**Data Appendix**

**Due: 11:59pm, March 19, 2021 (Friday)**

**15 points**

The first step for the final project is to find a data set that you are interested in working with. In general, you may choose a topic that is interesting to you. Try to pick something that is original and worth studying. Part of your overall grade will be based on originality.

1. Find a data set that meets these requirements:

Minimum requirements of the data set:

At least 500 observations with at least 5 variables (excluding Name, unique ID, datetime or something similar).

At least 3 quantitative variables (excluding datetime such as 2021-01-25 08:15:36)

At least 2 categorical variables. (It is okay if all the variables are quantitative, since they can be converted into categorical variables.)

No times series data (that is observations observed over time with no other information)

No summary data

Your raw data file(s) should ideally be in the CSV format (.csv). Excel data can easily be converted to .csv by using “Save As” or “Export” and choosing “comma separated values” as the format. If your data is in another format, talk to your professor about converting the data or using a different command for reading it in. It is okay if you have multiple data files.

1. Read your data into Python. Show the head of the data. If you have multiple files, do this for each file. You don’t need to combine them for this assignment.
2. Data: Write a short description of the data. What is the dataset about? How much observations does it have? Give the data source (e.g. URL of the data)
3. Questions to study: Think about some potential questions that you hope to answer about the dataset. Clearly state your research questions.

e.g. How can we predict housing prices in Connecticut?

How is the admission rate to four-year colleges affected by demographics and school characteristics?

What factors influence whether a person will get heart disease?

1. Variables: What are the response variables and the explanatory variables related to the questions you have in part 4?
2. Submit both your .ipynb file, the .html file and the csv data file on Moodle. Please change the name of your file(s) to something like “group-X-data” + file extension. Each group only needs to submit one copy.