Capstone Project 1
Final Report

Date Created : 5/1/2021

Contents

1. Problem Statement

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

2. 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

- NYPD 2019 Arrest Dataset NYC :

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

3. Data Wrangling

Data cleaning activities were performed on the Datasets

➔ In The Arrest Dataset – calculated the Percentage of missing Values for each of the key columns :

```
PD_CD  --->  0.02 %
PD_DESC  --->  0.03 %
KY_CD  --->  0.03 %
OFNS_DESC  --->  0.03 %
LAW_CAT_CD  --->  0.99 %
```

➔ Since the percentage was less than 1%, the rows which had at least one NaN value, were dropped to create a Clean dataset:

```
Count BEFORE cleaning - 140413
Count AFTER cleaning  - 138987
```

➔ For After School Dataset similar logic has been applied and rows which have at least 1 missing value were dropped from the Dataset.

➔ Cleaned up the BOROUGH_COMMUNITY to get the exact BOROUGH Name from the Zip Code

4. Data Storytelling:

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

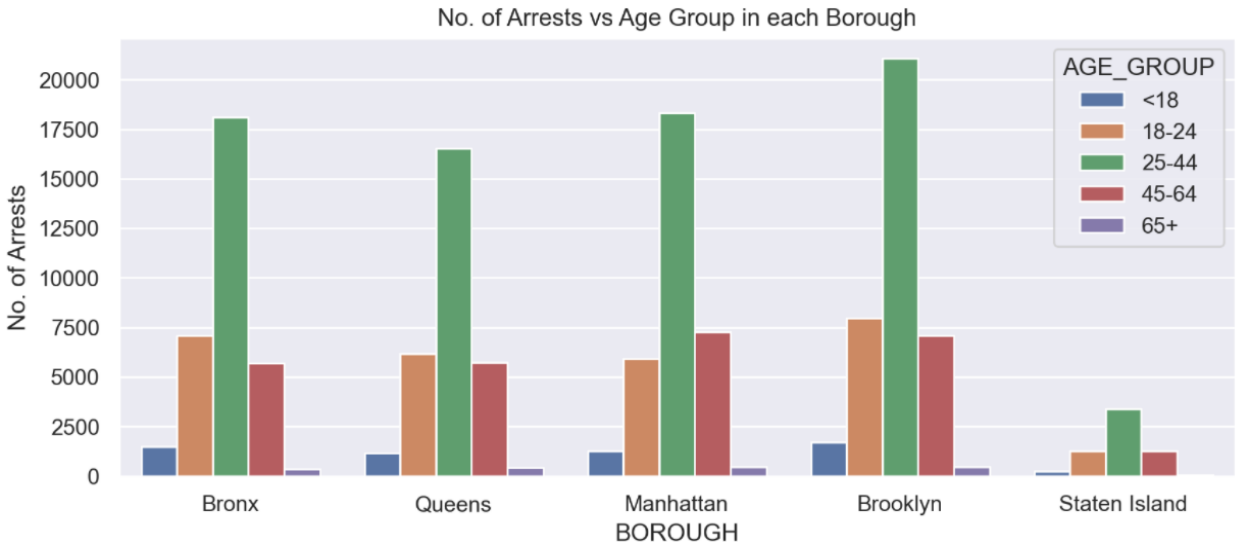

Fig 1: Across all the Age groups crime rate is significant higher among 25-44 age group

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

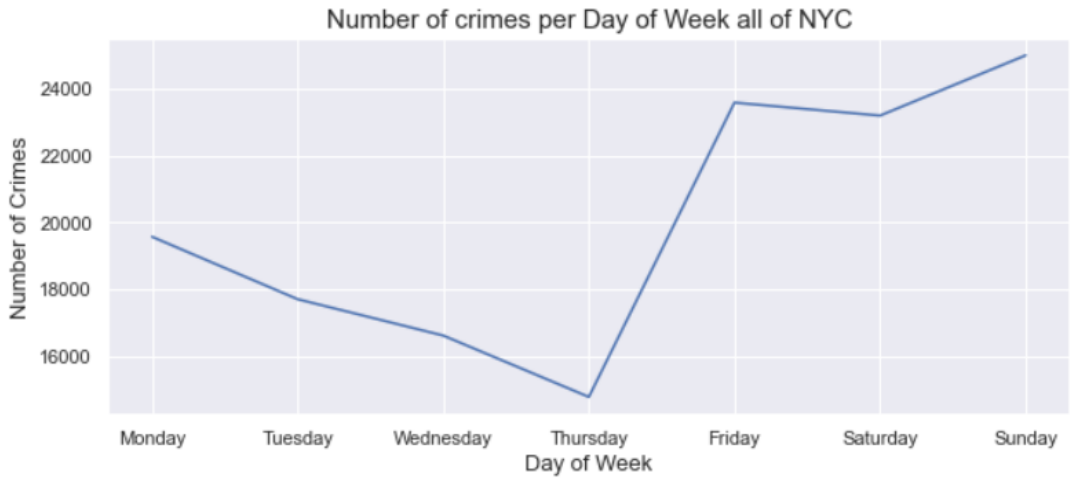

Fig 2: Crime rate picks up during the Weekends

Key Observations:

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

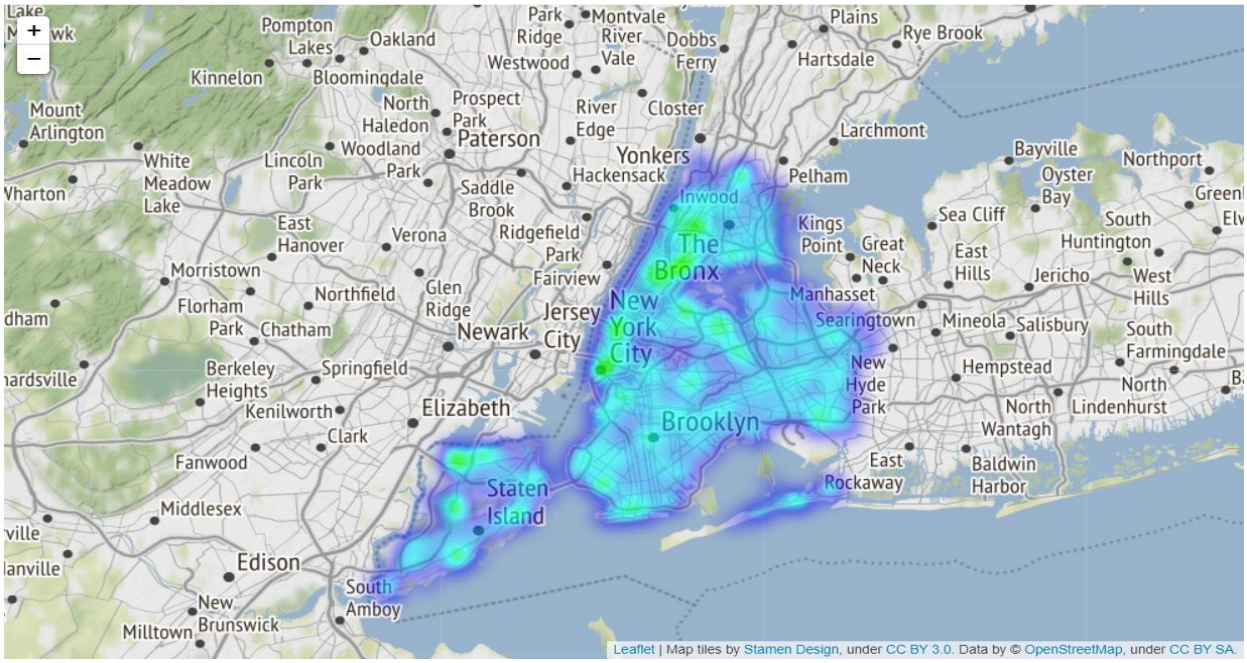

Fig 3: New York City map showing crime rate heatmap across all the Boroughs

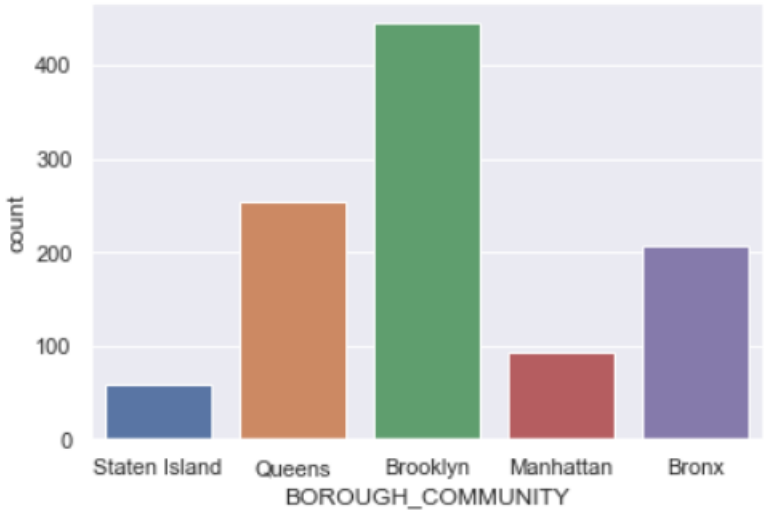

Fig 4: After school activities in each Borough

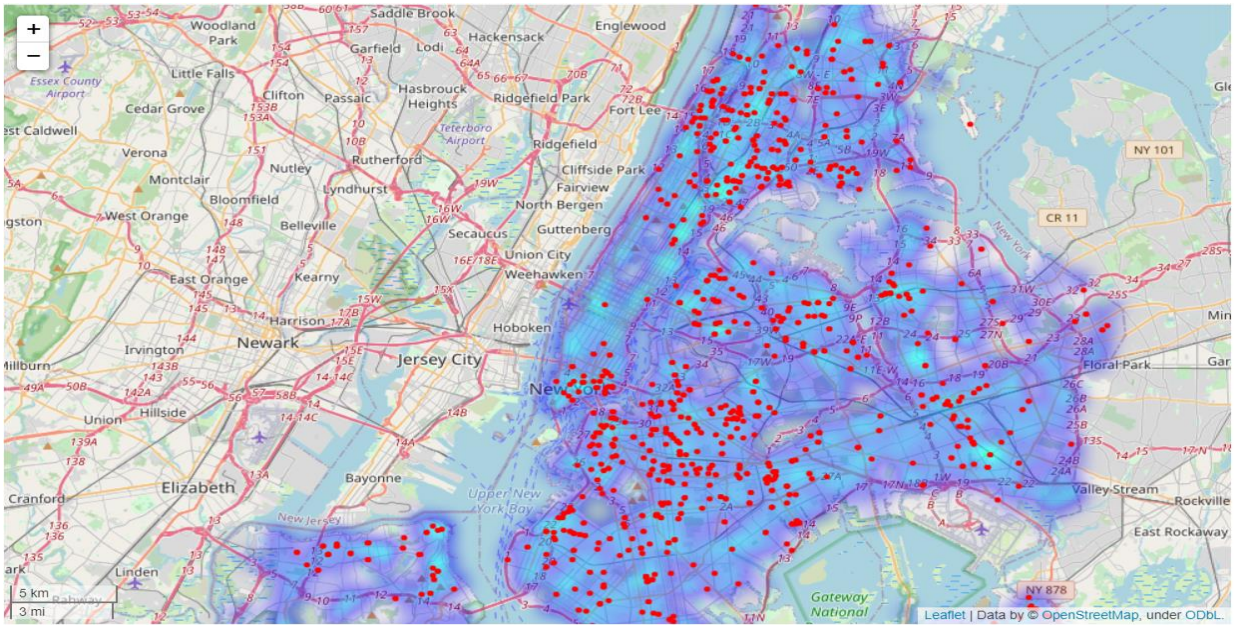

Fig 5: Heatmap show After Schhol Activities(In Red) vs Crime (In Blue)

5. 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).

6. Predicting Exact Crime Location based on Model:

The heatmap shows the crime rate heatmap we saw earlier

1. Blue indicates the location predicted by Gradient Boosting Regression
2. Red indicates the location predicted by Random Forest Regression