DATA 115: Final Project

This document provides details and guidelines for the group project, which accounts for 35% of your final grade. As this is the largest component of your grade, you should read the sections below carefully in order to be successful. The purpose of this project is to give you an opportunity synthesize all of the material that we have covered in the course and showcase your Python and data analytics skills by performing a case study on real data.

1 Project Outline

With a group of two to four other students, you will select a dataset from one of the resources below and analyze it in Python using the techniques we have discussed in class. Your main goal will be to provide an answer to a related 'Big Question' that you will generate based on the description of the contents of the data. Along the way, you will need to process the data, conduct exploratory data analysis, and justify the conclusions that you draw while constructing a satisfactory answer to your question.

2 Main Components

You group will summarize and describe your work in a final report, submitted as a .pdf by 10am on December 16, as well as in a 15 minute presentation to the class during the finals period that same day. The report and presentation will be evaluated on the criteria listed below with a particular focus on technical soundness and creativity:

- **'Big Question':** Is the question interesting, clearly stated, and specific? Is the chosen dataset a reasonable option for addressing this question?
- **Visualizations:** Are the visualizations used to represent the analysis effective and complete?
- **Analysis:** Are the methods that we used to analyze the data appropriate and carried out correctly? Is the analysis thorough and logically conducted?
- **Conclusions:** Do the final conclusions provide a satisfactory answer to the 'Big Question'? Are the conclusions supported by the analysis that was performed and presented?
- **Reproducibility:** Are the data processing and cleaning methods well-documented? Is the code used to analyze the data correct and easily interpretable?
- **Presentation:** Is the final report well organized and neat? Is the group presentation effective and informative? Are the plots designed with care for the presentation formats, including color choices, appropriate labelling, and other aspects of good visualization design?

2.1 Report

Your final report should thoroughly describe your experiences and analysis, as though you were reporting the results of this project to a manager or supervisor. The document should be submitted as .pdf by the beginning of the finals period for this class. The following information should be contained in the report, as well as appropriate figures from the analysis incorporated into the text and your final jupyter notebooks included as an appendix:

- Describe the dataset and why you selected it for this project.
- Describe any processing problems you identified with the data and how you overcame those issues.
- Describe your 'Big Question' and why the data is a good choice to answer it.
- Describe the results of your exploratory analysis and what preliminary conclusions you were able to draw based on this analysis.
- Describe how you selected the methodology for your analysis of the big question and the pros and cons of that method and any alternative methods you considered.
- Describe your final conclusions based on your analysis and support them with analytics on your dataset.
- Describe any additional analyses that you would have liked to carry out and any additional data that would have been needed in order to extend your analysis.

2.2 Presentation

During the finals period, your group will have 15 minutes to present the results of your analysis. Each member of the group should present (be speaking) for at least two minutes during this time. As this is not a large block of time, you will not be able to discuss every analysis and computation that you performed over the month that you were working on the project. Instead, you will need to prioritize the important pieces of your analysis decide how to efficiently present the data, question, and conclusions.

3 Data Sources

You should select the underlying dataset from one of the following sources. Note that these resources provide additional information about the data beyond the specific values and you should incorporate a discussion of the relevant pieces of this information (how the data was gathered, what was the original purpose of the data, what cleaning steps were performed before the data was uploaded, etc.) into your report and presentation.

- https://archive.ics.uci.edu/ml/datasets.php
- https://www.kaggle.com/datasets
- https://cseweb.ucsd.edu/~jmcauley/datasets.html
- https://datasetsearch.research.google.com/

4 Scheduling

There are only two formal deadlines for this project, a preliminary worksheet to be completed (through Blackboard) at the end of Week 12 and the final report submission and presentations on December 16. Demonstrating effective time management skills is part of this project. Along those lines there are some recommendations below for benchmarks that you might consider attempting to meet. In particular, please don't wait until the last minute to get started!

- **Week 12:** Download and import the data. Adress any preliminary cleaning concerns.
- **Week 13 (+ Thanksgiving):** Exploratory Data Analysis. Take a deep look at all columns and data properties.
- **Week 14:** Address 'Big Question', determine conclusions, decide on follow up questions to ask and answer.
- **Week 15 (+ Finals):** Compile final report, generate attractive figures, prepare presentation.