Lab 1G - What's the FREQ?

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

## Clean it up!

* In Lab 1F, we saw how we could *clean* data to make it easier to use and analyze.
  + Using the data you cleaned, we can start analyzing a small set of variables from the American Time Use (ATU) survey.
  + The process of cleaning and then analyzing data is *very* common in Data Science.
* In this lab, we'll learn how we can create frequency tables to detect relationships between categorical variables.
  + Use the data() function to load the atu\_clean data file to use in this lab.

## How do we summarize categorical variables?

* When we're dealing with categorical variables, we can't just calculate an **average** to describe a *typical* value.
  + (Honestly, what's the average of categories *orange*, *apple* and *banana*, for instance?)
* When trying to describe categorical variables with numbers, we calculate **frequency tables**

## Frequency tables?

* When it comes to categories, about all you can do is *count* or *tally* how often each category comes up in the data.
* Fill in the blanks below to answer the following: **How many more *females* than *males* are there in our ATU data??**

tally(~ \_\_\_\_, data = \_\_\_\_)

## 2-way Frequency Tables

* Counting the categories of a single variable is nice, but often times we want to make comparisons.
* Use a line of code, that's similar to how we facet plots, to tally the number of people with physical challenges and their genders.
  + **Does one gender seem to have a higher occurence of physical challenges than the other? If so, which one and explain your reasoning?**

## Interpreting 2-way frequency tables

* Recall that there were 1153 more women than men in our data set.
  + If there are more women, then we might expect women to have more physical challenges (compared to men).
* Instead of using *counts* we use *percentages*.
* Include: format = "percent" as option to the code you used to make your 2-way frequency table. Then answer this question again:
  + **Does one gender seem to have a higher occurence of physical challenges than the other? If so, which one and explain your reasoning?**
  + **Did your answer change from before? Why?**

## One final option

* It's often helpful to display totals in our 2-way frequency tables.
  + To include them, include margins = TRUE as an option in the tally function.

## On your own

* **Describe what happens if you create a 2-way frequency table with a numerical variable and a categorical variable.**
* **How are the types of statistical questions that 2-way frequency tables can answer different than 1-way frequency tables?**
* **Which gender has a higher rate of *part time employment*?**
* **Does one gender socialize more than the other? To answer this question first:**
  + **Create a subset of the ATU data that includes only people who socialized more than 0 minutes.**
  + **Create a histogram and include type = "percent" as an option in the function.**