Get the picture?

Lab 1B

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

## Where’d we leave off …

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

## Variable Types

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

## Which is which?

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

## Data Structures

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

str(cdc)

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

## Visualizing data

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

## We have options

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

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

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

## How often do people text & drive?

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

## Does texting and driving differ by gender?

* Fill in the blanks with the correct variables to create a side-by-side bargraph:
* *bargraph* (~ \_\_\_\_ , data = \_\_\_\_ , groups = \_\_\_\_ )
* **Write a sentence explaining how boys and girls differ when it comes to texting while driving.**
* **Would you say that most girls never text and drive? Would you say that most boys never text and drive?**
* **How did including the groups argument in your code change the graph?**

## Do males/females have similar heights?

* To answer this, what we’d like to do is visualize the distributions of heights, separately, for males and females.
  + This way, we can easily compare them.
* Use the groups argument to create a histogram for the height of males and females.
  + **Can you use this graphic to answer the question at the top of the slide? Why or why not?**
  + **Is grouping numeric values, such as heights, as helpful as grouping categorical variables, such as texting & driving?**

## Do males/females have similar heights?

* groups uses color to differentiate between groups.
  + **Why does this work for bargraphs but not for histograms?**
* Fill in the blanks with the correct variables to create a split histogram (The " | " symbol is usually between the delete and enter keys on a keyboard) to answer the questions below:
* *histogram* (~ \_\_\_\_ | \_\_\_\_ , data = \_\_\_\_ )
* **Do you think males & females have similar heights? Use the plot you create to justify your answer.**
* **Just like we did for the histogram, is it possible to create a *split* bargraph? Try to create a bargraph of drive\_text that’s split by gender to find out.**

## On your own:

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