Problem Set 1: Modules 1, 4, & 7

YOUR NAME HERE

due INSERT DATE HERE

Structural stuff:
1. Be sure to change the "author" above to your name. Also insert the due date for this term/assignment.
2. Save your .Rmd file as LastName_FirstName.Rmd (do this before you knit).
3. You need to submit your .Rmd code file AND a knit file (upload both simultaneously to the course webpage; you can't upload them one-by-one). You will only receive full credit if you upload both files.
4. Below I have set up the file for you with the libraries you'll need. I have also inserted code chunks for you (note, I won't do this every time).
5. I expect that the .Rmd file you submit will run cleanly, and that the knit file won't contain any errors (LOOK at the knit file after you create it - if questions/text are running into each other, if you see error messages, etc., you're not done).
6. You can use comments to tell me what you are doing either in text or in code chunks, but remove "old" code that didn't run/work.

```
knitr::opts_chunk$set(echo = TRUE)

library(yardstick)
```

```
## For binary classification, the first factor level is assumed to be the event.
## Use the argument `event_level = "second"` to alter this as needed.
```

```
library(haven)
library(tidycensus)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x readr::spec()   masks yardstick::spec()
```

This assignment is comprised of three parts: Part 1 (Mod #1), Part 2 (Mod #4), and Part 3 (Mod #7)

Part 1.

For this part, you'll be working with another piece of the College Scorecard dataset ("College DEBT dataset" from the course website) to predict the debt of college graduates using conditional means. You'll need to select the college-level characteristics that you think might be related to eventual debt.

1. Import/load the "sc_debt" R dataset here.

1a. Print (aka, display) the first 11 rows of the dataset below.

1b. Print (aka, display) the last 11 rows of the dataset below.

1c. How many observations are in this dataset?

1d. How many variables are in this dataset?

Part 2.

For this part, you'll need to open up, clean, and save datasets using the tools we've gone over in class. For each dataset, make sure that when you're done you have a nice, neatly labeled dataset that would be easy for you or another analyst to open and analyze. Some of the datasets may need no/minimal cleaning. Some may need cleaning. Save each dataset as an `RData` file. You will upload each saved R dataset to the LMS with your knit pdf and .Rmd file.

PLEASE NOTE: You can't just copy these links into your code. You need to figure out the format for the data. Take a look at the website to figure out what kind of data is there; then use the appropriate code in R to import it. Additionally, a CLEAN dataset means one that could be easily opened by another analyst and used immediately (e.g., has variable names, one observation per row, etc.).

2. Airline dataset: http://www.principlesofeconometrics.com/sas.htm

3. King county births: http://courses.washington.edu/b517/Datasets/datasets.html

Part 3.

For this part, you will be accessing the American Community Survey (ACS) data. You'll need to select the appropriate table and organize the data in such a way that you can calculate proportions.

4. Download data for all of the counties in Tennessee (TN) on highest education received.

4a. How many observations are in your dataset?

4b. What does each observation in this dataset represent?

5. Compute the proportion of the population that has an associate's degree or above in each TN county. Print the first several rows of your new dataset (hint: use the head() function). Do NOT print the whole dataset, please.

5a. Which county has the highest proportion of the population with an associate's degree or above?

6. Download data for all of the counties in TN on household income.

6a. How many observations are in this dataset?

6b. What does each observation in this dataset represent?

7. Compute the proportion of the population that has a household income above $50,000 in each TN county. Print the first several rows of your new dataset (hint: use the head() function). Do NOT print the whole dataset, please.

7a. Which county has the highest proportion of the population with a household income above $50k?