STS 2300 – Project 1

Data Investigation Project

# Background

The goal of this project is for you to apply skills you’ve learned in Notes 01 – 05 to a real-world data set of your choice and to be able to explain how you’ve demonstrated your command of STS 2300 topics. There will be two main products that you create.

The first product you will create is a .Rmd file (and knitted document) in which you do appropriate data management, data summary, and data visualization to explore a question of interest related to your data. The intended audience for this product is someone outside of this class who is interested in learning about your topic or question of interest.

The second product is a short paper explaining how your .Rmd file demonstrates your proficiency in the data analytics topics mentioned above. The intended audience for this product is your professor (or an interviewer) who is assessing your ability to write R code for data management and analysis.

**Note**: You may choose to work by yourself or with one partner. If you work with a partner, you will each need to submit a short paper describing how the product demonstrates your proficiencies. This means both of you should be involved in the coding and able to explain the parts of the first product.

# Task 1 – Project Proposal (Choose a Data Set)

To start, you should choose a dataset that you plan to use. One option is to pick one from the [Tidy Tuesday website](https://github.com/rfordatascience/tidytuesday)  (Note: Scroll down to see the list of datasets. You can click on the tabs for other years as well). You are also welcome to use data sets from other sources, but you should let the instructor know if you plan to do this.

In choosing a dataset, you should pick something that is complex enough to allow you to demonstrate your data management and data visualization skills sufficiently. Ideally you will have both quantitative and categorical variables for summary and graphing.

**Submission**: A .Rmd file (and knitted .html file) that demonstrates you can read your data into R and do one or two small things with the data (e.g. explore the structure, make a simple graph, summarize a variable, etc.). A template .Rmd file is available on the course Github page.

**Due Date**: Friday, September 20, by the start of class

**Audience**: The instructor.

# Task 2 – Main Product (R Markdown Report)

Once you have chosen your data set, you will write a report using R Markdown that includes the following sections:

* **Introduction**: Give background information about your data set and clearly state your question of interest related to it. Make sure the reader has a general understanding of the type of variables and information is in your data, how big the data set is, and where it comes from. This should include a citation of the dataset.
* **Data preparation**: Do data management that is appropriate for your question of interest (e.g. subsetting rows or columns, renaming variables, creating new variables, joining data frames, converting between wide and long format, etc.). The reader of this document won’t care about all the details (e.g., variable names), but you can explain the types of things you’re doing and why they are needed.
* **Descriptive Statistics**: Create well-formatted and visually appealing numeric summaries and tables appropriate for your question of interest (e.g., describing the center and spread of numeric variables for different groups, showing the count or proportion of observations within different categories). You should use the kableExtra or janitor packages to make this better formatted for a reader outside of this course.
* **Data Visualization**: Create at least two different kinds of visually appealing and informative graphs using the ggplot2 package that are related to your question of interest. Your graphs should have appropriate labels, themes, scales, etc. Since the goal is to demonstrate your data visualization proficiency, make sure to go beyond the basics when creating your graphs.
* **Conclusion**: State any overall takeaways as they relate to your question of interest. What new questions do you have? Are there limitations to how someone should use your report?

**Submission:** A knitted .html file (and its corresponding .Rmd file) via Moodle. A printed hardcopy of the .html file.

**Due Date (tentative)**: Wednesday, October 9, by the start of class.

**Audience**: Someone outside the course who is interested in your topic or question of interest.

# Task 3 – Companion Reflection (Paper Explaining How You Demonstrated Proficiency in STS 2300 Topics)

Write a short (approximately 2 page) paper explaining how your R Markdown file demonstrates your proficiency in the topics from Notes 01 - 05. For each topic, you will have flexibility in how you demonstrate your abilities, but you will need to make your case for why what you did demonstrates proficiency. Typically, this means more than a couple lines of code using only 1 or 2 things learned about that topic. Topics include:

* R Markdown coding (e.g., headers, links, formatting like bolding or italicizing, lists)
* Data management (e.g., subsetting data, variable creation, variable renaming, use of piping, converting from wide to long format, joining data frames)
* Data summary and formatting (e.g., appropriate numeric and table summaries, summaries by groups, kableExtra tables)
* Data visualization

**Submission:** A hardcopy of the paper. (No digital file needed)

**Due Date (tentative)**: Wednesday, October 9, by the start of class

**Audience**: The instructor or an interviewer assessing your ability to code and demonstrate data analytics knowledge and skills