DSC 540 Final Project

For this project I wanted to take a look at data that may influence or correlate with crime. In a personal context I believe that two of the things that contribute to the occurrence of crime is the environment and difficulties of daily life such as commuting time and adequate income for day to day life. The measures I aggregated were weather data from the Global Historical Climate Network(daily) sourced from the NOAA(flat file); Census data from the American Census Servey(1-Year) provided by the United States Census Bureau(API); Crime reporting from Crime in the United States yearly report provided by the United States Department of Justice(website scrubbing). The biggest challenge I faced was with the Global Historical Climate Network(daily) dataset.

The challenge with the Global Historical Climate Network(daily) primarily was due to the quantity of data. Originally, I went through the FTP download and obtained a 32GB file. I had 16GB of RAM at the time. Loading this file into a Pandas DataFrame was simply not feasible. I ended up upgrading to 64GB of RAM but this still was insufficient. I ended up loading the data into an SQLite Data Base in order to pair it down to manageable sizes. I later found that a “supercharged "version of the data was available. This version turned out to be over 100GB. I eventually was able to load chunks of the file in process it and then append it to a CSV file to overcome the RAM bottleneck. I was then able to clean the data. The next challenge was with the Census API.

The Census is even more extensive in size than the Global Historical Climate Network(daily). The problem however was not file size this time but finding the right information. The American Census Survey(1-Year) has an unfathomable number of variables. Each variable is encoded into a fixed width string with a jumble of numbers and letters. I was able to find the right information after locating a variable name to variable definition table. Even then it was a case of sifting through a massive table from a few grains of usable data. Once the variables were located, with the help of a US Census API specific library, I was able to pull the data that was of interest. After having working with an API previously, that portion was relatively trivial. The next data set was obtained via website scrubbing. This did not have many sticking points, so I won’t go into detail. The final portion of the project was uploading the data to a database.

Perhaps the most technically challenging step in the project was the SQL portion. Having previously worked with SQLite I decided to try out MYSQL. SQLlite can establish its’ own server through PyCharm, however I was unable to replicate this for MYSQL. I did eventually get this up and running. After that came the python to SQL work. It took a significant amount of time before I realized that Pandas had a to_sql() function. I had previously looked for MYSQL specific libraries. I had only found libraries that made the process more complicated than writing in actual SQL. Using SQLalchemy I was able to establish a connection to the MySQL database. From there I had some trouble with syntax errors and duplicate names on columns. After overcoming this hurdle, the data was finally joined and ready for use. I chose to work in Tableau for the visualization portion.

Tableau was an absolute pleasure to learn and work with after all the SQL issues. With a GUI that did the heavy lifting, the code that I had to write was little more than basic arithmetic for some of the variables to correctly apply in context. During the visualization portion of the project I learned quite a bit.

Overall, the project was very beneficial to my understanding of MYSQL, SQLite, API calls, Tableau and even some more unique python use cases. I had a fair share of difficulties during the project. After

taking some time to understand what an error was indicating and what may have caused it, my debugging skills were significantly elevated. This project was very helpful for my development as a Data Scientist.