Final Project Overview

Purpose:

To answer a data driven question by deploying the data science lifecycle. At the end of the process, you will have an appreciation for what it is like to develop your own questions, match these questions with data, test the ability of the data via a model to answer your question, develop a final solution and the articulation of the weaknesses and benefits of the approach. The goal is for you to pull all the skills together that we have worked on all semester, essentially living the life of a practicing data scientist! This is also a great way to start developing a data project portfolio on your repo or website.

Tasks:

Work with your lab groups to develop and answer a discrete question related to a dataset of your choosing. Some dataset resources are listed below for you to potential use, but you are also welcome to use a dataset you have used in the past or one from the class repo that we have not used as part of the course. Present a cleanly documented final presentation (power point or Juypter Notebook) that walks the reader through your project step by step. This means you need to reference the data science lifecycle and work through each stage deliberately.

General topics we have covered that can be a focus of the final project (feel free to combine topics or extend them)

- Linear Regression for Prediction
- kNN
- Clustering - Kmeans
- Decision Trees (Classification or Regression)

Include the following sections:

- Question and background information on the data and why you are asking this question(s). References to previous research/evidence generally would be nice to include.
- Exploratory Data Analysis – Initial summary statistics and graphs with an emphasis on variables you believe to be important for your analysis.
- Methods – Techniques you are using to address your question and the results of those methods.
- Evaluation of your model – Select appropriate metrics and explain the output as it relates to your question.
- Conclusions – What can you say about the results of the methods section as it relates to your question given the limitations to your model.

- Future work – What additional analysis is needed or what limited your analysis on this project.

Create and develop a github repo for the content, work collaboratively.

Submit: Code, Data, link to repo and presentation

Criteria for evaluation:

- Oral evaluation will count for 70% of the final grade
- Documentation and quality of the materials will be 15%
- Collaboration and teamwork will be 15%