DATS 6103 Data Mining 2019F – Project Instructions

Goal

The goal of this project is to use Python to obtain, pre-process, and clean dataset found online, and use it to present an analysis that includes EDA, various model building tasks that is applicable to the dataset. There will be an oral presentation, 15-20 mins for each team, and a written paper explaining your analysis, your results, as in either a 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 including any charts and graphs. You can have as many charts and graphs as you like. A chart is worth a thousand words.

Instructions

1. Each team should let me know what area of interest you would like to focus on.
2. Let me know if you already have a dataset in mind. Datasets should have 5000+ datapoints, and 5+ useful features/variables.
3. If you would like to use some API to get the data, please do let me know. Likewise, if you do not want to venture into APIs, tell me just that.
4. I will either approve your dataset if you have one, or let you know where to collect your data.
5. If your team have no objection, you should then start to explore questions that you would like to answer with the dataset.
6. Sometimes you might find additional datasets are needed to merge into the existing ones in order to answer certain questions. Your team can discuss with me regarding the suitability.
7. Every teammate should all contribute to the python codes. Each team should have their own repo on GitHub or other similar git-based network. We will briefly look at your repo and the commit history. If you have a private repo, you can simply share your screen of your repo with the TA and myself on webex at some point.
8. The questions that you develop should have at least some involving model building and/or predictive questions, and/or clustering. If you have only descriptive questions, you will not get a good grade.
9. Don't forget to include references and cite your sources.

Deliverables

- Area of interest of your choice, sent to your instructor in email by Oct 1.
- Proposals on Questions you would like to answer in your project (aka Project proposal on the syllabus) in email due by Oct 17.
- Team presentation on Dec 5.
- Team paper due Dec 10.
- Reminder: Final exam on Dec 12, 7:40pm.

Grading

1. 33.33% Summary Report
2. 33.33% Python coding, including data mining, EDA, models
3. 33.34% Presentation (Individually rated, not a team score)