Image Source: https://brooklyneagle.com/wp-content/uploads/2019/09/AP_17327743469429-1920x1278.jpg

New York Police Department Complaints Data Visualization and Prediction

A PROJECT REPORT for CS-GY 6313: INFORMATION VISUALIZATION

-Submitted By Nikhil Kishan Khaneja (nkk6190@nyu.edu)

TABLE OF CONTENTS

TABLE OF FIGURES

ABSTRACT:

The New York Police Department (NYPD) is the preliminary law enforcement agency in the City of New York which was established on 23rd May 1845. The NYPD manages all the New York City and with its vast population, the crime rate of the city is one of the major concerns for the NYPD. With the crimes happening all around the New York City, it is important to keep the record of all the data and create informatic and engaging visualization to effectively learn the crime pattern of the New York City so that constructive measures can be implemented to prevent the crimes in the city. The crimes are mainly classified in three categories namely Felony, Misdemeanor and Violation. It is important to implement the use of machines and try to form the strategies to prevent crimes. In this report I have tried to shed some light on some valuable insights from the explanatory visualizations and machine learning models. The models were trained on the data and compelling accuracy matrix was established.

DATA SOURCE:

The data used for this project is titled **NYPD Complaints Data Historic** [1-2] and was obtained from the New York City Open Data. NYC Open Data is free public data published by New York City agencies and other partners. This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of 2020.
The data contains over more than 7 million rows with over 35 features such as Complaint Type, Date, Geo Location of the crime, Suspect's, and Victim's attributes and many more. A detailed description of the data is mentioned below.

Field Name & Description:

'CMPLNT_NUM' (int)	Complaint Number
'CMPLNT_FR_DT' (date)	Complaint From Date
'CMPLNT_FR_TM' (date)	Complaint From Time
'CMPLNT_TO_DT' (date)	Complaint To Date
'CMPLNT_TO_TM' (date)	Complaint To Time
'ADDR_PCT_CD' (int)	Code of Precinct in which the Incident Occured
'RPT_DT' (date)	Report Date
'KY_CD' (int)	"Key Code": Offense Classification Code (3 digits)
'OFNS_DESC' (str)	Offense Description
'PD_CD' (int)	PD Code of Offense. More granular than Key Code
'PD_DESC' (str)	PD Description of Offense.
'CRM_ATPT_CPTD_CD' (str)	Whether Crime was Atempted or Completed (values: 'COMPLETED', 'ATTEMPTED')
'LAW_CAT_CD' (str)	Level of Offense (values: 'FELONY', 'VIOLATION', 'MISDEMEANOR')
'BORO_NM' (str)	Name of Borough in which Incident Occurred
'LOC_OF_OCCUR_DESC' (str)	Description of where the incident occurred with respect to the premises (values:'FRONT OF', 'REAR OF', 'OUTSIDE', 'INSIDE', 'OPPOSITE OF')
'PREM_TYP_DESC' (str)	Description of the type of premises in which the Incident Occurred
'JURIS_DESC' (str)	Description of Jurisdiction in which Incident Occurred
'JURISDICTION_CODE' (int)	Jurisdiction Code
'PARKS_NM' (str)	Name of Park in which Incident Occurred, if Applicable
'HADEVELOPT' (str)	Name of NYCHA Housing Development in which Incident Occurred, if Applicable
'HOUSING_PSA' (int)	Housing PSA
'X_COORD_CD' (int)	X-coordinate, New York State Plane Coordinate System
'Y_COORD_CD' (int)	Y-coordinate, New York State Plane Coordinate System
'SUSP_AGE_GROUP' (int)	Age Group of Suspect
'SUSP_RACE' (str)	Race of Suspect
'SUSP_SEX' (str)	Sex of Suspect
'TRANSIT_DISTRICT' (int)	Transit-District code
'Latitude' (float)	Global Latitude of Location where Incident Occurred
'Longitude' (float)	Global Longitude of Location where Incident Occured
'Lat_Lon' (str)	'Latitude' and 'Longitude' together
'PATROL_BORO' (str)	Patrol Borough
'STATION_NAME' (str)	Station Name
'VIC_AGE_GROUP' (int)	Age Group of Victim
'VIC_RACE' (str)	Race of Victim
'VIC_SEX' (str)	Sex of Victim

Fig1: The overview of the data

Since the dataset is very huge, for this project I have subsets of the dataset. For most of the project I have used first 1 million rows of the dataset and for some visualization I have used the subset data of the year 2018.

RELATED WORK:

I have referred two publications to gather the insights for this project. The first publication I referred was titled "**San Francisco Crime Classification**" [3] by Yehya Abouelaga of the American University in Cairo, Egypt. The paper publication aims to classify and predict the crimes happening in the city of San Francisco based on the geographical and time-based features. The publication employs some basic visualization and classification of the important features required for predicting the crimes and deploying the machine learning models and establishing an accuracy matrix for the models for selecting the best possible model. The paper classifies Latitude, Longitude, Address, Date, District and Day as the features while Category as the Label for the model. The paper uses five different models namely K-Nearest Neighbors Classifier, XGB Classifier, Random Forest Classifier, Decision Tree Classifier and Gaussian Naïve Bayes Classifier. The paper classifies Random Forest Classifier as the best model to predict the crime.

The second publication I referred was titled "**Crime Analysis and Prediction Using Data Mining**" [4] by Shiju Sathyadevan, Devan M.S, and Surya Gangadharan. S of Amrita Vishwa Vidyapeetha. The paper aims to develop a data mining-based approach to solve the crimes faster by predicting the crimes using the crime factors for each day. They are following 5 steps: Data Collection, Classification, Pattern Identification, Prediction and Visualization. For classification of the crimes, they have used Naïve Bayes Classification and for predication have implemented Decision Tree Classifiers. They have recorded the accuracy of over 90 percent for the Naïve Bayes Algorithm.

VISUALIZATION:

In this project, the first step was selecting a subset of data, as the size of data is very huge, compiling and visualizing the whole data will take a lot of time and is not feasible. As the crime rate has increased in the New York City in the past years, I have selected the first 1 million records comprising of the crime reports for the years 3 years (2020, 2019 and 2018). Next comes the data cleaning and profiling.

```
In [8]: results_df.drop(['cmplnt_to_tm','rpt_dt','pd_cd','pd_desc','juris_desc','jurisdiction_code','housing_psa','x_coord_cd',

In [9]: df.drop(['CMPLNT_TO_TM','RPT_DT','PD_CD','PD_DESC','JURIS_DESC','JURISDICTION_CODE','HOUSING_PSA','X_COORD_CD','Y_COORD

In [10]: results_df.dropna(subset=['cmplnt_fr_tm'], inplace=True)

In [11]: results_df['date'] = pd.to_datetime(results_df['cmplnt_fr_dt'],errors = 'coerce')
         df['date'] = pd.to_datetime(df['CMPLNT_FR_DT'],errors = 'coerce')

In [12]: results_df['month'] = results_df['date'].dt.month
         results_df['year'] =results_df['date'].dt.year
         results_df['day'] = results_df['date'].dt.day

         df['month'] = df['date'].dt.month
         df['year'] =df['date'].dt.year
         df['day'] = df['date'].dt.day

In [13]: df['SUSP_AGE_GROUP'].value_counts()

Out[13]: 25-44      28390
         UNKNOWN    27560
         18-24      10642
         45-64      10544
         <18         3412
         65+          849
         2018           3
         1018           2
         928            2
         938            1
         -2             1
         1967           1
         952            1
         920            1
         954            1
         Name: SUSP_AGE_GROUP, dtype: int64

In [14]: values = ['-958', '-43','-974', '-972', '972', '951', '-942']
         val = [np.nan, '928','952', '1018', '920', '2018', '938', '-2', '954', '1967']
         df = df[df['VIC_AGE_GROUP'].isin(values) == False]
         df = df[df['SUSP_AGE_GROUP'].isin(val) == False]
```

Fig 2: Data Cleaning and Profiling

I have removed all the NaN, NaT contain rows and filtered out the useful columns. Next, I have aggregated the crimes month wise and have visualized a horizontal bar plot which led to the result that September, October, November, and December have the highest crime rate while October tops the list.

```
In [20]: fig, ax = plt.subplots(figsize=(6, 5), dpi=100)
         sns.countplot(y='month', data=results_df,palette="Blues_d",ax=ax)
         plt.title('Count of all the Crimes month wise')
         plt.show()
```

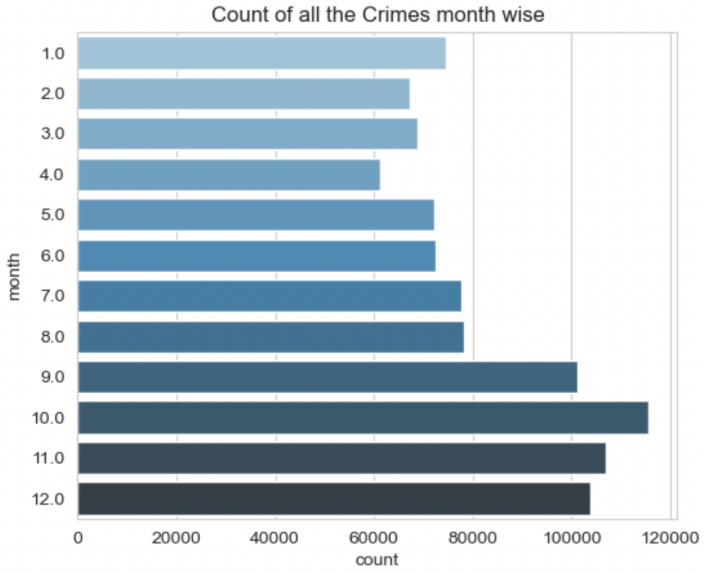

Fig 3: Count of all the crimes month wise

I then filtered the crimes by the borough of the New York City. Brooklyn stood out to the borough cumulating highest number of crimes with over 28% of the NYC crimes occur in the Brooklyn with Misdemeanor being the category with highest number of complaints. Manhattan comes second while the Staten Island counts for only 4% of the total crimes.

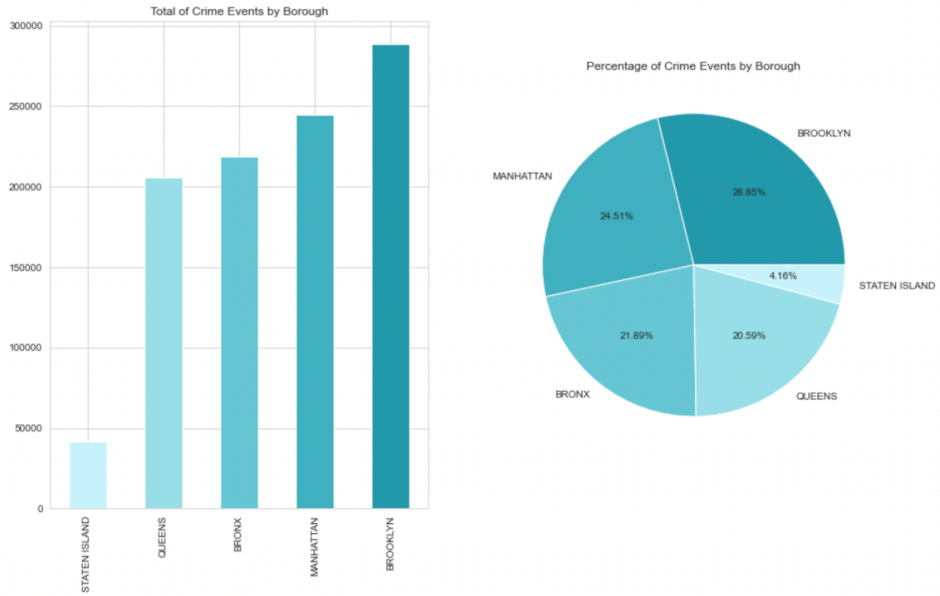

Fig 4: Classification of Crimes by Borough

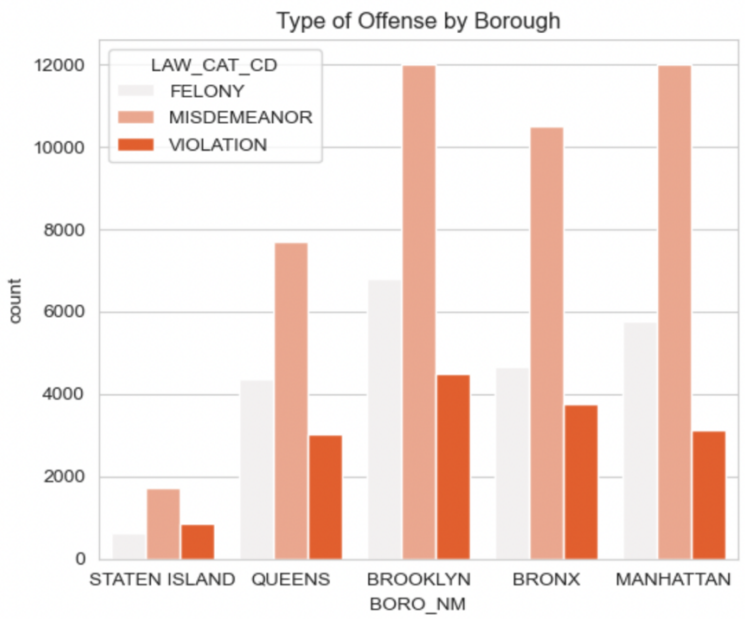

Fig 5: Types of Crimes by Borough

Victim and the Suspect are the two most important people related to the crime. So, there must be a high correlation between the victim's attributes and suspect's attributes. The premise came out to be right when I plotted the number of crimes for each of victim's attributes and suspect's attributes (Race, Age and Gender).

Majority of the victims are women who had chosen not to disclose their age and race due to many social factors. While most of the suspects are male residing in the age group of 25-44.

- The highest correlation for age feature is found for the features when the **suspect is male**, and the victim is female.
- he highest correlation for age race is found for the features when **both the victim and the suspect are African Americans**.
- he highest correlation for age race is found for the features when **both the victim and the suspect belong to age group 25-44**.

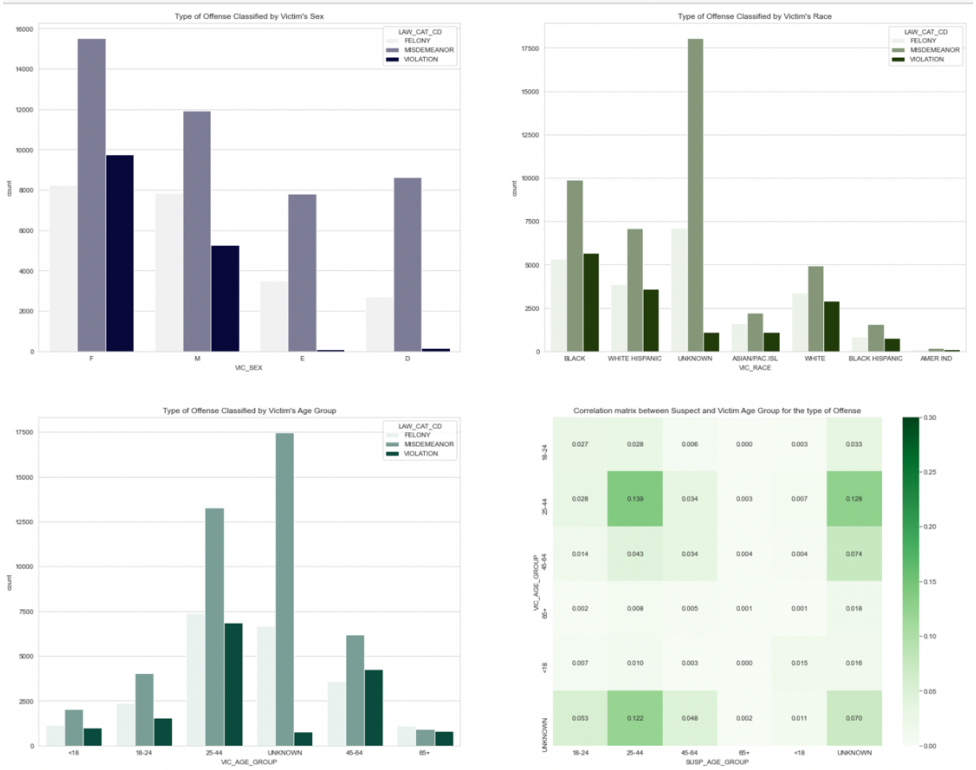

Fig 6: Visualization of Crimes by Victim's Attributes

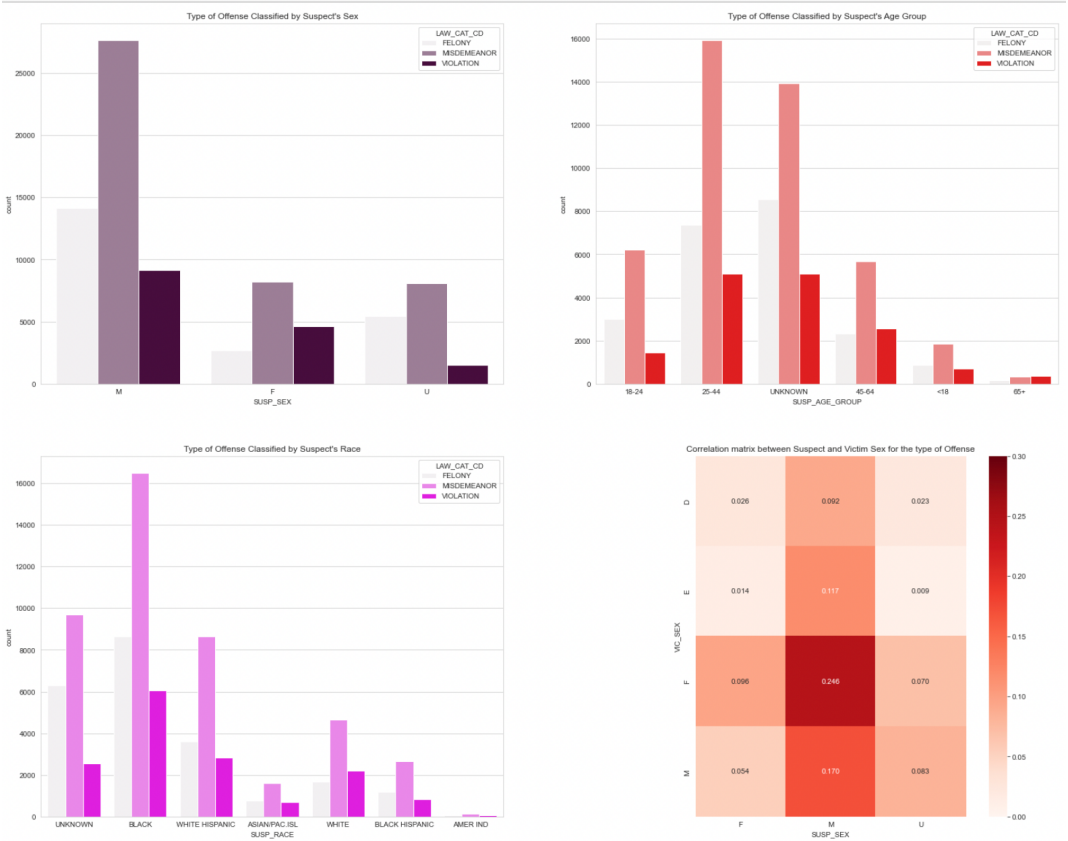

Fig 7: Visualization of Crimes by Suspect's Attributes

As the data acquired is Geospatial, I have plotted a heat map for the crimes happened in New York. Even the subset of the data I used for all other visualization was too big to plot, I took even a smaller subset of data consisting of the data just for the year 2018. The heat map visualization was done with the help of Kepler GL python library [5-6]. In the year 2018, Manhattan recorded majority of crimes as compared to Brooklyn.

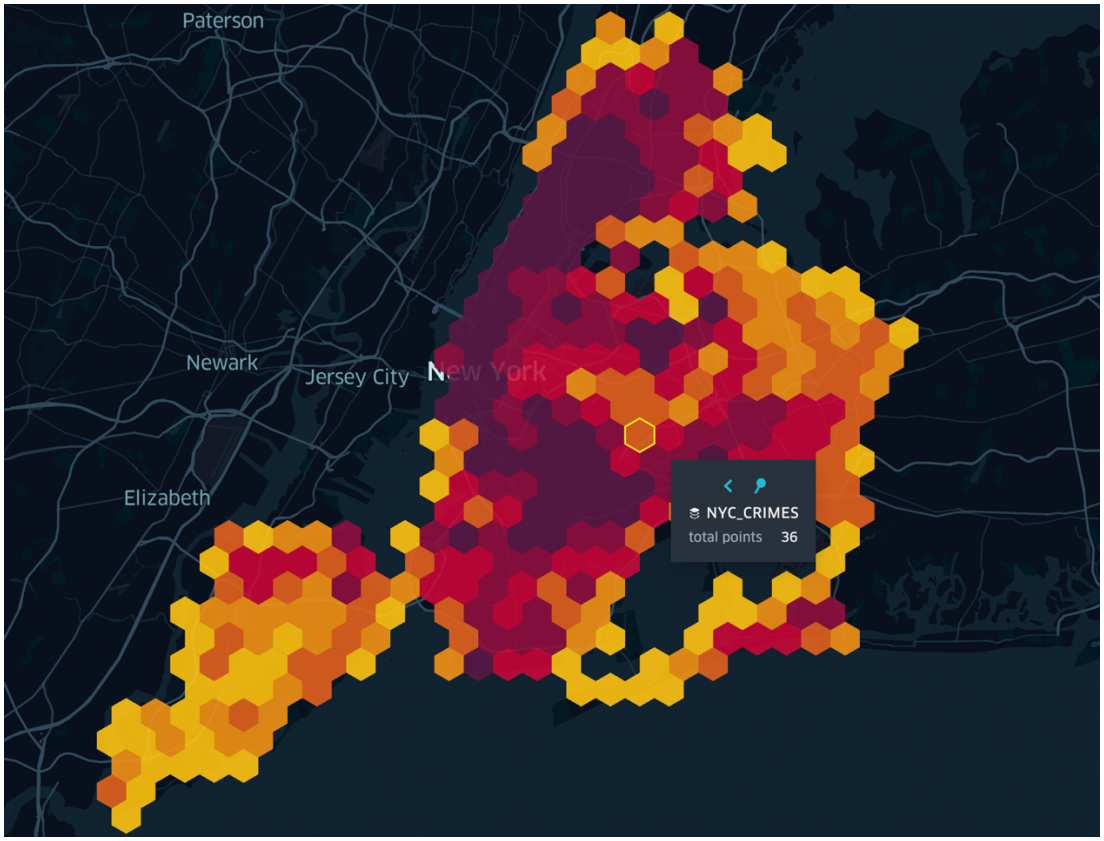

Fig 8: Heatmap of NYC's Crimes for the year 2018

MACHINE LEARNING MODELS:

As discussed in the Related work section I took guidance from the paper publication and have implemented their prediction models on my dataset to predict the crimes across NYC. I have used a subset of the data consisting only important features such as Longitude, Latitude, Date, Month, Borough and Description of type of premises where the crime took place and using the type of crime as Classification Label.
As the data is textual, it needed to be converted into the floating point via encoding so that it can be passed to the models. I have used original encoder to convert the data to floating point variables. This data was passed on to the models.
I have used four models to predict the crimes across NYC and have tabulated the recorded accuracy.

MODEL	ACCURACY
K-Nearest Neighbors (N=7)	43.5%
Decision Tree Classifier (Depth = 7)	53.16%
Gaussian Naïve Bayes	52.86%
Random Forest Classifier (Depth =7)	52.87%

Table 1: Accuracy Matrix

REFERENCES:

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

[2] https://dev.socrata.com/foundry/data.cityofnewyork.us/qgea-i56i

[3] https://arxiv.org/pdf/1607.03626v1.pdf

[4] https://www.researchgate.net/publication/280722606

[5] https://towardsdatascience.com/kepler-gl-jupyter-notebooks-geospatial-data-visualization-with-ubers-opensource-kepler-gl-b1c2423d066f

[6] https://www.kaggle.com/mariamingallon/kepler-gl-hex-data-demo