Data, Code & RStudio

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

Data Science & R

- The R programming language is one of the main tools used by actual data scientists.
- RStudio packages the R language into an easy to use interface.

So let's get started!

- Step one in any data science project is to load some data!
- Type these two commands into the your console:

```
data(cdc)
```

```
View(cdc)
```

What happened in RStudio after you ran these two commands?

Centers for Disease Control (CDC) Data

- The CDC is a federal insitution that studies public health.
  - **Why should we bother studying public health?**
  - **How might we study it?**
- Our data comes from a survey of high school aged Americans.
  - **How do you think the data were collected?**

Look again at our data

- Type `View(cdc)` into your console again and answer the following questions:
  - **What does each horizonal row of the data represent?**
  - **What information is the first vertical column telling us?**
  - **How are the rows of the data different from the columns?**

Data, Variables & Observations

- Data can be broken up into two parts.
  1. Observations
  2. Variables
- Answer the following questions about the CDC data

  - **Where are the observations and where are the variables in `View(cdc)`?**
  - **What are the differences between observations and variables?**
  - **How are variables and observations related?**

Uncovering our Data's Structure

- RStudio's main window is composed of four *panes*
- Find the pane that has a *tab* titled *Environment* and click on it.
  - **Can you find the number of people surveyed?**
  - **How many variables are there for each person?**
  - **What happens when you click on `cdc`?**

Uncovering our Data's Structure

- From the *Environment* tab, click on the blue arrow to the left of `cdc`
- Don't be overwhelmed! This is just some of the *structure* of our data:
  - We'll learn much more about this *structure* in the future.
  - Do you notice the names of the variables are listed?

Type the following commands into the console

```
dim(cdc)
```

```
nrow(cdc)
```

```
ncol(cdc)
```

```
names(cdc)
```

- **Write each *output* and what it tells us about the people in our CDC data**
  - The **output** is what gets printed after you hit *enter*

Baby Steps to Programming

- Typing commands into the console is your first step into the larger world of *programming* or *coding* (terms which are often used interchangeably).
  - *Programming* helps data scientist pull really useful information from the data.
- Coding is about learning how to send instructions to your computer.
  - We call the way we *speak* to the coding language, **syntax**.

R's most important syntax

**

function (y~x, data = _____)

**

- Look through the different panes for the *Plots* tab and click it.
- Then type the following commands into the console:

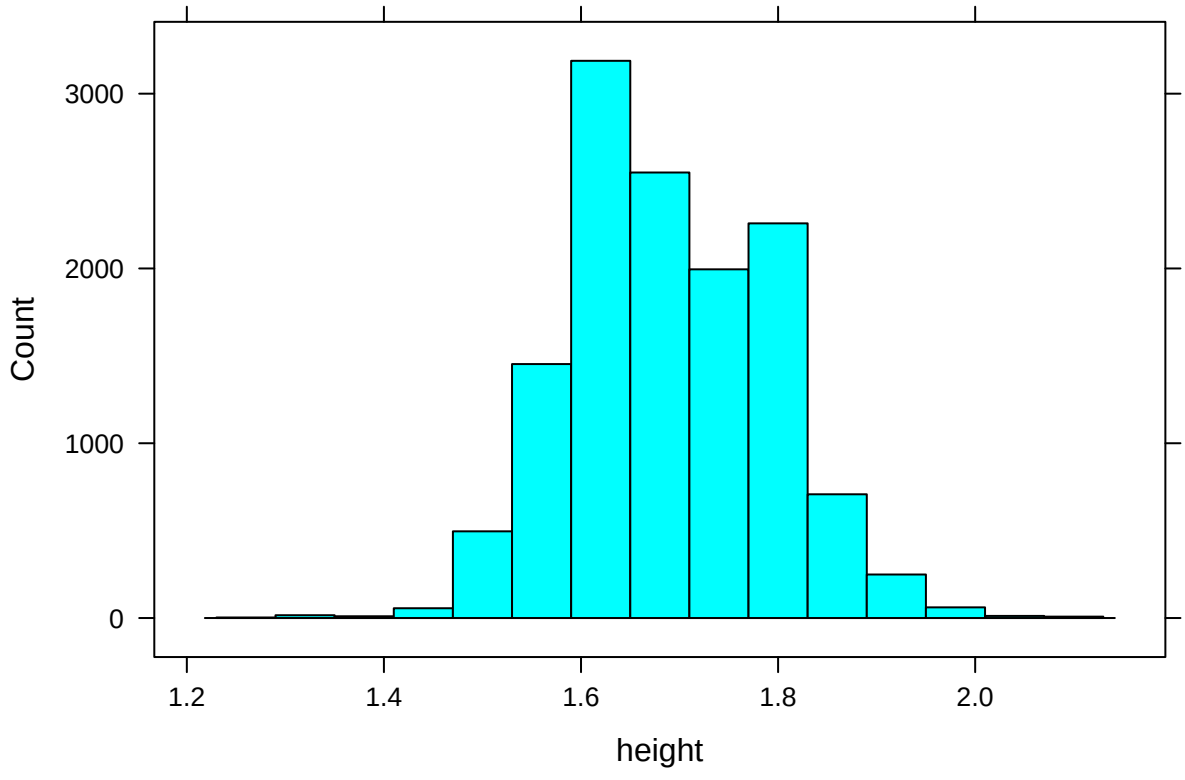
```
histogram(~height, data = cdc)
```

```
bargraph(~sunscreen, data = cdc)
```

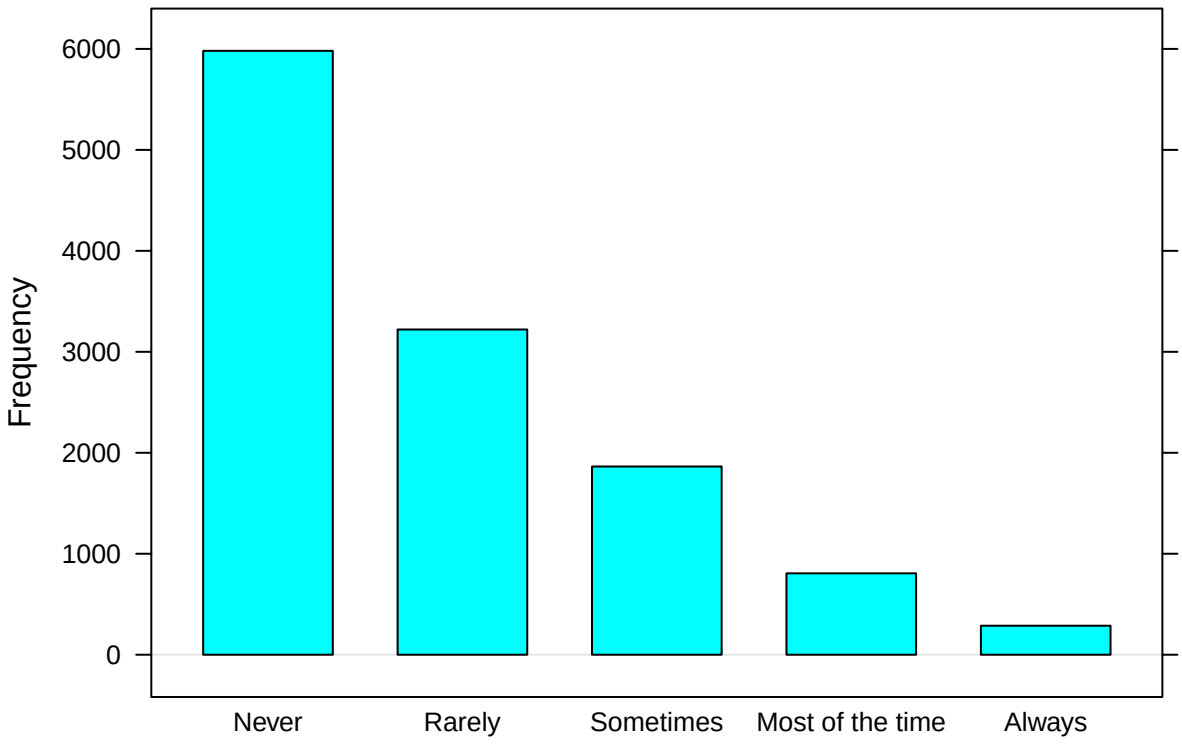
```
xyplot(weight~height, data = cdc)
```

Your First Plots

- **How are these two plots similar?**



- **How are these two plots different?**



Let's discuss

- In your teams:
  - Discuss the answers to the **red** questions you wrote down in your journals.
  - Agree on a single answer for each question.