, completing the questions in **blue** on your computer, and answering the questions in **red** in your journal.

### What is cross-validation?

- 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: training-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 `test` sets using the following two steps.**
- **First, fill in the blanks below to randomly select which rows of `arm_span` will go into the `training` set.**

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

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

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

- **Explain these lines of code and describe the `training` and `test` datasets.**

### Aside: `set.seed()`

- When we split data, we're randomly separating our observations into `training` and `test` 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 test, always use `set.seed` first.*

### Aside: training-test ratio

- When splitting data into `training` and `test` 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 datasets grow larger, we can use a larger proportion of the data to `test` with.

### Step 2: training 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_training`.**
- 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_training, 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 test 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

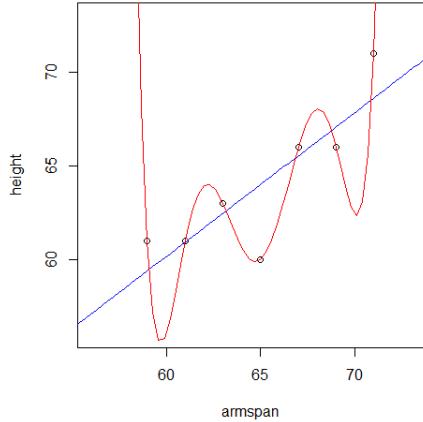
- Another way to describe the three steps is
  - Step 1: Split the data into `training` and `test` sets.
  - Step 2: Choose a slope and intercept that minimize training MSE.
  - Step 3: Using the same slope and intercept from step 2, make predictions on the `test` set, and use those predictions to compute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

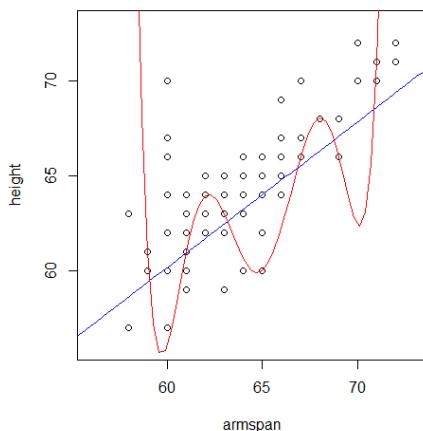
- The following example motivates cross-validation by illustrating the dangers of overfitting.
- We randomly select 7 points from the `arm_span` dataset and fit two models: a linear model, and a *polynomial model*.
  - You will learn how to fit a polynomial model in the next lab.
- Below is a plot of these 7 `training` points, and two curves 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 11: 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 of trends and interpret the slope and the intercept of a linear model in the context of the data.

### **Materials:**

1. *What's the Trend?* handout (LMR\_4.12\_What's the Trend)
2. *Predicting Values* handout (LMR\_4.13\_Predicting Values)

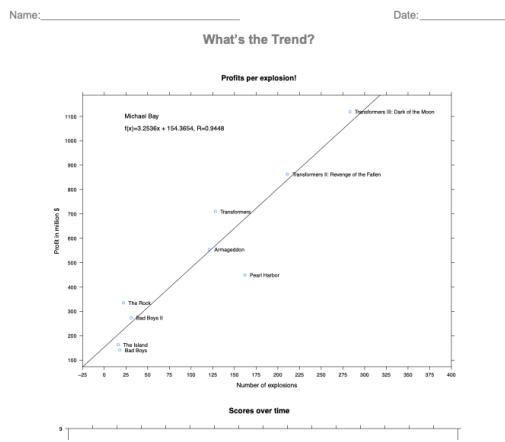
### **Vocabulary:**

trend, positive association, negative association, no association, linear, model

**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 errors can be made smaller.

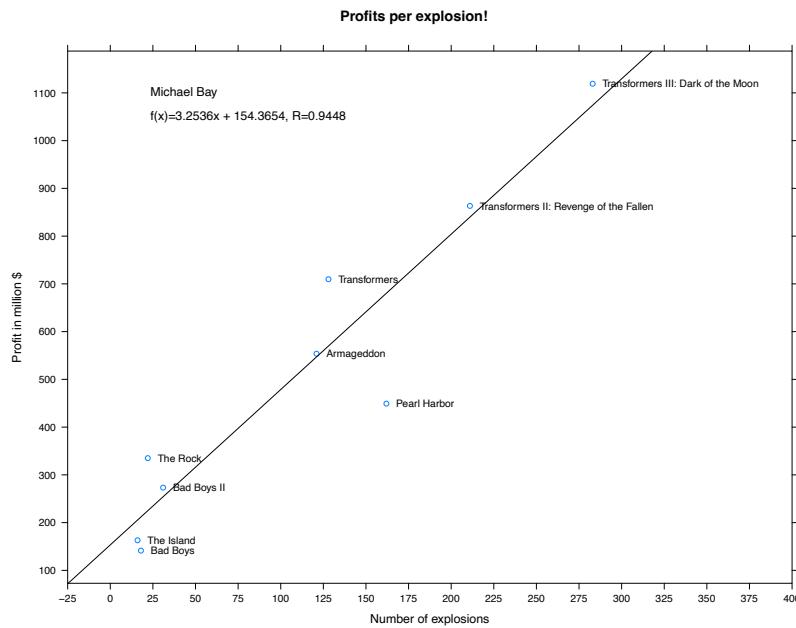
### **Lesson:**

1. Distribute the *What's the Trend?* handout (LMR\_4.12). 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 up to *The Last Airbender* in 2010 and their Internet Movie Database (IMBD) scores.

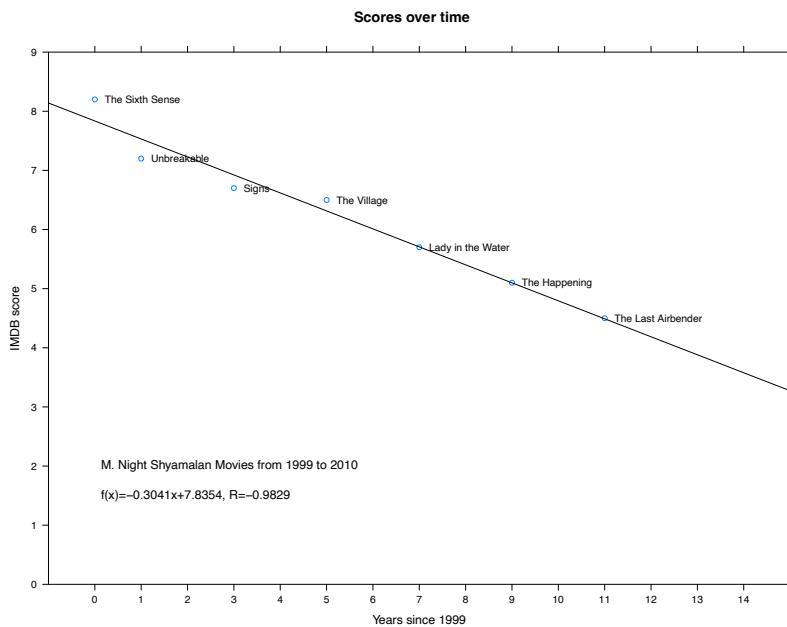


LMR\_4.12

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



- What kind of plot is this? **Answer: Scatterplot.**
- What do the numbers on the x-axis represent? What do the numbers on the y-axis represent? **Answer: The x-axis shows number of explosions and y-axis shows profit in millions of dollars.**
- 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.**



- What kind of plot is this? **Answer: Scatterplot.**
  - What do the numbers on the x-axis represent? What do the numbers on the y-axis represent? **Answer: The x-axis shows the number of years since 1999 and the y-axis shows the movie's IMDB score.**
  - 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.**
- 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:
  - a. 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.*
  - b. What does it mean to have an increasing trend and a positive association? *Answer: In Profits per Explosion, it means that as the number of explosions increase, the movie profits also increase.*
  - c. What does it mean to have a decreasing trend and a negative association? *Answer: 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. *Answer: 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 both plots follow a **linear** trend. The line then represents a **model** for the relationship between the two variables. The equations shown in the plots above represent the lines through the points. They provide a description of the data and the relationship between the variables.
9. Ask student teams to refer back to the *What's the Trend?* handout (LMR\_4.12). They should discuss the following questions and record their responses on the *Predicting Values* handout (LMR\_4.13):
  - a. What do you notice about where the points are and where the line is? *Answer: 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? *Answer: 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 the line? What does the slope mean in relation to the number of explosions? *Answer: 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 of 1 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? *Answer: 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. *Answer: 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? *Answer: 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. *Answer: 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.*

12. If time permits, have students answer the following questions about the Scores Over Time scatterplot in LMR\_4.12:
- What do you notice about where the points are and where the line is? *Answer: Some points are near the line, others are further away, and three points are exactly on the line.*
  - What do the x- and y-values represent in this equation? *Answer: The x-values represent the number of years since 1999 and the y-values represent the IMDB score.*
  - According to the equation, what is the slope of this line? What does the slope mean? *Answer: The slope is -0.3041. It is the rate of change between the number of years since 1999 and the IMDB score. It means that for every year since 1999 increase of 1 the IMDB score decreases by 0.3041.*
  - When the x-value is zero, what is the y-value? How do you know? What does this mean? *Answer: The IMDB score, the y-value, is 7.8354. Students may use the equation to show that they substituted zero for x, so the y-intercept is the IMDB score. It means that if M. Night Shyamalan were to make a movie in 1999, this would be his projected IMDB score, regardless of the movie type or other factors.*
  - What would the predicted value of the score be if M. Night Shyamalan released a movie in 2015? How do you know? *Answer: By substituting 16, because 2015 – 1999 = 16, in the value of x in the equation, the predicted IMDB score will be 2.9698, or by finding the point on the line or both.*

**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 finish answering the questions above about the Scores Over Time scatterplot in LMR\_4.12 referenced above.

## Lesson 12: 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. *Strength of Association* handout (LMR\_4.14\_Strength of Association)
2. *Correlation Coefficient* handout (LMR\_4.15\_Correlation Coefficient)

**Note:** Advance preparation required. This handout is the resource for the plot cutouts. DO NOT distribute as-is to students.

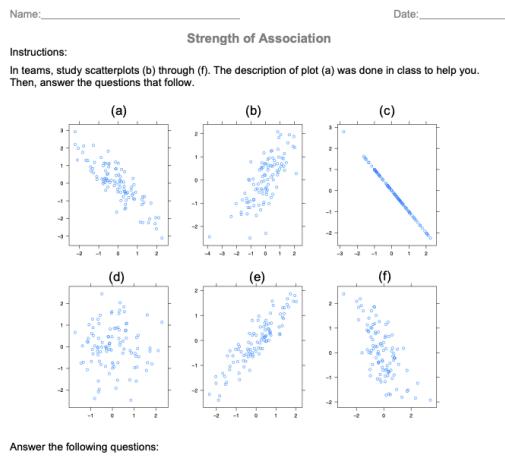
### **Vocabulary:**

correlation coefficient, strength of association

**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. Distribute the *Strength of Association* handout (LMR\_4.14). 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.



LMR\_4.14

2. 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.
3. 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.
4. Guide students to understand that a strong association has points closer to each other and a weak association has points more scattered.
5. 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\_4.15 (Part 1). Make envelopes according to the number of teams in the class. This process will be repeated for LMR\_4.15 (Part 2).

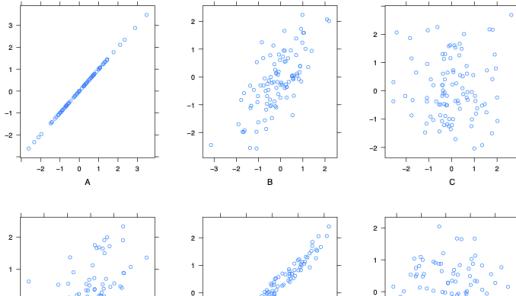
6. Distribute the envelopes to the teams. Students will examine the strength of association in each plot. Their task is to assign a correlation coefficient that corresponds to each plot and to explain why they assigned that correlation coefficient to that particular plot. The only piece of information 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 12 for details.

Plots for Part 1:



LMR\_4.15 (Part 1)

7. 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.
8. Using the *Voting Cards* strategy (see Instructional Strategies), the rest of the teams will show whether they A – approve , B – disapprove, or C – 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:  $r = 0.00$**

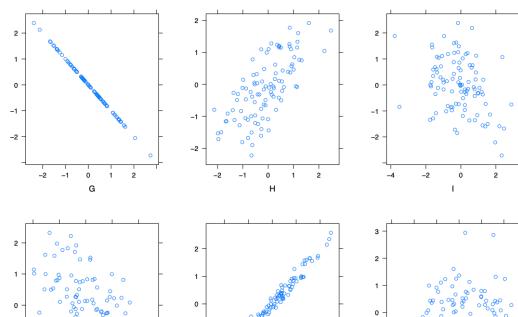
9. The last set of plots showed positive associations. Now students will assign the correlation coefficients for plots G-L for LMR\_4.15 (Part 2).

10. 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

Plots for Part 2:



LMR\_4.15 (Part 2)

-  11. 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.
12. Using the *Voting Cards* strategy, the rest of the teams will show whether they A – approve, B – disapprove, or C – 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:  $r = 0.00$*

-  13. 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 the journal entry for homework if not completed in class.

### **LAB 4D: Interpreting Correlations**

Complete Lab 4D prior to Lesson 13.

## Lab 4D - Interpreting correlations

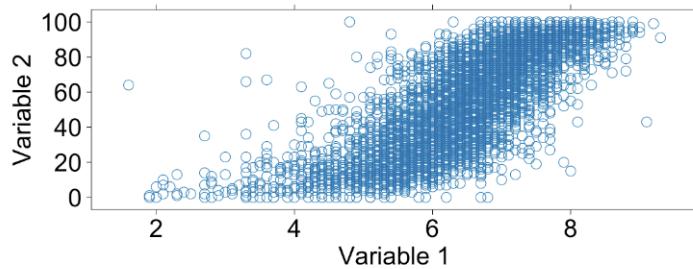
Directions: Follow along with the slides, completing the questions in **blue** on your computer, and answering the questions in **red** in your journal.

### 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 question:  
*Which variables are better predictors of a movie's `critics_rating` when the predictions are made using a line of best fit?*

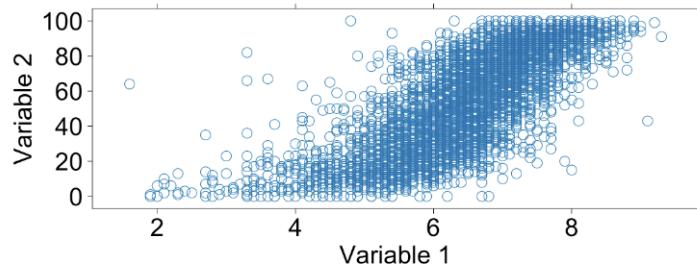
### Correlation coefficients

- 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.



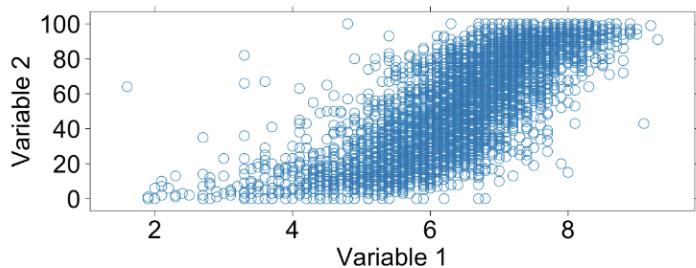
- **Are these variables linearly related? Why or why not?**

### 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 scatterplot 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 `xypplot()` 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?

## **Lesson 13: Improving Your Model**

### **Objective:**

Students will learn to describe associations that are not linear.

### **Materials:**

1. *Describe the Association* handout (LMR\_4.16\_Describe the Association)

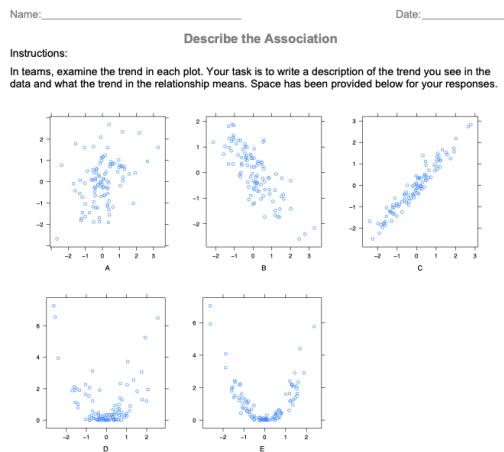
### **Vocabulary:**

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 the *Describe the Association* handout (LMR\_4.16). 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.



LMR\_4.16



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 to 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\_4.16):

- a. *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.*

- b. *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.*
  - c. *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.*
  - d. *Plot D: The trend is non-linear. There seems to be a weak association because there is scatter in the data. We 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.*
  - e. *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:
- a. What happens if we were to fit a linear model to these non-linear trends? Would it still make good predictions? *Answer: Fitting a linear model to a non-linear trend would not properly describe the trend of the data. Therefore no, it would not make good predictions.*
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:
- a. 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 shown in plots D and E can be described as **polynomial trends**. Other polynomial trends include cubic trends, quartic trends, etc. You may show students several choices of equations (linear, quadratic, cubic) along with their graphs and ask them which might be a good candidate to represent them.
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 4E: Some Models Have Curves**

Complete Lab 4E prior to Lesson 14.

## Lab 4E - Some models have curves

Directions: Follow along with the slides, completing the questions in **blue** on your computer, and answering the questions in **red** 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.
- 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 test that includes the remaining 25%.**
  - **Remember to use `set.seed`.**

### Problems with lines

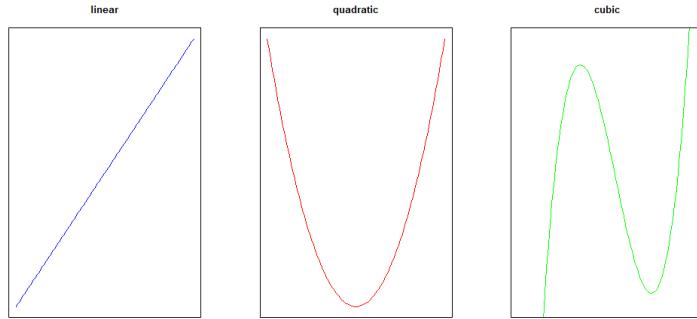
- Before learning how to fit curves, let's first fit a linear model for reference.
- **Train a linear model predicting audience\_rating based on critics\_rating for the training data. Assign this model to `movie_linear`.**
- **Fill in the blanks below to create a scatterplot with audience\_rating on the y-axis and critics\_rating on the x-axis using your test 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 test 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 to 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 test 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 test MSE to explain 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 test data?
- Use the difference in testing MSE to verify which model is the best at predicting the test data.

# Piecing it Together

Instructional Days: 4

## 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 than 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.
- 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 read informative texts to evaluate claims based on data.
2. Students will engage in partner and whole group discussions about how adding variables to a model will help or hinder its predictions.
3. Students will construct their own linear model using multiple variables to compare and contrast which model makes the best predictions.

## Data File or Data Collection Method

### Data File:

1. Movies: `data(movie)`
2. Cereal brands: `data(cereal)`

### Data Collection:

Students will collect data for their Team Participatory Sensing campaign.

## Legend for Activity Icons



Video clip



Discussion



Articles/Reading



Assessments



Class Scribes

## Lesson 14: More Variables to Make Better Predictions

### **Objective:**

Students will see that different variables can be used to make predictions about the same outcome (response variable) and consider whether combining these variables could improve prediction accuracy.

### **Materials:**

1. *Advertising Plots Part 1* handout (LMR\_4.17\_Advertising Plots 1)
2. *Advertising Plots Part 2* handout (LMR\_4.18\_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:**

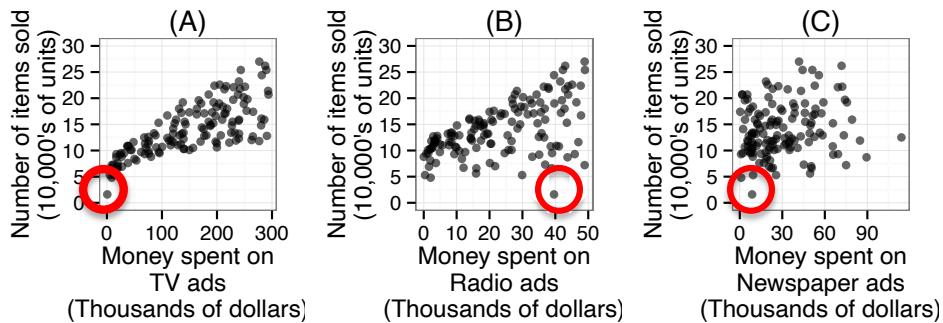
market

**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? **Answer: The success of a spinoff show.**
  - b. How many variables are used? **Answer: Four – the original show name, the number of episodes of the original show, the name of the spinoff show, and number of episodes of the spinoff show.**
  - c. What other variables might affect a spinoff? **Possible answers are budget or actors.**
  - d. The dotted line in the plot is not a regression line. How would you draw a regression line to make predictions? **Answers will vary but we would want to try to “fit” a line to the plotted data.**
  - e. What other information would you like to know to predict a spinoff’s success? **Answers will vary but may be similar to (c) above.**
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\_4.17) and inform the students about the data using the details below:



LMR\_4.17  
(Plots are presented separately in the LMR)

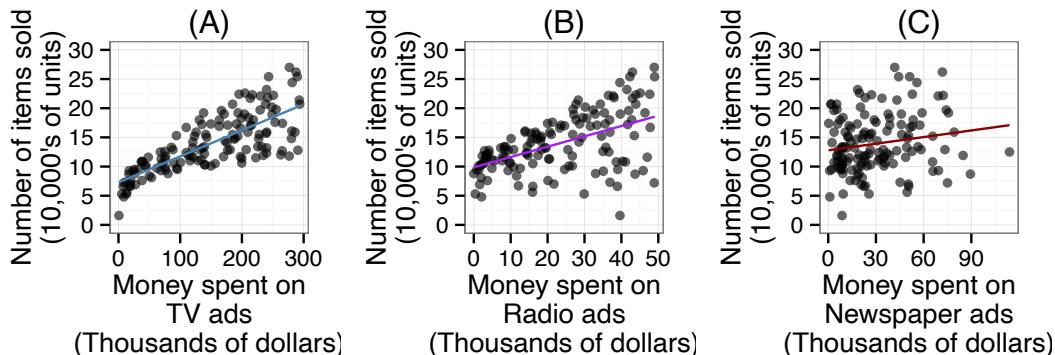
- a. These 3 plots show the number of items sold by a retailer (in 200 different markets) and the amount of money the company spent on *TV, Radio* and *Newspaper* advertisements.
  - b. The data has 200 observations, one for each different market. A **market** is simply a location where an item is sold. 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:
- a. How many items did this market sell? **Answer: About 20,000 items. The actual number of items sold was 1.6 (in 10,000's of units) which is 16,000 items.**
  - b. How much money did this retailer spend on TV ads in this market? **Answer: 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:
- a. How much money did the retailer spend on Radio ads in the same market? **Answer: 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:
- a. How much money did the retailer spend on Newspaper ads in the same market? **Answer: 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. **Answers will vary but we would expect the number of items sold to increase with increased advertisements.**

- b. Which type of advertisement is the most strongly correlated with the number of units sold? How can you tell? *Answer: Plot A, TV advertisements, appears to be the most strongly correlated with the number of units sold. We can tell because the points are more closely packed/together than in plots B or C.*
12. Distribute the *Advertising Plots Part 2* handout (LMR\_4.18) which contains plots A-C, but now include the line of best fit.



LMR\_4.18  
(Plots are presented separately in the LMR)

13. Ask students to recall from Lesson 7 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 7 – 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:

- 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.

- 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.
- 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 during class.

## Lesson 15: 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.

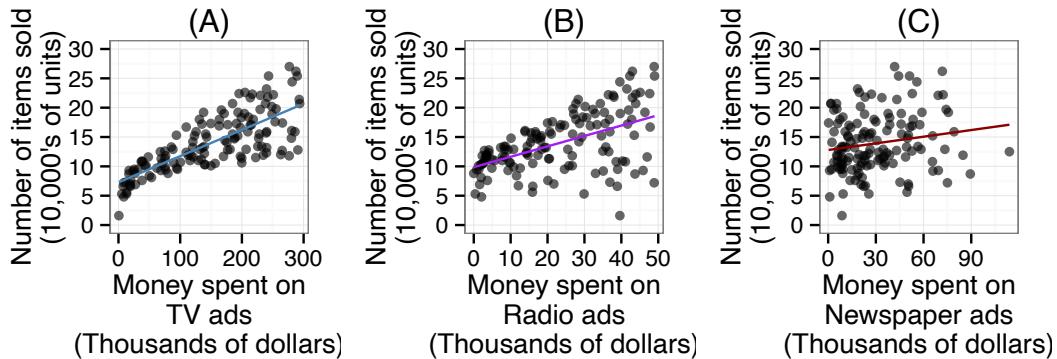
### Materials:

1. Advertising Plots Part 2 handout (LMR\_4.18) from Lesson 14

**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\_4.18  
(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 using the Agree/Disagree strategy.
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:

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

**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. *Student calculations 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? *Answer: 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\_4.18 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 (MAE)
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. Think back to our statements from the last lesson and the beginning of this lesson:

- 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.



Engage students in a discussion to check understanding of using multiple variables in a model.

- a. Has their statement selection changed or stayed the same? Why?

- b. What evidence do they have to support their statement selection? *Answers will vary because students can make an argument for either statement, both statements, or neither statements being true. There is evidence in step 7 that supports that more variables will lead to worse predictions (like TV-Radio-Newspaper) but there is also evidence that more variables lead to better predictions (like TV-Radio).*
9. 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 4F: This Model is Big Enough for All of Us**

Complete Lab 4F prior to Practicum.

## **Lab 4F – This model is big enough for all of us!**

Directions: Follow along with the slides, completing the questions in **blue** on your computer, and answering the questions in **red** in your journal.

### **Building better models**

- So far, in the labs, we've learned how to make predictions using the *line of best fit*, also known as *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).**
  - A set named **training** that includes 75% of the data.
  - A set named **test** that includes the remaining 25%.
  - Remember to use `set.seed`.
- **Create a linear model, using the training data, that predicts gross using runtime.**
  - Compute the MSE of the model by making predictions for the test 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 test 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 test 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 datasets: one training named cereal and one test named cereal\_test. Load both datasets. Write down the code you used.
2. Explore the training data. Which variable do you think is the best predictor of sugar? Choose at least 3 variables, make a plot for each one, and fit a linear regression line through each of 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 squared error of your model using your test data.
  - d. Now make a set of predictions with your test data. Calculate the mean squared error for the test 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 sugar, and describe, based on the test data, how far off your prediction might actually be.

# Decisions, Decisions

Instructional Days: 3

## Enduring Understandings

Decision trees are used to classify observations into similar groupings based on known characteristics. 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 be presented with the question about whether they would rather trust a doctor or a data scientist to diagnose them if they were having a chest pains. This will set the context for decision trees and how they are used to make predictions.

## 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 engage in partner and whole group discussions to express their understanding of classification trees.
2. Students will explain orally and in writing how to determine the accuracy of their non-linear model.
3. Students will make connections between decision trees and linear models in writing.

## Data File or Data Collection Method

Dataset:

1. USMNT/NFL dataset

Data File:

1. Titanic: data(titanic)

Data Collection:

Students will collect data for their Team Participatory Sensing campaign.

## Legend for Activity Icons



Video clip



Discussion



Articles/Reading



Assessments



Class Scribes

## Lesson 16: Football or Futbol?

### **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. *Decision Tree for Heart Attack Risk* graphic (LMR\_4.19\_CART Heart Attacks)
2. *CART Activity Player Stats* (LMR\_4.20\_CART Player Stats)
3. *CART Activity Round 1 Questions* (LMR\_4.21\_CART Round 1)
4. *CART Activity Round 2 Questions* (LMR\_4.22\_CART Round 2)

**Note:** Advanced preparation required for LMR\_4.20, 4.21, and 4.22 (see step 8 below).

### **Vocabulary:**

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

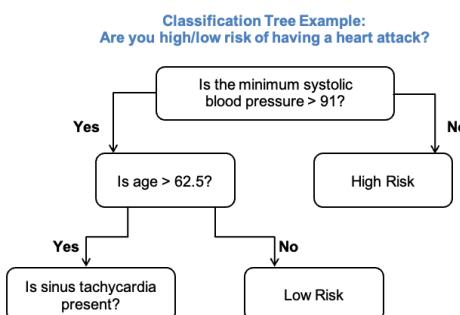
**Essential Concepts:** Some trends are not linear, so the approaches we've done so far won't be helpful. We need to model such trends differently. Decision trees are a non-linear tool for classifying observations into groups when the trend is non-linear.

### **Lesson:**

1. Ask students the following question:  
**If you were having chest pains, who would you trust more to diagnose you – a data scientist or a doctor?**
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 go 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. It was made to diagnose whether people were or were not having a heart attack. The accuracy of the decision tree compared to the accuracy of actual doctor diagnoses are shown below:
  - a. Correct diagnoses using the decision tree were above 95%.
  - b. Correct diagnoses based on individual doctors' expertise was anywhere between 75-90%.
4. Display the graphic from the *Decision Tree for Heart Attack Risk* handout (LMR\_4.19) and explain that this is one example of what the decision tree that Goldman developed might have looked like.

**Note:** This is NOT the actual tree Goldman developed.

Name: \_\_\_\_\_ Date: \_\_\_\_\_  
Decision Tree for Heart Attack Risk  
Directions to teacher:  
Display the following graphic during Step 4 of Lesson 16 in Unit 4.



LMR\_4.19



5. Using a *Pair-Share*, ask students to discuss the following questions using the graphic above.  
**Note:** Answers will vary. These questions are meant to gauge student thinking before defining decision trees in the following steps of the lesson.
    - a. What are decision trees?
    - b. How do they work at classifying data into groups?
  6. 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.
  7. Let students know that one method of modeling data in a non-linear way is with **decision trees**, like the one we saw with the heart attack classification. 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.
  8. 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.
  9. 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.
  10. 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).
  11. 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 an umbrella term to refer to the following types of decision trees:
    - a. Classification Trees: the leaves predict the values of a categorical variable
    - b. Regression Trees: the leaves predict a numerical value
  12. CART Activity: to get a sense of how classification 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 LMR\_4.20, 4.21, and 4.22 listed above (and previewed 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.

<b>Player 1</b> Name: Matt Besler Team: Kansas City Height (inches): 72 Weight (pounds): 170 Age: 28 League: USMNT	<b>Player 2</b> Name: Cam Newton Team: Carolina Height (inches): 77 Weight (pounds): 245 Age: 26 League: NFL	<b>Player 3</b> Name: Clint Dempsey Team: Seattle Height (inches): 73 Weight (pounds): 170 Age: 32 League: USMNT
<b>Player 4</b> Name: Steve Birnbaum Team: Washington, DC Height (inches): 74 Weight (pounds): 181 Age: 28 League: USMNT	<b>Player 5</b> Name: Jermaine Jones Team: New England Height (inches): 72 Weight (pounds): 179 Age: 34 League: USMNT	<b>Player 6</b> Name: Matt Cassel Team: Dallas Height (inches): 76 Weight (pounds): 230 Age: 33 League: NFL
<b>Player 7</b> Name: Russell Wilson Team: Seattle Height (inches): 71 Weight (pounds): 206	<b>Player 8</b> Name: Matt Hedges Team: Dallas Height (inches): 76 Weight (pounds): 190	<b>Player 9</b> Name: Robert Griffin III Team: Washington, DC Height (inches): 74 Weight (pounds): 223

LMR\_4.20

Name: \_\_\_\_\_ Date: \_\_\_\_\_

CART Activity Round 1

Directions for teacher:  
Create nodes by cutting out each question below.

<b>1 – Root Node</b> Is your team located in the United States? YES: Go to your right. NO: Go to your left.
<b>2 – Internal Node</b> Are you 33 years old or older? YES: Go to your right. NO: Go to your left.
<b>3 – Leaf Node</b> You play for the US Men's National Soccer Team (USMNT).
<b>4 – Leaf Node</b> You play for the National Football League (NFL).

LMR\_4.21

Name: \_\_\_\_\_ Date: \_\_\_\_\_

CART Activity Round 2

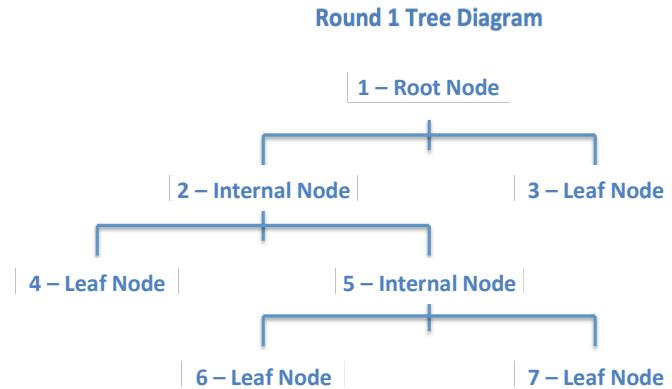
Directions for teacher:  
Create nodes by cutting out each question below.

<b>1 – Root Node</b> Are you 74 inches tall or taller? YES: Go to your right. NO: Go to our left.
<b>2 – Leaf Node</b> You play for the National Football League (NFL).
<b>3 – Internal Node</b> Do you weigh more than 200 pounds? YES: Go to your right NO: Go to your left.
<b>4 – Leaf Node</b> You play for the National Football League (NFL).

LMR\_4.22

13. Ask for 15 volunteers and hand each of them a data card from the *CART Activity Player Stats* handout (LMR\_4.20). 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 (inches)
  - e. weight (pounds)
  - f. league
14. The “players” will only be allowed to say “yes” or “no” in this activity. No other talking is permitted.
15. Now, ask for 7 additional volunteers to be the **nodes** on the classification tree. There are three types of nodes in a decision tree:
  - a. Root node – the initial question/characteristic that splits the population/dataset
  - b. Internal nodes – a question/characteristic that splits the data
  - c. Leaf nodes – an “end” where no more splitting is possible
16. Distribute one question/classification from the *CART Activity Round 1 Questions* handout (LMR\_4.21) to each node.

17. Arrange the 7 nodes in the room as depicted by the graphic below:

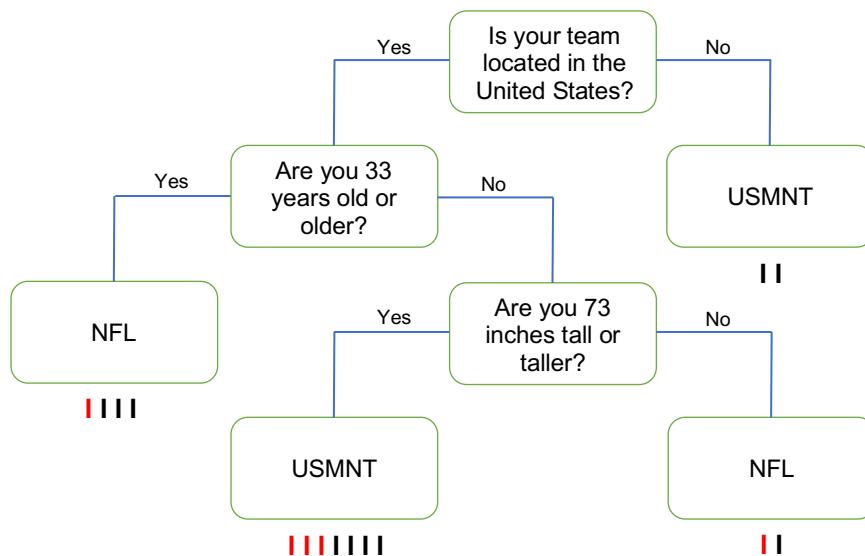


18. Now, each “player”, one at a time, will approach 1 – Root Node, who will ask the “player” the question listed on his/her card. Depending on the player’s answer, 1 – Root Node will direct the “player” to the next node.

19. 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).

20. Allow all the “players” to go through the nodes until each one is classified as either a soccer or football player.

21. After each player has been classified, record the classifications in the table below. The tally marks below correspond with classified incorrectly (red) and classified correctly (black), if all the player stats cards are used.

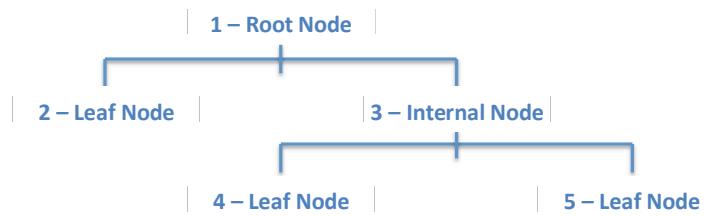


**Note:** This table will be referenced again later. A filled-out version is available in the next lesson.

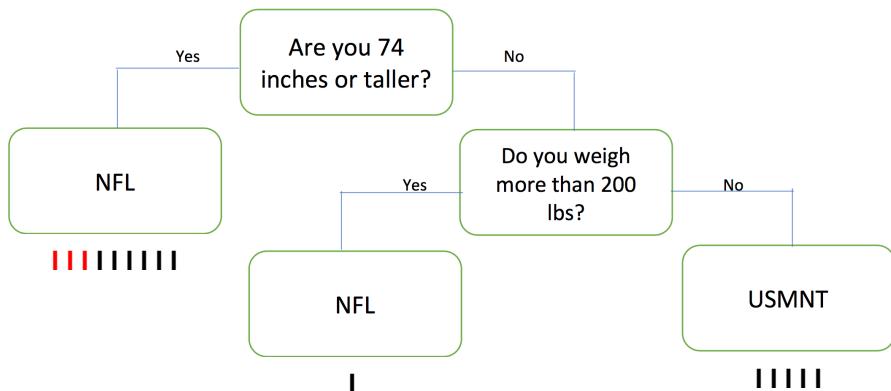
Round 1	Classified Correctly	Classified Incorrectly	Total Classified
3 – Leaf Node			
4 – Leaf Node			
6 – Leaf Node			
7 – Leaf Node			

22. Ask students, "How successful were we in classifying players correctly?" *Answers will vary but if all the activity player stats cards were used then 10/15, or 67%, were classified correctly. Students might measure success by saying that we only misclassified players 5/15, or 33%, of the time – this is called the misclassification rate (MCR) and will be defined in the next lesson.*
23. After proceeding through "Round 1", ask an additional 5 students to come up as more nodes, distribute the cards from the *CART Activity Round 2 Questions* strips (LMR\_4.22), and arrange the students like the diagram below:

**Round 2 Tree Diagram**



24. Have each "player" go through this new set of nodes until they are re-classified by these new rules.
25. After each player has been classified, record the classifications in the table below. The tally marks below correspond with classified incorrectly (red) and classified correctly (black), if all the player stats cards are used.



**Note:** This table will be referenced again later. A filled-out version is available in the next lesson.

Round 2	Classified Correctly	Classified Incorrectly	Total Classified
2 – Leaf Node			
4 – Leaf Node			
5 – Leaf Node			

26. Ask students, "How successful were we in classifying players correctly in this second round?" *Answers will vary but if all the activity player stats cards were used then 12/15, or 80%, were classified correctly. Students might measure success by saying that we only misclassified players 3/15, or 20%, of the time – this is called the misclassification rate (MCR) and will be defined in the next lesson.*



27. Once the activity has been completed, ask students the following questions:
- How do decision trees classify objects/people as being a member of a group? *Answer: By asking a series of questions, one at a time, and sending the participant down a particular path until he/she is classified.*
  - 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. If all activity player stats cards were used then we did better in Round 2, with 80% success, than Round 1, with 67% success, in correctly guessing which sport the “players” participate in.*
  - How can we figure out what questions to ask and in what order to minimize the number of incorrect classifications (also known as *misclassifications*)? *Answers will vary. This one might not be obvious but the point is for the students to wrestle with how they might think it can be done.*

**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 find a decision tree online that aligns with their interests and:

- identify the nodes – root, internal, and leaf
- describe the population that it applies to
- describe what the decision tree is predicting, i.e., what are the classifications at the leaves?

## Lesson 17: Grow Your Own Decision Tree

### **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. *Make Your Own Decision Tree* handout (LMR\_4.23\_Your Own Decision Tree)

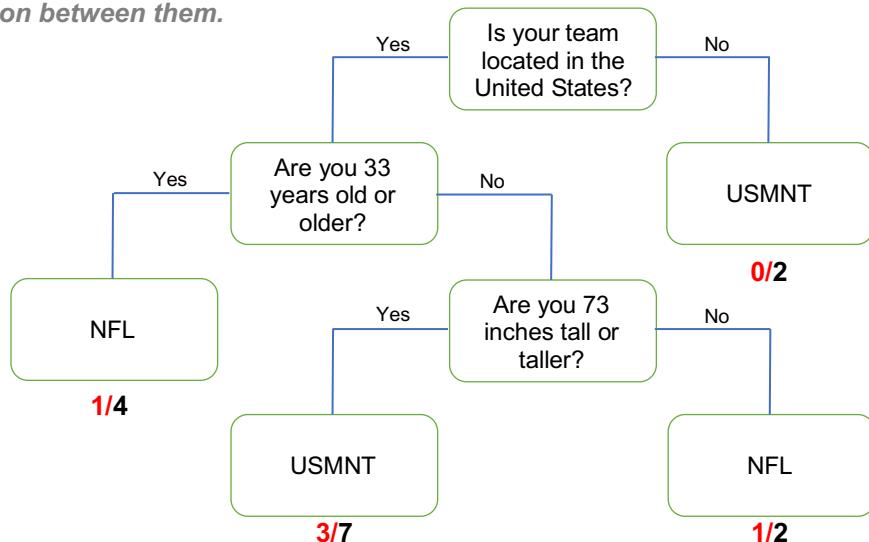
### **Vocabulary:**

misclassification rate (MCR), training data, test data

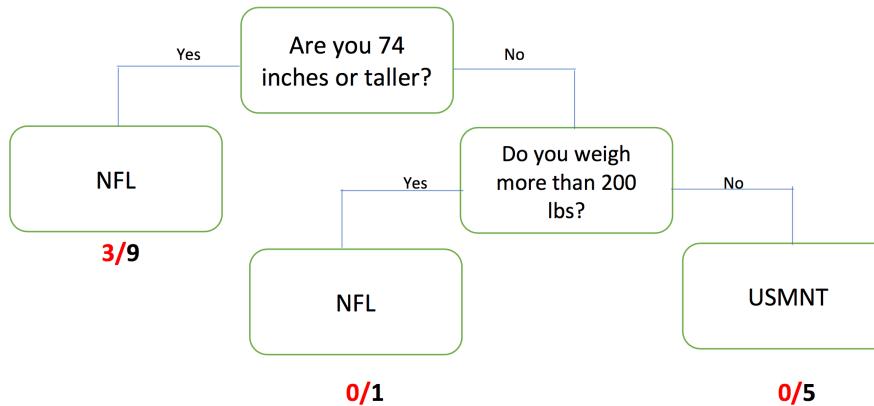
**Essential Concepts:** We can determine the usefulness of decision trees by comparing the number of misclassifications in each.

### **Lesson:**

1. Begin the lesson by asking the following question: *How did we assess whether a linear model made good predictions for a set of data? Answers will vary but so far in this unit, we have used Mean Squared Error (MSE) and Mean Absolute Error (MAE).*
2. Tell students that much like linear models, classification trees also have a method for determining how well they make predictions for a set of data.
3. Classification trees use a **misclassification rate (MCR)**, which is the proportion of observations who were predicted to be in one category but were actually in another.
4. Refer back to your tallied decision trees and correct/incorrect classification tables from the previous lesson. What were the overall proportion of incorrect classifications for Round 1? What about for Round 2? *Answers will vary. If all activity player stats cards were used then the MCR for Round 1 is 5/15, or 0.33, and for Round 2 is 3/15, or 0.20. You can reference the decision trees and tables from the previous lesson so that students can see the connection between them.*



Round 1	Classified Correctly	Classified Incorrectly	Total Classified
3 – Leaf Node	2	0	2
4 – Leaf Node	3	1	4
6 – Leaf Node	4	3	7
7 – Leaf Node	1	1	2
MCR		0 + 1 + 3 + 1	2 + 4 + 7 + 2 = 5/15 (or 0.33)



Round 2	Classified Correctly	Classified Incorrectly	Total Classified
<b>2 – Leaf Node</b>	6	<b>3</b>	<b>9</b>
<b>4 – Leaf Node</b>	1	<b>0</b>	<b>1</b>
<b>5 – Leaf Node</b>	5	<b>0</b>	<b>5</b>
<b>MCR</b>		<b>3 + 0 + 0</b>	<b>9 + 1 + 5</b> = <b>3/15 (or 0.20)</b>

- These proportions of incorrect classifications are our misclassification rates.
- Today we'll be creating our own classification tree and assess its prediction accuracy.
- 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

- Distribute the *Make Your Own Decision Tree* handout (LMR\_4.23) 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).

**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 **nodes/boxes**.
3. Continue to draw additional nodes/boxes to either write more questions or to classify a player.
4. Once you have completed your tree, sort the 6 players from the **test 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
Baltimore	Glen Newton	77	245	26	NFL
Chicago	Steve Johnson	75	217	26	USMNT
Dallas	Matt Catts	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	Michael Bennett	72	270	28	USMNT
New England	Tom Brady	76	225	39	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	29	NFL
Washington, DC	Steve Birnbaum	74	181	28	USMNT

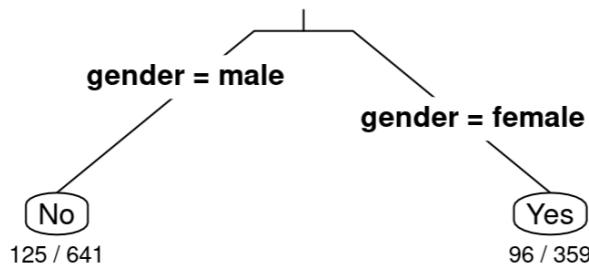


LMR\_4.23

9. Once the students have finished creating their classification trees, ask the following questions:
  - a. Will you be able to classify other players from a new dataset correctly using this particular classification tree?
  - b. Do you think this classification tree is too specific to the training data?
10. Inform the students that they should now use the **test 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 of the *Make Your Own Decision Tree* handout (LMR\_4.23).
11. Let the students compare their classification trees and league assignments with one another. Hopefully, there will be a bit of variety in terms of the trees and the classifications.
12. 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
Baltimore, MD	Justin Tucker	73	182	34	NFL
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

13. By a show of hands, ask:
  - a. How many students misclassified all of the players in the test 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 classification trees with the rest of the class.
14. 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 in the next lab.
- Note:** Included below is a front-load of the calculation/interpretation of MCR in RStudio.
15. In the next lab, 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 preview and break down one of the firsts outputs you'll see in the lab.



16. Project the image above and explain to the students that the lab uses the titanic data, so we are looking at a total of 1000 observations (in this case, people). Much like the ratios we saw for our Round 1 and Round 2 classification trees, RStudio gives us ratios for each of the leaf nodes where a classification was made. The denominator tells us how many observations ended up in that node and the numerator tells us the number of misclassifications. Also notice that the questions are not shown in the classification tree, but the characteristics are visible instead. Ask students:
- What does the output 125/641 represent? *Answer: The 641 tells us that six-hundred-forty-one people were classified as not surviving based solely on the fact that they were male. The 125 represents people who were misclassified (they actually survived), which means that 516 people were classified correctly (they indeed did not survive).*
  - 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.  $(125+96)/(641+359)=221/1000$  or 0.221.*

#### 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 record their responses to the following discussion questions:

- How is a decision tree/CART similar to or different than a linear model? *Answers will vary but a similarity is that they both make predictions, and a difference is that decision trees can make predictions for both numerical and categorical data while linear models only make predictions for numerical data.*
- Why is a decision tree considered a model? *Possible answer: We consider a decision tree a model because it still represents relationships between variables.*
- Describe the role that training data and test data play in creating a classification tree. *Possible answer: We use training data to build a classification tree and then use test data to see how well our tree makes predictions – we shouldn't use the entirety of our data because then the model will be too specific/overfitted and will not make good predictions when presented with new data.*

## LAB 4G: Growing Trees

Complete Lab 4G prior to Lesson 18.

## Lab 4G - Growing trees

Directions: Follow along with the slides, completing the questions in **blue** on your computer, and answering the questions in **red** 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`?**

## Predictions and Cross-validation

- Just like with *linear models*, we can use cross-validation to measure how well our *classification trees* perform on unseen data.
- First, we need to compute the predictions that our model makes on test data.
  - **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`.**

Note: the argument `type = "class"` tells the `predict` function that we are predicting a categorical variable and not a numerical variable.

```
titanic_test <- mutate(___, prediction = predict(___, newdata = ___, type = "class"))
```

## Measuring model performance

- 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.
- **Run the following command to see a side-by-side comparison of the actual outcome and the predicted outcome:**  

```
View(select(titanic_test, survived, prediction))
```
- **Where does the first misclassification occur?**

## Misclassification rate

- In order to tally up the total number of misclassifications, we need to create a function that compares the actual outcome with the predicted outcome. The **not equal to** operator (`!=`) will be useful here.
- **Fill in the blanks to create a function to calculate the MCR.**
- Hint: `sum(____ != ____)` will count the number of times that the left-hand side does not equal the right-hand side.
  - We want to count the number of times that actual does not equal predicted and then divide by the total number of observations.

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

- 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 determine which points in a plot should be grouped as football players and which points should be grouped as swimmers based on clustering of 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 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 write, in their own words, an explanation of k-means clustering.
2. Students will describe the differences between time spent on videogames and time spent on homework, from their own class data.
3. Students will create visualizations and numerical summaries to explain and justify, orally and in writing, a recommendation to better their community.

## Data File or Data Collection Method

Data File:

1. USMNT and NFL: `data(futbol)`
2. Students' *TimeUse* campaign data

Data Collection:

Students will collect data for their Team Participatory Sensing campaign.

## Legend for Activity Icons



Video clip



Discussion



Articles/Reading



Assessments



Class Scribes

## Lesson 18: 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

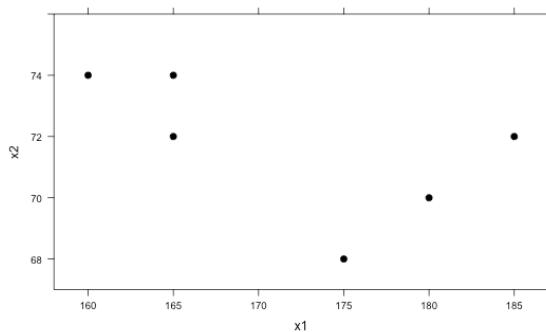
**Essential Concepts:** We can identify groups, or “clusters”, in data based on a few characteristics. For example, it is easy to classify a group of people into football players and swimmers, but what if you only knew each person’s arm span? How well could you classify them into football players and swimmers 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/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:

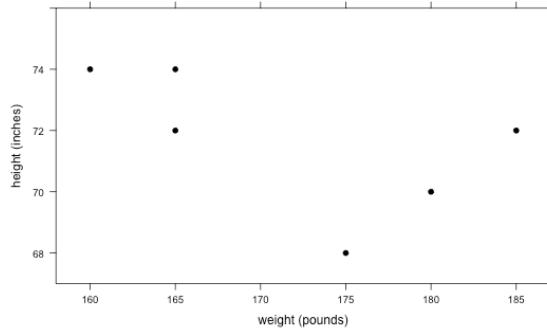
Obs	X <sub>1</sub>	X <sub>2</sub>
1	160	74
2	165	72
3	165	74
4	175	68
5	180	70
6	185	72

5. Plot the X<sub>1</sub> and X<sub>2</sub> points on a scatterplot either on the board or on poster paper (X<sub>1</sub> can be on the horizontal axis and X<sub>2</sub> can be on the vertical axis). The graph should look like the one below:

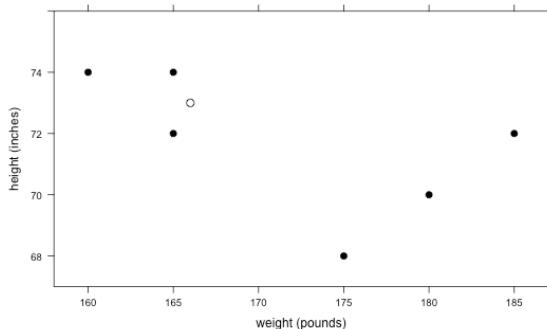


6. Ask students if they think there are any clusters, or groups, that stand out to them. It is likely that they will say there are 2 clusters in the graph: the top left corner 3 points, and the bottom right 3 points.

7. Now pose the following scenario that further describes the data:
- A doctor provides yearly physicals to the men's football and men's swimming teams at a local high school.
  - 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.



- Unfortunately, the doctor never recorded what sport each person played.
8. Using the information about height and weight, ask the students to decide:
- Which group of points most likely represents players from the swimming team? **Answer: The points in the upper left corner are probably swimmers because swimmers are usually tall (and have large arm spans) and thin.**
  - Which group of points most likely represents players from the football team? **Answer: 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. **Answer: 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 0" graph.

Name: \_\_\_\_\_ Date: \_\_\_\_\_

Find the Clusters

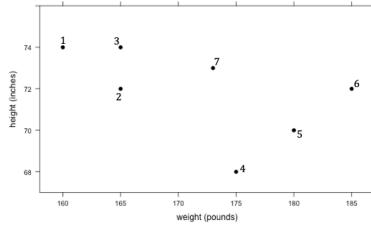
**Round 0 – Initialization**

Pick random points for your initial cluster centers.

Cluster A center: ( \_\_\_\_\_, \_\_\_\_\_ )

Cluster B center: ( \_\_\_\_\_, \_\_\_\_\_ )

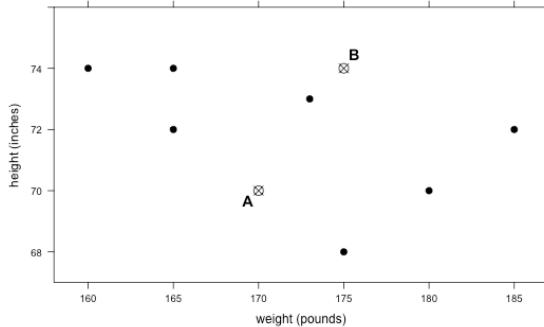
Plot your two points for A and B on the graph below.



LMR\_4.24

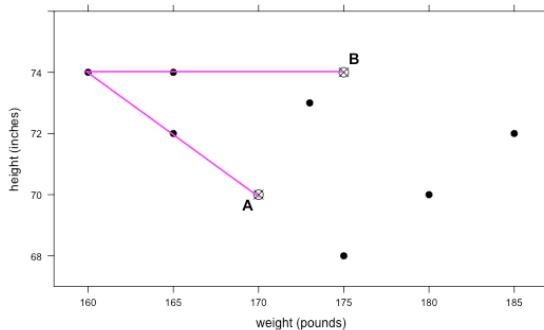
12. Ask students:

- a. On which team do you think this person plays? *Answer: 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 a process 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 clusters change. To begin, we will pick 2 points, A and B, to represent the center of each cluster. We will be referring to this as the initialization step. 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 near the other cluster and label it as “B”.
16. **Initialization:** For now, we will start with the following two points as our initial cluster centers of each group: A: (170, 70) and B: (175, 74). In the “Round 0” plot on the *Find the Clusters* handout, each student should plot and label these two points.

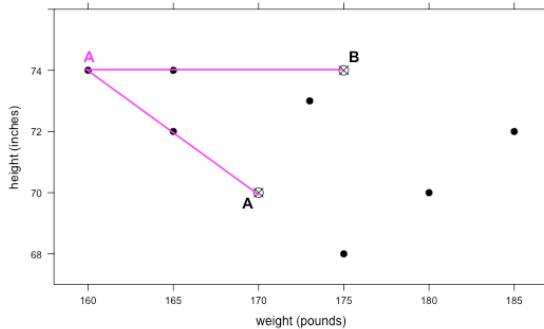


17. **Assignment Step:** Inform the students that they will be determining the distance between the 7 observations and point A and point B. Then, they will decide if the point is closer to cluster center A or cluster center B and label the point with that letter.

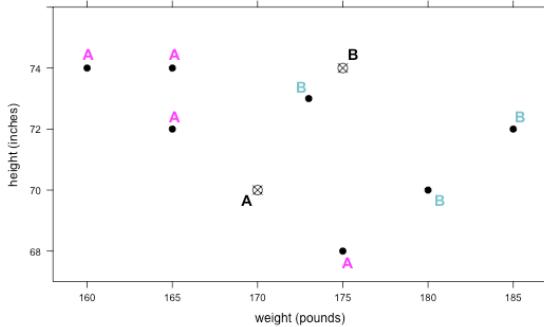
- a. Lines have been drawn from the top left point to the cluster centers in the plot below as a guide. You can draw this on the board as a reference for the students as well.



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



- c. The students may draw similar lines for every point on the graph, or they can simply eyeball it, to make a decision as to which cluster each point belongs in. Even if they guess incorrectly, the algorithm should be able to find the correct groups after some time. The correct classifications for Round 0 are as follows, using our points from step 16:



18. **Update Step:** Once the class has agreed on the Round 0's cluster classifications, they should compute new values for points A and B by using the clustered point means. For point A, they simply need to find the mean x-value for the 4 points and the mean y-value for the 4 points. Repeat this process to find the new point B – these will be our new cluster centers as we move onto Round 1. Calculating means to derive cluster centers, like points A and B, for grouping data points is part of the **k-means** algorithm.

- a. The new points for A and B have been calculated below. The students should be calculating these on their own and recording their new cluster centers on the handout.

$$\text{x-value for A} = (160 + 165 + 165 + 175)/4 = 166.25$$

$$\text{y-value for A} = (74 + 72 + 74 + 68)/4 = 72$$

$$\text{x-value for B} = (173 + 180 + 185)/3 = 179.3$$

$$\text{y-value for B} = (73 + 70 + 72)/3 = 71.67$$

$$\text{new A} = (166.25, 72)$$

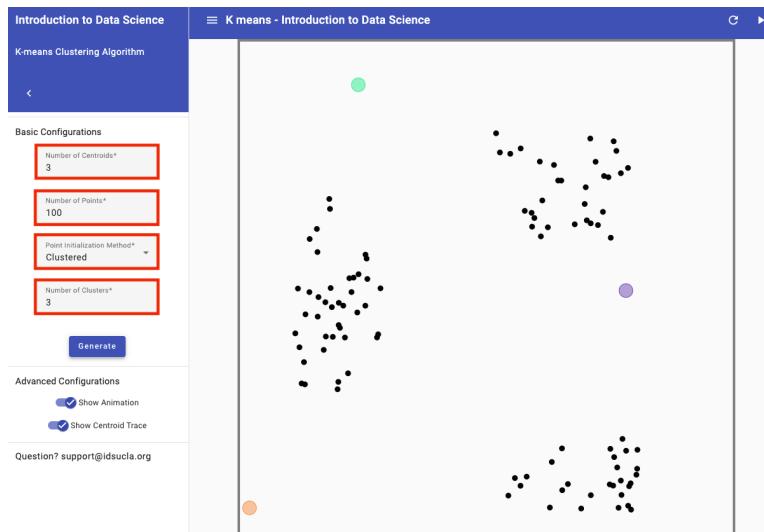
$$\text{new B} = (179.3, 71.67)$$

19. 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.

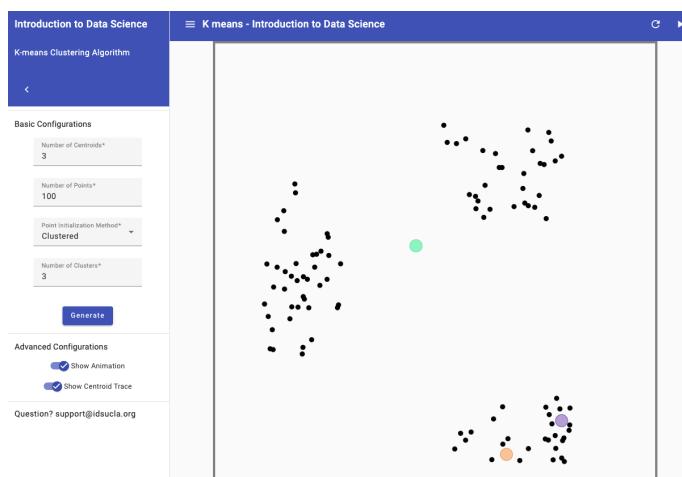
**Note:** If students used the initial cluster centers from step 16, they would finish clustering in Round 2 with points 1, 2, and 3 belonging to cluster A and points 4, 5, 6, and 7 belonging to cluster B

20. Where you choose your initial points matters in determining which points end up in which clusters. Demonstrate this to the class using the K-means Clustering App, located on the Applications page on Portal under Explore (<https://portal.idsucia.org/#curriculum/applications/>).

- In the app, “centroids” is the academic term for cluster centers. For this example, we will use 3 centroids and choose a “Clustered Initialization” with 100 points and 3 Clusters. See image below for how to adjust the settings.



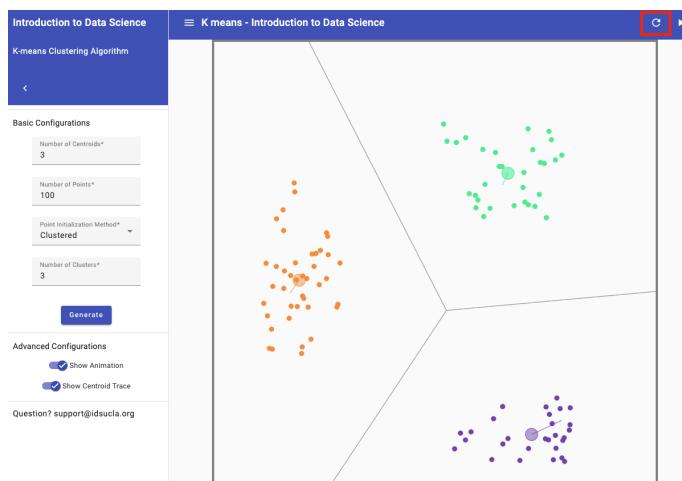
- Move two of the centroids so that they are close to/within one cluster. Move the remaining centroid in between the other two clusters. An example is shown below.



- c. Click “Next Step” until no points change from the previous update. See image below of end of clustering for the example from (b).



- d. While we still have three clusters, we can clearly see that our initial points have influenced the end result of those clusters. You can click “Restart and play again” to move the centroids to the center of the clusters and click “Next step” until no points change from the previous update to illustrate this point (see image below of correctly clustered groups).



**Note:** Clicking "Restart and play again" allows you to move your cluster centers to new positions while still using the same 100 points. You can repeat the same process as above to create different additional groupings.

#### 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, completing the questions in **blue** on your computer, and answering the questions in **red** 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 `filter` 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 19: 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* handout (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:**

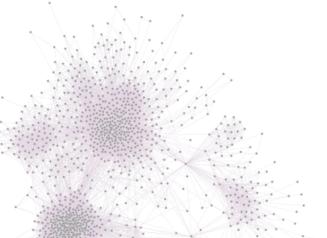
1. Display the *Friend Network Graphic* handout (LMR\_4.25), which shows a WolframAlpha visualization of someone's Facebook friends. Inform the students that this type of model is called a **network**, which is simply a group of people or things that are interconnected in some way.

Name: \_\_\_\_\_ Date: \_\_\_\_\_

**Friend Network Graphic**  
Prepared by WolframAlpha

Instructions for teacher:  
Display the Friend Network graph during Step 1 of Lesson 19 in Unit 4.

Friend network:



LMR\_4.25

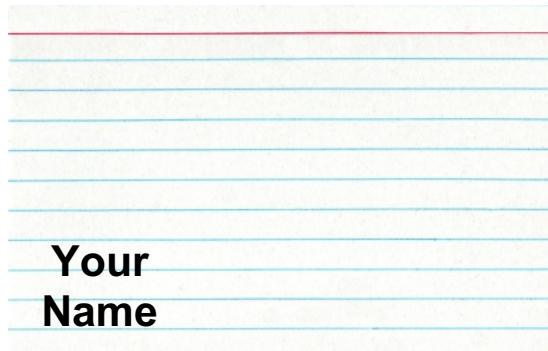
2. Ask the following questions about the graphic:



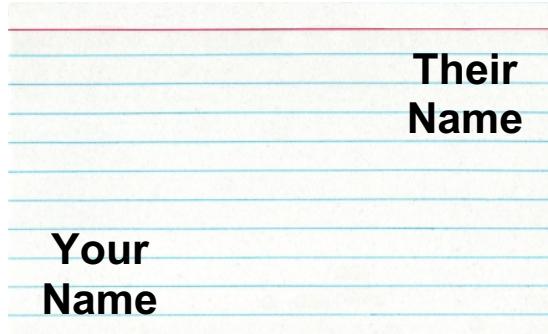
- a. What does each dot represent? **Answer: Each dot represents one person.**
  - b. What does each line represent? **Answer: Each line represents a friendship between two people.**
  - c. How are all the people in this graphic connected to each other? **Answer: They are all friends with the person whose Facebook this is.**
  - d. Why are some areas denser than others? **Answer: 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)? **Answer: 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.

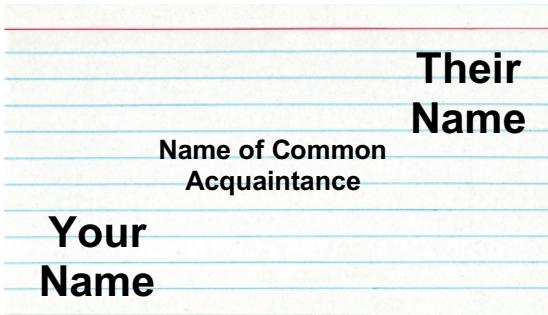
4. Next, inform the students that networks can be as big or as small as we want. We can even determine our own class's social network and create visualizations from it!
5. Network Activity:
  - a. Distribute index cards to students. Each student will need enough cards to make a connection with every 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 3<sup>rd</sup> 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.

**Note:** 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.



**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 RScript file can be opened and viewed in the "source" pane of RStudio after it has been uploaded as a file into your RStudio project. There are 2 places where the code needs to be edited by the teacher:

- Be sure to change the file name when reading in the .csv file in Line 7 of the code.
- 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 Team Participatory Sensing 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 4 Modeling Activity Project and Presentation**

### **Objective:**

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

### **Materials:**

1. Computers
2. *IDS Unit 4 – Project and Presentation (LMR\_U4\_Modeling Activity Project)*

### **End of Unit 4 Project and Presentation: Community Issue**

At the beginning of this unit, you explored data from the Trash Participatory Sensing campaign as well as from drought.gov (U.S. drought). You also created a Participatory Sensing campaign to investigate an issue in your community.

For this assignment, you will use the results of your Participatory Sensing campaign, and possibly an official dataset, to apply what you have learned in unit 4 and to answer the research question you chose with your group at the beginning of the unit.

Your assignment is as follows:

1. Research which government entity is responsible for the community issue that your team chose – it could be local, state, federal, or even an international entity.
2. Propose one or two recommendations to the government entity in step 1 to help raise public awareness and/or alleviate the issue. Specifically, you will create a presentation in which you answer the following questions:
  - a. What is/are the specific recommendation(s) you are proposing to increase public awareness and/or alleviate the issue?
  - b. Why do you think this will work? What evidence do you have to support this? Include any necessary plots and analysis.
3. Your 5-minute presentation comprising 4-5 slides should include:
  - a. An introduction – Who are you? Introduce yourself and your team.
  - b. The issue – What is the issue? Why is this issue important? Why should we care about it?
  - c. The Participatory Sensing campaign – Explain your campaign. What was your research question? What statistical investigative question(s) were you hoping to answer?
  - d. Your recommendation – What is your recommendation? How will it raise public awareness and/or alleviate the issue? Why do you think this will work? This is where you include your evidence:
    - i. How does your article connect to your recommendation?
    - ii. How does your Participatory Sensing campaign data support your recommendation? Or how does it show that there is a lack of something needed?
    - iii. If you have an official dataset, how does it support your recommendation?
    - iv. Include visualizations, numerical summaries, and/or statistics.
  - e. A closing – Summarize your point into a few closing sentences.

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.