Final Project Guidelines -

The deadline for your Final Project is the Sunday after Finals Week - 12/17.

Final Project should have sufficient scope. That is, there should be **1** *primary question*, and at least **2** *secondary questions* to ask about the data. If we can't agree on that scope in your project, we may want to adjust it by changing the data set or adding some secondary data.

In an ideal world, you would make use of different types of methods, for example classifiers and clustering to answer these questions. But, this is not a requirement. If it turns out that different classifier approaches are best suited to your problem, then that is fine.

The Final Project has 3 parts: (1) A Github repository for your project. (2) One or more notebooks that incorporate all of your data analysis, including data preparation, exploratory data analysis, data modeling, and figure generation. (3) A final project paper in pdf format that is included in your repository. This paper is submitted when partially completed and then at the end of the project as detailed below.

The 3-5 page paper (maximum, including figures) that summarizes the analysis of a data set. I recommend you use either Google Docs, or Overleaf to collaborate on the paper, and share with me access to make comments as you are working on it.

The paper should include -

(1) A description of the data set and how/ where you obtained it.
(2) A table that summarizes the different variables and their data types. Identify any missing data, or other issues with the data organization (for example, unwanted observations and how you plan addressed it (removing the data, imputation, etc.).I
(3) Exploratory Data Analysis. Summarizes the main features of the data as they relate to your data analysis goals. Not every exploratory data analysis figure you make in the python notebook needs to be included in the final paper. You should include the figures that help you set up your answer.
(4) A statement of your data analysis goals and the hypothesis supported by EDA.
(5) Results of data modeling, supported by graphs, tables, and statistical measures.
(6) A summary paragraph.

You will sign up as a group for a Github repository for the class project. Please note that I will be archiving your class repository by January 2023, so I would recommend that each of you make a clone of the repository in your own accounts. Its good to have this stuff, because people (potential grad school advisors) look for this stuff.

The data analysis should be carried out in Jupyter notebooks, which should contain well organized, easy to read, and well explained code. Please make use of markdown boxes to

explain what you are doing, and break up your notebooks into subsections that carry out different analysis. Don’t carry around a lot of junk code.

Some recommendations of intermediate deadlines:

(1) EDA by November 29.
(2) Draft write up of 1-4 by December 3.
(3) Complete Jupyter notebooks with results of data modeling by Dec 10
(4) Finalize paper by Dec 17

Submission, including a pdf of your paper, jupyter notebooks, and a copy of the data should be in your github repository.