Shrina Parikh
NetID: sp4681
Assignment 3

URL to data:
https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh/data

The dataset lists details about legally operating businesses in New York. It includes information about the license number, status, creation and expiration dates. It also includes information about the business name, address, and some property and geographic information (BBL code - borough,block,lot and coordinate points). A lot of the fields are empty in the original dataset. Additionally, I eliminated a number of fields from the original dataset. I have attached my python script in the zip file as well as the cleaned version of the data.

From running my SQLite queries, I learned a lot about the businesses in NYC. One thing I found particularly interesting was the range of addresses signed on these licenses. There are so many businesses operating in NYC in which the business operator signs an address from outside NYC. A lot of NJ addresses are signed, for example, which is expected. However, some other places listed within the Address_State column are Costa Rica, Metro Manila, India, and Mumbai.

From the data results, I learned that Brooklyn is the borough with the most DCA licenses. However, many of the borough fields were left empty, and many were listed as “Outside NYC,” so I am not sure if that is actually true. I also learned that, at least in this dataset, the industries with the most licenses are Home Improvement and Tobacco. I found this interesting, and I am curious to look into if the number of tobacco licenses being given out has decreased over time, or if a lot of them are inactive since people smoke less now.

I do not know how reliable this data is. Although the source is reliable, there is so much data missing that it is hard to tell if my conclusions based on queries are actually true or not. There are also many typos. For example, query 8 helps me identify some of the problematic zip codes. The frequency of these errors and irregularities is small given the size of the dataset, but still worth questioning and investigating further. Moreover, there are 73,758 entries with an empty Address_Borough field. Thus, I think it is unfair to make conclusions based on this data.

I was not expecting to find some of the results I did, however. For example, I did not expect so many of the addresses listed on the licenses to be from outside of the tri state area.