|  |
| --- |
|  |
| Capstone Project 1 In Depth Analysis |
|  |
| **Date Created : 5/1/2021** |

Contents

[1. Problem Statement 3](#_Toc76430389)

[2. Data Set 3](#_Toc76430390)

[3. Data Wrangling 3](#_Toc76430391)

[4. Data Storytelling: 3](#_Toc76430392)

[5. Multi-test Regression Model 6](#_Toc76430393)

[6. Predicting Exact Crime Location based on above Model: 7](#_Toc76430394)

# Problem Statement

Analyzing New York City Crime Data and Predicting Future Crime Location.

# Data Set

Dataset for this Project has been taken from City of New York Public Data portal.

* After school activities NYC :

[https://data.cityofnewyork.us/Education/DYCD-after-school- programs/mbd7-jfnc](https://data.cityofnewyork.us/Education/DYCD-after-school-%20%20%20programs/mbd7-jfnc)

* NYPD 2019 Arrest Dataset NYC :

<https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc>

# Analysis and Modeling

In this document I will be covering a few topics that involve learning from data and developing a predictive model for crime rates. Specifically I will,

* Perform exploratory analysis on temporal and geospatial crime data to find where and when specific crimes happen the most.

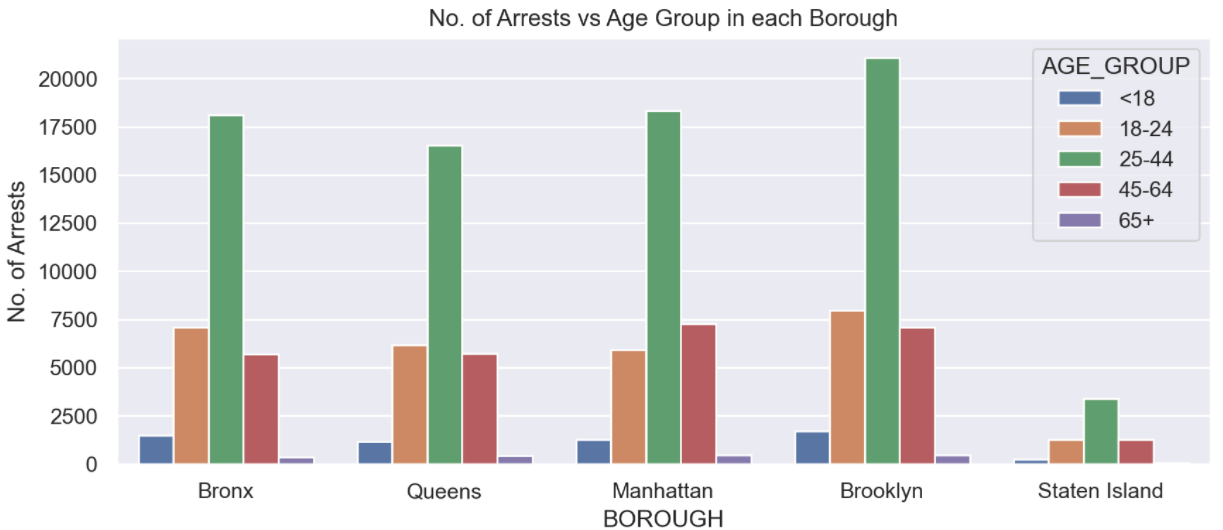
We will see that different crimes affect different areas in the Manhattan and that monthly crime rates peak at different points of the year depending on the type of crime. We then,

* Develop a time series model that uses historical data to predict monthly number of crimes in two neighborhoods that have the highest crime rates in Manhattan.

Once we can predict future montly crime rates on a neighborhood level, police can distribute their resources to the right neigborhoods at the most appropriate time.

# Data Storytelling:

The Arrest dataset gave us significant data insights on the demographics:

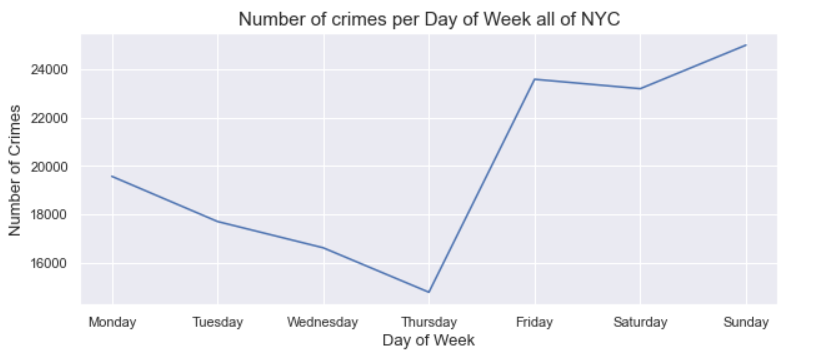


**Fig 1: Across all the Age groups crime rate is significant higher among 25-44 age group**

Key Observations:

i) In each Borrow, Criminal activities are highest among \*\*25-44\*\* Age group

ii) Among all the Borrows, \*Brooklyn\* has the Highest Crime Rate across all the Age Groups



**Fig 2: Crime rate picks up during the Weekends**

Key Observations:

1. Overall crime numbers in NYC is highest on Sundays and lowest on Thursdays
2. Trend shows that Crime numbers are lower between Mondays and Tursdays and spikes up on Fridays.
3. Except for Staten Island, all other Broughs have highest Crime Numbers on Sundays while Staten ISland highest numbers on Saturdays

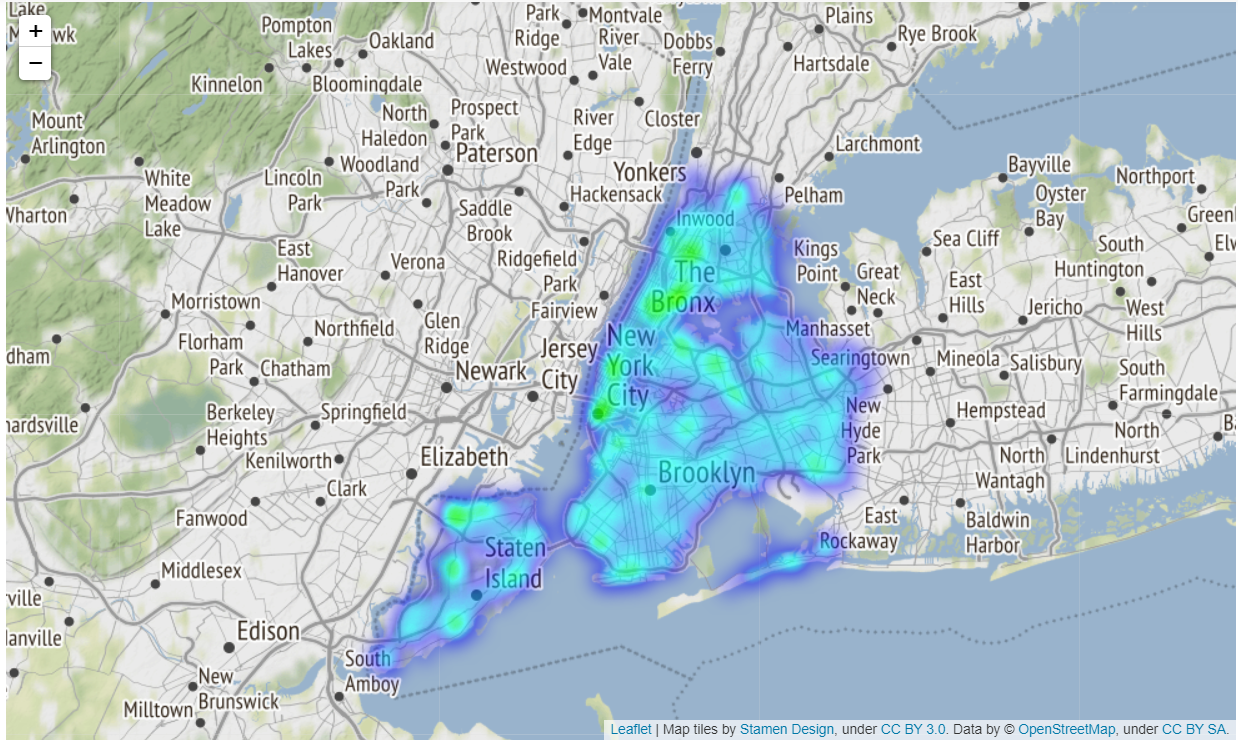
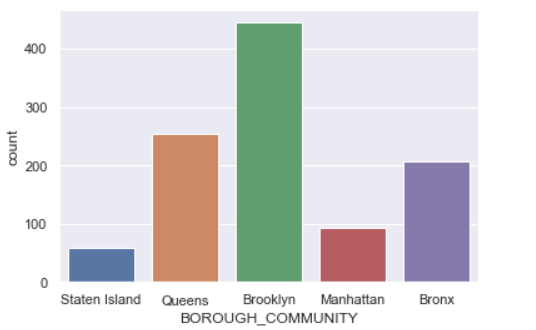
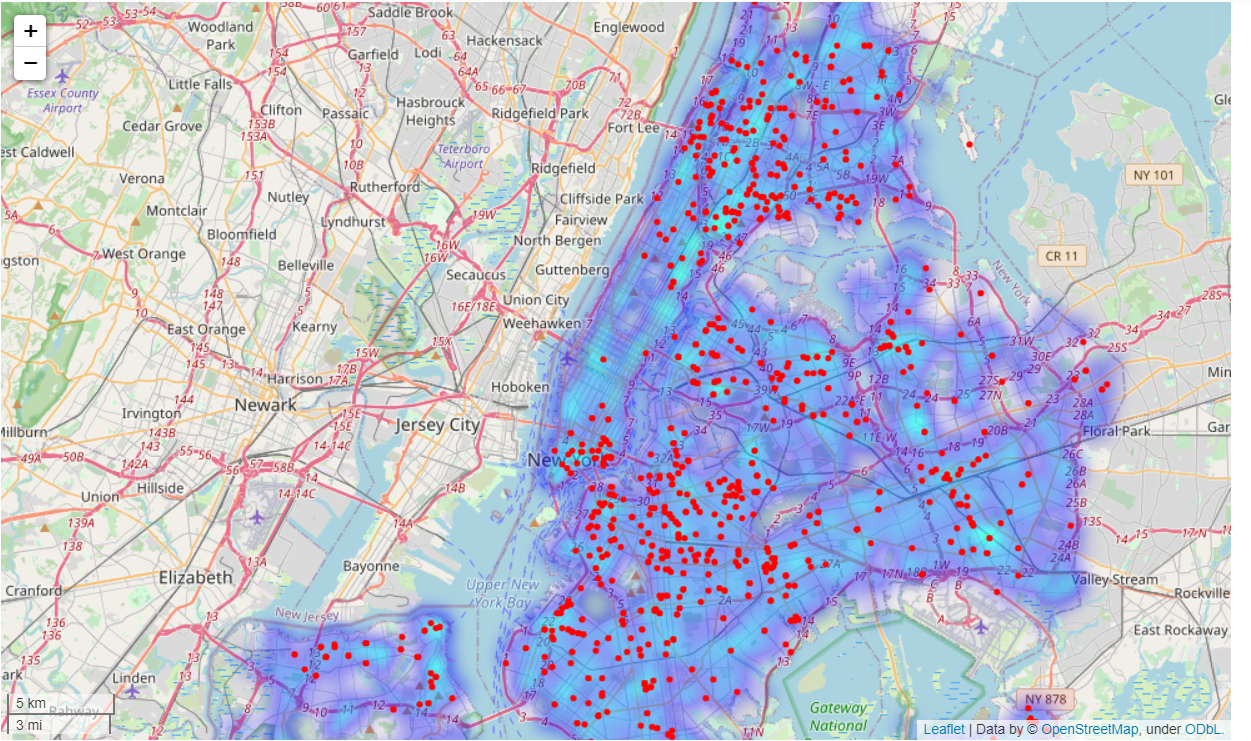


Fig 3: New York City map showing crime rate heatmap across all the Boroughs



**Fig 4: After school activities in each Borough**



**Fig 5: Heatmap show After Schhol Activities(In Red) vs Crime (In Blue)**

# Multi-test Regression Model

Create MTR model to predict the loaction of a crime base on some basic features and then using the data lets try to forcast the next month. Random forest and Gradient Boost Regressors will be the base estimators for the MTR.

The Mean Squared Error of RF = 0.0014327743867625468

The Mean Squared Error of GBR = 0.0013144329347024096

Gradient Boosting Regressor(GBR) showing better result than Random Forest(RF).

# Conclusion:

In this document we have looked into using exploratory data analysis to come up with a data driven strategie to reduce crime in Manhattan. We saw that larceny and assault are two crimes that effect different parts of Manhattan and also peak at different times of the year. Larceny effects midtown and peaks at the end of the year, while assault effects East Harlem and peaks in summer time. We looked into time series analysis to model monthly crime rates in both Midtown and East Harlem and used them to predict monthly crime rates into the future. This is just an small example with two neighborhoods, but if police are better able to predict monthly crime rates on a neighborhood level they can more effectively distribute their resources at the appropriate place and time.