Project Presentation, STAT 450

Due: Last week of class

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 the `tidyverse` (e.g., `ggplot2`, `dplyr`, `readr`), which have been the focus of the class.

- You may work in a group of 2-3 students, or individually. Because of time constraints, please try to work in group.

- Presentations will take place during the **last week of class (Monday, December 2 and Wednesday, December 4)** and should be between 5-10 minutes long.

- For your presentation, prepare 3-5 slides that include the following:

  1. **Title**: title of project and names of group members.
  2. **Data Description**: provide the data source, dimensions (number of rows and columns), and descriptions of relevant variables.
  3. **Results**: present the main results of your data analysis. This should be some kind of compelling visualization(s) and/or table of summary statistics. Be selective about the results you choose to include. A single high quality visualization is preferable to a large number of mediocre visualizations.

- By **Friday, December 6** each group should also submit the following two files to Canvas: (1) presentation slides in PDF format, and (2) Quarto document with R code rendered to HTML or PDF format.

Grading: A list of specific expectations are provided below.

- The source of the data set is provided, and relevant variables are listed and described.
- The selected results (plots, tables) illustrate important aspects of the data set.
- Figures and tables are well-formatted with appropriate labels.
- Each group member makes a contribution to the project.
- Your presentation is not exceedingly long (under 10 minutes, please).
- R code is provided in a Quarto document.

Projects that meet these expectations will receive an A. Projects with minor flaws, that mostly address the above expectations, will receive an A-. Projects that fail to meet several expectations in significant ways will receive a B or C. Projects 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. However, **do not reuse a data set that has already been used in lecture or homework**.

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

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 `fivethirtyeight` package:

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