DATS 6103: Data Mining Project – 2024 Fall

Goal

The goal of this project is to use Python to procure, pre-process, and clean a dataset found online, and use it to present an analysis that includes EDA and various model-building tasks that are applicable to the dataset. The skills that all data scientists treasure the most would be all on display. They include technical analysis, critical thinking, teamwork, communication, and visualization skills.

There will be a 15-minute in-person presentation for each team. The instructor, TA, and your fellow classmates might give you feedback and comments on your presentation. Based on all your work as well as others' comments, your team will submit a final paper explaining your analysis, and your results, as if you were submitting the formal report to your supervisor, or an article to be submitted to peer-reviewed journals. The paper should be about 10-15 pages long (NOT counting any charts and graphs), and less than 5000 words in any case. You can have as many charts and graphs as you like. A chart is worth a thousand words (not included in word count, however).

We require datasets to have at least 4000 observations (i.e. 4000 rows of data), and eight or more relevant variables/features (i.e. 8+ columns of useful data).

Steps for Completion

This project has these components (100 points total, which is 25% of your course grade):
1. Topic proposal (5 points, team grade*)
2. Presentation slides (10 points, team grade*)
3. Presentation (25 points, your in-person presentation, individually graded)
4. Git usage (10 points, individually graded, from the git repo history log online)
5. Codes (25 points, team grade*, your technical and coding skills)
6. Final Write Up, any format (25 points, team grade*, your communication skills)

* Even though these items are graded as a team, I reserve the right to make adjustments if there is strong evidence of unequal contributions among team members. I hope I never need to use this special clause.

Key Due dates

Due Date	Assignment
Dec 5th	Topic Proposal
Dec 12th	Presentation Slides
Dec 12th	Code Files
Dec 14th	Discussion forum comments
Dec 17th	Final Write up
Dec 17th	Git Usage

Part 1: Topic Proposal

Your topic proposal (one per team) is a 150-200 word description of
1. the research topic your team comes up with
2. the research question(s)
3. the source of your data set(s) and
4. the link to your team's GitHub repo
5. the modeling methods you propose to use (you can change afterward)

Part 2: Presentation Slides

You can use Powerpoint, Google Slides, Canva 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 needed on slides. Just bullet/list of the main points you want to deliver. Use charts, graphics, and animations to capture the audience's attention. Keep in mind some of your audience might be watching your presentation on a low-res screen, or a mobile device, or in general could be real-time with low bandwidth. Cramping a lot of words onto the slides would be unwise.

Using succinct bullet points is one way to thoughtfully ensure readability and accessibility for others. Another step to take is using appropriate contrast (e.g., a very dark blue background with white or off-white text) and San Serif fonts.
To fully check the accessibility of a PowerPoint presentation, you can use the Tools > Accessibility Checker option.

Part 3: Presentation

Each team will prepare a 15-minute presentation delivered on our presentation day. Each member needs to be a part of the presentation. Although we do not have strict rules on the time/duration of appearance for each member, roughly equal time allocation is desired. Any order of appearance and combinations are allowed. Making the best impression on the audience is the goal.
The presentation score is awarded individually to each member.

Part 4: Git Usage

While you are working on the project and collaborating with your teammates, you all should be using the git source-control-management (SCM) system. Remember to commit often, and fetch/pull regularly, so that the teamwork goes smoothly.

This part is graded individually. As long as you are doing your part, and updating your codes regularly on the shared repo, you will get full credit for this part.

Parts 5: Codes

Submit a separate Python file (.py) that includes your codes. The codes need to be working. Indicates what packages you are importing, so that I can follow and run your codes. Also, remember to either include the dataset or include a link to the dataset either online or on GitHub.

Parts 6: Final Write Up

The final write-up can be prepared in any word processing format and style your team chooses. There is a 5000-word limit (max) for the document. There is no minimum, as long as all the information about the research and the result is conveyed. If your team has to go over the limit, please contact me to get pre-approved.

These are some of the items for the write-up, if applicable. Use your organizational skills to make the logical summary of your research:

1. Overview of your project (if applicable):
    - Why did your team choose this topic?
    - What prior research and analysis have been done on this topic?
    - Information about the dataset(s) you used.
    - Any unusual EDA results that are worth mentioning?
2. Your research questions, and how did they come up?
3. Your models or machine learning algorithms. Score or evaluate your models.

4. Interpret your results.
5. What predictions can you make from your models? Examples?
6. Draw conclusions. How do these answer the research questions?
7. References (APA style, see for example GW Himmelfarb: APA Citation Style, 7th Edition)

Grading Criteria

Your final paper will be graded according to the given rubric document.