Overview

25% of your course grade will depend upon successful, **on time completion** of a project, to be done in groups of 2-3 students. You need to turn in a thorough but concise professional quality technical report of **at least 8 pages** (includes relevant tables and figures but not appendices) and presentation of key findings.

The final project is an analysis of a dataset using the methodology (models/algorithms) we have discussed in DATA-445. Your team is to find some "real-world" data, possibly of personal and/or professional interest to members of your team. Then, you and your team will use this data to perform a through analysis, and present your findings to the class.

Sourcing Data

Your data set should contain at least **500 observations and 8 variables** (if your data set has categorical variables, that is fine.) Your team will need to find a public dataset to analyze using AWS SageMaker. To collect data you may download files containing tabular data (such as .txt and .csv files) online, copy/paste from tables online or in pdf files, or perform some other data collection methodology (reading in a URL directly, manually aggregating Excel files, etc.).

As you are searching for data, feel free to help your classmates by sharing some of the websites where you found free/open data (sources that you may or may not have used.) In this project, students are allowed to use public datasets. However, **your analysis must be novel and not just a copy of what others have done. Copying results and presenting them as your own will result in a 0 on the project**. If you don't already have a dataset to use, listed below are a few suggested websites where raw data can be obtained, but you are welcome to use whatever data you find interesting.

- Health Data
- General Government Data
- General Social Survey
- University of California @ Irvine Machine Learning Repository
- Kaggle
- FiveThirtyEight
- Awesome-Public-Dataset-on-Github

Deliverables

Project Brainstorm: Due on October 15, 2021

Each team should meet with Oscar Aguilar between **October 11, 2021** and **October 15, 2021**. The purpose of this meeting to develop a plan for your project. Your team is expected to have at least two datasets that could be used for the project. All team members are expected to attend to this meeting.

Project Proposal: Due on November 5, 2021

Each team should submit a project proposal (pdf file or word document). This file should contain a description of the data set that you have chosen in a 2-3 pages summary (should include data source). An initial exploratory analysis and graphics of the data will be expected in this summary. A project proposal example is posted on Blackboard. This report should include at least:

- Dimensionality: number of rows and columns
- Description of what each row represents (e.g. a person, a location, etc.)
- Description of each column: possibly in the form of a formatted table containing each column name, column data type (or class), and a description that includes the meaning of values that a certain column may take.
- In case of missing values, a count of the number of missing values by column.
- Some graphs for numerical and categorical variables are expected (including description of them).

Project Report Feedback: Due on November 29, 2021

For this part of the project, you need to submit a preliminary and complete report and review another groups' preliminary and complete report with the goal of providing feedback. Before you write your final report, review the feedback given to you by your classmates. If you do not incorporate the feedback given to you, you will not get full credit for this portion of the project. Notice that your group has to submit the preliminary and complete report on **November 28, 2020 by 11:59 pm**.

Project Code: Due on December 5, 2021

Each team should also submit a finalized **.ipynb** file. Exclude earlier versions of your work. The program files should contain section headers and other comments to help with readability. The comments should make the reader aware of the main tasks associated with each piece of code. The project code is going to be divided as follows:

- Data Cleansing.
- Data aggregation if needed.
- Data summarization.
- Data visualization.
- Modeling

Report: Due on December 5, 2021

A pdf or word file that addresses at least the required components listed below along with a commentary explaining what was was done analytically and discussing the results.

- This document should not detail the code itself but should have notable visualizations such as formatted tables (e.g. a cross-tabs showing summary dataset information) and charts/plots. There is no need to include every visual produced in the report (a report example is posted on Blackboard).
- Your report should include at least:
  - A title for the work with contributors' names listed.
  - An abstract (goals and major findings).
  - A table of contents.
  - A description of the reason for your project/study.
  - What you did and how you did it.
  - A statement of the subject matter implications of your project.
  - A discussion of further questions raised by your project.

Presentation Slides: Due on December 5, 2021

You are required to create a separate file (such as PowerPoint or pdf slides) for your presentation to the class you may submit that as well. The most interesting visualizations (tables and charts/plots) should be included in your presentation. Your presentation as a group should last 15 minutes from which every single team member has to present for at least 3 minutes, and the last 3 minutes should be devoted to answer questions.

Grading

- Project Brainstorm (20 points)
- Project Proposal (30 points)
- Project Report Feedback (20 points)
- Project Code (50 points)
- Report (80 points)
- Presentation (50 points)