## **DSCI 591 Final Project Readme**

This project was built by Raj Patel, Hong Son, and Kunal Sharma

Drexel University, Fall Quarter 20/21, DSCI 591 Data Science Capstone I

**Program Operations (Follow along in .ipynb)**

1. Data Loading & Pre-Processing

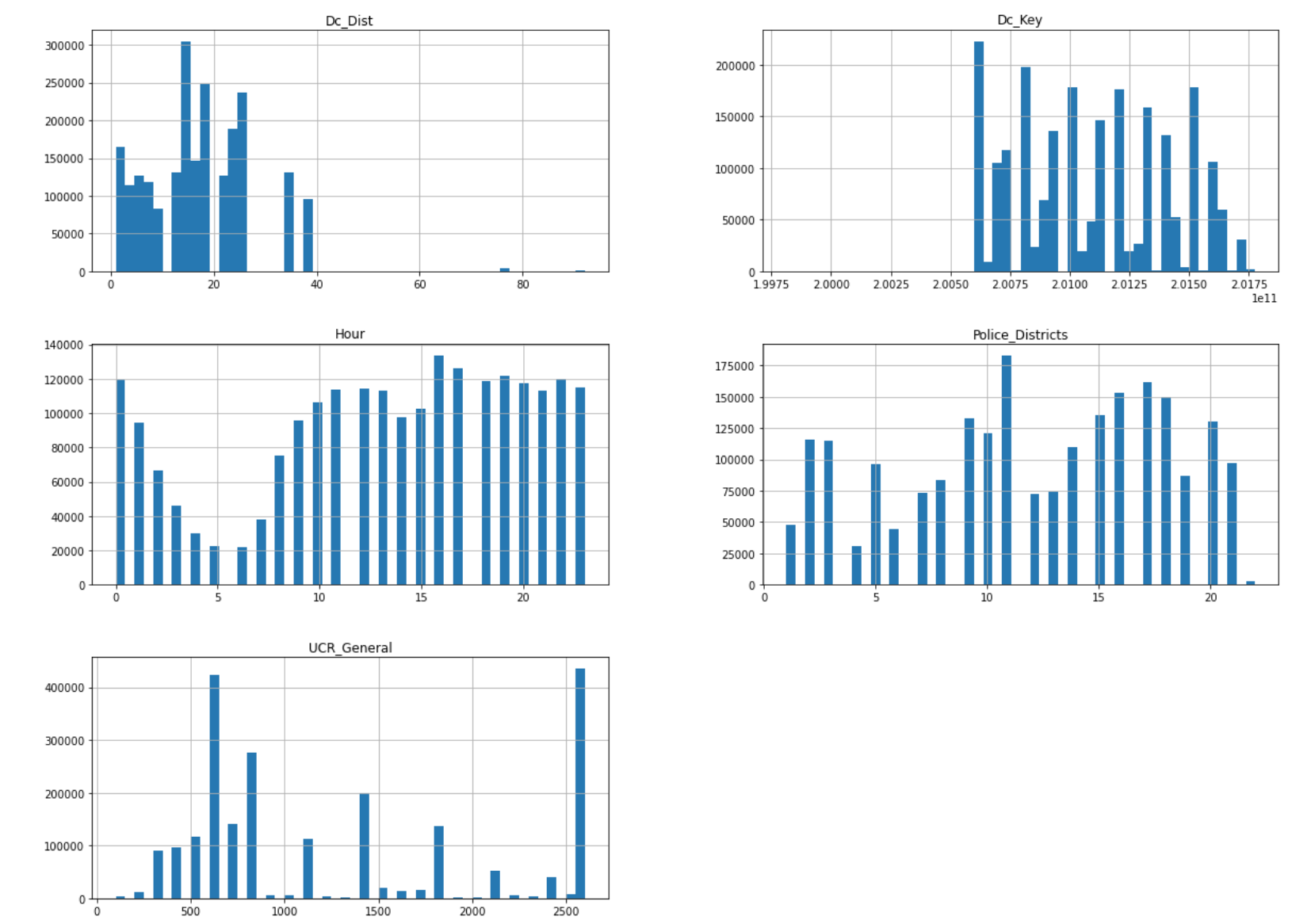
The following packages will be used in our analysis of the SBA Loan dataset:

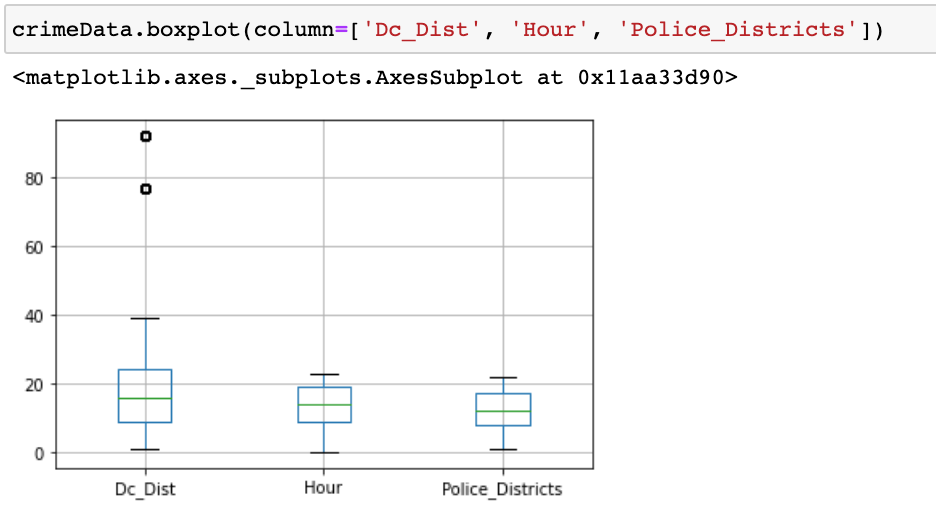
* import pandas as pd
* import numpy as np
* from matplotlib import pyplot as plt
* import seaborn as sns
* import os
* from scipy.stats import f\_oneway
* import plotly.express as px
* import datetime
* import googlemaps
* import folium

We first read in the file and review the first 10 rows. Next, we look at the data types for the columns. After making sure that the correct data types were used, we looked to see if there were any missing values. For missing values on the columns, we mainly just removed the rows with missing values [1]. They make up a small portion of the entire dataset (<1%). We finally created a column called events that lets each row be an event.

2. Descriptive Statistics

We looked at the summary statistics for the numeric columns ['Dc\_Dist', 'Hour', 'Dc\_Key', 'UCR\_General', 'Police\_Districts']. We plotted histograms and boxplots for each respective column.



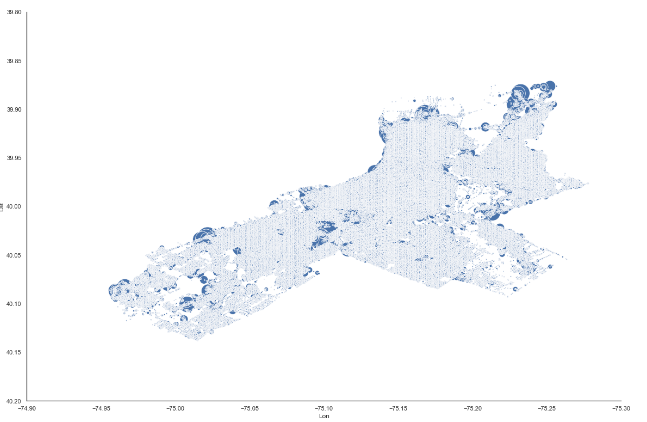


3. Visualizations

We created a dataframe called ‘groupCrimeLocations’ by grouping the data by the locations and counting the number of events for each location.

We created a dataframe called ‘groupLonLat’ by grouping the data by the latitude and longitude and counting the number of events.

We first plotted a graph through geolocation data to get a sense of areas/regions with high numbers of crimes.

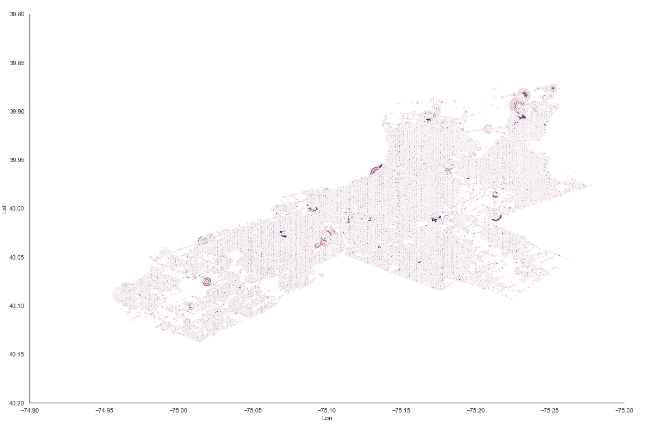


Findings: The above graph is created by grouping the geo locations from our data set where the crimes have occured. We found that crimes have occurred throughout Philadelphia in the past however, the larger blobs in the graph above like the ones we see in northern Philadelphia are the locations where the crimes have occured the most.

We created the column ‘year’ by retrieving the year from ‘Dispatch\_Date’

We created a dataframe called ‘groupLonLatYear' by grouping the data by the latitude, longitude, and year and counting the number of events.

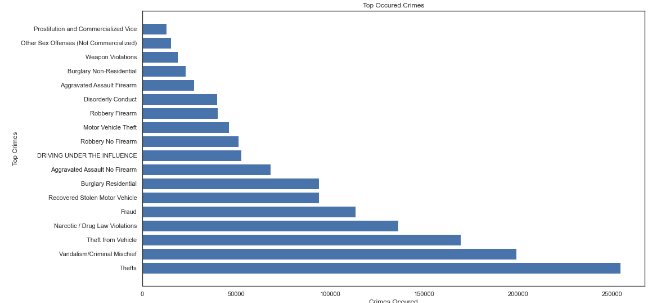
We plotted a graph through geolocation and year data to get a sense of areas/regions with high numbers of crimes through the years.



Findings: Here we have grouped the data set by the location and year and the events are more distributed in the graph above and findings are similar to the once we found above.

We created a dataframe called ‘topCrimes’ by grouping the data by the General Code and counting the number of events.

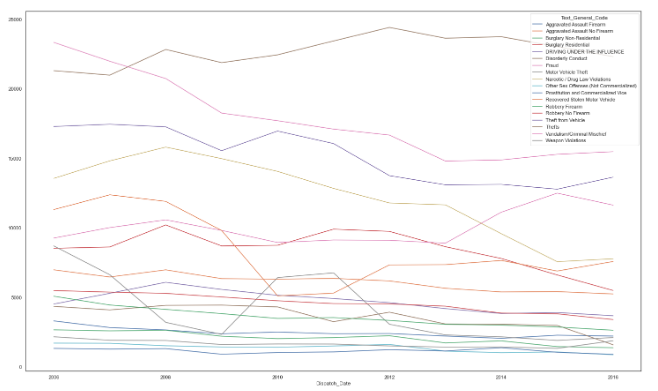
We produced a graph to visualize each crime occurred each day over last 10 years



Findings: We found the crimes that have occured the most in the past from our data set. The top 5 kinds of crimes being Theft, Vandalism/Criminal Mischief, Theft from Vehicle, Narcotic / Drug Law Violations and Fraud.

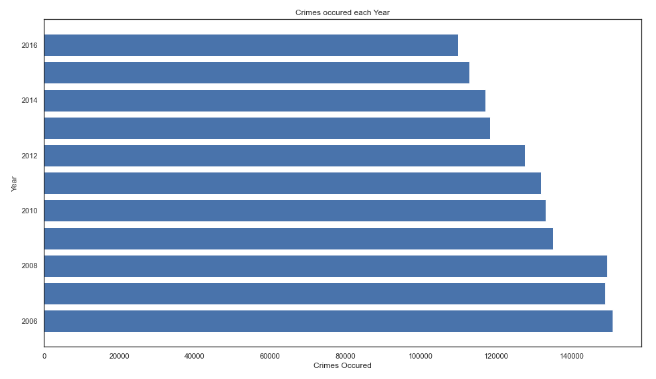
We created a dataframe called ‘topCrimeCasesByYear’ by grouping the data by the Dispatch Date and counting the number of events.

We produced a graph to show the number of crimes for each type over a 10 year period.



Findings: The graph above shows the top crimes in philadelphia and the total number of times it has occurred in the past years from the data set. We found that crimes that have inclined in numbers of occurrence like Thefts, frauds etc and the crimes that declined in numbers occurrence like criminal mischief and motor vehicle theft.

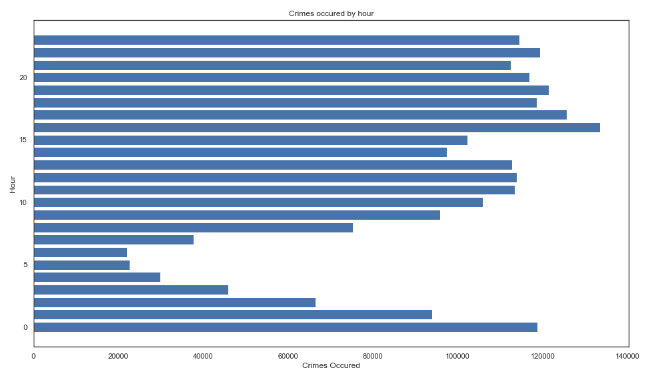
We pivoted the data by summing the events for each year to produce a graph to visualize the number of crimes per year.



Findings: The graph above shows the total number of crimes that have occurred each year in the past from our dataset. We found that the number of crimes occured in the past years have declined.

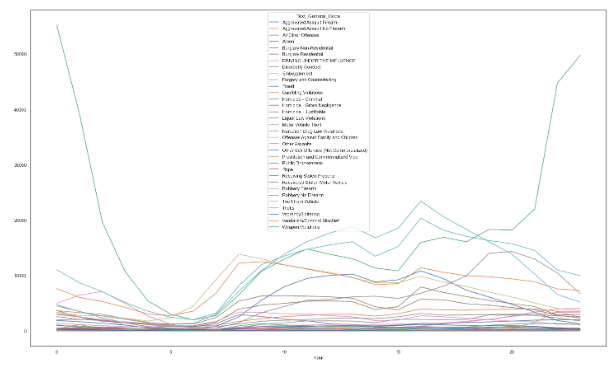
We created a dataframe called ‘crimeDataHour’ by grouping the data by hour and summing the rows

We produced a graph to show the number of crimes occurred by hour over last 10 years



Findings: The graph above shows the occurrence of crime by the hour in the past years from our dataset. We found that the least number of the crimes that have occured in the past is between 3am-6am and after that the numbers go up and are at its peak at 4pm.

To see hours that crimes happen, we constructed a graph that shows the hours with the types of crimes and the counts.



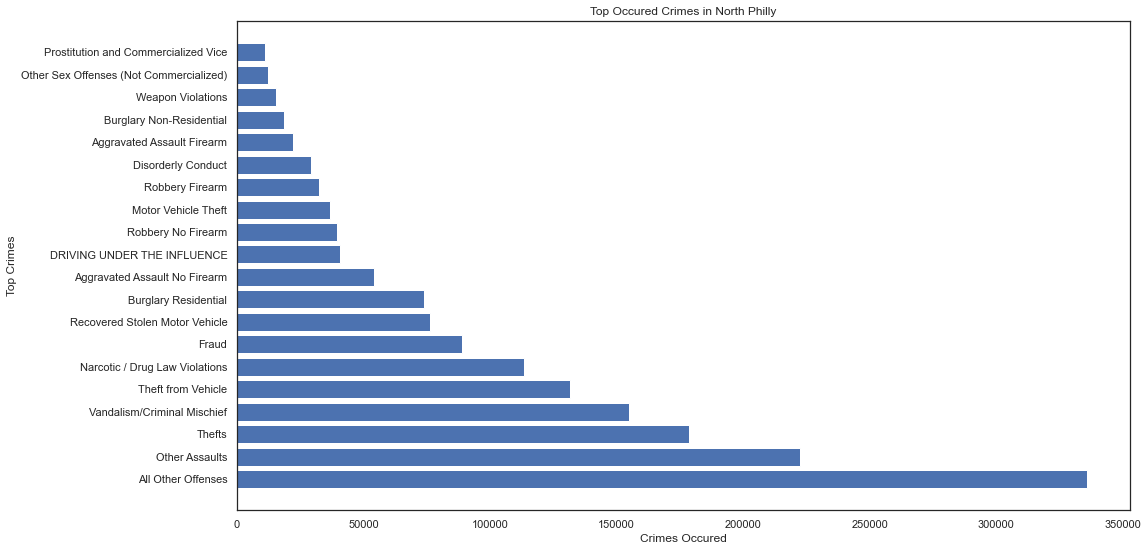
Findings: The graph above shows different types of crime that have most occured in Philadelphia by the hour in the past years from our dataset. We see a similar pattern for most of the top crimes. We found that the least number of crimes that have occured in the past is between 3am-6am and after that the numbers go up for every crime.

Comparing data from different parts of Philly

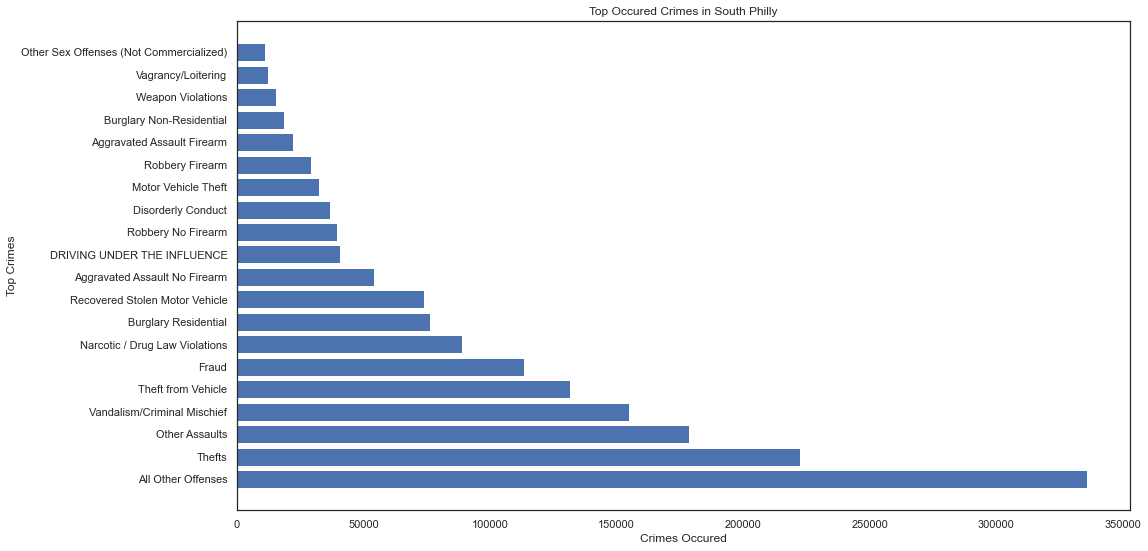
We then use lat coordinates to split the data for North, South, and West Philly

We looked at the crimes per hour, the types of crimes by hour, and top occured crimes

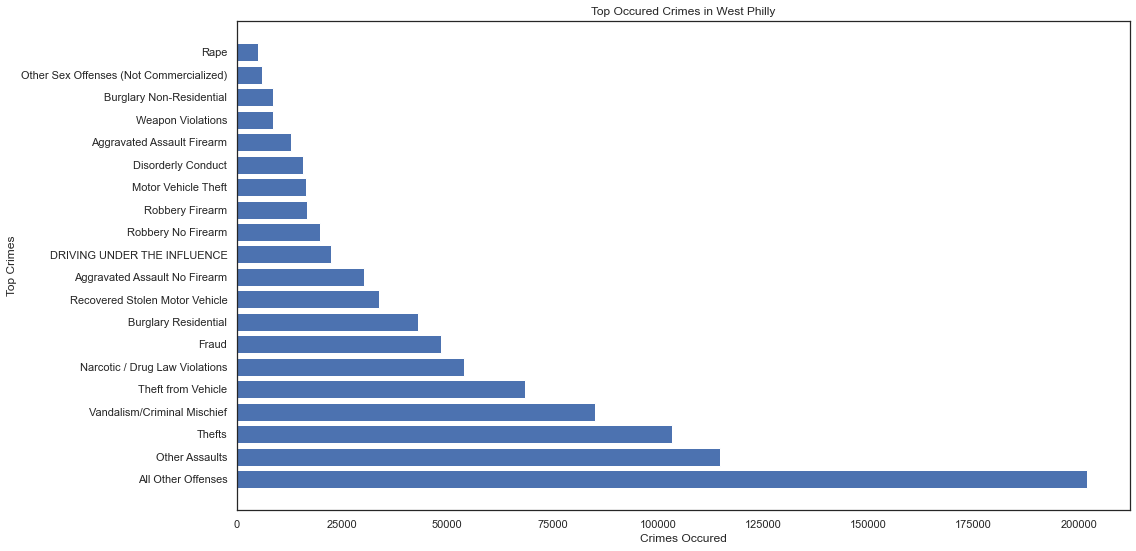
1. North Philadelphia



2) South Philadelphia



3) West Philadelphia



Findings: We found similar pattern in the crimes that have occured in all three region

4. Statistical Analysis: Anova Test

For this section, we wanted to use the ANOVA test to compare the crime counts among north, west, and south Philadelphia.

First we totaled the number of cases for each religion,

Our null hypothesis (Ho) is that the Average crime count is the same across regions

Our alternative hypothesis (Ha) is that the Average crime count is not the same across regions. Our alpha value we used is 0.05.

Our F value came out to be 36641.862962362226. P-value is 0.0. From the results, we rejected the null hypothesis as p < 0.05.

We also visualized he result using a CDF plot



We can see that there is a higher probability for crimes in North Philly compared to South and West Philly. South Philly is less likely to have 200 crimes. This graph also shows that Average crime counts are not the same across regions.

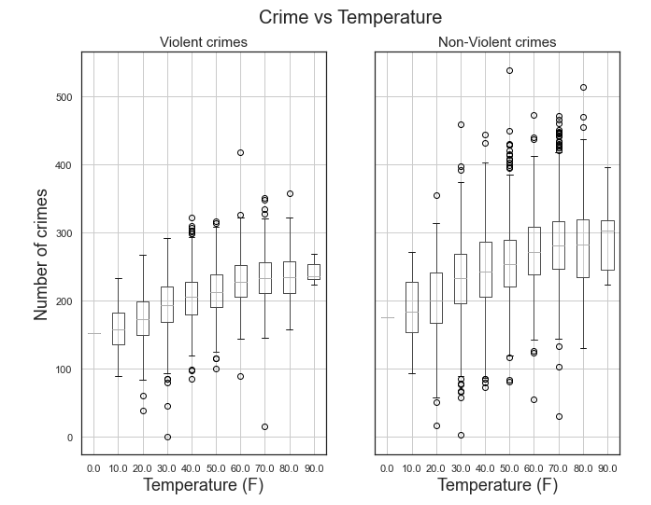
5. Weather and Seasonal Crime Analysis

To assist in getting more out of our analysis, we will be scraping Philadelphia weather data for Philadelphia from 2007 to 2018. The dataset includes the temperatures in Fahrenheit, the max and min temperature, environment information, wind, and precipitation. We hope to look into the seasonal forecast and see if there are weather trends in crimes.

For our weather data, we were missing values for total precipitation but that made sense as missing values indicated no precipitation. Values of TRACE were converted to the value 0.001 to indicate that there was very little precipitation but not that there wasn’t any.

We created a dataframe called ‘ts’ which counted the number of violent crimes, Non-violent crimes and the binned temperature. Violent crimes are crimes where the UCR code is greater than 800. Non-violent crimes are crimes where the UCR code is less than 800. An example of a binned temperature is the temp 85 would be counted as 80.

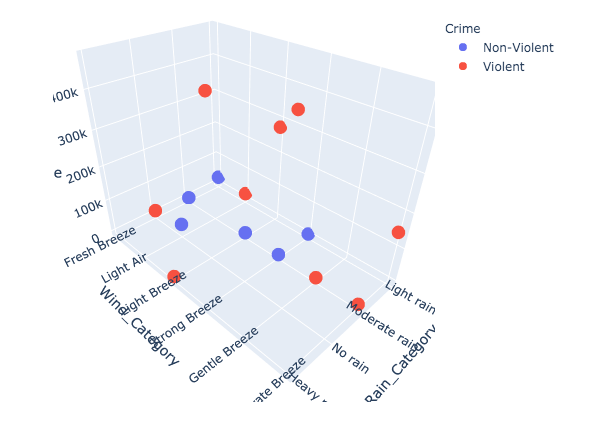
We created two plots w/ Temperature and Number of Crimes separated by type of crime



The mean of the crimes is increasing with temperature. Seasonal changes can affect the crime rate (crimes are high during summer season - we will confirm in Seasonal Analysis).

We then categories the avg wind speed and total precipitation into their own classes (example - Avg Wind Speed >= 0 and Avg Wind Speed <= 3 is light air)

We then plotted a graph that looks at the number of crimes based on the wind and precipitation categories.

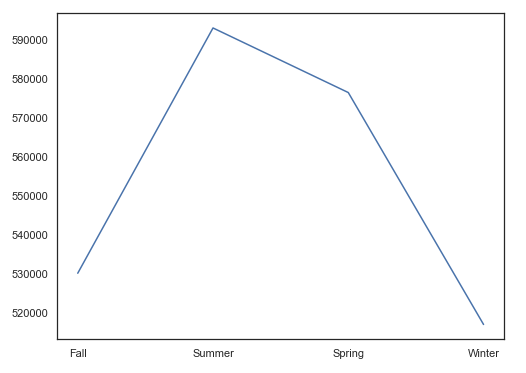


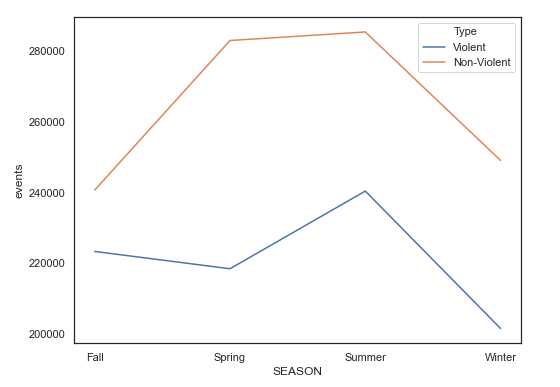
What’s interesting about this visualization is that for violent crimes, they mostly occur when it is either little rain or no rain and little wind. For Non-violent crimes, they occur when it is lightly raining.

For our seasonal analysis we first grouped the rows based on the season they were in by converting the dates into the seasons they are associated with.

We created a pivot table based on the counts by seasons dn type.

We produced the following visualization based on the counts:

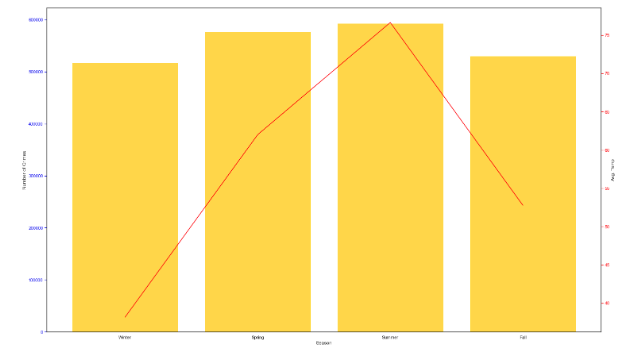




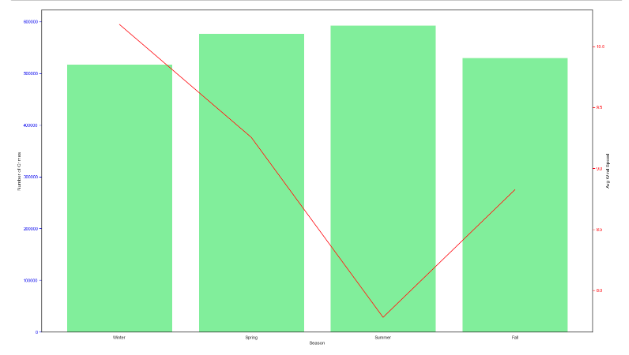
We see that the number of crimes is the most in the Summer, followed by the Spring, Spring, and then Winter.

After we group the data by the seasons and took the aggregate for the following columns - Avg Temp (mean), Avg Wind Speed (mean), Total Precipitation (mean), Events (sum)

We produced the following graph based on season, number of crimes and the respective weather columns (Avg Temp, Avg Wind Speed, and Total Precipitation)



We see that the trend in temperature can be seen in the seasons.



We don’t see much of a trend with wind and crime counts

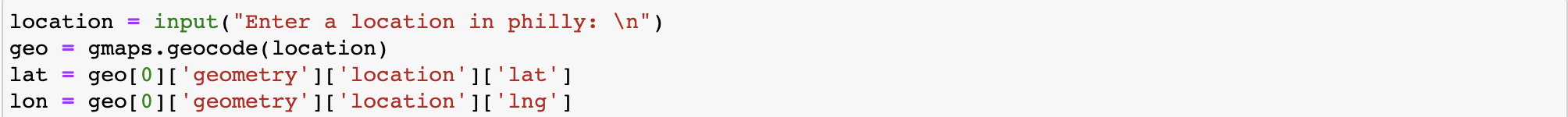


We don’t see much of a trend with wind and crime counts

6. Machine Learning

To gain a probability that a crime might occur given a user specified location and day of the year we have performed the following.

We are using Google’s cloud API to gain the geolocations of any user specified location within Philadelphia in real time.



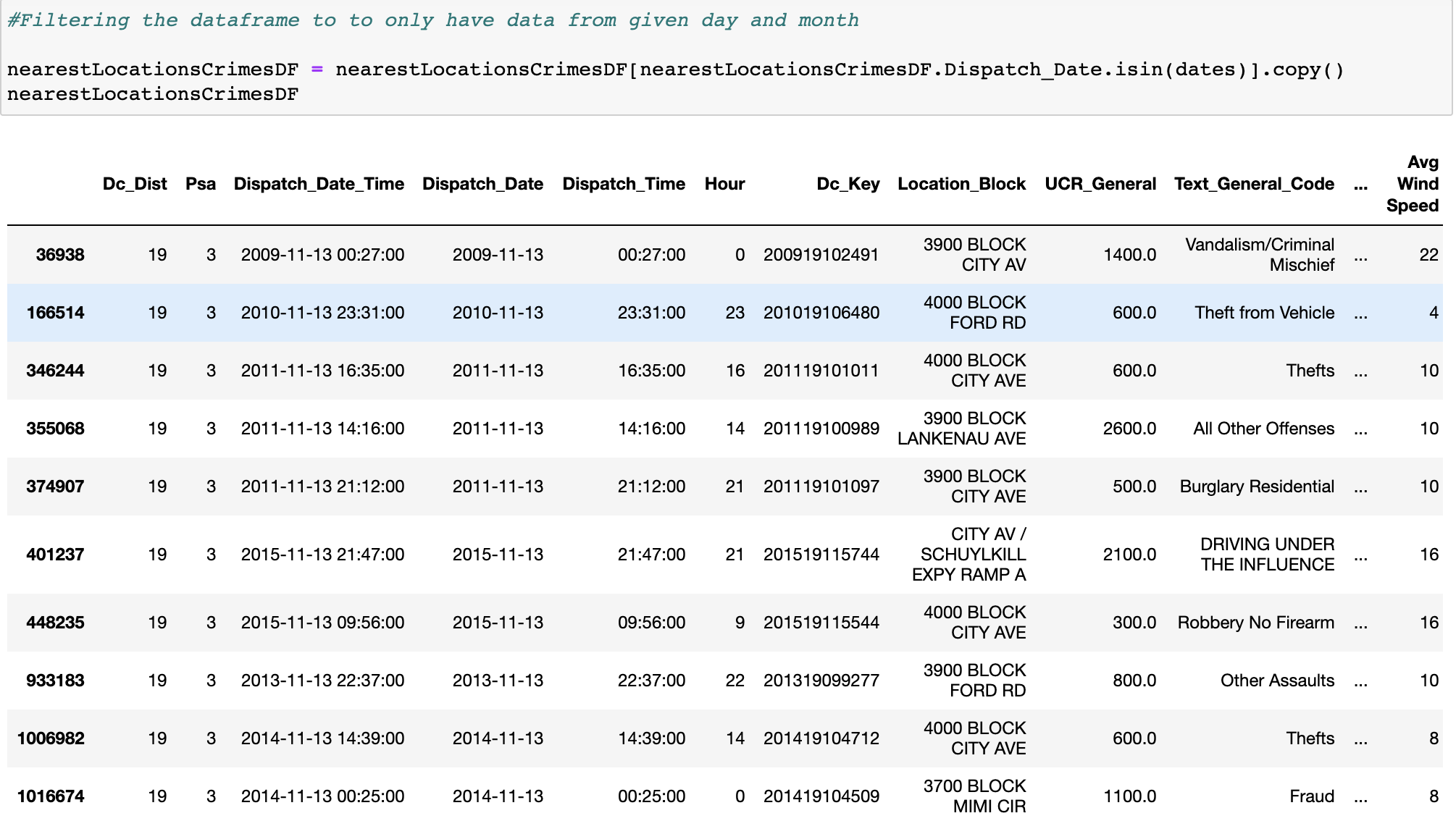
Philadelphia has more than 500,000 different points of geo-location and therefore once we gained the geo-location above we are also considering 500 nearest locations from the crime data set. To gain the nearest locations we are using k nearest neighbor from sklearn. [4]

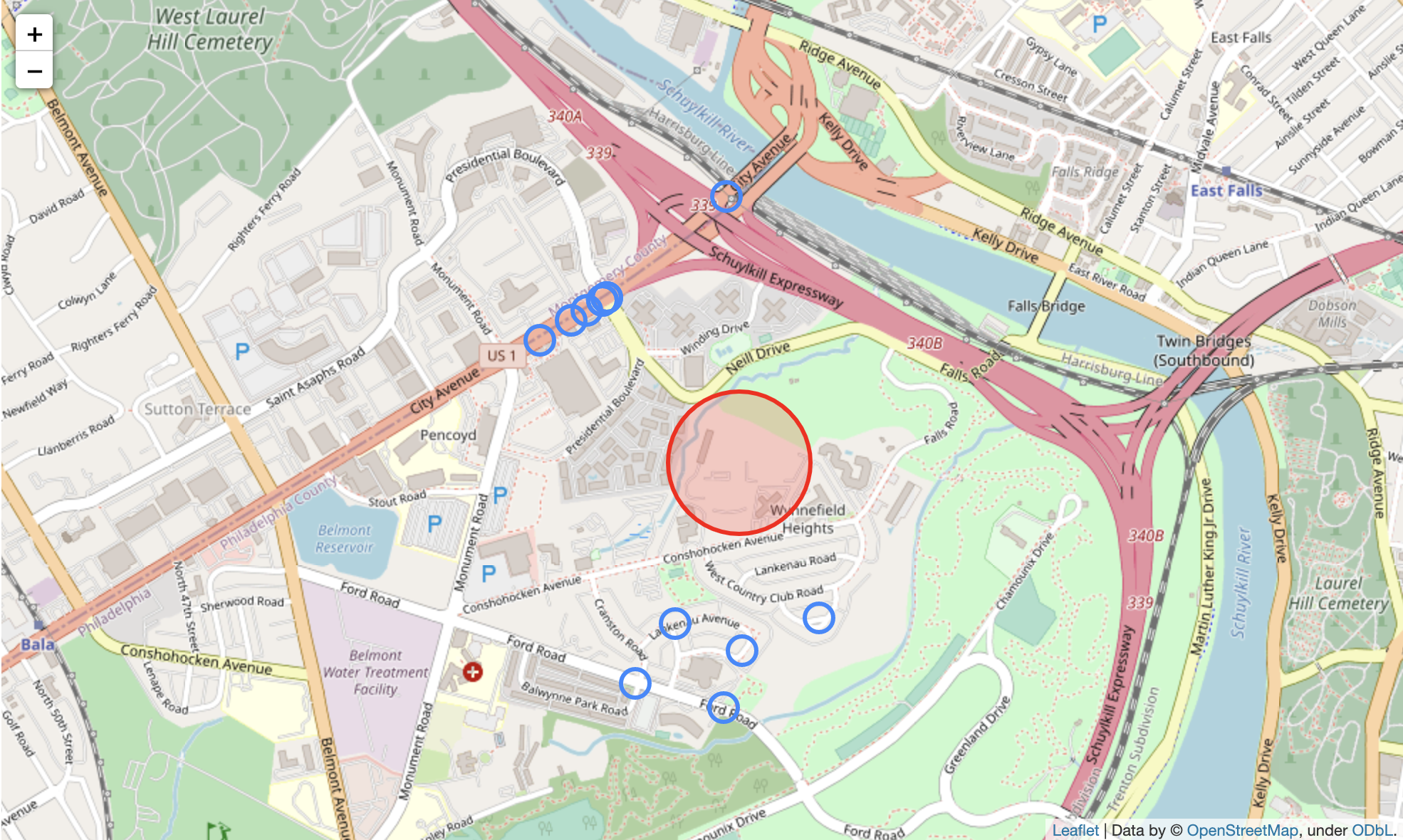




The above image contains the location that we focused on with a red marker and the nearest 500 different locations where crimes have occured in the past years from our dataset.

Since we want to focus only on the user specified day of the year, we filtered our dataset to have only the day (eg: Nov 11th) for each year in the dataset that we are using and the nearest locations from above.





The image above contains only the nearest locations where the crimes have occurred on a user specified date (Nov 11th in this case). In the filtered data frame above based on user specified day and location, we have also achieved essential data about the weather, wind speed, type of crime (violent or non-violent).

Finally, to gain a probability that a crime will occur on a user specified data and location we get the total number of years when a crime has occured in our dataset and divide it with the total number of years from the dataset.

