FOUNDATION OF DATA SCIENCE
CSGY-6053

PROJECT REPORT:
FORECASTING NYC CRIME DATA

Harsh Sonthalia – hs4226
Jacob Fishman – jf4322
Ujjwal Vikram Kulkarni – uk2011

Introduction

We have all moved from far and wide to live in the greatest city in the world. With all of the excitement from moving comes worries about safety, more importantly in terms of crime. There are plenty of statistics that show that crime in New York city is concerning, from the elevated violent crime rate to crimes per square mile in New York City totalling 1,549 compared to the national average of 28.3.[1] The goal of this project is to model the time series of crime trends in New York City, and forecast the crimes for the next year, classifying these predictions to better help with preventive measures for the police department and local citizens.

Pre Processing

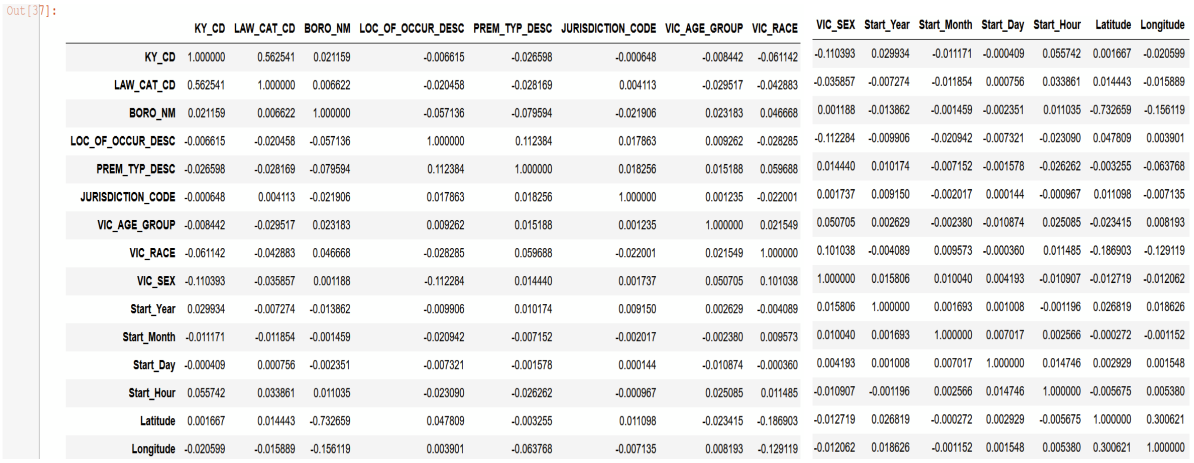
We are using the data set from NYC Open Data Website:

https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

This dataset provides 7.38 Million instances of crimes, all of the crimes that have been reported from 2006 onwards, with 35 features. We have gone through the dataset and removed any duplicate or unnecessary features, which has reduced our features to 28. For the rest of the preprocessing we will use the data exploration to determine the factors that are relevant to our predictions.

Data Exploration

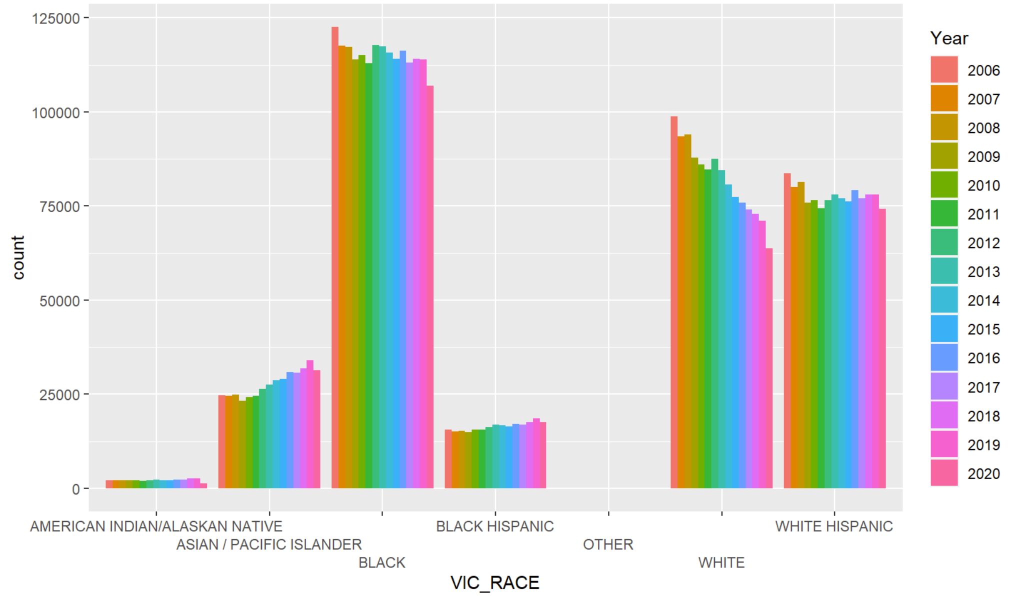
1) Here is the Pearson correlation matrix for all of our features that we deemed to be helpful in understanding crime data. You can see that there are no features that are strongly correlated. The highest absolute value is between Key ID and Law Category ID, both of which should be correlated since they deal with the classification of the crime itself.



Out[37]:

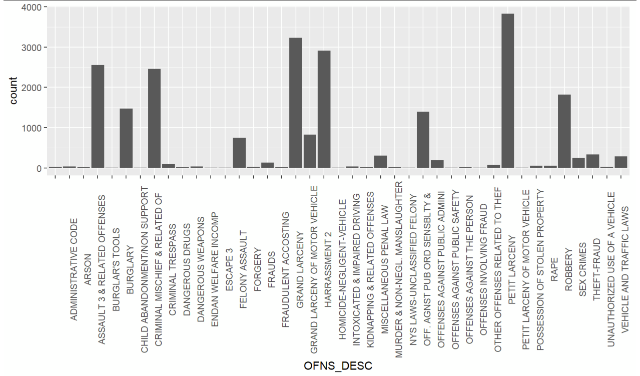
	KY_CD	LAW_CAT_CD	BORO_NM	LOC_OF_OCCUR_DESC	PREM_TYP_DESC	JURISDICTION_CODE	VIC_AGE_GROUP	VIC_RACE	VIC_SEX	Start_Year	Start_Month	Start_Day	Start_Hour	Latitude	Longitude
KY_CD	1.000000	0.562541	0.021159	-0.006615	-0.026598	-0.000648	-0.008442	-0.061142	-0.110393	0.029934	-0.011171	-0.000409	0.055742	0.001667	-0.020599
LAW_CAT_CD	0.562541	1.000000	0.006622	-0.020458	-0.028169	0.004113	-0.029517	-0.042883	-0.035857	-0.007274	-0.011854	0.000756	0.033861	0.014443	-0.015889
BORO_NM	0.021159	0.006622	1.000000	-0.057136	-0.079594	-0.021906	0.023183	0.046668	0.001188	-0.013862	-0.001459	-0.002351	0.011035	-0.732659	-0.156119
LOC_OF_OCCUR_DESC	-0.006615	-0.020458	-0.057136	1.000000	0.112384	0.017863	0.009262	-0.028285	-0.112284	-0.009906	-0.020942	-0.007321	-0.023090	0.047809	0.003901
PREM_TYP_DESC	-0.026598	-0.028169	-0.079594	0.112384	1.000000	0.018256	0.015188	0.059688	0.014440	0.010174	-0.007152	-0.001578	-0.026262	-0.003255	-0.063768
JURISDICTION_CODE	-0.000648	0.004113	-0.021906	0.017863	0.018256	1.000000	0.001235	-0.022001	0.001737	0.009150	-0.002017	0.000144	-0.000967	0.011098	-0.007135
VIC_AGE_GROUP	-0.008442	-0.029517	0.023183	0.009262	0.015188	0.001235	1.000000	0.021549	0.050705	0.002629	-0.002380	-0.010874	0.025085	-0.023415	0.008193
VIC_RACE	-0.061142	-0.042883	0.046668	-0.028285	0.059688	-0.022001	0.021549	1.000000	0.101038	-0.004089	0.009573	-0.000360	0.011485	-0.186903	-0.129119
VIC_SEX	-0.110393	-0.035857	0.001188	-0.112284	0.014440	0.001737	0.050705	0.101038	1.000000	0.015806	0.010040	0.004193	-0.010907	-0.012719	-0.012062
Start_Year	0.029934	-0.007274	-0.013862	-0.009906	0.010174	0.009150	0.002629	-0.004089	0.015806	1.000000	0.001693	0.001008	-0.001196	0.026819	0.018626
Start_Month	-0.011171	-0.011854	-0.001459	-0.020942	-0.007152	-0.002017	-0.002380	0.009573	0.010040	0.001693	1.000000	0.007017	0.002566	-0.000272	-0.001152
Start_Day	-0.000409	0.000756	-0.002351	-0.007321	-0.001578	0.000144	-0.010874	-0.000360	0.004193	0.001008	0.007017	1.000000	0.014746	0.002929	0.001548
Start_Hour	0.055742	0.033861	0.011035	-0.023090	-0.026262	-0.000967	0.025085	0.011485	-0.010907	-0.001196	0.002566	0.014746	1.000000	-0.005675	0.005380
Latitude	0.001667	0.014443	-0.732659	0.047809	-0.003255	0.011098	-0.023415	-0.186903	-0.012719	0.026819	-0.000272	0.002929	-0.005675	1.000000	0.300621
Longitude	-0.020599	-0.015889	-0.156119	0.003901	-0.063768	-0.007135	0.008193	-0.129119	-0.012062	0.018626	-0.001152	0.001548	0.005380	0.300621	1.000000

2) The next figure shows the relation between total crimes and the victim's reported race, grouped by year. Here we can see that there is a declining crime rate for most of the race groups, which follows the trend we see with the crimes per year. However the "Asian/Pacific Islander" group have seen an increase in being victims in crimes:

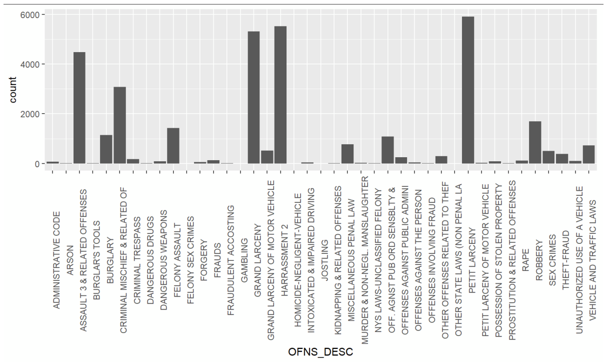


We evaluated the data for the years 2009 (the lowest rate) and 2019 (the highest Rate) to dig in deeper regarding this upward trend:

2009

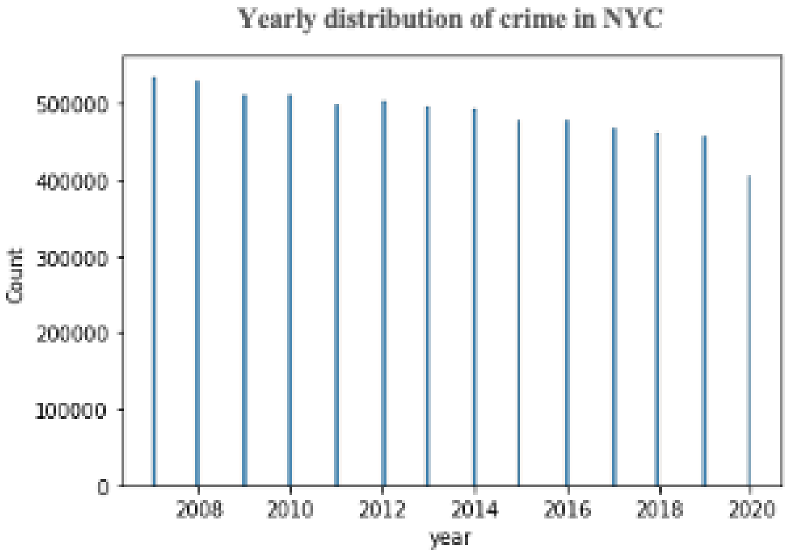


2019

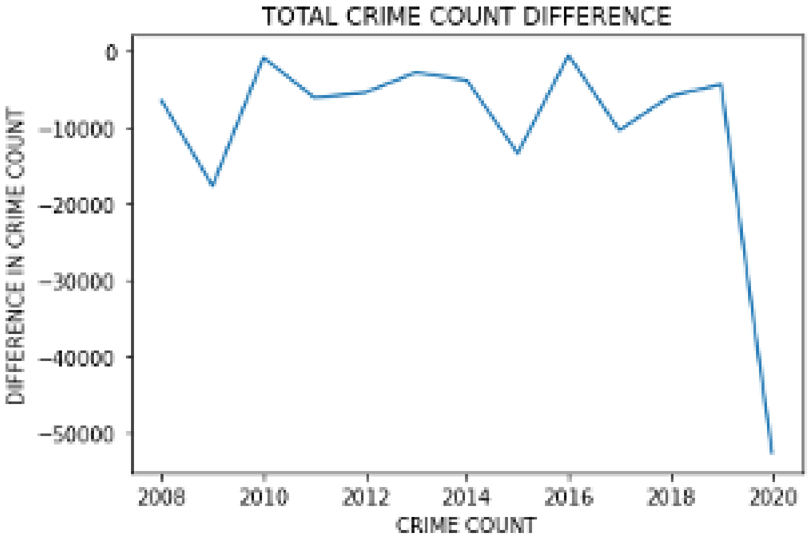


The largest increase in total crimes in specific offense categories from 2009 to 2019 is Harassment 3 and Assault 3, both of which are physical crimes. These results indicate that victim race would be an important feature in determining classification of crimes.

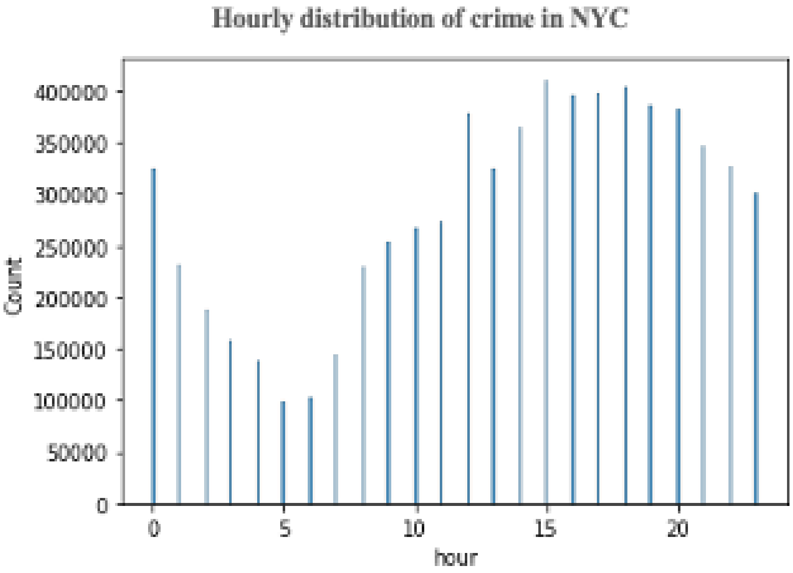
By looking at the total crimes per year, we see a downward trend from 2007 to 2020. This trend will be useful to address in terms of our forecasting.



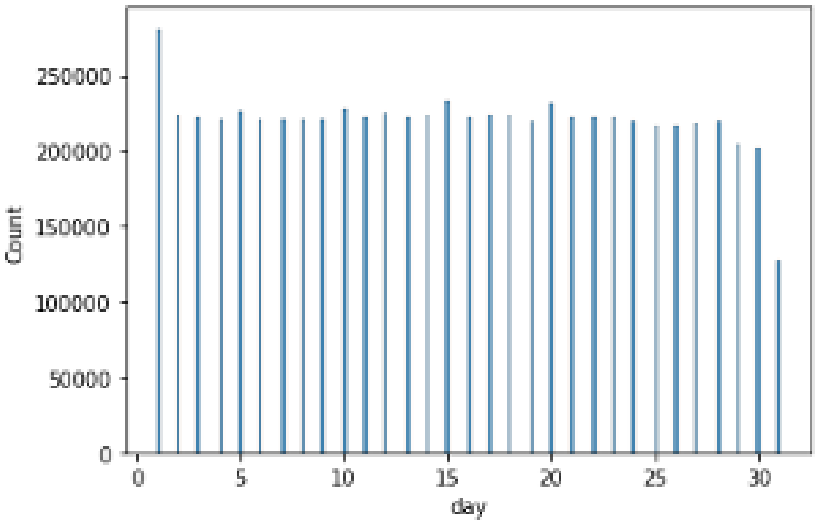
Even though the number of crimes have been decreasing since 2008, the rate of change has not been constant and does not follow any evident pattern. This indicates we do not need to use the year as a feature in our KNN Model. The sudden drop in 2020 is obvious because of the restrictions imposed due to COVID-19.



Prior to our analysis we expected the preponderance of crimes to occur late at night. Surprisingly the data shows that crime rates per hour gradually increases from 5AM and reaches its peak in the evenings.

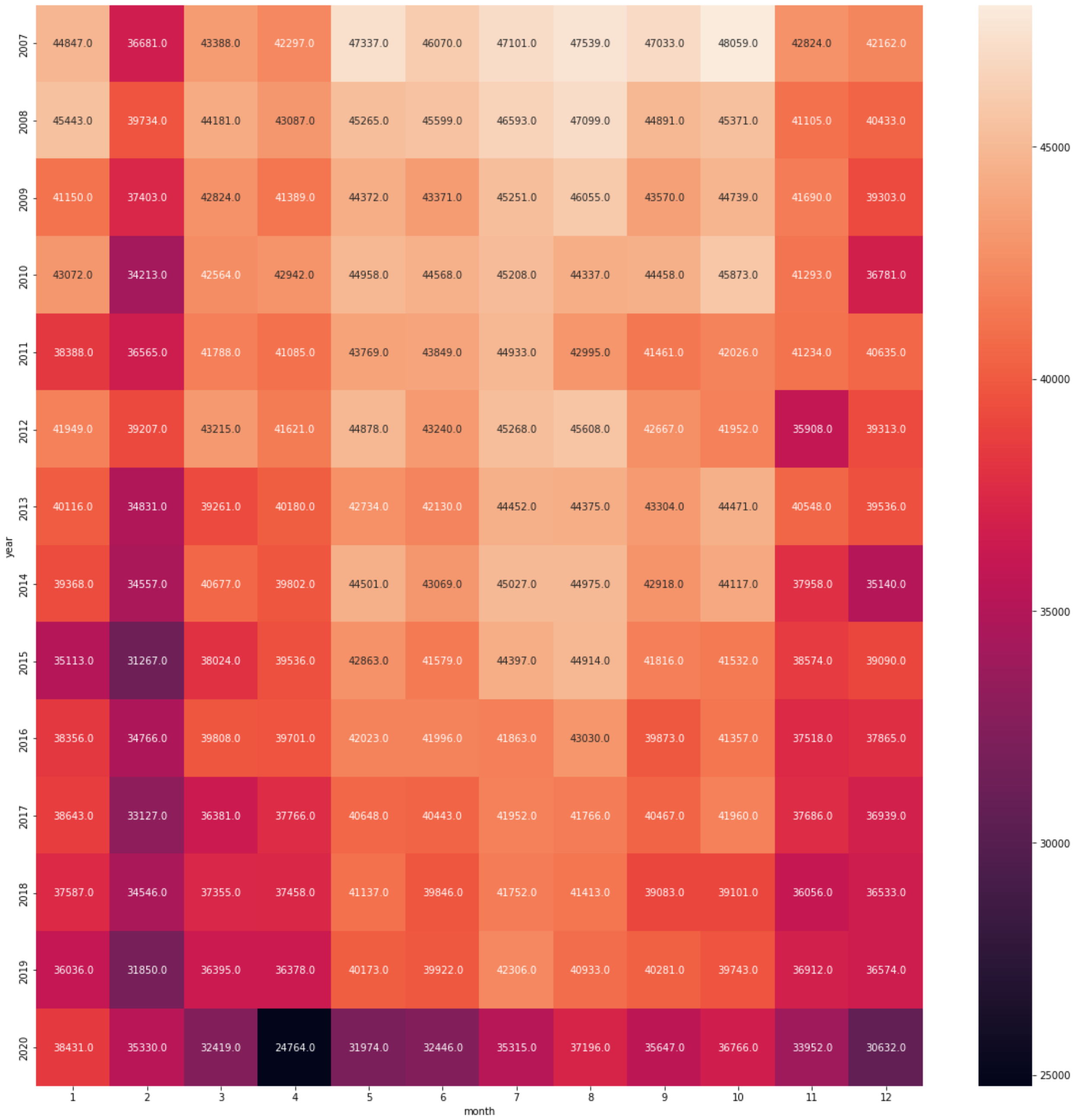


This is an intriguing plot as it provides a clear indication that the number of crimes on the 1st day of the month is the highest when compared to the rest of the days.



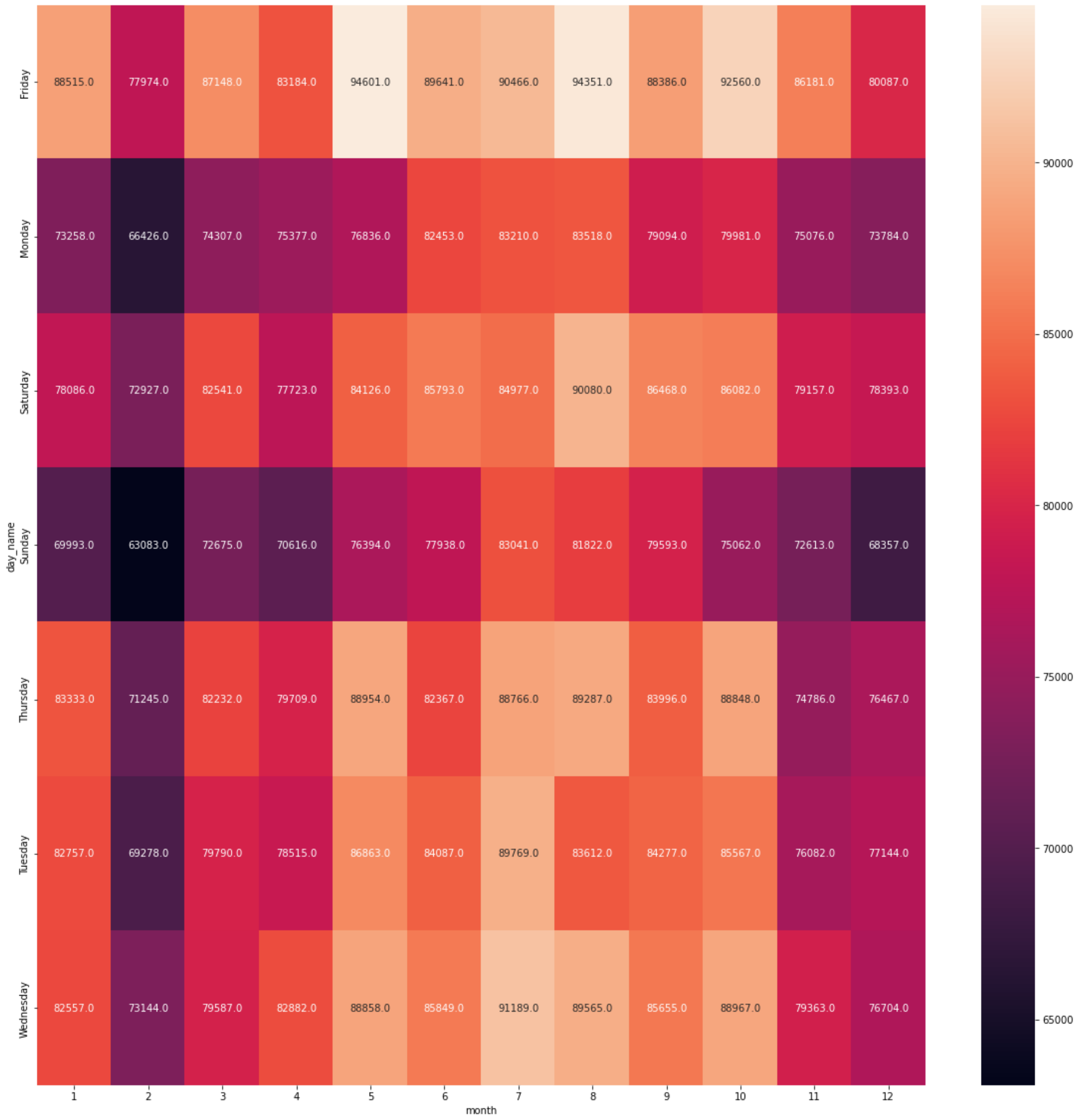
Heatmaps:

Year vs Month



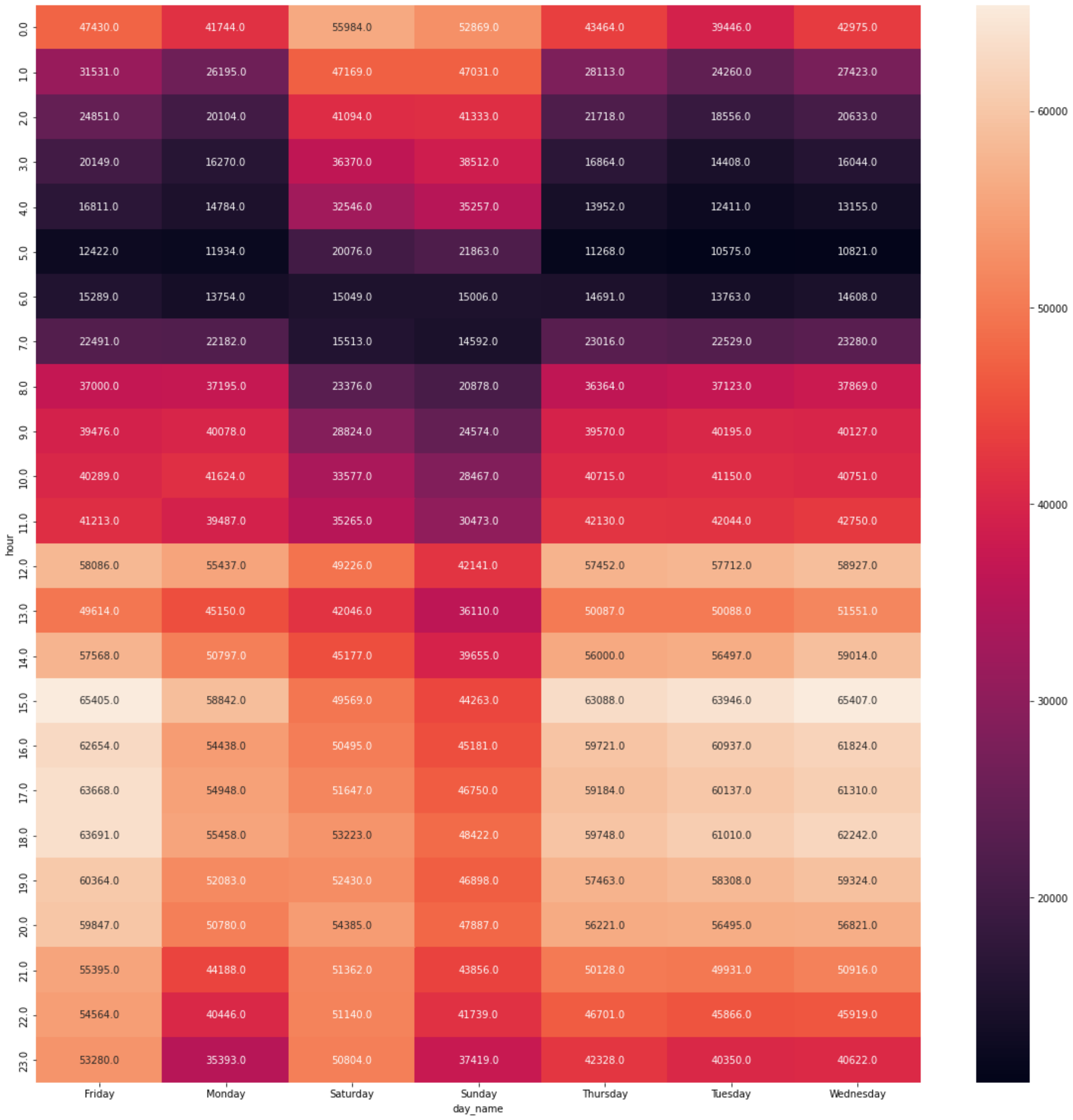
- There is a higher frequency of crimes from May to October most likely because more people venture out in the summer than in winters.
- April 2020 had the lowest crime count seemingly due to the COVID Lockdown.

Month vs Day of the week



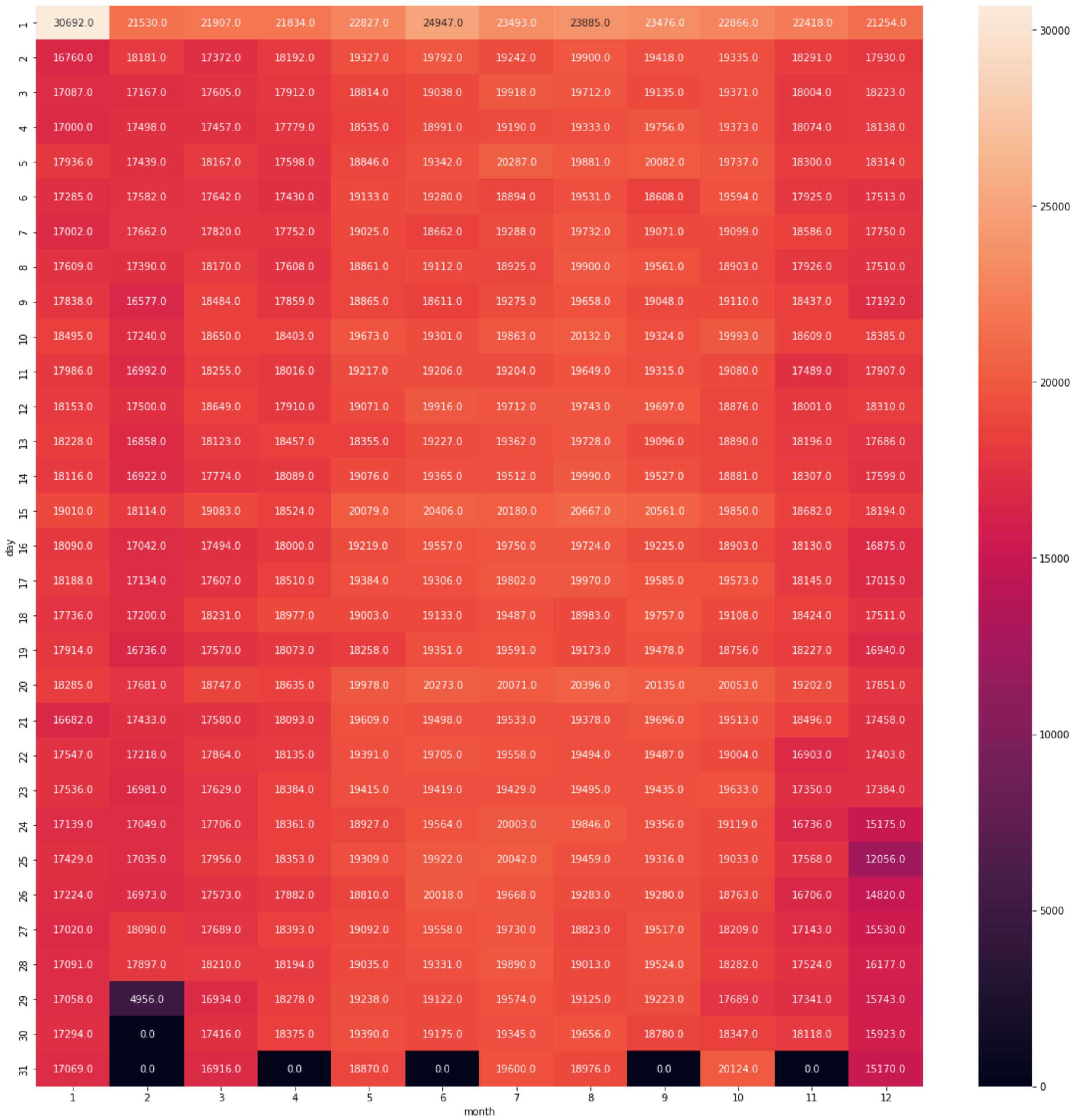
- Most crimes happen on Friday compared to Sunday where the crime count is the lowest.

Hours vs Day of the week



- This chart, combining daily and hourly crime counts, indicates more crimes occurring when we would expect more people to be out at night, which is dusk to night on the weekends.

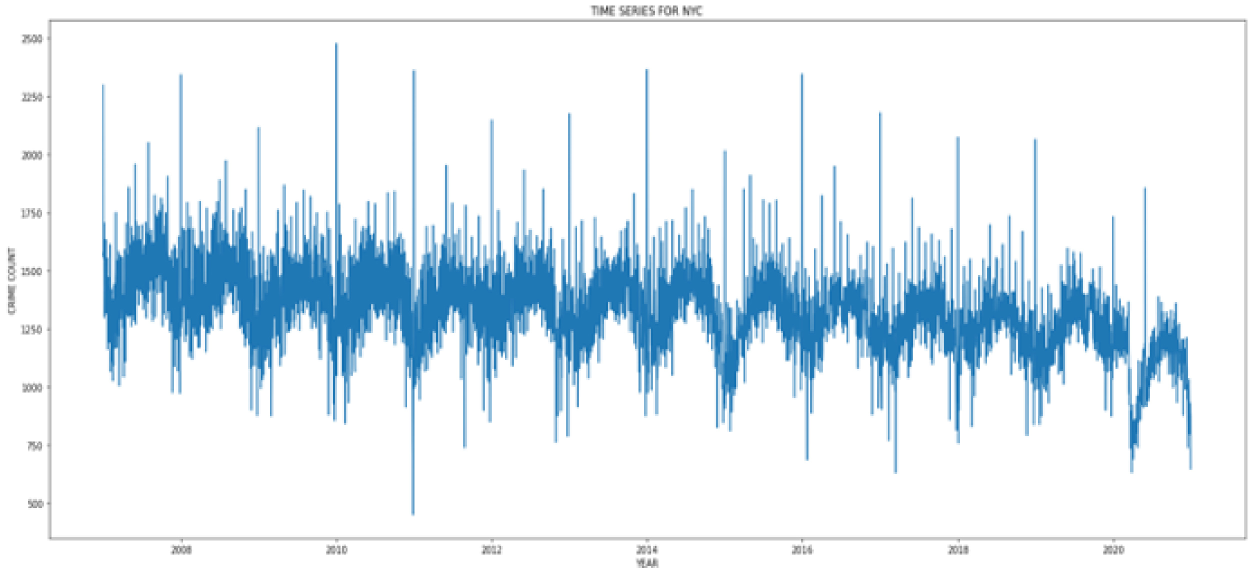
Days vs Months



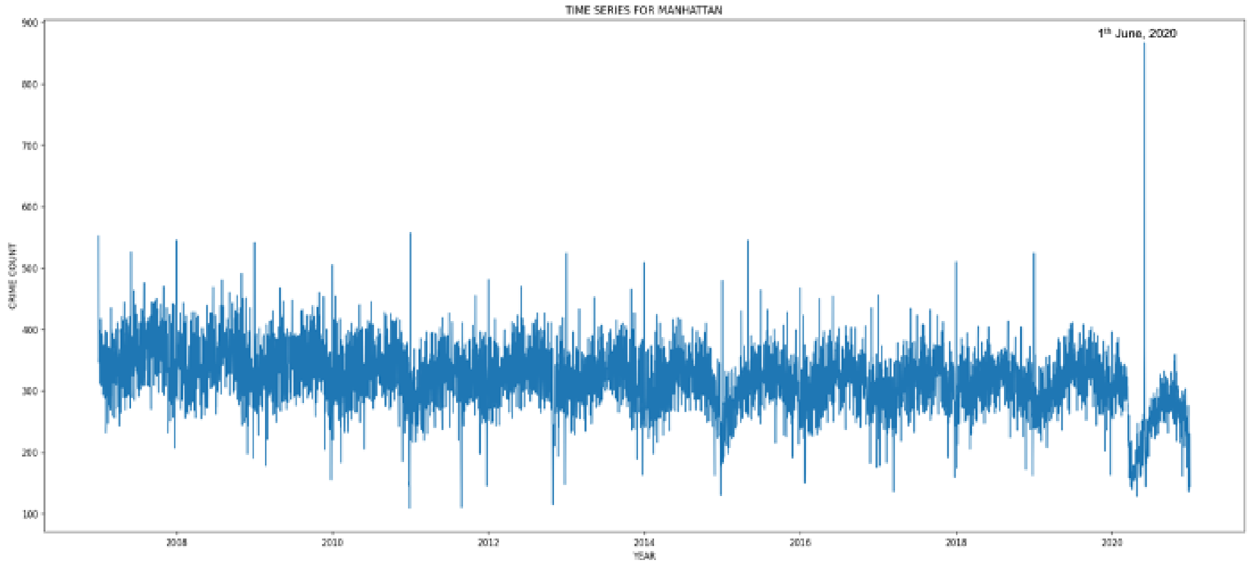
- January 1st is the least safe day of the year.
- The 1st day of every month is the least safe day of the month because this is usually the day when people are required to pay their rent and other recurring bills. These financial stressors often lead to commiting crimes.

Time Series Visualizations:

To model an ARIMA for the given data, it was imperative to visualize the time series of the crime count in NYC.



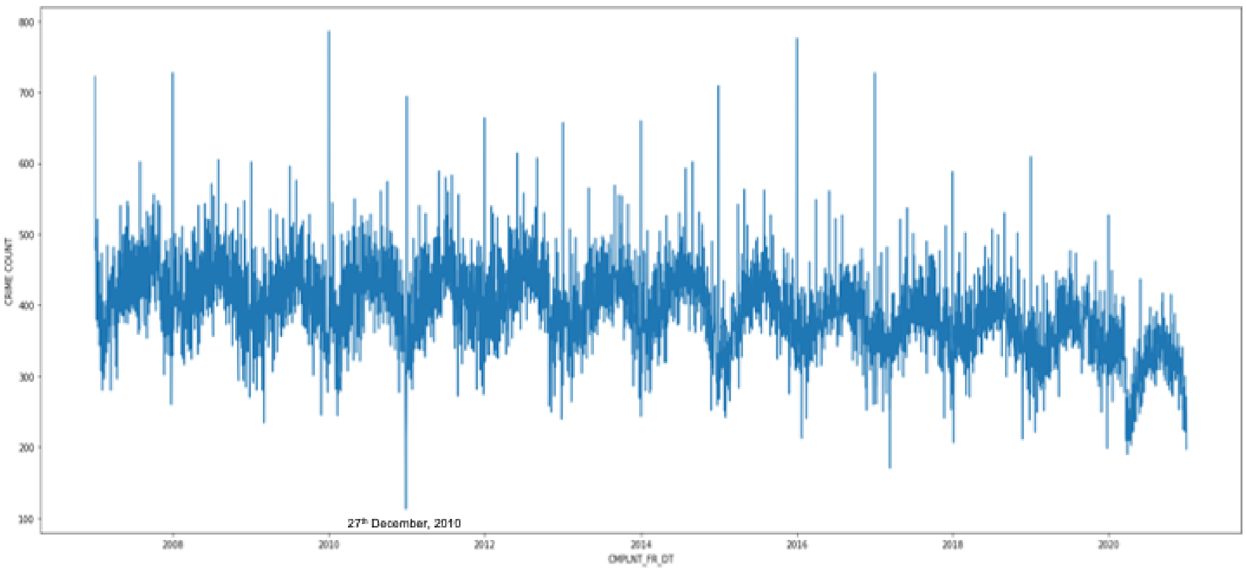
We also wanted to check whether similar patterns hold true when the data set is split borough-wise, and to analyze the trends in each borough.



- The peak in the graph denotes 1st June, 2020. This was a result of the widespread protests and unrest caused by the death of George Floyd. Many shops were looted and some sites also experienced protest violence. The table below shows the crimes which were committed on 1st June, 2020.

OFNS_DESC	COUNT
BURGLARY	313
CRIMINAL MISCHIEF & RELATED OF	185
POSSESSION OF STOLEN PROPERTY	65
PETIT LARCENY	59

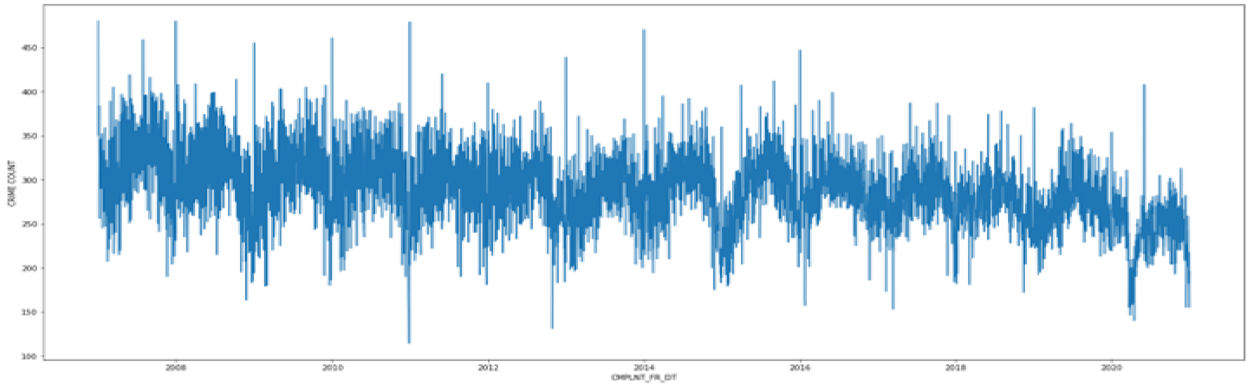
TIME SERIES FOR BROOKLYN



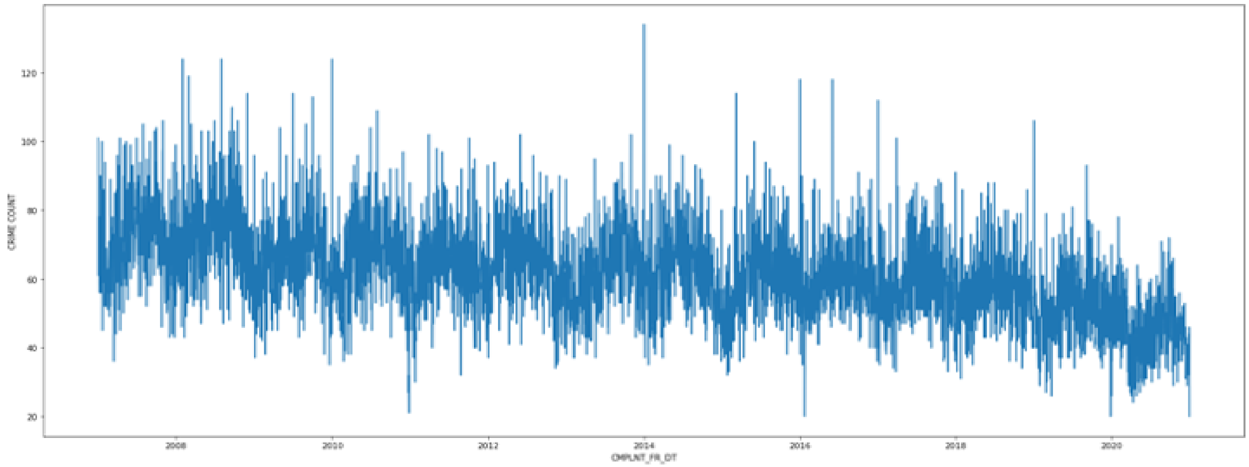
- The drop in crime count on 27th December, 2010 is common throughout all the boroughs. New York City witnessed one of its worst snow storms in its history on this day which reduced the mobility of the people.

From the three time series above, it can be observed that the time series of crimes per day is very similar across all boroughs. There are a few outliers but in general the shape of the graph is similar.

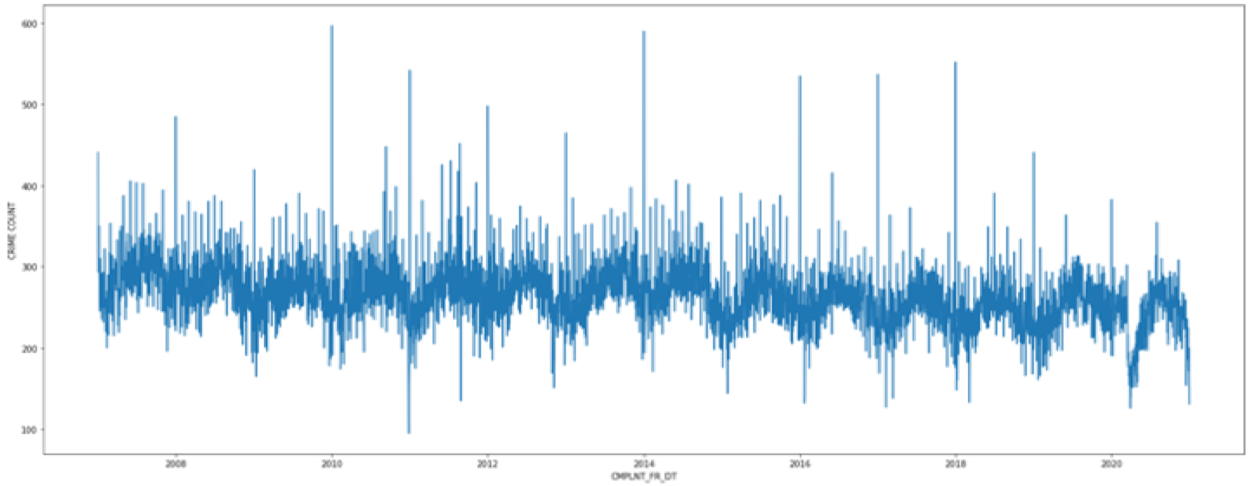
TIME SERIES FOR BRONX



TIME SERIES FOR STATEN ISLAND



TIME SERIES FOR QUEENS



AutoRegression Integration Moving Average (ARIMA) Model

ARIMA is a class of models that explains a time series of events based on its own past values. These values are used to forecast the future values. One of the main reasons we went with the ARIMA forecasting model is because of its robustness. The model requires a long historic horizon of data. Our dataset had complaints dating all the way back to 2006 which were over 6 million entries. This allows the algorithm to have a historic demand which helps in a more accurate forecasting.

The ARIMA is a combination of two different models - AutoRegression (AR) and Moving Average (MA) model. The blend of these two models do not overfit the data during training. The ARIMA model is characterized by 3 terms: p, d, q.

- p is the order of the AR term. The p term for the AR model can be determined by looking at the Partial Autocorrelation (PACF) plot since the PACF can be seen as a correlation between the series and its own lag.
- d is the number of differences required to make the time series stationary.
- q is the order of the MA term. It refers to the number of lagged forecast errors that should go into the model. Just like we determine the p value by analyzing the PACF plot, we can find the q value by looking at the Autocorrelation (ACF) plot.

There are two basic types of ARIMA models that are used for forecasting. They are Non-Seasonal ARIMA and Seasonal ARIMA (also known as SARIMA). The non-seasonal ARIMA consists of just the above three characteristics whereas the SARIMA is used with seasonal data.

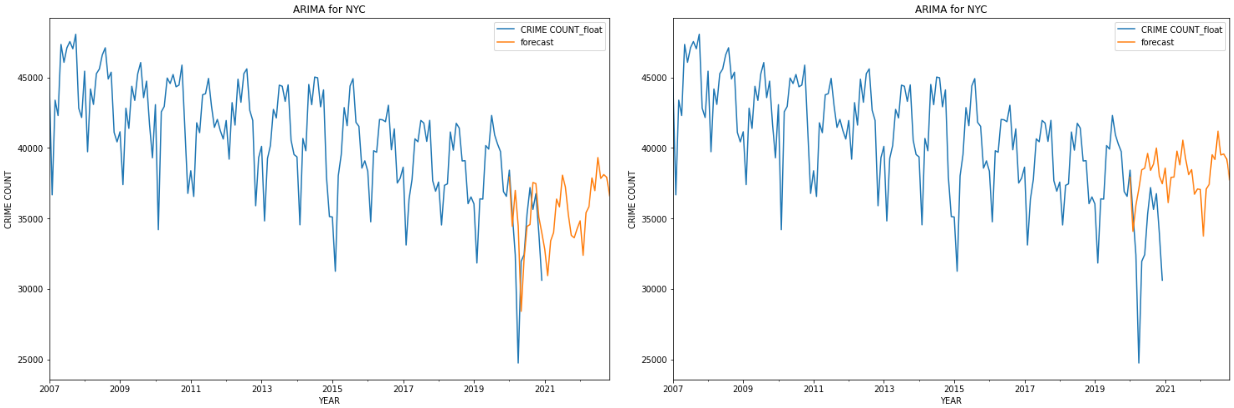
We add three new hyperparameters to the AR, I and MA components of the model as well as an additional parameter to determine the seasonality of the period of seasonality. We then configure the following four parameters:

- P: This is the seasonal autoregressive order of the model
- D: Seasonal difference order
- Q: Seasonal moving average order
- m: The number of time steps that we want to make for our model. We could select m to be 3 (for quarter of a year), 12(one year) or maybe more depending on how we want to analyze our time series.

Once these parameters and hyperparameters are determined, we start to fit our data into the model and train it for forecasting. Then the results for each borough were observed for the years 2021 and 2022.

Results

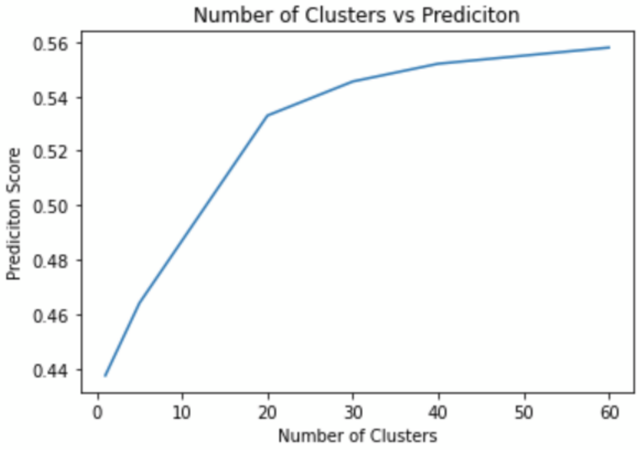
- Due to the spread of the COVID-19 virus, there was a significant drop in crime rate in the year 2020. This also led to a Mean Squared Error Percentage of 7.5%. If this pandemic had not occurred, the Mean Squared Error Percentage for this model would have been 2.5% giving us an overall accuracy of 97.5%.
- After a significant drop in crime numbers in 2020, we see that the number of crimes that are expected to take place in the next two years will increase in all five boroughs of New York City.



Difference between forecasting due to presence and absence of COVID-19 respectively

Clustering with KNN

We decided to further our forecasting research by evaluating if we could categorize our future predictions. We have decided to use a KNN model to provide these classifications due to its low training time and classification accuracy given the robustness of our data set. With the number of features in the data set, we decided to do some data preprocessing and limit the number of features used in the KNN, as described in the Data Exploration section. After not finding anything significant in the Pearson Correlation Matrix, we tried running the KNN model on all of the features. This approach did not work given our limited computational power. To overcome this we started to remove features from our subsequent trials based on the cardinality of each feature. In our many tries to train a predictive KNN model, we noticed that the classification accuracy of our test data hovered around 20%. For our final model, the features we settled on using were those associated with the ARIMA data; time features and location features. Given all of this we trained our final data set over a variety of k clusters, giving us a graph of prediction percentages:



Using the 60 clusters here gave us a predictions rate of 55%, which we deem is fairly accurate given our data set. Trying to use this model with our prediction from ARIMA led to our main pitfall.

Pitfalls

ARIMA has proven to do a good job in forecasting crime trends in each of the boroughs for the next year, however our goal of using this information for preventive measures did not come to fruition. In order to use the KNN that we trained, we need to have specific instance predictions instead of projections. Our expectation was we would be able to get the data needed to predict not only the total number of crimes per month, but also have instance data that would fit into our KNN Model, giving us the predicted classification of crime for time and location. With advancements in modeling techniques for time series data, we expect to complete all of our goals for this project.

Conclusion

Given the robustness of our dataset, we have successfully projected an accurate forecast of crime in New York City. Through our analysis, we verified that one of the most accurate predictors of future crime is the historical crime rates. We have seen that the outside factors, such as blizzards and protests, have significant effects on crime rate that are hard to predict. We furthered our analysis by producing a KNN Classification model that efficiently and somewhat confidently predicted crime classification given a certain amount of features. This work lays the groundwork for future research into this problem, starting with research into different time series analysis techniques, and if they would be able to predict instances rather than features. Papers have already been published on “Predictive Policing”, leaving us with future routes of taking this work and furthering it for the safety of others. Given the success we have had so far, we believe that this is the start of great predictive models that can help with the prevention of crime.

References:

(1) = https://www.neighborhoodscout.com/ny/new-york/crime#data