SPRINGBOARD
DATA SCIENCE CAREER TRACK

CAPSTONE PROJECT 1

ANALYSIS OF CRIMES IN THE CITY OF NEW YORK

MICHAL CZAPSKI

JULY 2018

Table of Contents

Introduction

New York City is the most populous city in the USA. Its population is estimated for around 8.5 million people and keeps on growing. The city encompasses five different administrative divisions called boroughs (Fig. 1). Each borough historically was characterized by different demographics, wealth and lifestyle. They also differ significantly in areas they cover. Population dense cities often have to deal with increased crime rates, which usually are unevenly distributed within a city. The main goal of this study is to understand how crime rate changed throughout the decade between 2006 and 2016 in the city of New York and find out if there is any major differences between boroughs in crime rate on its own and in relation to demographics and housing market.

Figure 1. New York City Boroughs (1 – Manhattan, 2 – Brooklyn, 3 – Queens, 4 – The Bronx, 5 – Staten Island)

Databases

Data sources used for this project were:

- NYPD Complaint Data Historic (https://data.cityofnewyork.us/PublicSafety/NYPD-Complaint-Data-Historic/qgea-i56i) (data from 2006 till 2015) and Current (https://catalog.data.gov/dataset/nypdcomplaint-data-current-ytd) (2016 and 2017). It contains information about type of offence, time of occurrence, specific location and borough.
- NYU Furman Center (https://data.cityofnewyork.us/Public-Safety/NYPD Complaint-Data-Historic/qgea-i56i) provides a lot of information on neighborhood demographics between 2010 and 2015.
- NYC Department of Finance (http://www1.nyc.gov/site/finance/taxes/property-annualized-salesupdate.page) - Homes Sales Data from 2005 through 2016, provides information on property prices (means, medians) and number of sales per year per borough per type of home.
- NYC Department of Planning(https://www1.nyc.gov/site/planning/data-maps/nycpopulation/current-future-populations.page) provides information on population size in 2010 and 2016.

Other potential databases that could be used are:

- NYC Department of Finance (http://www1.nyc.gov/site/finance/taxes/property-annualized-salesupdate.page), Detailed Annual Sales Reports by Borough is the same datasource as Neighborhood Sales Data listed above, but with detailed prices instead of aggregated information.
- United States Census Bureau (https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045216) provides census information on NYC, but the data is difficult to collect, and basically NYU Furman Center provides exactly the same information in easier to read and collect form.
- New York City Police Department (http://www1.nyc.gov/site/nypd/bureaus/patrol/find-yourprecinct.page) also provides crime statistics per precinct, but the data should be consistent withthe data from the other NYPD source used in this analysis.

Client

There are several clients potentially interested in the outcome of this investigation:

- New York Police Department (NYPD) should be interested in understanding how to better deploy police resources in the city in order to decrease the crime rate differences between boroughs. Additionally, they can compare the outcome of this study with the policies and means that were introduced during the last decade and their impact on crime rates in the city.

- New York City authorities work on urban planning, and their own policies that can improve the quality of life in the city. Therefore offence analysis will allow the city to focus on projects that can improve the situation in certain boroughs and make them more attractive to the potential real estate developers and new citizens.

- Housing market is influenced by many factors that can have impact on home prices. This analysis can be useful for companies in the housing business to better future pricing in relation to crime situation.

Data Wrangling

NYPD Complaint Data - Historic and Current

- The historic and current parts of NYPD Complaint database were concatenated into one database. All columns not needed for further analysis were dropped (e.g. park name or jurisdiction responsible for the offence, which is more NYPD internal information).
- Columns with significant number of NaN values were dropped too, since the information they contained would be only useful if NaNs were only 10 percent or less of the all rows.
- The original database had 3 date columns (offence start date, offence end date and offence reporting date). Since only approximate date for the crime for statistical reasons is needed, the start date of the offence was kept and NaN values were filled based on the missing values either from the offence end date column or offence reporting date, which had no NaN entries. Both dates generally should be similar to the offence start date.
- Column names were changed from NYPD abbreviations to more descriptive names.
- Rows with erroneous year entries were also removed from the database (e.g. year 1015).
- Number of crimes after 2006 is significantly bigger and consistent till 2017. The only rationale for state is lack of consistency in updating the database earlier than 2006; therefore the analyzed period starts from 2006 till the end of the database.
- Index was changed into datetime index, which only possible after removing erroneous years, since pandas can only represent the time span limited to around 586 years.
- Numerical entries representing cardinal numbers were converted into strings.

Neighborhood Sales Data

- The information about NYC property sales was available in a few excel files with several sheets per file. All spreadsheets were concatenated into one database.
- Columns representing name and borough that were originally only given in the file and sheet names were added as separate columns to the database.
- Column representing total number of properties sold was dropped since it had significant number of NaNs.
- Irregular names for type of houses were changed into three categories (one, two and three family home).

NYU Furman Center Data

- The data from NYU Furman Center was available on the website in numerous excel spreadsheets.
- All spreadsheets were concatenated into one database.
- Irrelevant to this analysis columns were removed.
- Part of the database had to be melt since the data for separate years was given in separate columns.
- Entries for years below 2006 were remove since they are not part of this analysis.
- A few type transformations were performed since there were many entries that can normally can be easily handled by excel but were imported as strings (including percentage, and currency signs i.e. '%' ,'$').
- There were NaN values encoded in two ways. i.e. as regular numpy NaN recognized by pandas and as string NA (recognized by Excel) that caused a lot of problems since pandas couldn't convert those and they couldn't be filtered out easily from the database for user-specified conversion.
- There is also some missing data for 2016, which could be filled out with extrapolation values based on earlier years in the further analysis.

Data Exploration

Crime Rates

There are 6,029,561 complaints registered in the NYPD dataset between 2006 and 2016. Tab. 1 explains in detail the features of this dataset. Additionally, borough area information and information about each borough population from the Census of 2016 were added to the database (Tab. 2).

Column	Description
ComplaintID	Randomly generated persistent ID for each complaint
Date	Exact date of occurrence for the reported event (or starting date of occurrence)
Offence Code	Three digit offense classification code
Description	Description of offense corresponding with key code
Internal Code	Three digit internal classification code (more granular than Offence Code)
Internal Description	Description of internal classification corresponding with Internal code (more granular than Description)
Status	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
Level of Offence	Level of offense: felony, misdemeanor, violation
Borough	The name of the borough in which the incident occurred
Neighborhood	The precinct in which the incident occurred
Premise Description	Specific description of premises; grocery store, residence, street, etc.
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

Tab. 1 NYPD Complaint Dataset (cleaned)

Borough	Area (sq. mi)	Population (2016)
Manhattan	22.82	1,643,734
Brooklyn	69.5	2,629,150
Queens	108.1	2,333,054
Bronx	42.47	1,455,720
Staten Island	58.69	476,015

Tab. 2 Boroughs' area and population (2016)

In the first part of this analysis the crime rate for the city and for each borough throughout the years 2006 till 2017 was analyzed.

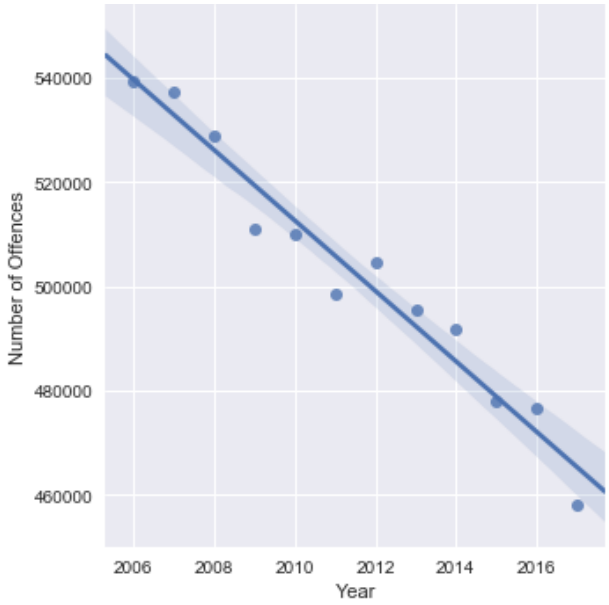

Figure 2. Yearly crime rate for the city of New York

One can see that generally there is steady drop of around 7000 crimes per year. And around year 2012 the number of crimes decreased to less than 500000 crimes per year on average.

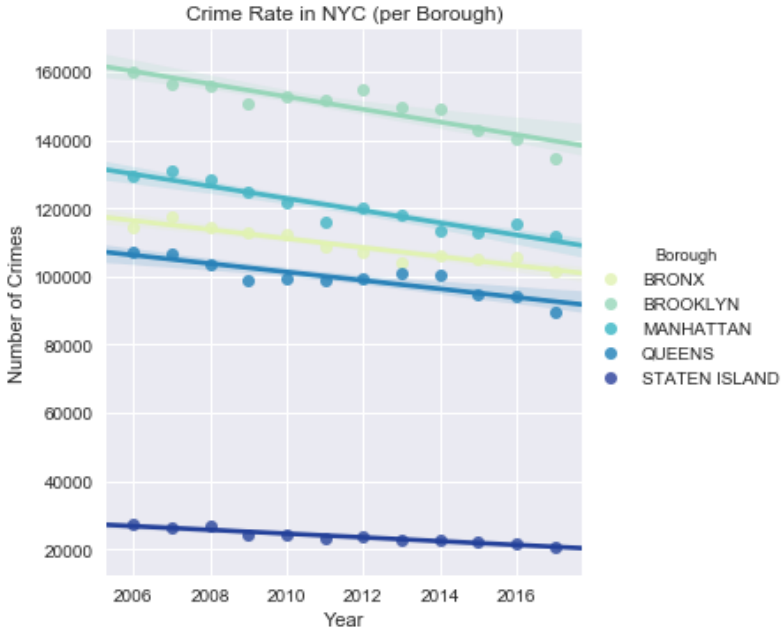

Figure 3. Yearly crime rates for the boroughs

Crime rate decline is observed in all the boroughs, with the fastest drop for Brooklyn (around 1858 crimes per year) and slowest for Staten Island (551 crimes per year).

Crime Density

Number of crimes might not be the best indicator that compares the boroughs. As shown in Tab. 2 and in Fig. 3 the boroughs differ in area size significantly e.g. Queens is almost five times bigger than Manhattan. The same number of crimes can be considered less problematic if scattered over a bigger area.

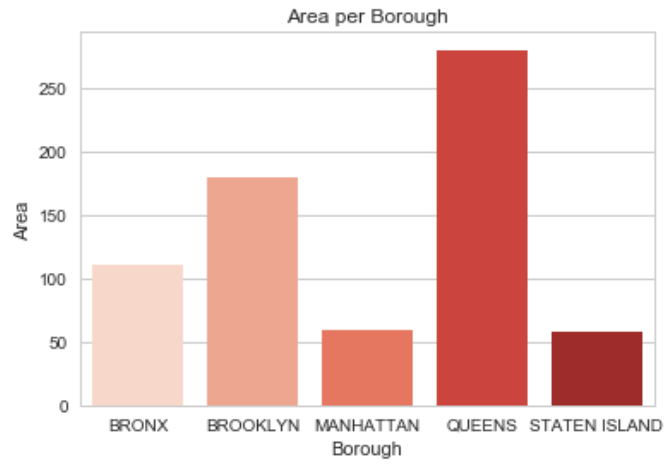

Figure 4. Area per borough (sq. km)

By dividing the number of crimes per area a new crime indicator was introduced called 'crime density'. The denser in crimes the borough is, the more crimes one can expect in their neighborhood.

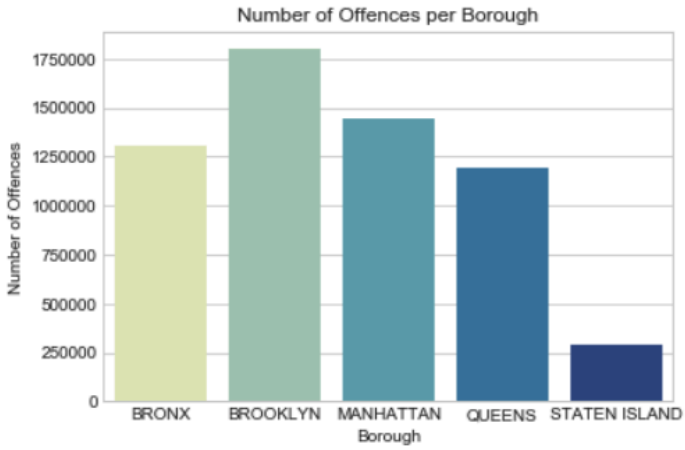

Figure 5. Number of crimes per borough

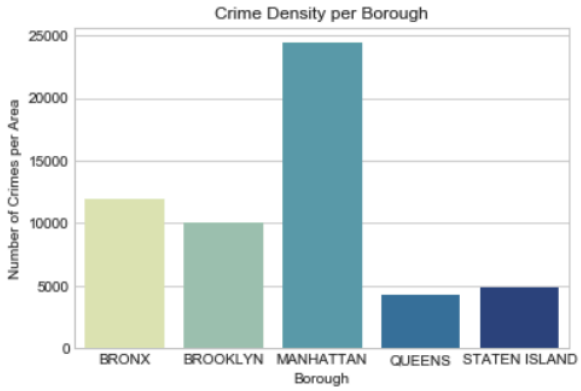

Figure 6. Crime density per borough

One can see from Fig. 5 that Brooklyn has the most crimes of all the boroughs, however when crime density is taken into account (Fig 6.) Manhattan becomes a ranking leader, with significantly higher crime density than other boroughs. Heat map plotted for precincts (Fig. 7) indeed confirms that Manhattan is the most crime dense borough (intense yellow, green coloring). Staten Island in comparisons has very low values of number of crimes and of crime density, respectively.

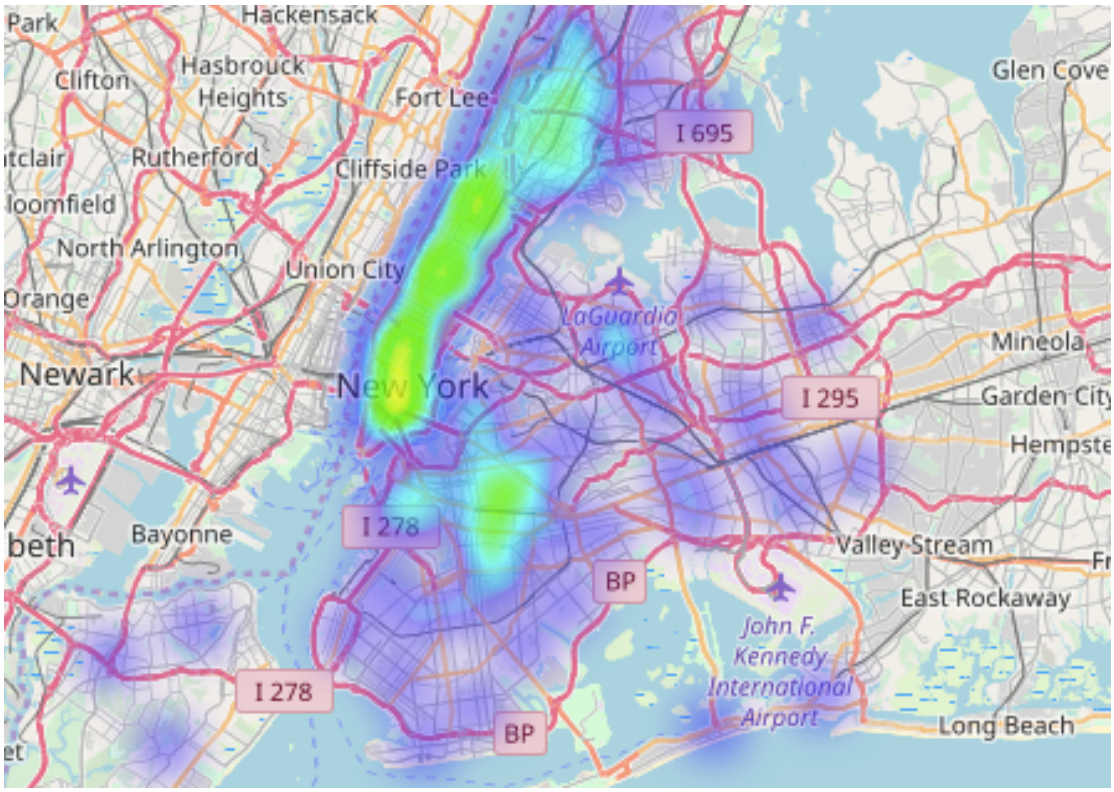

Figure 7. Crime heat map of New York City (with intensity decreasing in the following order: yellow – green – blue - purple)

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.

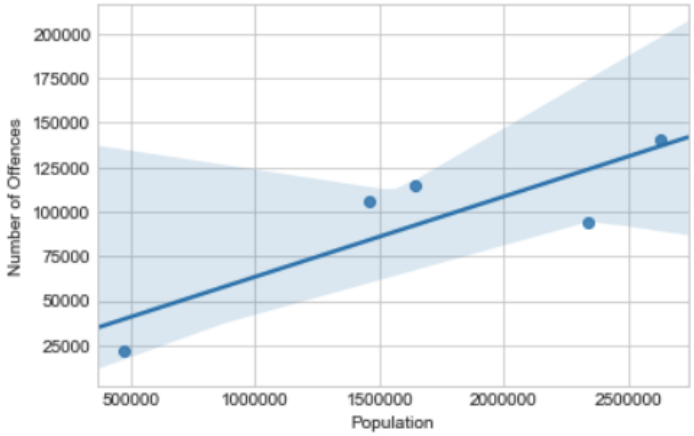

Figure 8. Correlation between boroughs' population and number of crimes for 2016.

Generally, there is a positive correlation between population of boroughs and the number of crime (variance in explained in 85%).

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 (Fig. 9). In this comparison, Bronx followed closely by Manhattan has the highest reported crime rates per 1000 citizens.

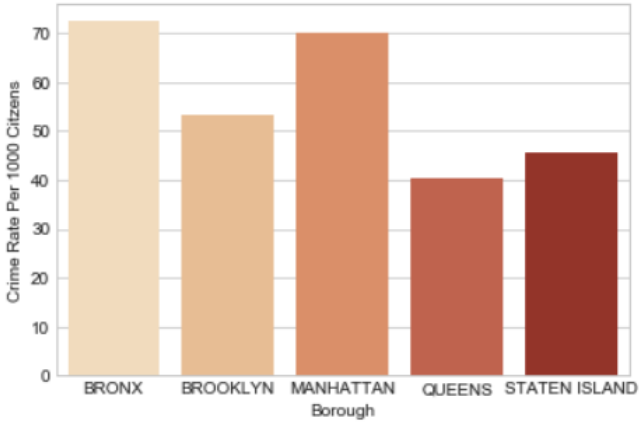

Figure 10. Crime rate per 1000 citizens for the boroughs

Crime Homogeneity

To summarize the differences between the boroughs with respect to the crime rates and the influence of area and population, fractions of recorded complaints were compared with the population fraction and the area fraction of respective boroughs. The Fig. 10 shows there are disproportions in crime homogeneity of the boroughs.

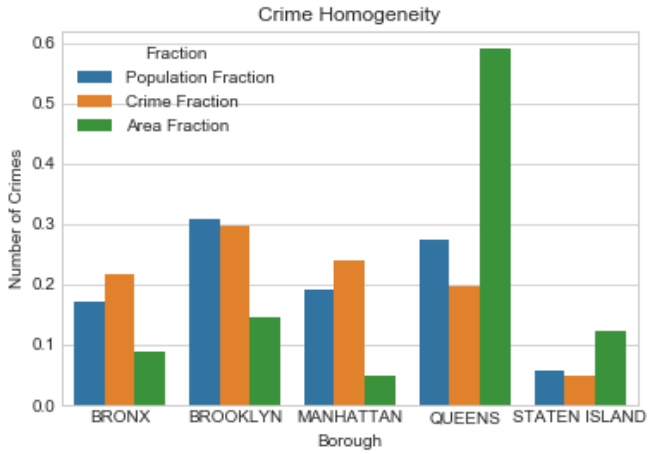

Figure 11. Crime homogeneity in New York City

Bronx and Manhattan have both higher crime fraction than population and area fraction. In contrary, Queens and Staten Island has the lowest crime fraction in comparison to area and population fractions. Queens' and Manhattan's area fractions are disproportionately higher or lower than other two fractions, respectively.

Crime Status Analysis

There are different types of crimes:

- **Felony** is a serious crime, typically one involving violence, regarded as more serious than a misdemeanor, and usually punishable by imprisonment for more than one year or by death, including such crimes as grand larceny, robbery, felony assault, burglary and others.
- **Misdemeanor** is a minor wrongdoing, and includes such offences like petit larceny, assault, criminal mischief or dangerous drugs.
- **Violation** is the latest category in the NYPD dataset and includes such offences like harassment, disorderly conduct, loitering and others.

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.

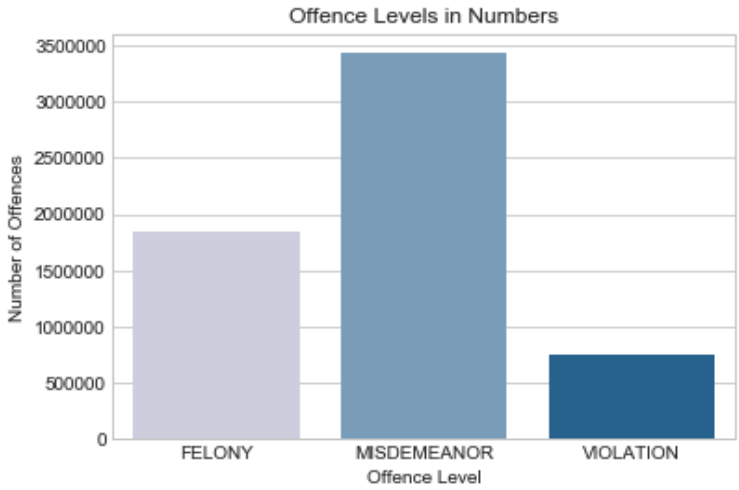

Figure 12. Number of offences of different level

In the case of the New York City misdemeanor is the most reported crime (Fig. 12). The number of misdemeanors is almost twice higher than the number of felonies and almost 5 times higher than the number of violations.

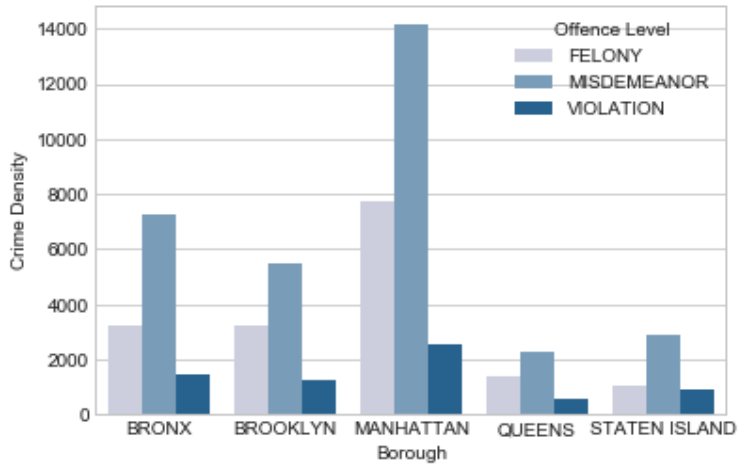

Figure 13. Crime density of crimes of different level per borough

The analysis per borough showed that, generally distribution of different level crimes is similar in all three boroughs and follows general trend that misdemeanors are almost twice more frequent that felonies and around five times more frequent than violations. Manhattan has the highest density of all three offences types, whereas Queens has the lowest.

Crime Rates and Housing Market

Many quantitative indicators can be used to study housing market. In this analysis number of sales and average median housing prices were used to investigate relationship with crime rates.

Number of Sales

Queens is the leading borough in terms of number of sales followed by Brooklyn (Fig. 14) whereas Manhattan is the borough with lowest number of sales.

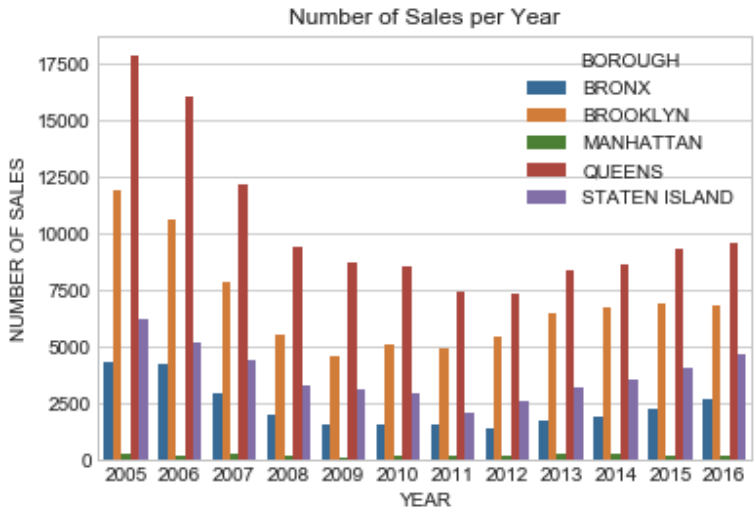

Figure 14. Number of sales of houses per borough

Similarly to crime density, sales density (number of sales per area) can be a better comparison indicator due to significant area differences between boroughs (Fig. 15). In that comparison, Staten Island has reached the highest sales densities for the last decade, Manhattan, on the other hand has the lowest number of sales per area.

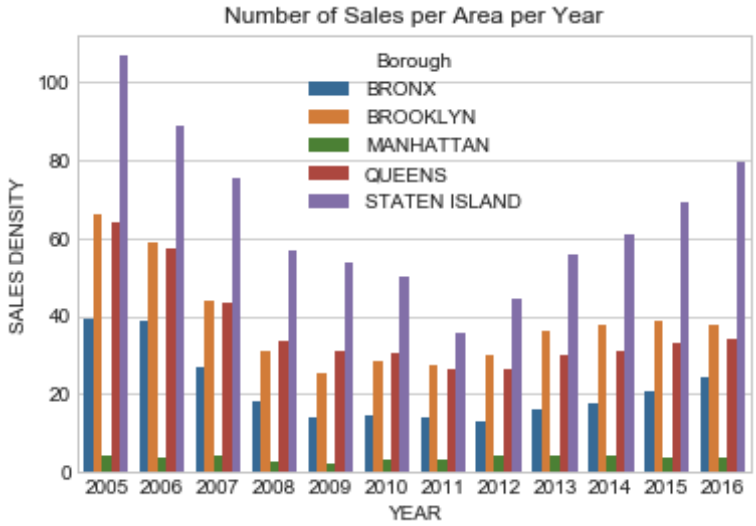

Figure 14. Houses' sales density per borough

Housing market in the whole city of New York seems to follow a huge sale decrease between 2010 and 2012, probably can be explained by the Great Recession period at beginning of 2010. There is a positive correlation observed between number of sales and crime density for Queens, which might be due to the fact that the number of sales didn't increase significantly after the recession

therefore the decrease over the period of 11 years between 2005 and 2016 is more pronounced that in case of other boroughs.

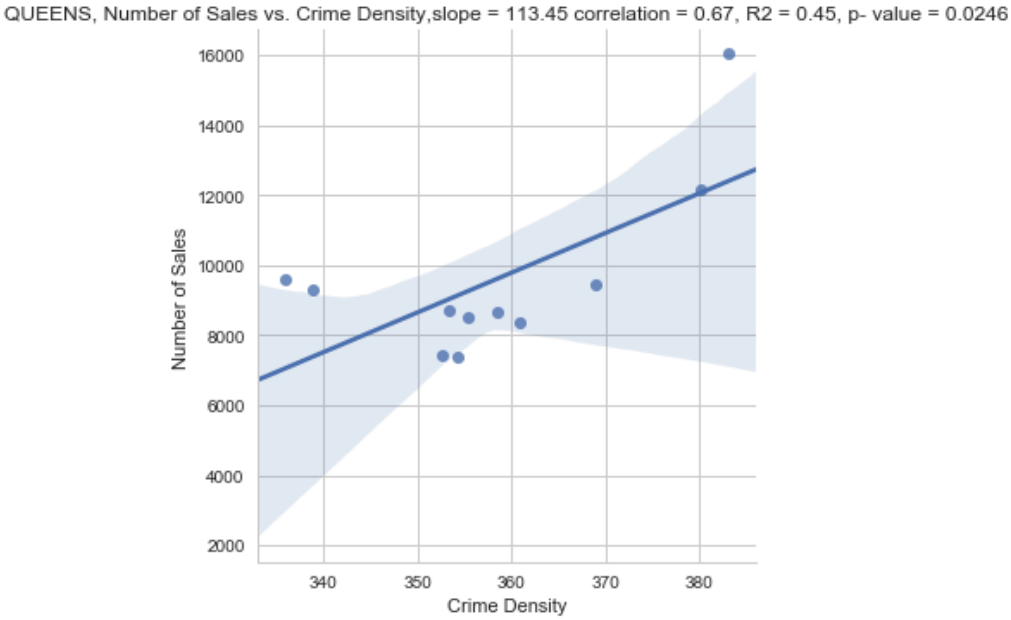

Figure 15. Correlation between number of sales and crime density for Queens

Average Median Sales

Average median sales vary significantly in order of magnitude between Manhattan the other boroughs (Fig. 16). In general, a dip in price ranges can observed during the Great Recession period. There are two negative correlations between average median sales and crime densities for Manhattan and Brooklyn. Those two boroughs apart from the recession period demonstrate a steady increase in property prices.

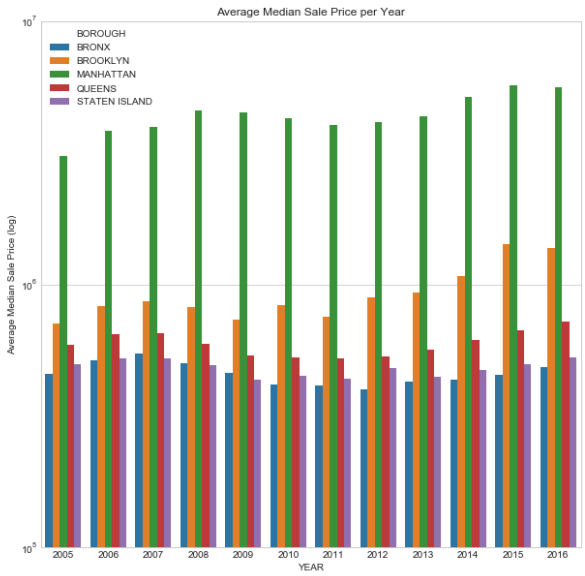

Figure 16. Average median sale price per year per borough

Queens and Staten Island don't correlate significantly with crime densities and their median prices on average remained steady during the period of 2005 – 2016(Fig. 17 and 18). The Bronx didn't recover after the recession and the prices levels stayed lower than before the recession period, because of which, housing prices correlate negatively with crime densities (Fig. 19).

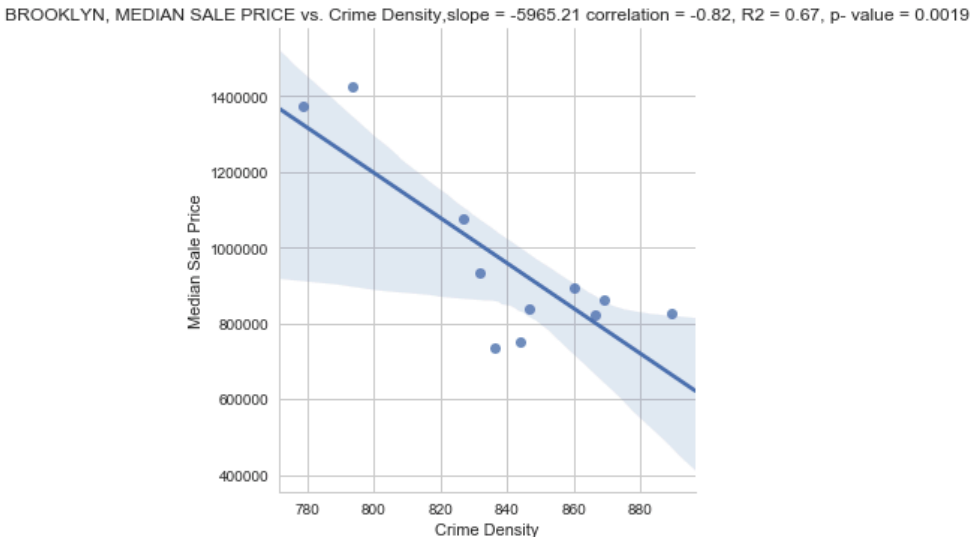

Figure 17. Correlation between median sale price and crime density for Brooklyn

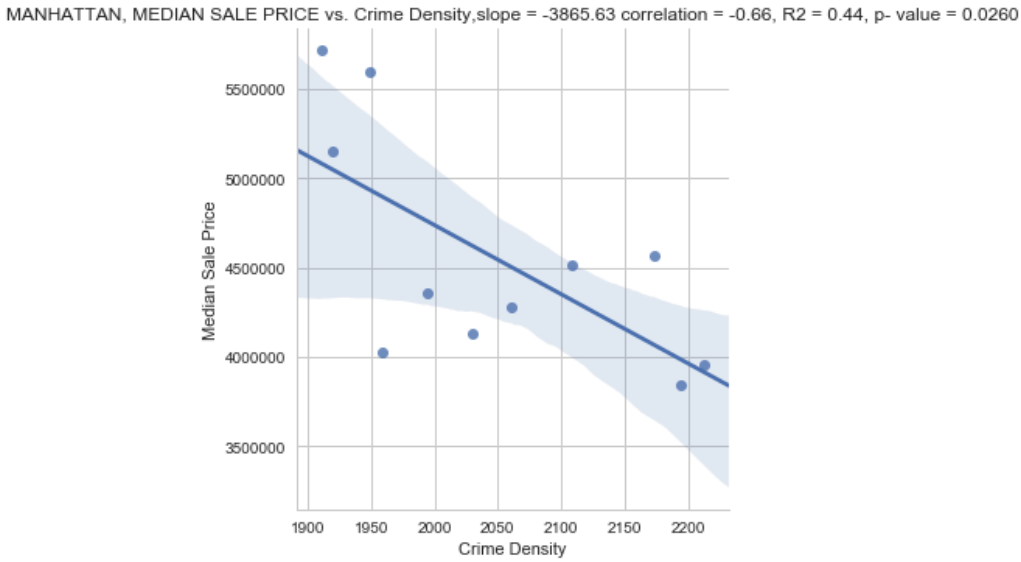

Figure 18. Correlation between median sale price and crime density for Manhattan

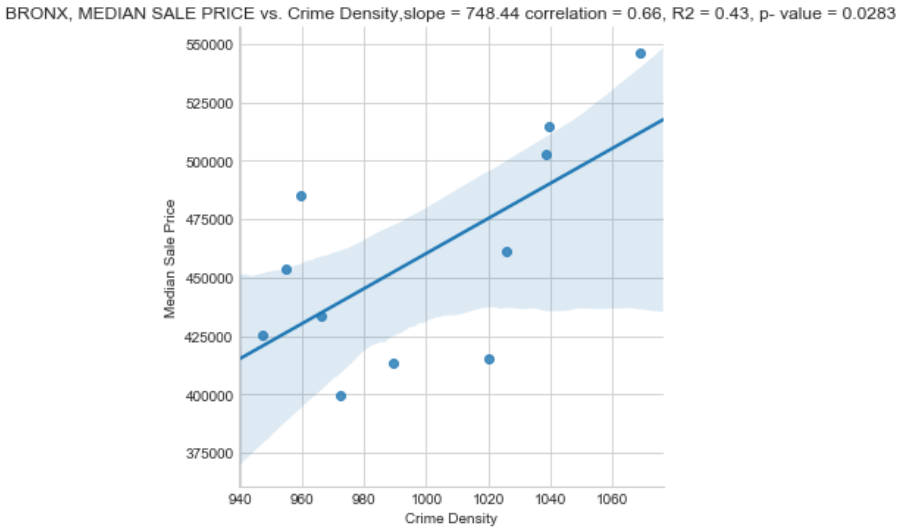

Figure 19. Correlation between median sale price and crime density for Bronx

Data Modeling (Supervised Learning)

Crime Rates and Demographics

Linear Regression Model

NYU Furman Center Data was used to create a model predicting serious crime rates (per 1,000 residents) based on different demographic metrics at the neighborhood level of granularity. Among different predictors the ones that were initially selected for the regression model are shown in Tab. 3.

Indicator	Description
Borough	Borough
Year	Year
Single-person households	The share of households that include only one person.
Income diversity ratio	The income earned by the 80th percentile household divided by the income earned by the 20th percentile household for a given geographic area, excluding all households without positive income.
Poverty rate	The number of people below the poverty threshold divided by the number of people for whom poverty status was determined.
Poverty rate, population aged 65+	The number of people aged 65 or older below the poverty line divided by the total population of that age group for whom poverty status was determined.
Poverty rate, population under 18 years old	The number of people under 18 years old below the poverty line divided by the total population of that age group for whom poverty status was determined.
Disconnected youth	The percentage of people aged 16 to 19 who were neither enrolled in school nor participating in the labor force.
Labor force participation rate	The number of people aged 16 years and older who are in the civilian labor force, divided by the total number of non-institutionalized people aged 16 years and older.

Population aged 25+ with a bachelor's degree or higher	The percentage of the population aged 25 and older who have attained a bachelor's degree or higher.
Population aged 25+ without a high school diploma	The percentage of the population aged 25 and older who have not graduated from high school or received a GED.
Unemployment rate	The number of people aged 16 years and older in the civilian labor force who are unemployed, divided by the total number of people aged 16 years and older in the civilian labor force.
Born in New York State	The percentage of city residents who were born in New York State.
Disabled population	The percentage of the adult population who have disabilities that impair hearing, vision, ambulation, cognition, self-care, or independent living.
Foreign-born population	The share of the population that is born outside the United States or Puerto Rico.
Population	All people, both children and adults, living in a given geographic area.
Population aged 65+	The percentage of residents who are aged 65 years and older.
Percent Asian	The percentage of the total population that identifies as Asian (non-Hispanic).
Percent black	The percentage of the total population that identifies as black (non-Hispanic).
Percent Hispanic	The percentage of the total population that identifies as Hispanic (of any race).
Percent white	The percentage of the total population that identifies as white (non-Hispanic).
Racial diversity index	The probability that two randomly chosen people in a given geographic area will be of a different race.
Population density (1,000 persons per square mile)	The geographic area's population divided by its land area and is reported in thousands of people per square mile.

Tab. 3. Demographic metrics

All the predictors but Borough are quantitative variables, therefore Borough had to be transformed into dummy variable first. The initial regression model was examined for outliers (using studentized residuals) and high leverage points.

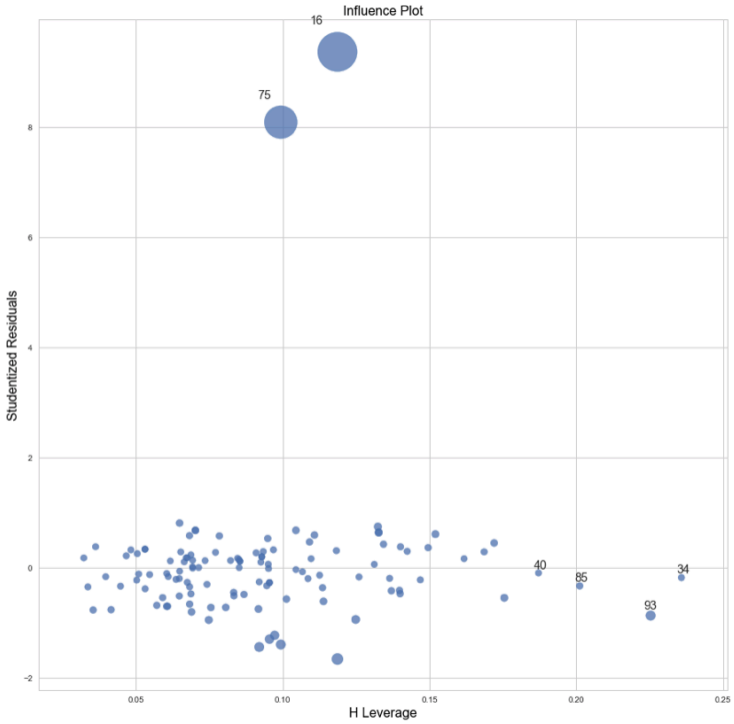

Figure 20. High Leverage vs. studentized Residuals plot for demographic indicators vs. severe crime rates linear regression model

The plot in Fig. 20 shows plot high leverage points vs. studentized residuals. Two major outliers at position 16 and 75 were observed. Those points belong to Manhattan neighborhoods and will be removed from the model. After the process of backward elimination predictors that p-value is less than 5%, the final 9 variables were selected:

- Year (A2)
- Born in New York State (A3)
- Foreign-born population (B1)
- Income diversity ratio (C1)
- Percent Asian (C2)
- Percent Hispanic (D1)
- Population aged 25+ with a bachelor's degree or higher (E1)
- Population aged 65+ (E3)
- Single-person households (G2)
- Population density (1,000 persons per square mile) (Q2)

OLS Regression Results

Dep. Variable:	S1	R-squared:	0.680
Model:	OLS	Adj. R-squared:	0.653
Method:	Least Squares	F-statistic:	25.04
Date:	Fri, 06 Jul 2018	Prob (F-statistic):	1.68e-22
Time:	09:10:08	Log-Likelihood:	-288.60
No. Observations:	116	AIC:	597.2
Df Residuals:	106	BIC:	624.7
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-598.9521	234.742	-2.552	0.012	-1064.350	-133.554
A2	0.3344	0.117	2.859	0.005	0.103	0.566
A3	-41.7807	8.148	-5.128	0.000	-57.935	-25.626
C1	-38.7320	6.554	-5.910	0.000	-51.725	-25.739
D1	-4.2909	2.029	-2.115	0.037	-8.313	-0.269
B1	-19.3411	6.295	-3.072	0.003	-31.822	-6.860
C2	-12.2188	3.322	-3.678	0.000	-18.806	-5.632
E3	-84.4775	12.571	-6.720	0.000	-109.400	-59.555
G2	22.7605	6.967	3.267	0.001	8.947	36.574
Q2	-0.0788	0.016	-4.883	0.000	-0.111	-0.047

Omnibus:	5.096	Durbin-Watson:	1.522
Prob(Omnibus):	0.078	Jarque-Bera (JB):	4.602
Skew:	0.473	Prob(JB):	0.100
Kurtosis:	3.236	Cond. No.	1.67e+06

Tab. 4. Regression Results for serious crime rate (per 1,000 residents) vs. different demographic metrics

Cross validation average score	0.5077
R-squared (R2)	0.6
Residual standard error (RSE)	8.48
Average response	12.6
RSE / Average response	0.67

Tab. 5. Regression parameters for serious crime rate (per 1,000 residents) vs. different demographic metrics

Fig. 21 and 22 show that residuals are equally and normally distributed, with slight heteroscedasticity observed.

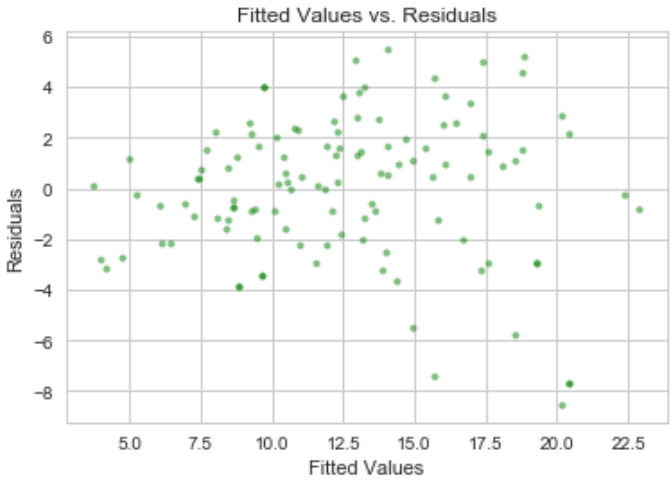

Figure 21. Residuals plot for the linear regression model

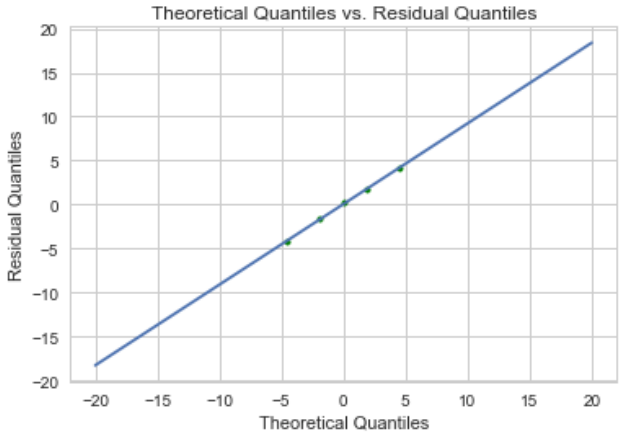

Figure 22. Correlation between theoretical quantiles of the normal distribution and residual quantiles of the linear regression model

Regression coefficients (Tab. 4):

- Almost all predictors are negatively correlated with the number of serious crimes rates but year and single-person households.
- The population aged 65+ has the strongest negative effect on the crime rates
- Interestingly, high population of native New-Yorkers also was negatively correlated with the number serious crimes. Similarly, crime rates decrease with increase of population of foreign-born residents.

Regression parameters (Tab. 5):

- The regression results showed that the model explained around 70% (Pearson R2) of the variance of the train data.
- The cross validation test showed on average one can expect that the model will explain around 50% (Pearson R2) of the variance of the independent data.

- The RSE / average response ratio indicates percentage error of roughly 67%, which quite significant

Regression parameters in general indicate that model it is not the most efficient in predicting values, but it can help to understand qualitatively and quantitatively demographic factors that contribute the number of severe crime rates in the NYC neighborhoods.

Data Modeling (Unsupervised Learning)

The goal of unsupervised learning is to check how strong are the similarities between the 77 NYC precincts and if there are in any homogenous groupings among precincts that could be investigated further. The analysis started with calculation of frequencies of standard offence codes per precinct from NYPD dataset. So created normalized offence code vectors were used for further investigation.

Similarities between boroughs

In the first step, the cosine similarities between borough crime records were investigated. The borough vectors were created through summation of precinct vectors belonging to each borough and normalized. Cosine products of the borough vectors were calculated for each pair of the boroughs (Tab. 6). The result closer to 1 means greater similarity between borough crime patterns.

	BRONX	BROOKLYN	MANHATTAN	QUEENS	STATEN ISLAND
BRONX	1.000	**0.977**	0.894	0.936	0.935
BROOKLYN	0.977	1.000	0.944	**0.986**	0.954
MANHATTAN	0.894	**0.944**	1.000	0.940	0.872
QUEENS	0.936	**0.986**	0.940	1.000	0.958
STATEN ISLAND	0.935	0.954	0.872	**0.958**	1.000

Tab. 6. Cosine similarities between borough crime rate patterns (max scores marked in bold)

The closest similarity can be observed between Brooklyn and Queens in both directions (98.6%). Bronx and Manhattan are also the most similar to Brooklyn, however Manhattan differs the most from other boroughs (the lowest cosine scores among all boroughs). Staten Island resembles the most almost equally Brooklyn and Queens. On the other hand, Manhattan and Staten Island differ the most from all boroughs. All the results were plotted visually on a heat map (Fig. 23). Generally, all scores are pretty high and prove that differences should be investigated on more granular level.

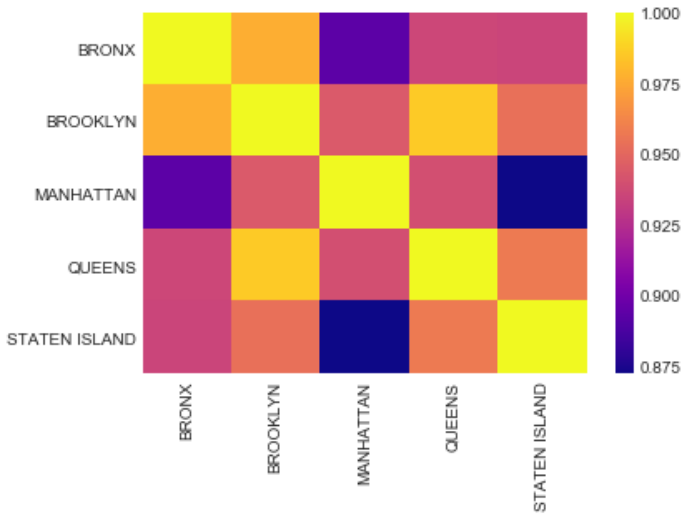

Figure 23. Heat map of cosine similarities between boroughs' crime rate patterns

Precinct Segmentation

In the next step, more intricate relationships between precinct crime rate patterns were investigated. Visual inspection of offence code frequencies was performed in low dimensional space. T-Distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA) methods were used for two dimensional space transformations of the data.

T-Distributed Stochastic Neighbor Embedding

T-SNE transformation preserves well nearness of the samples. Fig. 24 shows the result of applying t-SNE on precinct dataset. One can see that generally most Manhattan precincts are grouped together. Bronx and some Manhattan precincts also seem to be grouped far from other points. Queens and Brooklyn generally are mixed together. Staten Island precincts are close to each other and enclosed by other points.

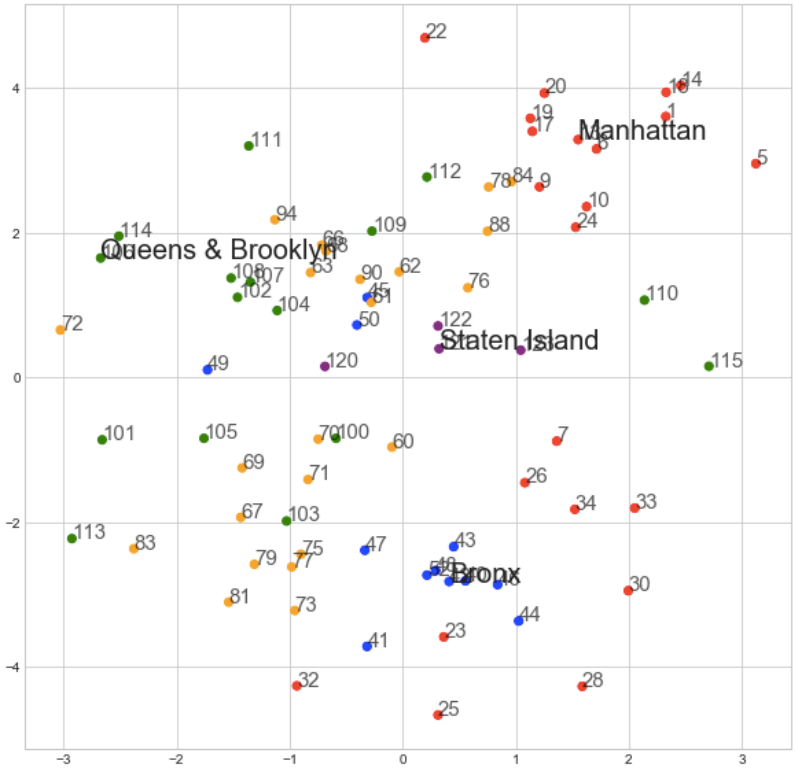

Figure 23. T-SNE method applied to precinct crime frequency dataset (red – Manhattan precincts, blue – Bronx precincts, orange – Brooklyn precincts, green – Queens precincts, purple – Staten Island precincts)

Principal Component Analysis

Second method to find low dimensional representation of the crime data is PCA. The number of intrinsic dimensions (i.e. number of features needed to approximate the dataset that have significant variance) was investigated first. From Fig. 24 one can see that the two first dimensions explain most of the variance. Fig. 24 (right) shows that they explain almost as much variance (32 % variance explained) as the next eight dimensions together (65% variance is explained with all 10 components), therefore the dataset can be reduced to two principal components and still a lot of information should be preserved.

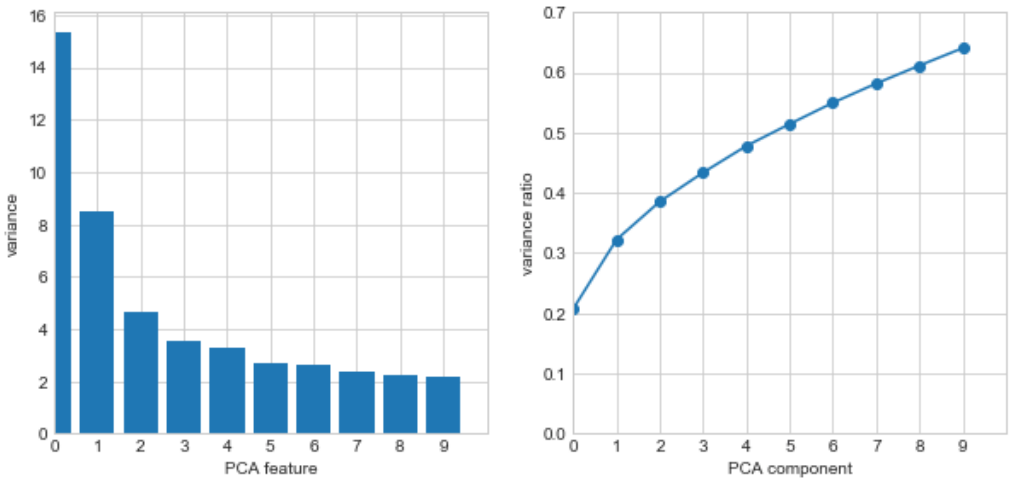

Figure 24. Amount of variance explained by each component (right), cumulative percentage of variance explained by each component.

In Fig. 25 one can PCA transformation of the crime frequency vectors into two-dimensional space. Similarly to t-SNE, one can observe a cluster of Manhattan precincts as well as a grouping of some Manhattan precincts and Bronx precincts together. The rest of points i.e. Queens, Brooklyn and Staten Island form the third cluster with some clear overlap.

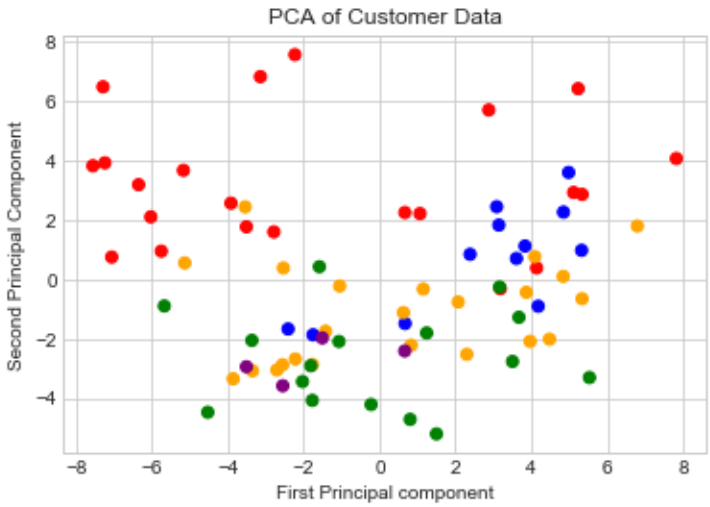

Figure 25. PCA method applied to precinct crime frequency dataset (red – Manhattan precincts, blue – Bronx precincts, orange – Brooklyn precincts, green – Queens precincts, purple – Staten Island precincts)

Clustering Methods

Clustering methods were used to verify if there is any precinct clustering i.e. if precincts can be grouped in homogenous subgroups among the observations based on their magnitude of their offence frequencies. Two clustering methods were used:

1. K-Means

2. Hierarchical clustering

K-Means

The first clustering method used to find clusters was K-Means, which requires specifying the number of clusters. Optimal number of clusters investigated through three methods:

- Inertia plot – elbow method
- Silhouette score
- PCA visual inspection

Fig. 26 shows inertia plot for different number clusters. The inertia plot seems dropping steadily, however the fastest drop is around K = 3 (i.e. elbow of the plot).

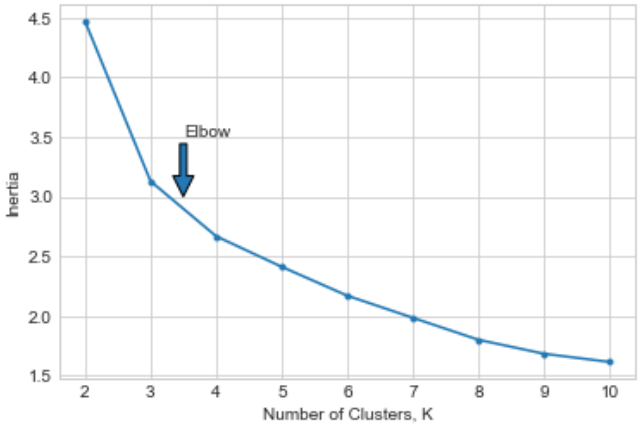

Figure 26. Inertia plot for K-Means

Silhouette method uses mean intra-cluster distance and the mean nearest-cluster distance to calculate the clustering score. Fig. 27 shows the highest scores are at K between 2 and 3.

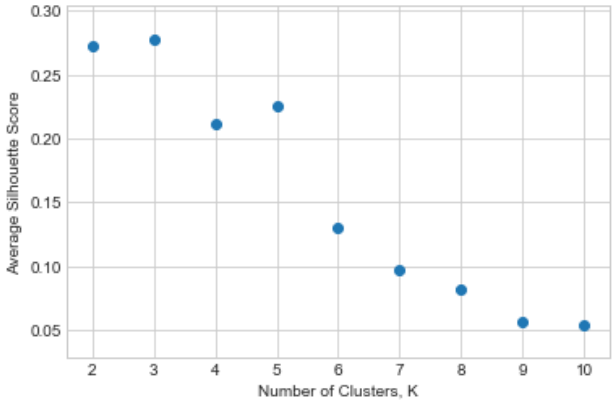

Figure 27. Silhouette plot for K-Means method

Finally, the dataset was decomposed into two principal components through PCA and the plots were inspected visually for optimal number of clusters. One can see

in Fig. 28 that there are around two, three clusters that can be visually separated. Four and five cluster results present more overlapping.

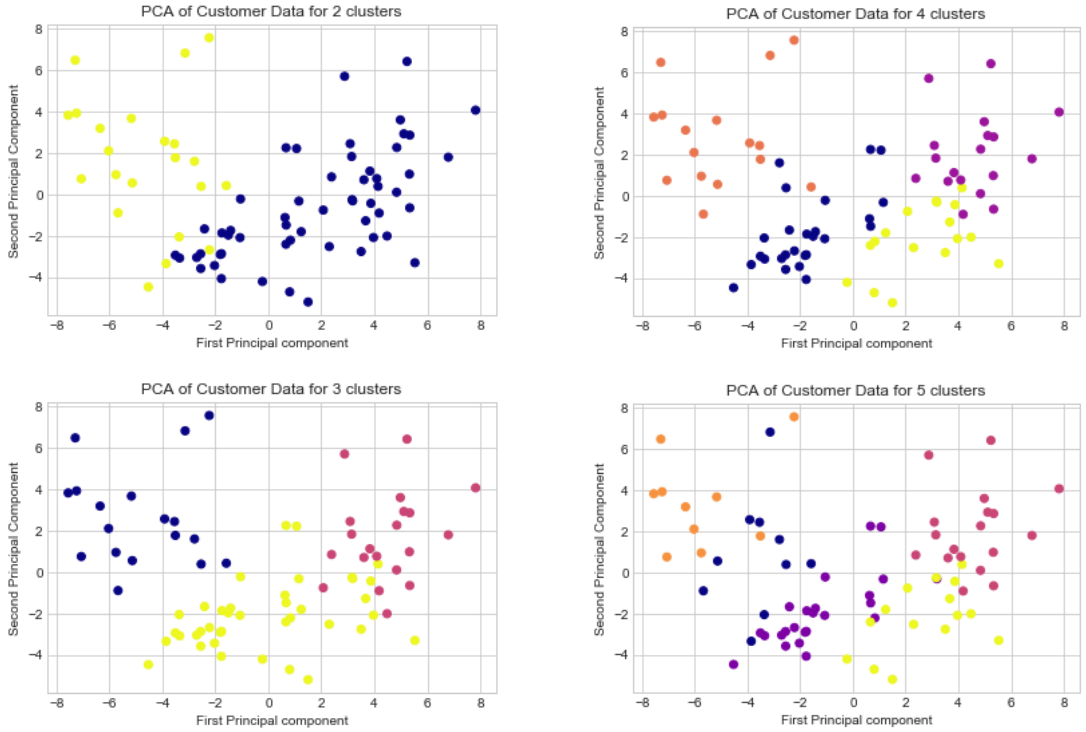

Figure 28. Principal Component Analysis for NYC precincts. Colors mean clusters calculated through K-Means method.

After comparison of the results of the three methods used, it was decided to continue the analysis for K = 3. Tab. 7 shows the cross table of K-Means labels and the NYC boroughs.

Label	BRONX	BROOKLYN	MANHATTAN	QUEENS	STATEN ISLAND
0	0	0	8	1	0
1	5	6	5	8	0
2	0	7	0	2	1

Tab. 7. Cross table of boroughs vs. labels of K-Means method

Fig. 29 shows New York City precinct centers plotted on the map, with numbers corresponding to the cluster to which they were assigned. Fig. 28 (lower) shows the same result, but the precincts were colored according to their cluster label.

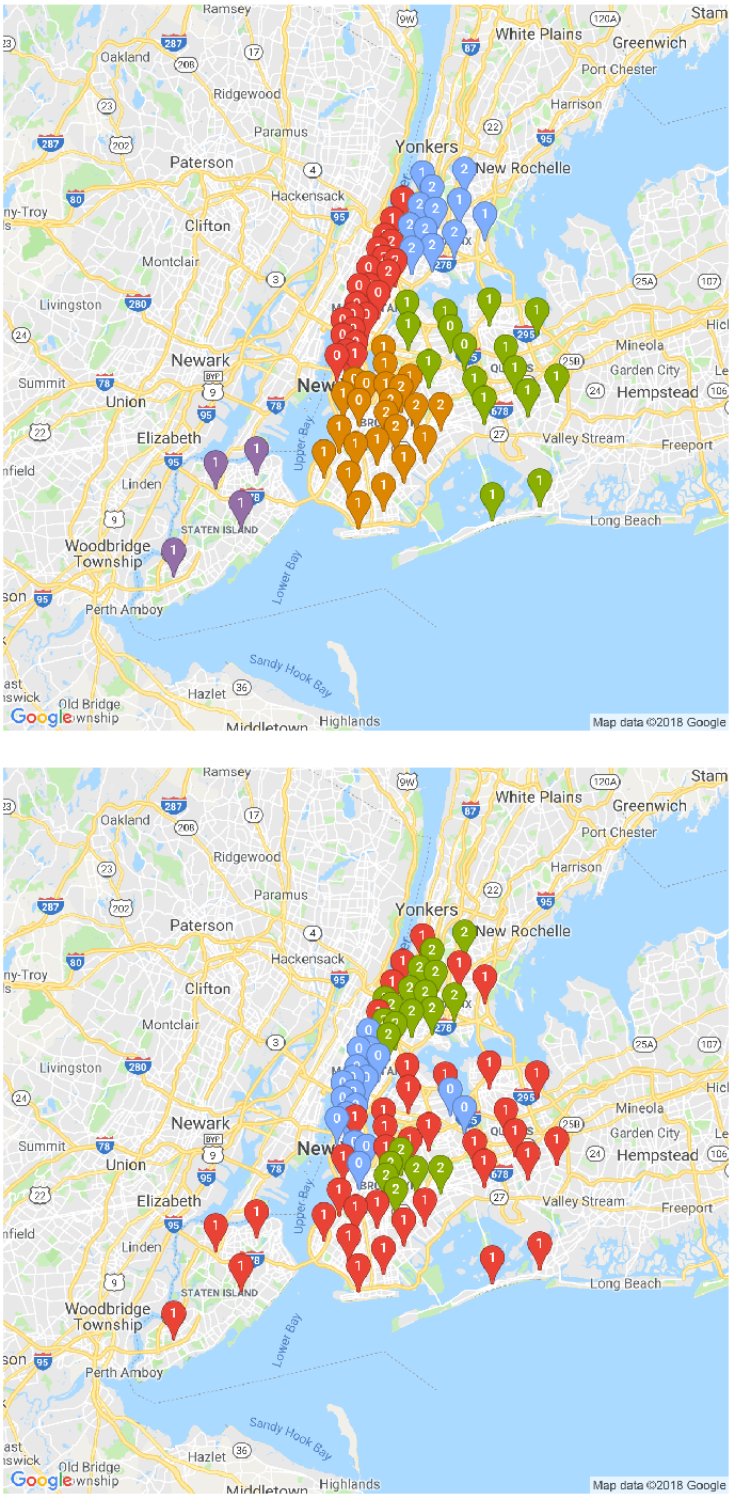

Figure 29. Result of K-Means applied for K = 3. Upper plot shows precincts colored according to the boroughs they belong to (red – Manhattan, blue – Bronx, green – Queens, orange – Brooklyn, purple – Staten Island). Lower plot shows precincts colored according to the label they were assigned to (0 – blue, 1 – red, 2 – green)

Generally, one can see that labels:

- **'0'** grouped mostly Manhattan precincts and some Manhattan neighboring Brooklyn precincts with two additional precincts from Queens.
- **'1'** grouped mostly Brooklyn, Queens and Staten Island with a small number of Bronx precincts.
- **'2'** grouped mostly Bronx and Bronx neighboring Manhattan precincts, some Brooklyn precincts and a few Queens's precincts.

When one looks into the most frequent crimes for those labels (Tab. 8) that generally there is are similar offence codes for all the labels (e.g. petit larceny is on the first position for all three labels), with some important differences i.e. Label '1' has no felony in the 5 most popular crimes versus label '2' that has felony on the second position. Felony is ranked fifth for precincts labeled '0'. That is consistent with borough similarity results, and subtlety of differences.

Offence Code	Description	Offence Level	Label
341	PETIT LARCENY	MISDEMEANOR	0
109	GRAND LARCENY	FELONY	0
578	HARRASSMENT 2	VIOLATION	0
344	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	0
351	CRIMINAL MISCHIEF & RELATED OF	MISDEMEANOR	0
341	PETIT LARCENY	MISDEMEANOR	1
578	HARRASSMENT 2	VIOLATION	1
344	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	1
351	CRIMINAL MISCHIEF & RELATED OF	MISDEMEANOR	1
235	DANGEROUS DRUGS	MISDEMEANOR	1
341	PETIT LARCENY	MISDEMEANOR	2
578	HARRASSMENT 2	VIOLATION	2
351	CRIMINAL MISCHIEF & RELATED OF	MISDEMEANOR	2
344	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	2
109	GRAND LARCENY	FELONY	2

Tab 8. Most frequent crimes for each label (colored by label according to Fig. 28 lower)

Hierarchical clustering

The second method used to examine clustering in the dataset was hierarchical clustering, which doesn't require any particular choice of number of clusters a priori. In Fig. 29 a dendrogram is shown for all the precincts using correlation-based distance as dissimilarity measure in place of Euclidean distance as the goal of the model is to investigate similar profiles and not only physical proximity of the points.

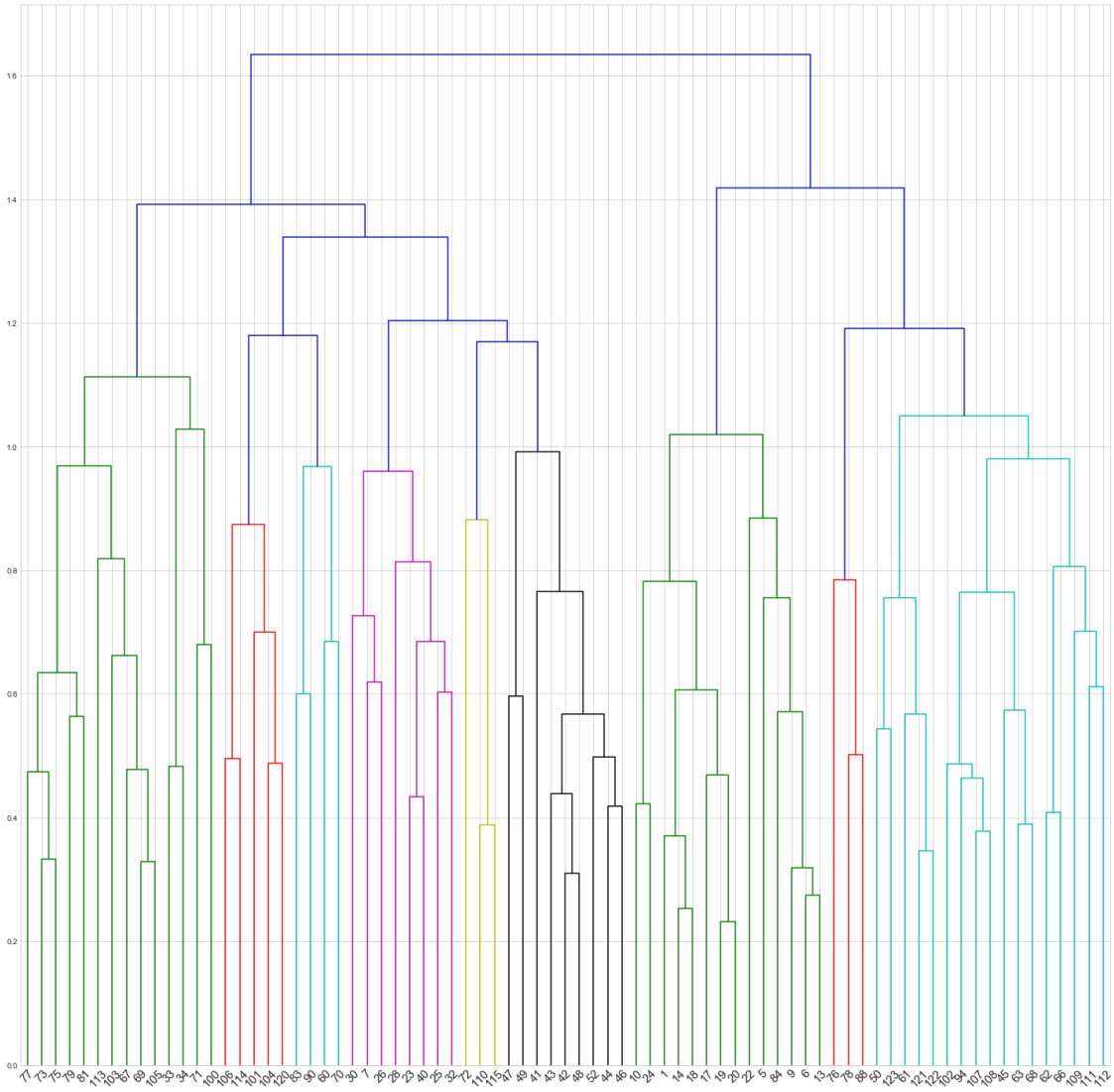

Figure 30. Dendrogram result of hierarchical clustering of precinct crime rates dataset

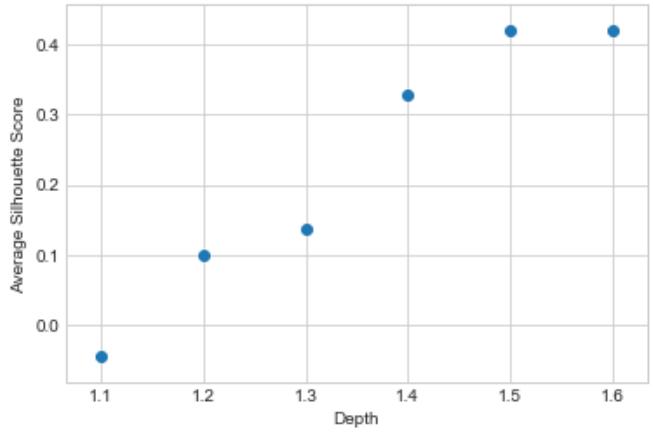

Figure 31. Silhouette plot for the hierarchical clustering method for the precinct offence frequency dataset

The silhouette score plotted for different depths shows that the highest scores are achieved for greater depths. To compare two clustering scores depth of 1.4 was selected that corresponds to 3 clusters on the dendrogram. Cross table between labels and boroughs (Tab. 9).

Labels	BRONX	BROOKLYN	MANHATTAN	QUEENS	STATEN ISLAND
1	10	13	9	10	1
2	0	1	13	0	0
3	2	9	0	6	3

Tab. 9. Cross table of boroughs vs. labels of Hierarchical clustering method

Generally label:

- **'1'** grouped mostly Bronx, Bronx –neighboring Manhattan precincts, most Brooklyn and Queens precincts and one Staten Island precinct.
- **'2'** grouped mostly Manhattan precincts and one Manhattan neighboring Brooklyn precinct
- **'3'** grouped mostly Brooklyn, Queens and Staten Island with a small number of Bronx

Generally, labels 1, 2 and 3 of Hierarchical clustering have similar the most frequent crimes per label structure (Tab. 10) correspond to labels 2, 0, 1 of K-Means method with some small differences.

Offence Code	Description	Offence Level	Label
341	PETIT LARCENY	MISDEMEANOR	1
578	HARRASSMENT 2	VIOLATION	1
344	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	1
351	CRIMINAL MISCHIEF & RELATED OF	MISDEMEANOR	1
235	DANGEROUS DRUGS	MISDEMEANOR	1
341	PETIT LARCENY	MISDEMEANOR	2
109	GRAND LARCENY	FELONY	2
578	HARRASSMENT 2	VIOLATION	2
344	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	2
351	CRIMINAL MISCHIEF & RELATED OF	MISDEMEANOR	2
341	PETIT LARCENY	MISDEMEANOR	3
578	HARRASSMENT 2	VIOLATION	3
351	CRIMINAL MISCHIEF & RELATED OF	MISDEMEANOR	3
109	GRAND LARCENY	FELONY	3
344	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	3

Tab 10. Most frequent crimes for each label (colored by label according to Fig. 32 lower)

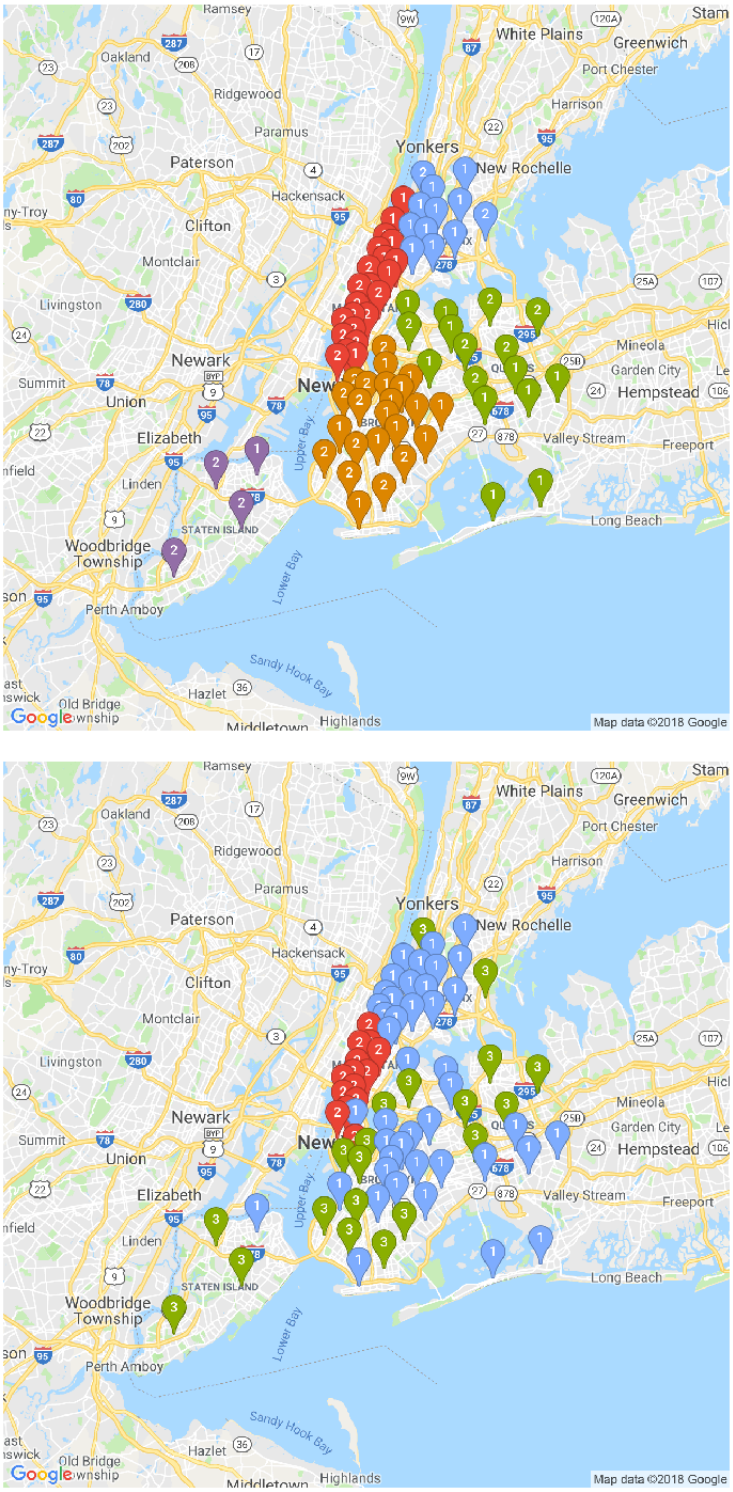

Figure 32. Result of hierarchical clustering for depth corresponding to three clusters. Upper plot shows precincts colored according to the boroughs they belong to (red – Manhattan, blue – Bronx, green – Queens, orange – Brooklyn, purple – Staten Island). Lower plot shows precincts colored according to the label they were assigned to (0 – blue, 1 – red, 2 – green)

Similarities between precincts (Tool)

To better understand why certain precincts are clustered together Non-Negative Matrix Factorization (NMF) method was used, which is, similarly to PCA, another dimension reduction technique, but its components are interpretable. NMF was used to create a model that can calculate cosine similarities between precincts (Tab. 11).

Precinct	Map order	Cosine product
30	1 (selected precinct)	1.0000
43	2	0.9923
46	3	0.9881
42	4	0.9812
23	5	0.9781

Tab 11. Example of cosine products (cosine similarity) of precincts for precinct 30

By selecting a precinct, one can see which other precincts are most similar to it and in what order as well as visualize their centers on the NYC map (Fig. 33). By analyzing the component that contributes the most to the cosine product, one can select the largest elements it consists of. Those elements will be offences that influence the most similarities between selected precincts (Tab. 12).

Offence Code	Description	Offence Level
235	DANGEROUS DRUGS	MISDEMEANOR
344	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR
105	ROBBERY	FELONY
351	CRIMINAL MISCHIEF & RELATED OF	MISDEMEANOR
106	FELONY ASSAULT	FELONY

Tab 12. The most frequent crime rates that contribute the most to cosine similarities to precinct 33

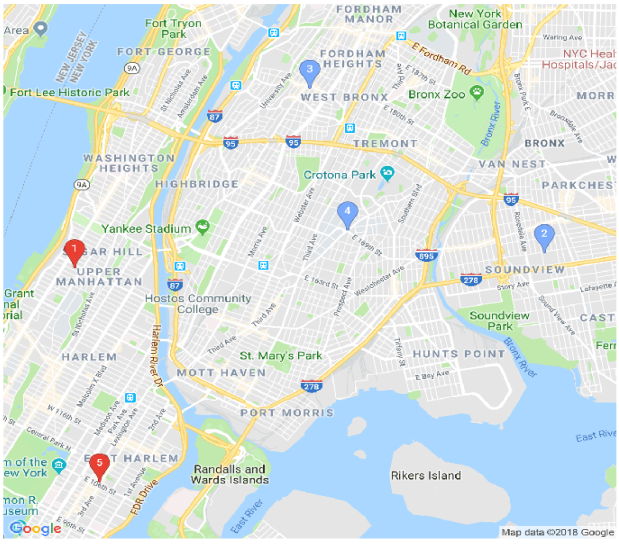

Figure 33. (red – Manhattan precincts, purple – Bronx precincts)

Assumptions and Limitations

- The indicators used to compare crimes assume homogenous distribution of citizens within boroughs. This generalization doesn't take into account areas of lower population density e.g. cemeteries, stadiums, airports, etc. Even though some of those areas aren't crime free, still their contribution might be significantly lower and therefore should be weighted out appropriately.

- Two datasets used in this study operate on two different granularity levels, which are below borough level i.e. precinct and neighborhood. Even though their respective sizes and number are comparable, still it makes it difficult to fully merge datasets on below borough level. In order to keep more granular level (and therefore not to loose information through aggregation), the learning models were developed on separate datasets. Nevertheless, the source of information about crime rates is always NYPD, therefore it assumed that crime rates in both datasets are equally reliable.

- Little is known about the collection of the information about complaints. It is assumed that generally all the precincts have similar police resources and work ethics. All complaints were digitally recorded for the period studied. It was noticed during the data-cleaning step that during some earlier years, there was significantly less crimes in the NYPD dataset. There is no reason to believe that a sudden jump in number of crimes is due to other factors than technical, administrative issues that were resolved later. Still one can't be certain if the data after 2006 is complete. Other than missing data, the number of police resources could influence the number of recorded complaints e.g. Manhattan is the richest and the worldwide famous borough, therefore there might be more resources deployed there and more officers to report/react to incidents.

Recommendations and Future Work

- More information is needed on NYPD dataset. It was unknown during this analysis what were the police resources used in respective precincts, how recordings are made, if there could be any known bias. This information could be important for the validity of this analysis

- Since neighborhoods of respective boroughs often have very distinctive character (e.g. Manhattan's Lower East Side is more hipster, relaxed

neighborhood with some China Town influences, Upper East Side on the other hand is very posh, and expensive), it would be recommended to work on the neighborhood level to get better sense of crime analysis and its causes and influences on housing market.

- Generally not all complaints necessarily are indicators of level of safety. Selective crime type study e.g. (murders) could be studied separately to get a better understanding of safety. The tools developed in that study can be easily adapted to that purpose.

Conclusions

- In this study, crime analysis of NYC was performed using primarily NYPD dataset. Generally, the number of recorded offences decreases every year. There are significant differences between the boroughs in terms of size and population i.e. they share different fractions of the population and area; therefore different crime indicators were proposed to better compare the boroughs (i.e. crime density and crime rate per 1000 citizens).

- Housing market numbers where influenced by the Great Recession period and therefore no strong correlations between crime rates and housing metrics were observed on the borough level. Neighborhood level study could give better results.

- A linear regression model was developed to understand relation between serious crime rates and some demographic factors. Generally, crimes rates mostly decrease with the increase of the fraction of the old population, native New Yorkers or foreign-born citizens.

- There are a lot of similarities between boroughs in terms of crime rate patterns. The analysis of precincts showed that there is some clustering observed. Three clusters were analyzed using two different clustering methods that gave comparable results. Clusters bear subtle differences but often include borough – neighboring precincts. A tool was also developed to compare similar precincts.