Lab 3

Christy Lei

Math 241, Week 4

```
# Put all necessary libraries here
library(tidyverse)
```

Due: Thursday, February 27th at 8:30am

Goals of this lab

1. Practice using GitHub.
2. Practice wrangling data.

Data Notes:

- For Problem 2, we will continue to dig into the SE Portland crash data but will use two datasets:
  - `CRASH`: crash level data
  - `PARTIC`: participant level data

```
# Crash level dataset
crash <- read_csv("/home/courses/math241s20/Data/pdx_crash_2018_CRASH.csv")

# Participant level dataset
partic <- read_csv("/home/courses/math241s20/Data/pdx_crash_2018_PARTIC.csv")
```

- For Problem 3, we will look at chronic illness data from the CDC along with the regional mapping for each state.

```
# CDC data
CDC <- read_csv("/home/courses/math241s20/Data/CDC2.csv")

# Regional data
USregions <- read_csv("/home/courses/math241s20/Data/USregions.csv")
```

- For Problem 4, we will use polling data from FiveThirtyEight.com.

```
# Note I only want us to focus on a subset of the variables
polls <- read_csv("/home/courses/math241s20/Data/generic_topline.csv") %>%
  select(subgroup, modeldate, dem_estimate, rep_estimate)
```

Problems

Problem 1: Git Control

In this problem, we will practice interacting with GitHub on the site directly and from the RStudio Server. Do this practice on **your repo**, not your group's Project 1 repo, so that the graders can check your progress with Git.

a. Let's practice creating and closing **Issues**. In a nutshell, **Issues** let us keep track of our work. Within your repo on GitHub.com, create an Issue entitled "Complete Lab 3". Once Lab 3 is done, close the **Issue**. (If you want to learn more about the functionalities of Issues, check out this page.)

b. Edit the ReadMe of your repo to include your name and a quick summary of the purpose of the repo. You can edit from within GitHub directly or on the server. If you edit on the server, make sure to push your changes to GitHub.

c. Upload both your Lab 3 .Rmd and .pdf to your repo on GitHub.

Problem 2: `dplyr` madness

Each part of this problem will require you to wrangle the data and then do one or both of the following:

- Display the wrangled data frame. To ensure it displays the whole data frame, you can pipe `as.data.frame()` at the end of the wrangling.
- Answer a question(s).

Some parts will require you to do a data join but won't tell you that.

a. Produce a table that provides the frequency of the different collision types, ordered from most to least common. What type is most common? What type is least common?

b. For the three most common collision types, create a table that contains:

- The frequencies of each collision type and weather condition combination.
- The proportion of each collision type by weather condition.

Arrange the table by weather and within type, most to least common collision type.

c. Create a column for whether or not a crash happened on a weekday or on the weekend and then create a data frame that explores if the distribution of collision types varies by whether or not the crash happened during the week or the weekend.

d. First determine what proportion of crashes involve pedestrians. Then, for each driver license status, determine what proportion of crashes involve pedestrians. What driver license status has the highest rate of crashes that involve pedestrians?

e. Create a data frame that contains the age of drivers and collision type. (Don't print it.) Complete the following:

- Find the average and median age of drivers.
- Find the average and median age of drivers by collision type.
- Create a graph of driver ages.
- Create a graph of driver ages by collision type.

Draw some conclusions.

Problem 3: Chronically Messy Data

a. Turning to the CDC data, let's get a handle of what is represented there. For 2016 (use `YearStart`), how many distinct topics were tracked?

b. Let's study influenza vaccination patterns! Create a dataset that contains the age adjusted prevalence of the "Influenza vaccination among noninstitutionalized adults aged >= 18 years" for Oregon and the US from 2010 to 2016.

c. Create a graph comparing the immunization rates of Pennsylvania and the US. Comment on the observed trends in your graph

d. Let's see how immunization rates vary by region of the country. Join the regional dataset to our CDC dataset so that we have a column signifying the region of the country.

e. Why are there NAs in the region column of the new dataset?

f. Create a dataset that contains the age adjusted influenza immunization rates in 2016 for each state in the country and sort it by highest immunization to lowest. Which state has the highest immunization?

g. Construct a graphic of the 2016 influenza immunization rates by region of the country. Don't include locations without a region. Comment on your graphic.

Problem 4: Tidying Data Like a Boss

I was amazed by the fact that many of the FiveThirtyEight datasets are actually not in a perfectly *tidy* format. Let's tidy up this dataset related to polling.

a. Why is this data not currently in a tidy format?

```
polls
```

```
## # A tibble: 1,529 x 4
##    subgroup  modeldate dem_estimate rep_estimate
##    <chr>     <chr>            <dbl>        <dbl>
##  1 All polls 9/18/2018         48.8         39.8
##  2 All polls 9/17/2018         49.0         39.9
##  3 All polls 9/16/2018         49.0         39.9
##  4 All polls 9/15/2018         49.0         39.9
##  5 All polls 9/14/2018         48.9         39.8
##  6 All polls 9/13/2018         48.8         39.7
##  7 All polls 9/12/2018         48.8         39.6
##  8 All polls 9/11/2018         48.5         39.9
##  9 All polls 9/10/2018         48.4         39.9
## 10 All polls 9/9/2018          48.4         39.9
## # ... with 1,519 more rows
```

b. Create a tidy dataset of the `All polls` subgroup.

c. Now let's create a new untidy version of `polls`. Focusing just on the estimates for democrats, create a data frame where each row represents a subgroup (given in column 1) and the rest of the columns are the estimates for democrats by date.

d. Why might someone want to transform the data like we did in part c?

Problem 5: YOUR TURN!

Now it is your turn. Pick one (or multiple) of the datasets used on this lab. Ask a question of the data. Do some data wrangling to produce statistics (use at least two wrangling verbs) and a graphic to answer the question. Then comment on any conclusions you can draw about your question.