DATS 6101-11 Introduction to Data Science

First Project Outline (Spring 2023)

Description and Purpose:

The goal of this (first) project is to better understand the initial stages of data-focused research by conducting background investigation, developing one or more research-driven SMART questions, completing exploratory data analysis (EDA), and performing appropriate statistical testing to help answer the SMART questions.

Each team will choose their own research topic and question(s). We are not collecting data ourselves. Instead, we will look for available data online or from other sources your team might have access to. For this "big data" class, we require datasets to have at least 4000 observations (i.e. 4000 rows of data).

Details:

I. **Topic Proposal, 5%, due February 15 before class**.

Each team will submit one proposal on Blackboard. Use 150-200 words to describe a) the research topic, b) the SMART question(s) of your research (you can still change them afterwards), c) the source of your data set(s) and how many (roughly) observations, and d) the link to your team's GitHub repo.

II. **Presentation slides, 10%, due March 8 by noon**.

You can use Powerpoint, Google Slides, or any other appropriate products. Typical slides should not be too "wordy". It's not the purpose of slides to be read like an article. Even though this following point is not universally agreed on, I personally do not feel complete sentences are appropriate on slides. Just bullet/lists of the main points you want to deliver. Use charts, graphics, animations to capture the audience's attention.

If you need a copy of Office 365 PowerPoint, currently-enrolled GW students can get a free version from GW Information Technology.

III. **Presentation, 25%, March 8 in class, individually graded**.

Develop a **15-to-20-minute** presentation for the team that effectively communicates your research question(s), data, and important parts of your data analysis. Each team member must present.

IV. **Summary paper / Write-up, 50% total, due March 22 before class**.

Write a roughly 10-page (definitely no more than 4000 words, charts do not count) summary of the research and EDA process of your project. The summary should be prepared in ***Rmarkdown*** and knitted into ***HTML*** – *you must use the code folding option.* You may make modifications to your analysis based on feedback from your presentation.

(a) Exposition and Curation, 25%.

This summary is to be presented to your boss, your client, or to be submitted for publication in journals. *Note: you shouldn't include code in such a report, but I need to check the correctness of your code, so use the code folding option.* It should be well-organized, well-written, and contain the important parts of your data analysis that efficiently, yet completely tell the story of your project. The write-up should also provide thoughtful insight into the process of the analysis, including its strengths and weaknesses. Potential area of topics to address in this summary may include:

- What do we know about this dataset?
- What are the limitations of the dataset?
- How was the information gathered?
- What analysis has already been completed related to the content in your dataset?
- How did the research you gathered contribute to your question development?
- What additional information would be beneficial?
- How did your question change, if at all, after Exploratory Data Analysis?
- Based on EDA can you begin to sketch out an answer to your question?
- References (APA style preferred)

(b) Technical Analysis, 25%.

You will also be graded on the technical quality of your analysis as presented in your write-up. You will be assessed for proper and sensical usage of R functions and statistical tools such as:

- Summary of the dataset
- Descriptive statistics
- Graphical representations of the data
- [Proper usage when applicable] Measures of Variance / sd
- [Proper usage when applicable] Normality tests
- [Proper use when applicable] Initial correlation / Chi Square tests / ANOVA analysis / Z-test or Z-interval / T-test or T-interval etc.

V. **Git usage, 10%, individually graded**.

After all the submissions, submit your team's git activity report. Using git is a daily routine. Don't wait until the last week or days before the project is due to use git. All your participation and commits are recorded in the git system. You will be graded on frequency and timeliness.

All grades except for II and V are team-based. But I reserve the right to award different grades to team members if there is evidence of unfair contribution within the team.