CS-GY-6513 Big Data
Project Report

NYC Crime Analysis

Meixuan Chen(mc7146), Jiaqi Li(jl9555), Shuyi Lu(sl6736),
Zhenghan He(zh1158), Xinchao Min(xm644)

1. Introduction

New York city is the most populous city in the United States divided into 5 different boroughs, and every borough differs in demographics, wealth, and lifestyle. In other hand, the New York city is known as the "City Of Crime", so we want to to do analysis based on crime data to investigate crime rates, finding some correlation in crime factors and to make some prediction, which can help the police department to decrease the crime rate. The main focus is to investigate the changes in crime rates between 2005 and 2018, and the differences between boroughs in crime rates.

Data:

NYPD Complaint Data Historic (https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i) contains information about type of offense, time of occurence, specific location and borough (6.04 Millions Rows, 35 Features)

NYC Population (https://www1.nyc.gov/site/planning/data-maps/nyc-population/current-future-populations.page)

NYC Borough Area (https://en.wikipedia.org/wiki/Boroughs_of_New_York_City)

Data Cleaning:

Merged historic and current part of the NYPD Complaint Data. Removed columns with significant number of NaN. Dropped all the irrelevant columns. Removed erroneous data (e.g. year = 1015). Kept the record with crime happened between 2005 and 2018. Changed categorical variable to cardinal descriptive name.

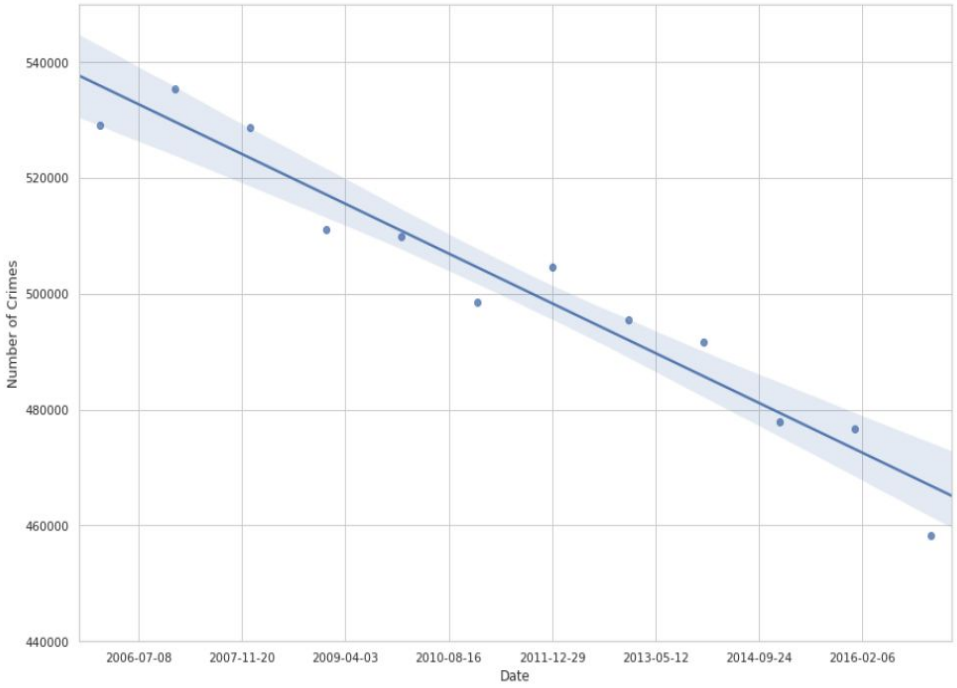
Framework:

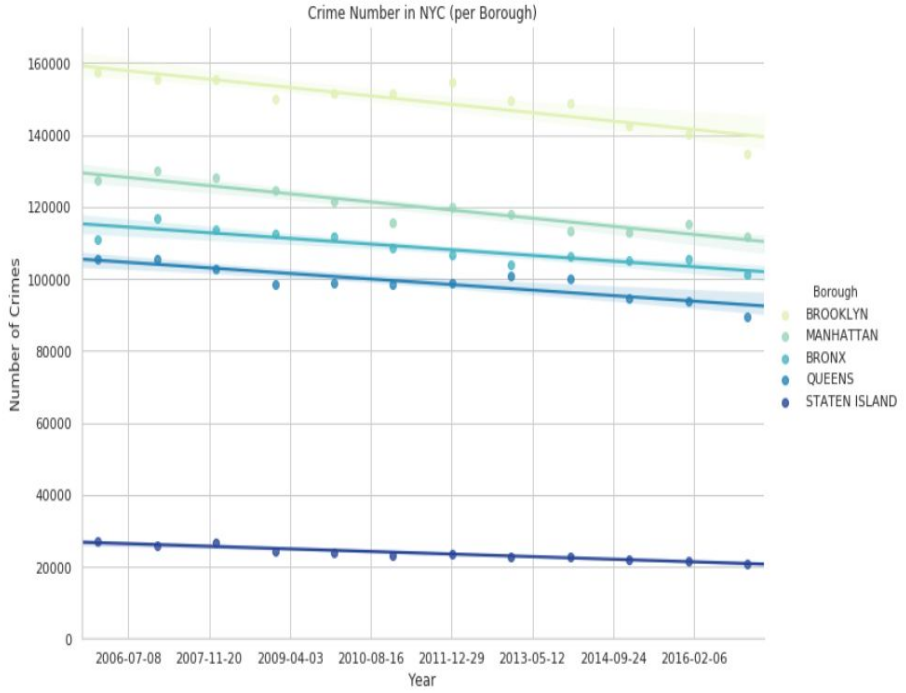
Using PySpark to cleaning and processing all data.
Using Python libraries to visualize the result.
Using Spark ML to do clustering and predicting.

2. Basic Data Exploration

Crime rates:

In first part, it shows that on average the number of reported crimes decline every year on the city and borough level.

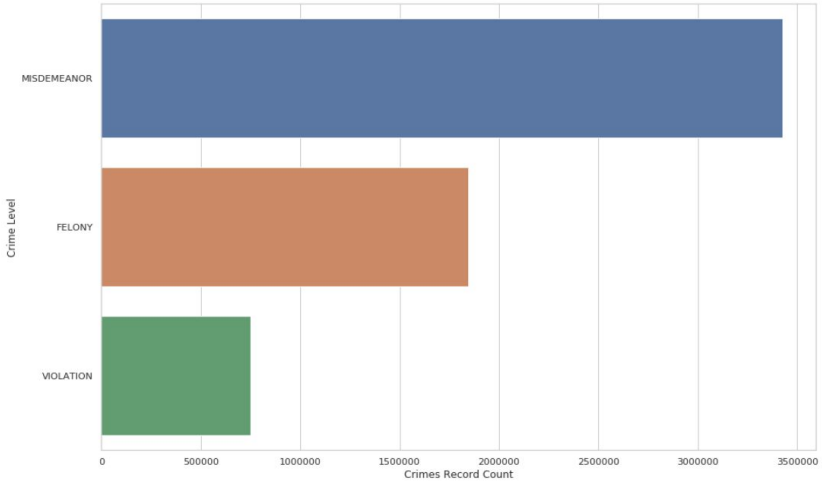


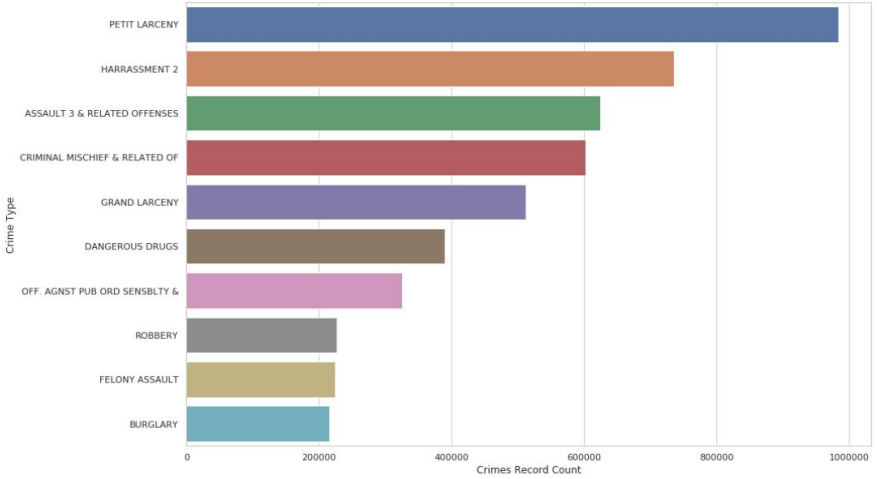


We can see that generally there is steady drop of around 7000 crimes per year
Crime rate decline is observed in all the boroughs, with the fastest drop for Brooklyn and slowest for Staten Island.

Distribution of crime levels and crime types:

Then we try to analyze the distribution of crime levels and crime types. The results are shown below.

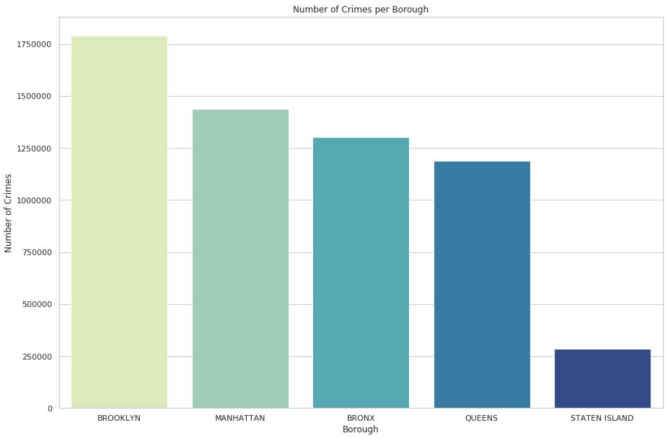


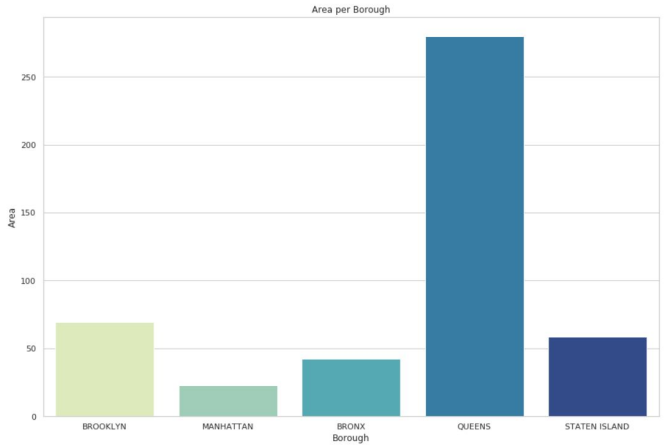


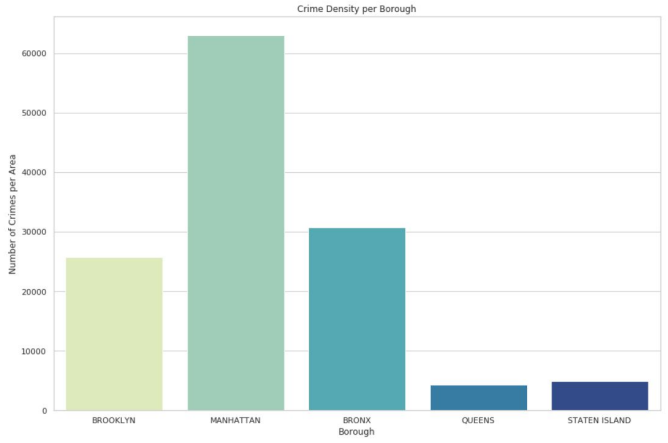
It is generally better to live in a city with where the number of felonies is lower than the number of misdemeanors and the number of misdemeanors is lower than number of violations.

Crime density:

Number of crimes might not be the best indicator that compares the boroughs. As shown in this picture the boroughs differ in area size significantly. The same number of crimes can be considered less problematic if scattered over a bigger area. We try to visualize the area of each borough and the number of crimes per borough. By diving the two number, we get the crime density of each borough.

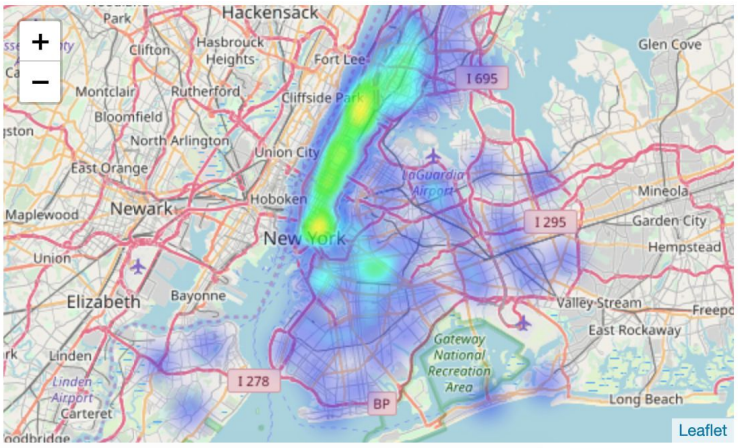






One can see from this picture that Brooklyn has the most crimes of all the boroughs, however when crime density is taken into account. Manhattan becomes a ranking leader.

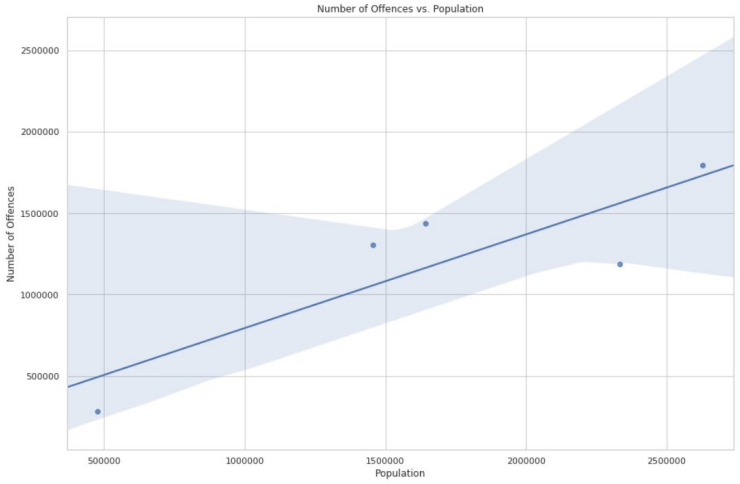
We also use folium to get heat map to visualize the result.



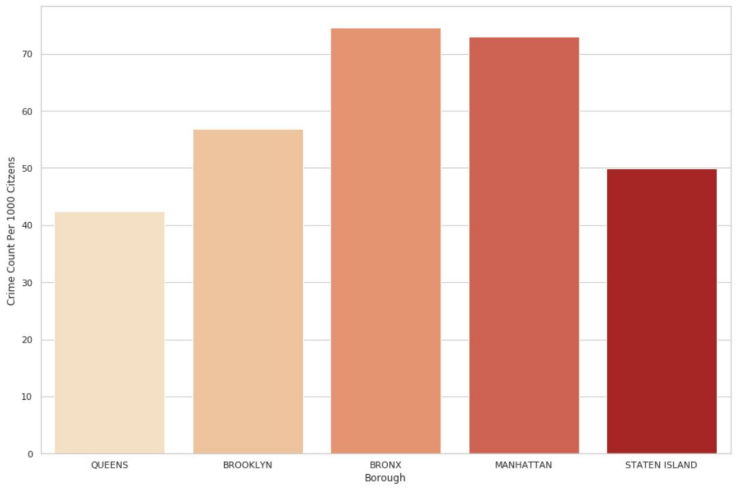
This Heat map plotted confirms that Manhattan is the most crime dense borough. Staten Island in comparisons has very low values of number of crimes and of crime density, respectively.

Crime rates and population

Population also can impact crime rates. The bigger population generally the likelihood of crime should be generally greater. The relationship between borough population and number of recorded complaints was investigated for 2016.

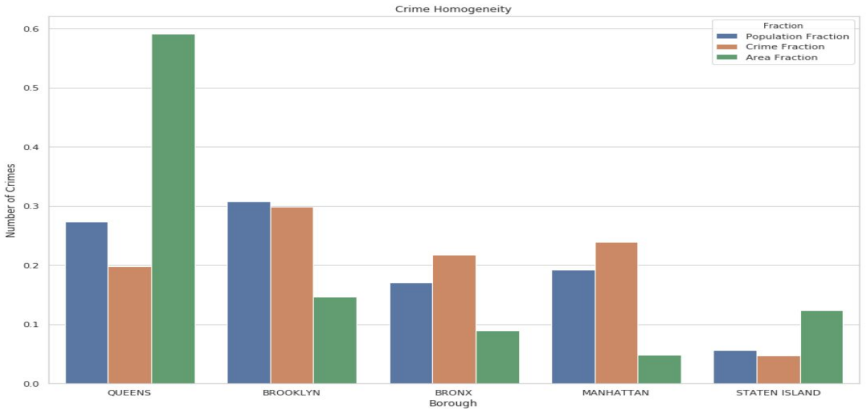


Generally, there is a positive correlation between population of boroughs and the number of crimes. To compare the crime in population, scale an indicator of Crime Rate per 1000 Citizens was calculated, which shows the number of crimes that on average is recorded among 1000 citizens in a given borough.



In this comparison, Bronx followed closely by Manhattan has the highest reported crime rates per 1000 citizens.

3. Crime & Area Analyze:



This is a straightforward statistical illustration, from which we can see that population, crime and area fraction in these different boroughs. Intuitively, the area of each borough is not necessarily related to the crime rate. For example, Queens has the most population fraction but a small crime rate. However, the population is quite related to the crime rate, almost proportional, because every borough has nearly the same height of population fraction and crime fraction. Then let's do a more detailed analysis.

Step 1:

Here's the original data set which is about numbers of crime between boroughs and offence_codes.

Offence_Code	102	103	104	105	106	107	109	110	111	112	...	364	365	455	571	572	578	675	676
Borough																			
BRONX	15.0	31.0	3513.0	52850.0	59449.0	38388.0	67294.0	24730.0	1779.0	10590.0	...	3292.0	2333.0	18.0	13.0	244.0	159105.0	238.0	1.0
BROOKLYN	30.0	52.0	4730.0	78531.0	74822.0	74023.0	135613.0	37015.0	3281.0	17458.0	...	1041.0	4403.0	15.0	116.0	323.0	221006.0	351.0	9.0
MANHATTAN	18.0	24.0	3388.0	43899.0	40448.0	39924.0	198592.0	11857.0	3389.0	18613.0	...	1804.0	1214.0	13.0	82.0	250.0	147117.0	609.0	19.0
QUEENS	26.0	26.0	3568.0	46261.0	42798.0	55676.0	96883.0	35844.0	2084.0	12514.0	...	469.0	3157.0	16.0	17.0	74.0	153596.0	303.0	5.0
STATEN ISLAND	10.0	4.0	604.0	5184.0	6802.0	7836.0	13573.0	4140.0	229.0	3834.0	...	182.0	561.0	1.0	3.0	40.0	53582.0	54.0	1.0

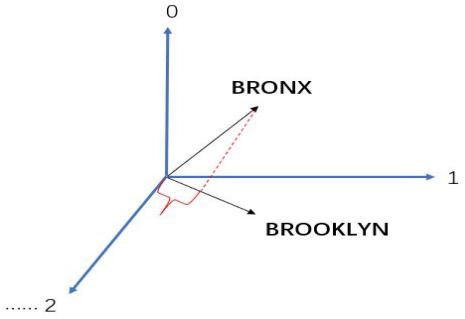
5 rows × 60 columns

We first normalize it into the range from 0 to 1.

	0	1	2	3	4	5	6	7	8	9	...
BROOKLYN	0.000042	0.000087	0.009828	0.147851	0.166312	0.107393	0.188259	0.069184	0.004977	0.029626	...
BRONX	0.000060	0.000104	0.009438	0.156694	0.149294	0.147700	0.270591	0.073857	0.006547	0.034834	...
QUEENS	0.000040	0.000054	0.007595	0.098415	0.090678	0.089503	0.445212	0.026582	0.007598	0.041727	...
MANHATTAN	0.000077	0.000077	0.010530	0.136525	0.126305	0.164311	0.285920	0.105783	0.006150	0.036931	...
STATEN ISLAND	0.000113	0.000045	0.006802	0.058384	0.076607	0.088252	0.152864	0.046626	0.002579	0.043180	...

5 rows × 60 columns

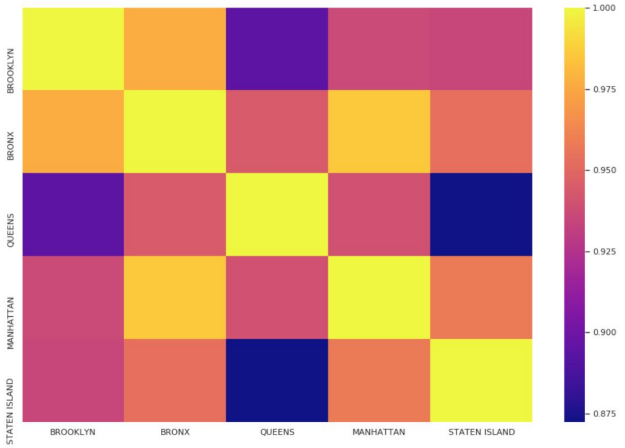
Next, we turn these data into a high dimensional space, where every offence_code will become an attribute, which means 70 dimensions totally. We only draw three dimensions but it's the same meaning.



In order to analyze the similarity of boroughs, we take advantage of dot product. The dot product will be the projection of one vector say "BROOKLYN" onto another vector like "BRONX". We use a loop to computer the dot product of every possible pair of boroughs and the result is shown below.

	BROOKLYN	BRONX	QUEENS	MANHATTAN	STATEN ISLAND
BROOKLYN	1.000000	0.976945	0.893947	0.936941	0.935516
BRONX	0.976945	1.000000	0.944328	0.985771	0.953748
QUEENS	0.893947	0.944328	1.000000	0.939772	0.872447
MANHATTAN	0.936941	0.985771	0.939772	1.000000	0.958130
STATEN ISLAND	0.935516	0.953748	0.872447	0.958130	1.000000

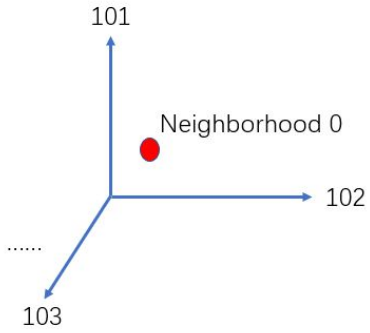
After visualization, we can see the similarity between different boroughs on a general and high level. For example, "BRONX" and "BROOKLYN" both have vivid colors, which indicates they have high similarity by crime patterns we define.



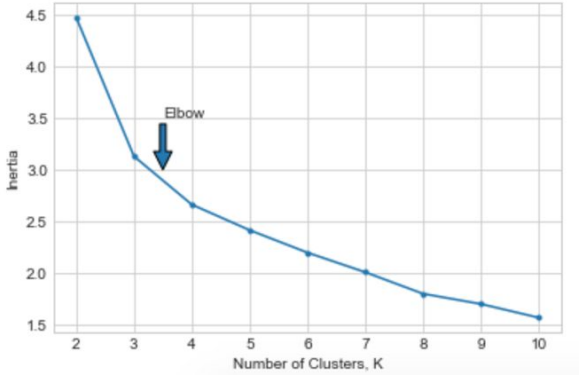
Step 2:

We want to classify neighbors into different types, remember neighbors are smaller

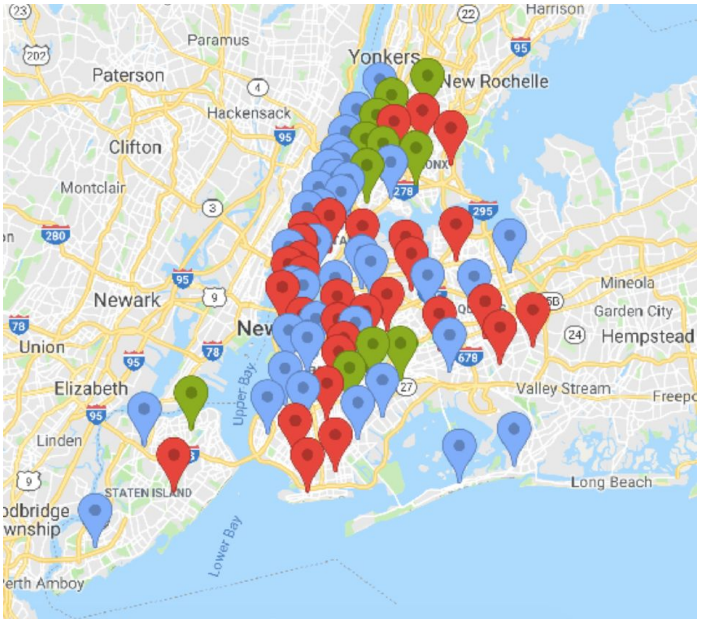
concepts than boroughs. So, this time different neighborhoods will be a point in such high-dimensional space based on its value on each property.



Next, we use K-means to classify neighborhoods into three types. We first find the relation between inertia and number of clusters and determine that k=3. The relation is shown below.



After visualization, we can see three different colors in the map. The green points are at the location of Bronx, Upper Manhattan and somewhere in Brooklyn. If you look at data set carefully, these green dots represent places where more severe crimes occur, like murder, dangerous drugs, felony assault, robber and so on.



4. Prediction:

We try to use the data to predict whether a crime will happen and the level of the offense by giving the day of week, latitude, longitude, neighborhood borough. The algorithm we choose: K nearest neighbor and Random Forest

Data preprocessing in prediction:

- For non-sequential attributes such as day of the week, neighborhood, borough, we do the one-hot encoding.
- To prevent the model relying on the dimension that have large scale, we do normalization on the data before we feed them to our model.
- We also find that there are some missing values in our original data. To fill those blanks in our dataset, we use KNN algorithm.

Our Model and test result:

KNN:

For the choice of k in KNN, we tried k equals to 10, 20 and 30 and do cross-validation on it, and finally, we choose k equals to 30. And the accuracy and confusion matrix are shown below.

```
from sklearn.metrics import accuracy_score,confusion_matrix
accuracy_score(y_test,predictions)
```

```
0.5590719119142067
```

```
cm = confusion_matrix(predictions,y_test)
cm
```

```
array([[930265, 475748, 200106],
       [ 90133,  72412,  21366],
       [  3464,   1600,   2063]])
```

Random Forest:

We used the same method to choose the number of estimators in random forest, we tried k equals to 20, 50 and 100 and finally we choose k equals to 100. And the accuracy and confusion matrix are shown below.

```
from sklearn.metrics import accuracy_score,confusion_matrix
accuracy_score(y_test,y_predict)
```

```
0.4181326394967162
```

```
cm = confusion_matrix(y_predict,y_test)
cm
```

```
array([[453069, 200116,  71561],
       [295154, 205867,  59460],
       [275639, 143777,  92514]])
```

5. Summary

Result:

Number of offenses decreases every year on the city level as well as in borough level. Crime rates show huge differences if studied per area or per population of given borough. Neighboring precincts show the similarity of offence type and number of offences.

Limitation:

There is no information about police resources deployed in each borough so we can not find the relation between the crime rate and police resources. The population data is based on borough level but it varies among precincts, so the overall crime density could varies by precincts or neighborhood. It is not reliable to predict crime solely based on time and location, we need to find more stronger features connect or relate to the crime rate.

Future Work:

We can study the relation of the housing market with the crime rate (NYC Dept. of Finance provides information on property prices and number of sales per year per borough per type of home). It is very useful to predict the rental or sale price with the influence of crime rate, and it should have a reliable prediction.

Study the relation of the neighborhood demographics with the crime rate (NYU Furman Center provides a lot of information on neighborhood demographics between 2010 and 2015). The demographics information could help us to improve the prediction of crime furthermore.