Problem Set 4

To help you brainstorm for your final projects, complete the following questions. Please document your work throughout when you complete these questions. Journal what you do throughout completing the problem set in a Word document, for example, and submit a PDF document on Brightspace.

Your answers depend on how well-developed your final project idea is. If you already have a well-developed topic, you can write just one topic (for question 1) and one dataset (for question 2), and so forth, but then I want more detailed elaboration on each step. If you do not already have a well-developed topic, I want to see attempts to list several possibilities in each question and work on sorting through them. This is understandably an open-ended and somewhat ambiguous problem set, but it is important for you to complete it with decent effort so that we can provide feedback, and then you can make progress toward completing the final project for the class.

Note: what you submit for this problem set will be shared with other students for in-class group discussions.

1. Consider one or more general topics of interest to you that could plausibly be studied with large data ("large" means 100 rows or more). Examples: "real estate sales prices and related variables," "stock price growth/decline by company during March 2020," "lobbying by the energy sector in federal government," "market concentration in specific industry X over time."

2. For each topic that you listed in question 1, search the internet and library website to identify a dataset that you have access to that is related to the topic. Discuss the variables that are (at least potentially) available in the dataset. The dataset could require scraping (e.g. if it is on tables on various websites) or it could be in a flat file (e.g. csv, excel file). If you cannot find relevant data for your topic, explain what makes it difficult. If possible, modify your topic in order to better match it to available data. If you encounter insurmountable dead-ends, say so, but try to find data for as many of the topics you wrote in question 1 as possible.

For example, if my topic is "real estate sales prices and related variables," and I search, I can find data on New York City sales transactions for the last 12 months https://www.nyc.gov/site/finance/taxes/property-rolling-sales-data.page, which has sale price, information about the building class, year built, tax class, land area, and the number of residential units in the building, at the time of sale.

3. For each topic, modify the topic and elaborate it, informed by the variables in the dataset.

Following the real estate example: Looks like the dataset has sale prices, year built, and building classes. I checked out the building classes, finding that with some data work I can separate condos, coops, and one family dwellings. I'm going to focus on dwellings with 1 residential unit (or blank for residential unit, which appears to be the case of coops). Then I will ask: what is the relationship between year built and sale price for the three kinds of residential properties – condos, coops, and other residential? I am curious which of these different types of property depreciate faster or slower (in my report, I can cite references on the business concept of “depreciation”). I might expand my analysis to rental apartment buildings, too, asking whether they deprecate faster or slower than owner-occupied properties.

4. For each topic and dataset, discuss how you would load the dataset into python. If you do not have access to the tools to load it yet, say so and discuss.

For my residential property example, I am able to download the data from the government website as Excel files, and load it using the following code,

```
import pandas as pd

sales_manhattan =
pd.read_excel('rollingsales_manhattan.xlsx',skiprows=4)
```

I can load the data for each borough separately, then stack them using `pd.concat`.

5. Make an informed judgment about which topic and dataset seems most fruitful for you to use going forward. Explain your choice.