Lab 1A - Data, Code & RStudio

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

## Welcome to the labs!

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

## So let's get started!

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

data(cdc)

View(cdc)

* **Describe the data that appeared after running View(cdc):**
  + ***Who* is the information about?**
  + **What sorts of information about them was collected?**

## Data: Variables & Observations

* Data can be broken up into two parts.
  1. Observations
  2. Variables
* If need be, re-type the command you used to View your data. Then answer the following:
  + **How are our *observations* represented in our data?**
  + **What does the first column tell us about our observations?**
  + **How often did our first observation wear a seatbelt while riding in a car?**

## Uncovering our Data's Structure

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

## Type the following commands into the console

dim(cdc)

nrow(cdc)

ncol(cdc)

names(cdc)

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

## First Steps

* Typing commands into the console is your first step into the larger world of *programming* or *coding* (terms which are often used interchangeably).
* Coding is all about learning how to send instructions to your computer.
  + We call the way we *speak* to the coding language, *syntax*.
* *Capitalization*, *spelling* and *punctuation* are REALLY important.

## Syntax matters

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

Names(cdc)

NAMES(cdc)

names(cdc)

names(CDC)

## R's most important syntax

\*

*function* (y~x, data = \_\_\_\_ )

* Search through the different panes. Find and then click on the *Plots* tab.
  + To get back to the slides, find and then click on the *Viewer* tab.

## Syntax in action

\*

*function* (y~x, data = \_\_\_\_ )

* **Which one of these plots would be useful for answering the question: *Is it unusual for students in the CDC dataset to be taller than 1.8 meters?***

histogram(~height, data = cdc)

bargraph(~drive\_text, data = cdc)

xyplot(weight~height, data = cdc)

* **Do you think it's unusual for students in the data to be taller than 1.8 meters? Why or why not?**

## On your own:

* After completing the lab, answer the following questions:
  + **What is *public health* and do we collect data about it?**
  + **How do you think our data was collected? Does it include every high school aged student in the US?**
  + **How might the CDC use this data? Who else could benefit from using this data?**
  + **Write the code to visualize the distribution of weights of the students in the CDC data with a histogram. What is the *typical* weight?**
  + **Write the code to create a barplot to visualize the distribution of how often students wore a helmet while bike riding. About how many students never wore a helmet?**