Project, STAT 650

Due: Friday, October 7

Instructions:

- For the project you will find a data set of interest, and use methods learned in this class to analyze that data set. You should primarily use packages from `tidyverse` (e.g., `ggplot2`, `dplyr`, `readr`), which have been the focus of the class.

- You may work individually, or in a group of 2-3 students.

- Your paper should have the following sections:

  1. **Introduction**: Describe the main research questions or goals of your data analysis. (1 paragraph should be sufficient.)
  2. **Data Description**: Briefly describe your data set. What is the source? What is the dimension (number of rows and columns)? What are the variables of interest?
  3. **Results**: Present your main results. This should be some kind of compelling visualization(s) of your data. But you may also present a table of summary statistics, or the output of a statistical model (with clearly defined response and predictors). Be selective about the results you choose to include. A single high quality visualization is preferable to a large number of mediocre visualizations. This section should also include some written interpretation of your results.

- The paper should be about 2-4 pages with figures and tables, and submitted to Canvas in PDF format. Make sure to include a title and the names of all members in your group. (If working in a group, only one member needs to make a submission on Canvas.)

- Your R code should be in a separate R Markdown file. You can either submit your code as an attachment on Canvas, or as a link to GitHub repository.

Grading: A list of specific expectations are provided below.

- The research questions and goals of the analysis are clearly described.
- The source of the data set is provided, and the relevant variables are listed and described.
- The selected results (plots, tables) illustrate important aspects of the data set.
- The paper is well-formatted and organized. There are very few typos or grammatical mistakes.
- Figures and tables are well-formatted with appropriate labels.
- The R code is easy to follow and reproducible.

Papers that meet these expectations will receive an A. Papers with minor flaws, that mostly address the above expectations, will receive an A-. Papers that fail to address several of the above expectations in critical ways will receive a B or B-. For example, papers that have poor formatting, organization, and/or writing will receive a B or B-. Papers that are incomplete, plagiarized, and/or demonstrate little interest or effort will not receive a passing grade.

Data Sources:

Here are some potential sources for data sets. You do not need to limit yourself to these.

- Tidy Tuesdays: `https://github.com/rfordatascience/tidytuesday`
- Kaggle: `https://www.kaggle.com/datasets`
- FiveThiryEight: `https://data.fivethirtyeight.com/`
  R package: `library(fivethirtyeight)`
- UCI Machine Learning Repository:
  `https://archive.ics.uci.edu/ml/datasets.php`
- DataSF: `https://datasf.org/opendata/`
- Awesome Public Datasets:
  `https://github.com/awesomedata/awesome-public-datasets`
- Google data set search: `https://datasetsearch.research.google.com/`

You can also use a data set from one of the textbooks cited in this class. However, **do not reuse a data set that has already been used in lecture or homework**.

- *Modern Data Science with R*:
  `https://mdsr-book.github.io/mdsr2e/ch-prologue.html#datasets`
- *R for Data Science*: `https://r4ds.had.co.nz/index.html`

To get a list of the data sets in an R package run the command `data(package = "name")`. For example, run the following command to get a list of data sets in the `mdsr` package:

```r
data(package = "mdsr")
```