Data Analytics Programming (COMP 162/COMP 293A) Python Project Description

Purpose of the project:

- Apply the data analytics tools covered in class to a dataset of interest to you.
- Practice using Python to explore, visualize, and analyze data.
- Gain experience communicating your results in a presentation.
- Learn how to develop your own hypotheses about a dataset and test them.

Due dates:

- Project presentation: Tuesday, April 25th; Thursday, April 27th; or Tuesday, May 2nd in class. You can find your date here:https://docs.google.com/spreadsheets/d/16r0qW6DAWt5m9lvgy0ochA4Q6c7fRJtdmIUmImP6q3A/edit?usp=sharing
- Project write-up: Tuesday, May 9th at 5 PM

Presentation requirements (approx. 5 minutes per group)

If you have a partner, both of you should speak during the presentation.

- Introduce your dataset, and describe your motivation for analyzing it.
- Description of what each observation represents.
- Description of the variables that you'll analyze in the presentation.
- State at least two questions you had about your data.
- Show one plot to address each question.
- Explanation of at least two statistical analyses you performed, and your interpretation of the result of each one. These can include:
    - Correlation
    - T test
    - Linear regression
    - Ridge regression
    - Lasso regression
- Your overarching conclusions from your analysis.

Write-up requirements

- Submit your write-up as a Jupyter Notebook file with all text, code, output, and plots for the project.
- *From Homework 5:*
    - A link to your dataset.
    - Description of what each observation of the dataset represents.
    - Description of what each variable in the dataset represents. If your dataset includes many variables, you can just describe the most relevant ones (at least eight).
    - Description of your motivation for analyzing this dataset.
- *From Homework 6:*
    - At least five questions that you could try to answer with your data.
    - A plot addressing one of these questions. Include your interpretation of the plot.

    - **Extra credit for all students**: Check your dataset for columns with outliers that you may want to filter out. Note down any columns that you plan on filtering on, and the values that you will filter out.
- *From Homework 7:*
    - At least one t test on your dataset. Include your interpretation of the result.
    - At least one linear regression on your dataset. Include your interpretation of the result.
    - Choose one quantitative variable in your dataset. Try to predict its value based on all other quantitative variables in your dataset:
        - Split the data into a train and test set.
        - Train a linear model using either `LinearRegression()`, `Ridge()`, or `Lasso()`.
        - Find the "score" of the prediction on the training and test set.
        - Find the coefficient corresponding to each variable in your model. Which variable had the biggest effect on the prediction?
        - Would you recommend your model to be used for this prediction in the future?
- *Additional requirements:*
    - Include at least one more plot addressing one of the five questions you brainstormed.
    - Train a logistic regression model on your data. You can create a binary column yourself if one does not exist in your dataset. Try to predict this binary column based on all the other quantitative variables in your dataset (not including any variables used to create the binary column):
        - Split the data into a train and test set.
        - Train a Logistic Regression model.
        - Find the "score" of the prediction on the training and test set.
        - Find the confusion matrix for your model.
        - Find the coefficients corresponding to each variable in your model. Which variable had the biggest effect on the prediction?
        - Would you recommend your model to be used for this prediction in the future?
    - Include a conclusion, summarizing what you have learned about your dataset.
    - **Challenge problem (required for 293A, bonus for 162):** A third statistical test (correlation, linear regression, or t-test) and your interpretation of the result.