DATS 6101 Introduction to Data Science

Midterm Project Outline (Spring 2022)

Description and Purpose:

The goal of this (midterm) project is to better understand the initial stages of a data focused research by conducting background investigation, developing one or more SMART questions, completing the Exploratory Data Analysis (EDA), and performing appropriate statistical testing to help answer SMART questions.

Each team will choose their own research topic and question. 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 3000 observations (ie. 3000 rows of data).

Your team should submit a topic proposal one week before the presentation is due. After the presentation, you will receive feedback from your fellow classmates, TA, and instructor within the next 2 days. You can then make modifications to your final write up before submitting. Your team will prepare and submit a research paper, no more than 5000 words (in the html format, not counting TOC, figure caption, etc). Even though each picture/chart/graph might be worth a thousand words, they do not count towards the word-limit.

Details:

I. (Due: Feb 23rd, Wednesday, End of Day) Topic Proposal. Each team submit one proposal on Blackboard, in 150-200 words 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. Development of a **research driven question (SMART)** focused on a dataset either inside of R or one of your choosing from any online sources. Acceptable dataset for this "big data" class requires 3000+ observations (that is, 3000+ rows of data for the data frame).

III. (Due: March 9th, Wednesday, End of Day) Provide an **R-markdown file**, knitted into **HTML**, which shows the R-code and brief explanations as well as the rationale of the **Exploratory Data Analysis** of your project. (**Also** submit your **data file** or give the online source url.) This document shows a technical person the math/stat/codes that you used in your analysis. It should include:
    - Summary of the dataset
    - Descriptive Statistics
    - Graphical representations of the data
    - [When applicable] Measures of Variance / sd
    - [When applicable] Normality tests
    - [When applicable] Initial correlation / Chi Square tests / ANOVA analysis / Z-test or Z-interval / T-test or T-interval etc.

IV. (Due: March 2nd, Wednesday, class time) Develop a **15-to-20-minute** presentation for the team that effectively communicates the results of these initial stages of a data science project to be presented during class.

V. (Due: March 16th, End of Day) Write a roughly 10-page (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 **R-markdown** and knitted into **HTML**. You may take some of the work in part II (such as graphs and results) to include here. They can overlap. This summary is to-be presented to your boss, your client, or to-be submitted for publication in journals. 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)

Grading:

I. Topic Proposal, 5%
II. SMART Question (see V)
III. R-codes (RMD), Technical analysis, 25%
IV. 25% Individual presentation score + 10% team presentation doc (pptx or google slides)
V. Summary paper, together with SMART question in part II, total of 25%
VI. Git usage – 10%, individually graded, based on your git activity in repo history)

NOTE: Grades for parts I, II, III, V, and part of IV, 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. A peer evaluation form will be submitted individually by all students after the completion of the project.