Analysis of Crimes in the City of New York

Capstone Project

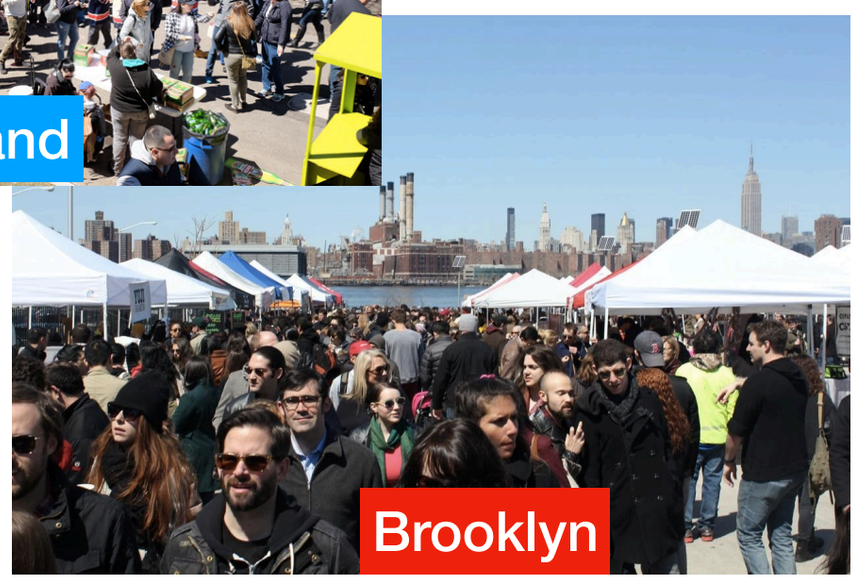
Michal Czapski
czapski.michal@gmail.com

Contents

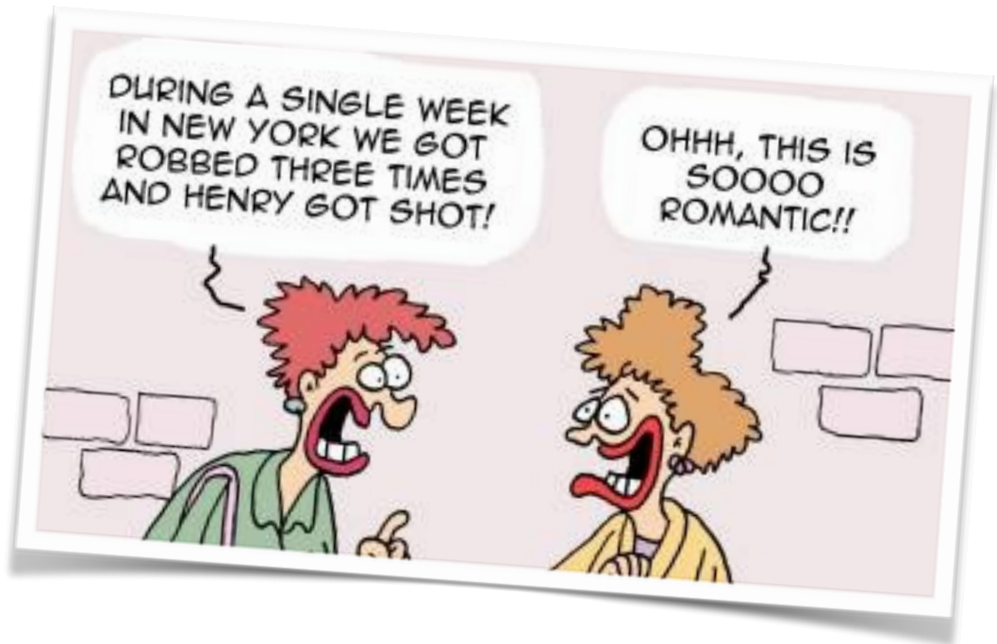
Introduction

New York City is the most populous city in the USA divided into 5 different boroughs

Every borough differs in demographics, wealth and lifestyle



The goal of this study is to investigate crime rates in NYC



The main focus is to investigate:

- Changes in crime rates between 2006 and 2016
- Differences between boroughs in crime rates
- Crime rates and demographics / housing market relationships

There are three potential clients that can be interested in this study

NYPD

Housing Market

NYC Authorities

Three main datasets were used in this analysis

NYPD Complaint Data
contains information about type of offense, time of occurrence, specific location and borough.
6,050,000 records / 24 features (CSV)

NYU Furman Center
provides a lot of information on neighborhood demographics between 2010 and 2015
*(1219 = 53 neighborhoods * 23 records) / 8 features (Excel)*

NYC Dept. of Finance
provides information on property prices (means, medians) and number of sales per year per borough per type of home.
7180 / 7 features (Excel)

All the data provided in separated files was merged into three data sets, that were cleaned and filtered out for entries between 2006 and 2016

NYPD Complaint Data

- Merged historic and current part
- Removed significant NaN columns
- Transformed all dates into one date column chosen from start date or end date or report date
- Removed erroneous data removed (e.g. year = 1015)
- Kept only year > 2005 entries (significantly fewer entries before 2006)
- Changed categorical variable for cardinal numbers (e.g. offence no.)

NYU Furman Center

- Merged all the Excel files for different neighborhoods
- Added borough column
- Pivoted demographic indicator column into a separate column for each indicator
- Melted different year columns into one year column
- Removed entries before 2006 (consistancy with NYPD data base)
- Removed %, $ signs from statistics and transformed values into numeric
- Unified NaN values (two types of NaN - NA (string) and NaN (object))

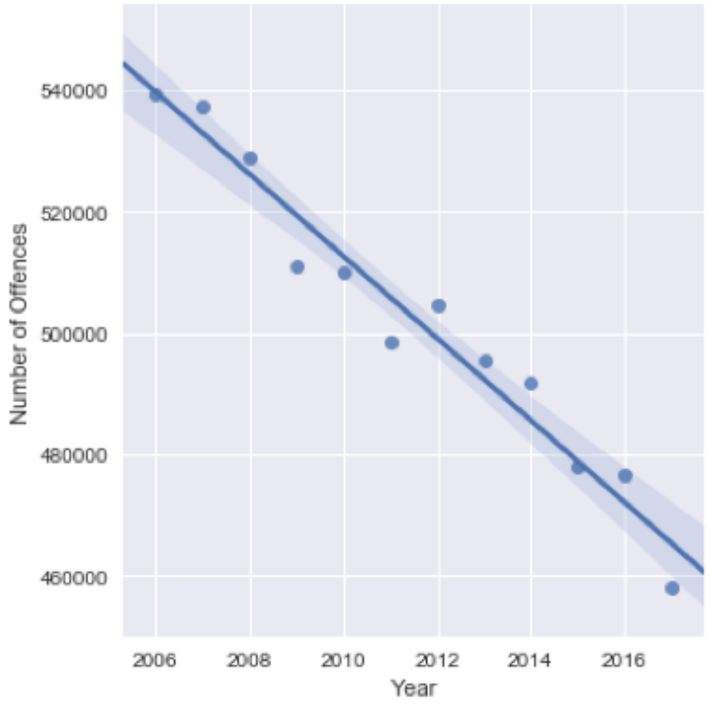
NYC Dept. of Finance

- Merged all Excel files for different boroughs into one
- Added borough column
- Unified houses types names (removed extra spaces from strings)
- Removed significant NaN columns

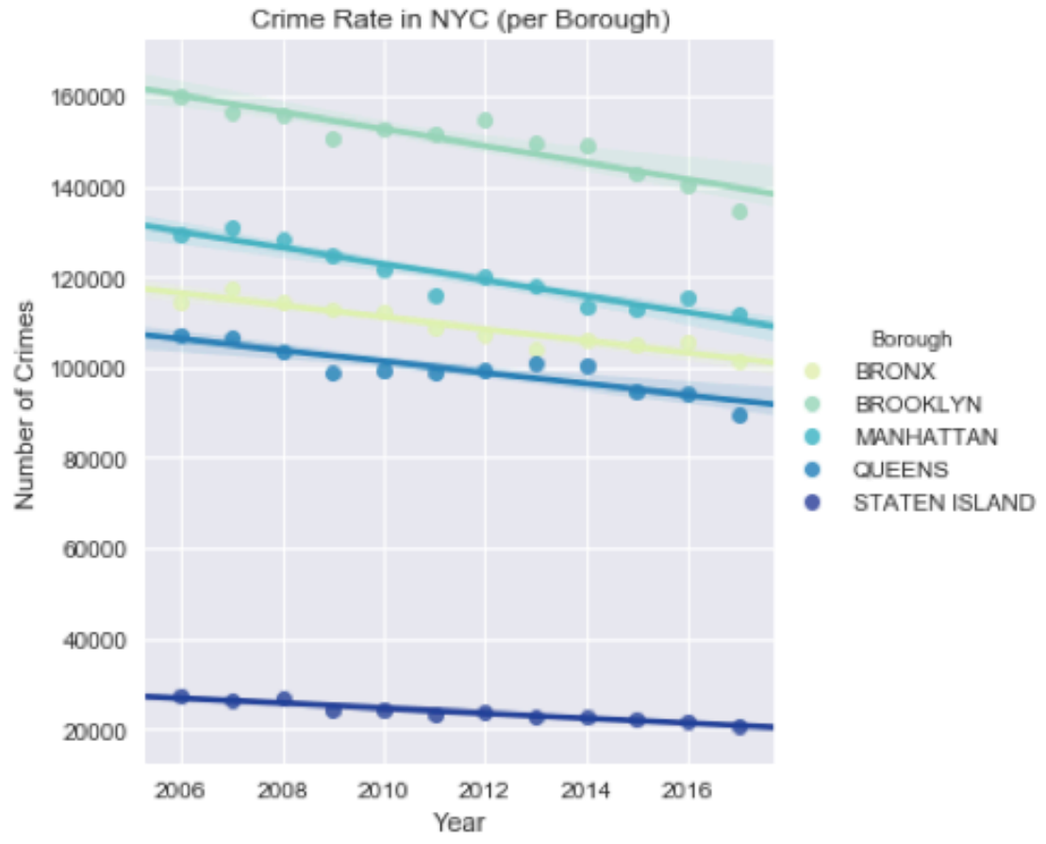
Data Exploration

NYPD data set

On average the number of reported crimes decline every year on the city and borough level

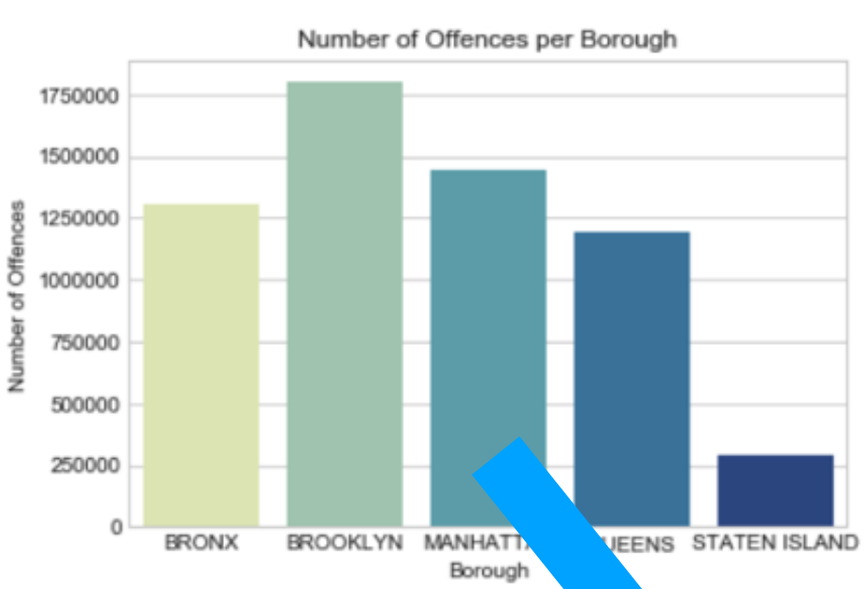


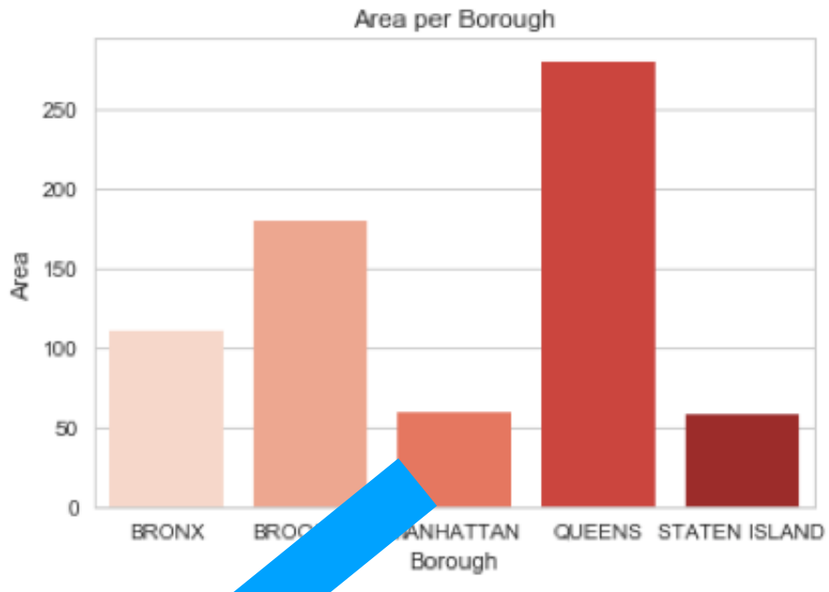
NYC
-7000 crimes / year

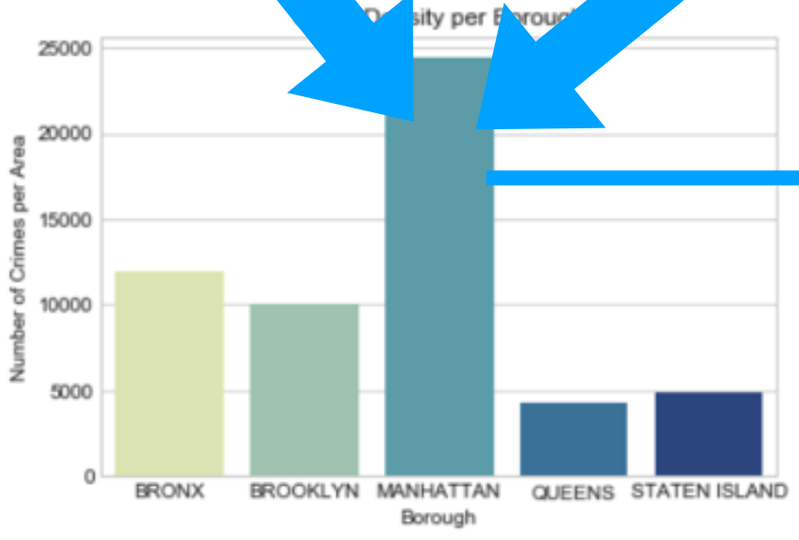


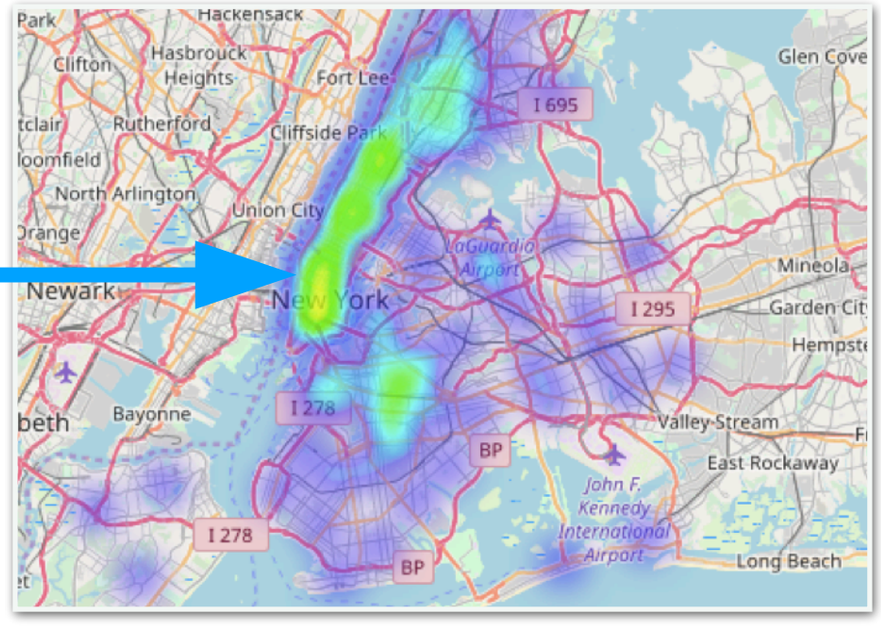
max (Brooklyn)
-2000 crimes per year
min (Staten Island)
-550 crimes per year

Boroughs differ significantly in areas, (reported) crime density can better compare borough-level crime rates



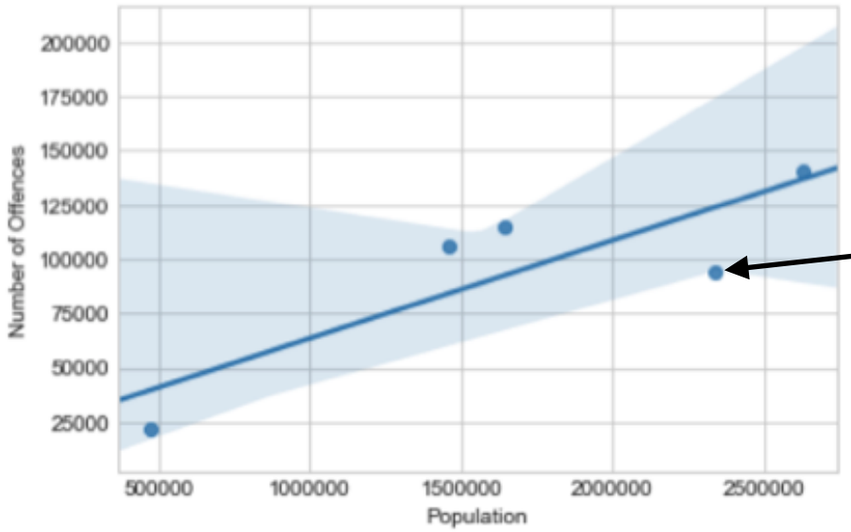






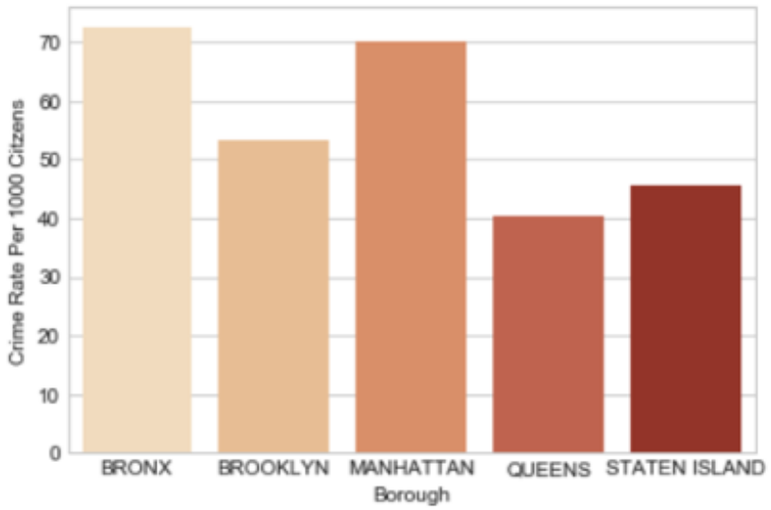
Manhattan outdistanced other boroughs

Population also can impact crime rates, the bigger population generally the likelihood of crime should be generally greater

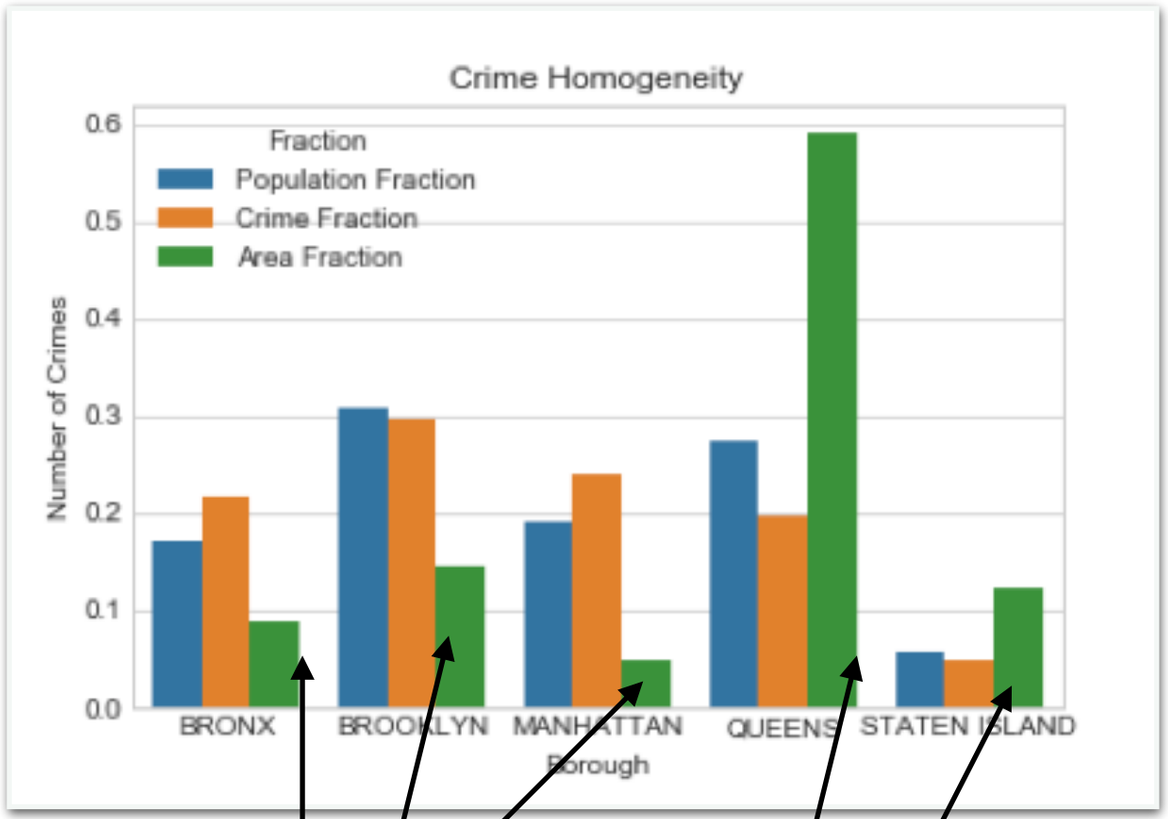


Points represent data for boroughs

Crime Rate per 1000 Citizens
number of crimes that on average is recorded among 1000 citizens



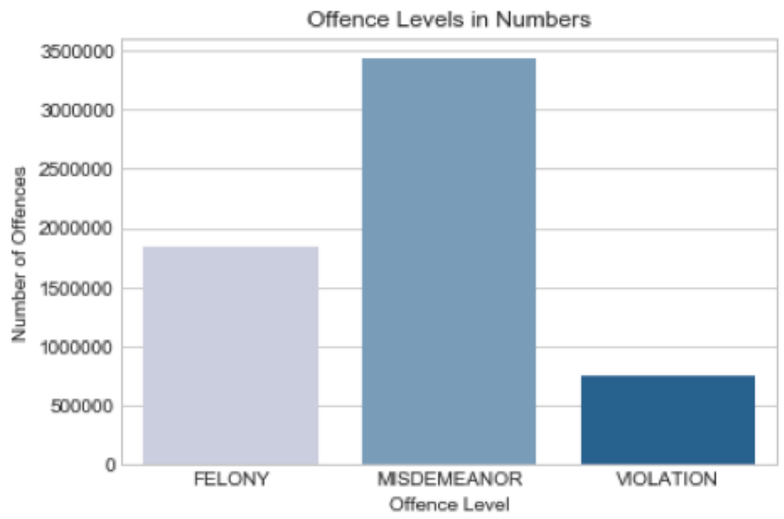
There are disproportions in boroughs' homogeneity in terms of population, area and crime rates share

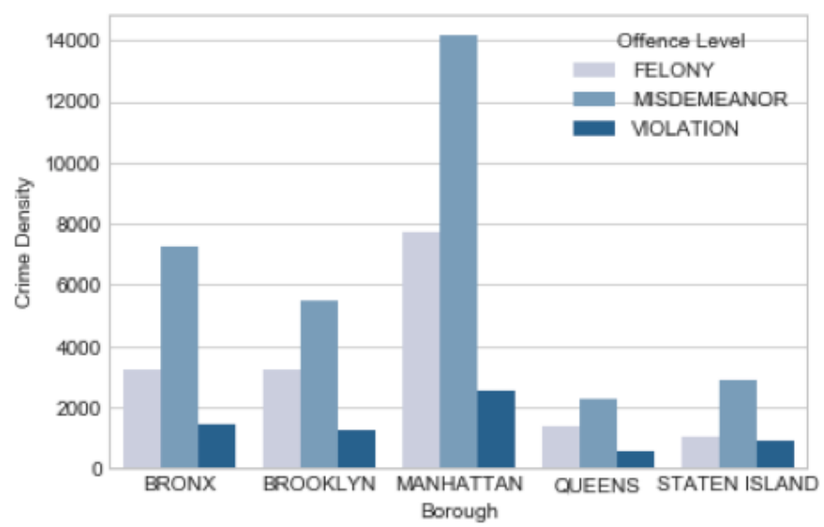


Area is the smallest fraction bigger fraction

Area is the largest fraction bigger fraction

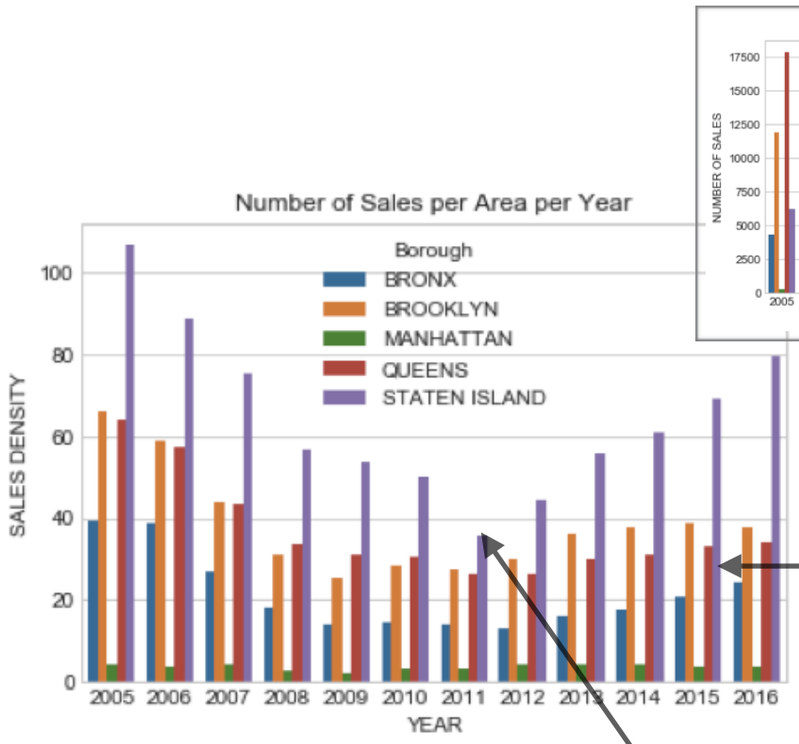
Generally distribution of different crime level is similar in all three boroughs i.e. misdemeanors are twice more frequent that felonies and five times more frequent than violations

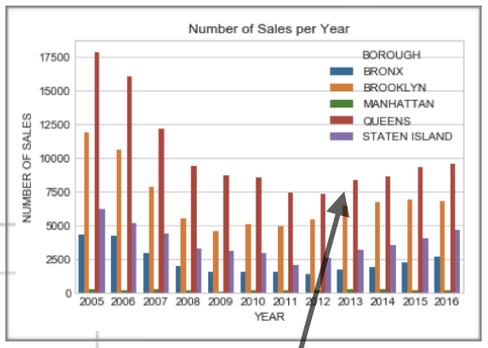




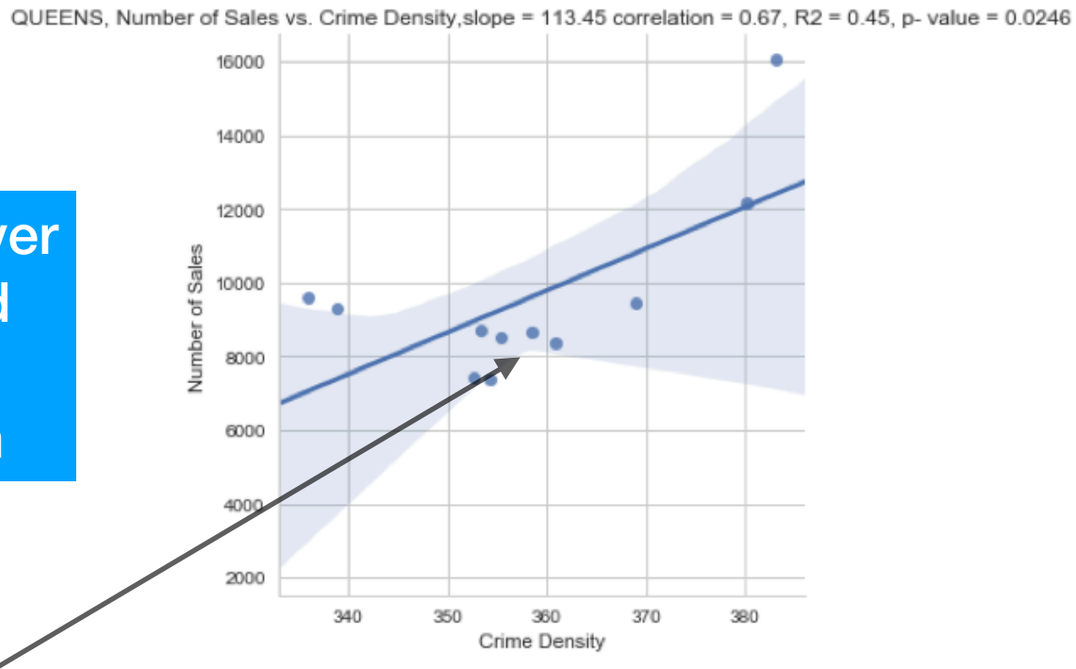
Housing Market vs. Crimes

Sales density is a better indicator to compare number of sales, the Great Recession had significant impact on sales



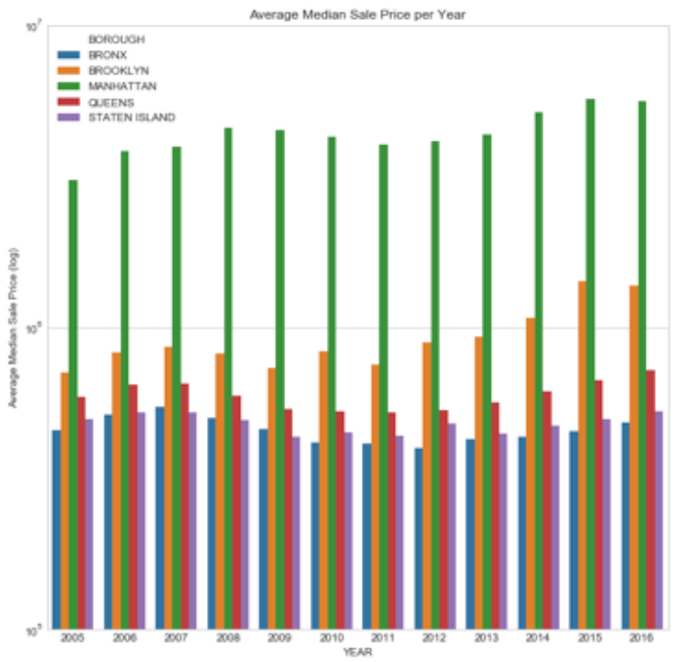


Queens never recovered after recession

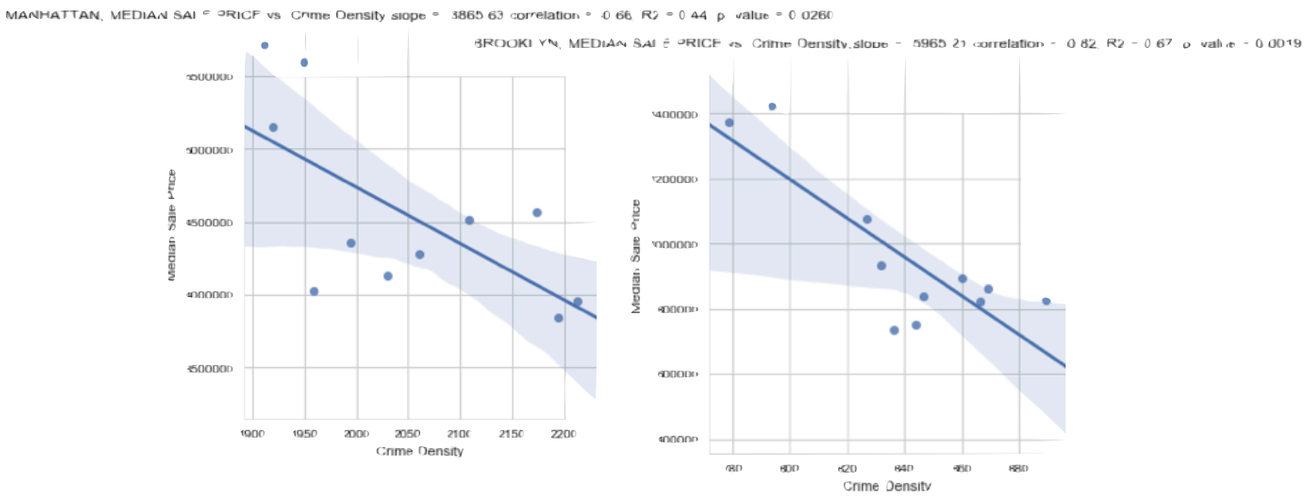


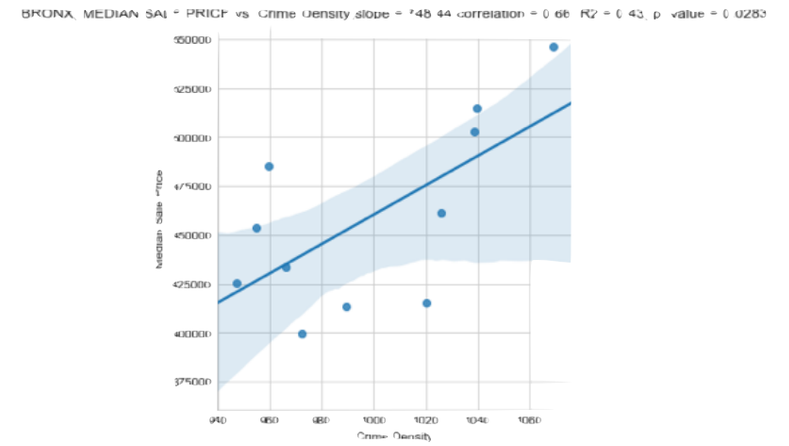
the Great Recession 2010 - 2012

Average median sale prices also were affected by the Recession after which Manhattan and Brooklyn prices went up and in the Bronx they never recovered



the Great Recession 2010 - 2012



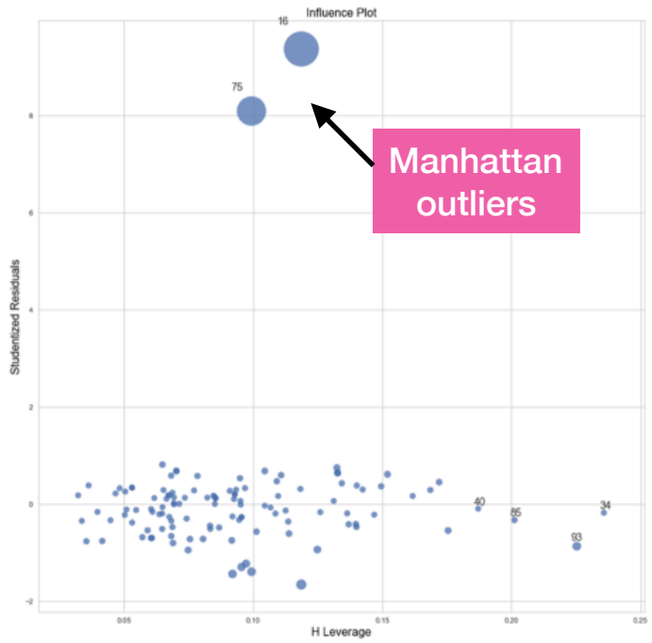


Data Modeling

Supervised Learning

Different regression models were optimized to select the best model that can describe demographic factors that contribute to severe crime rates

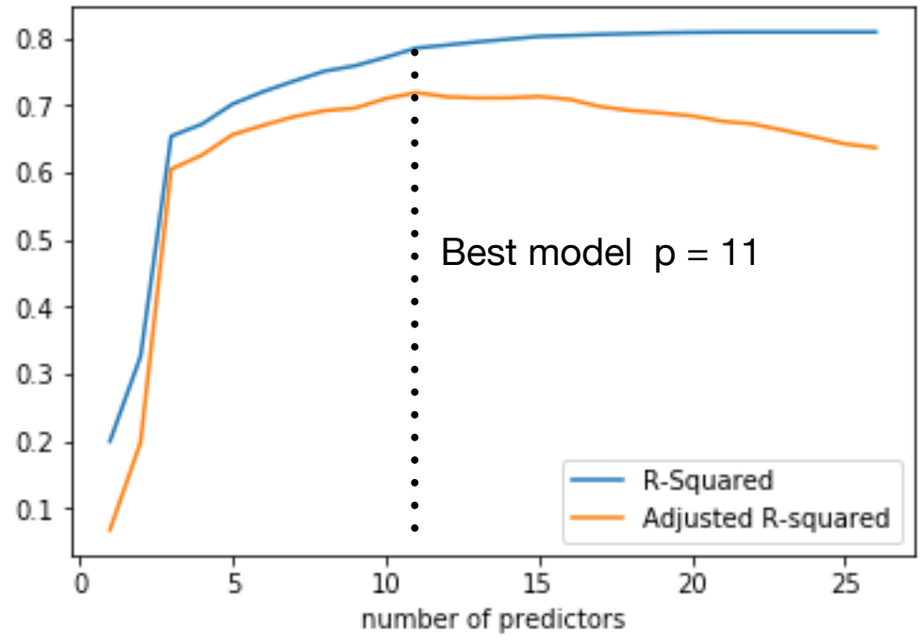
Response Variable:
Severe Crime Rates
(per 1000 citizens)

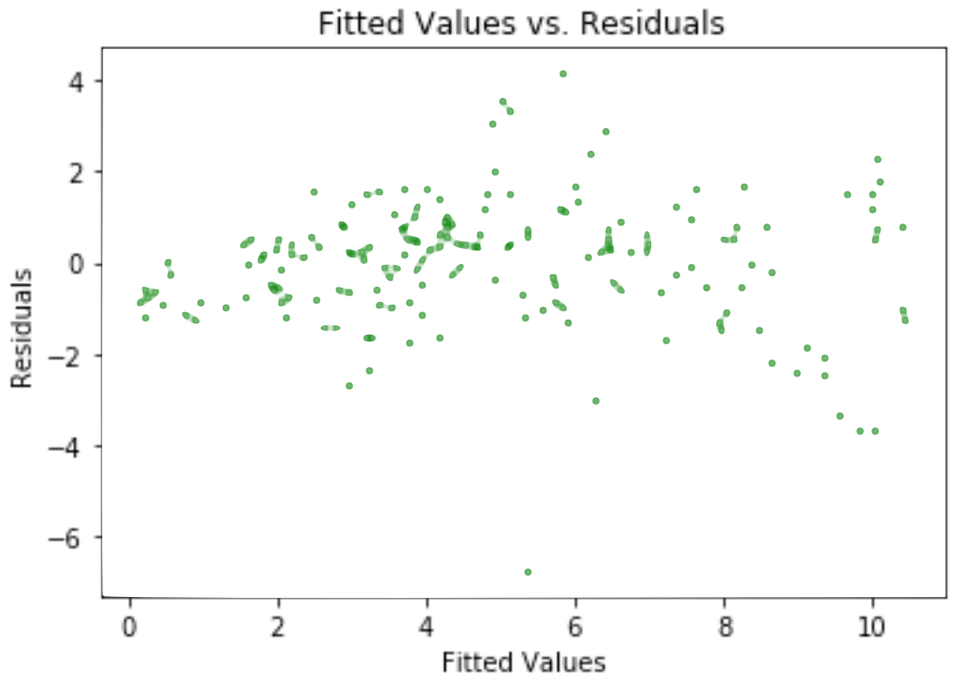


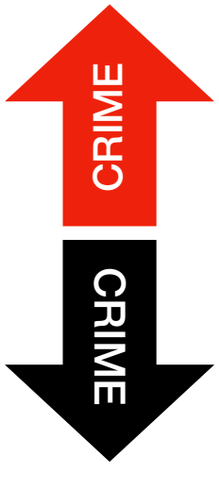
A1	Labor force participation rate
A2	Percent Asian
A3	Population aged 65+
B1	Foreign-born population
B2	Percent white
C1	Mean travel time to work (minutes)
C2	Single-person households
C3	Poverty rate, population aged 65+
D1	Income diversity ratio
D2	Poverty rate
D3	Percent black
E1	Population
E2	Born in New York State
E3	Racial diversity index
F1	Poverty rate, population under 18 years old
F2	Population aged 25+ without a high school diploma
G1	Percent Hispanic
G2	Car-free commute (% of commuters)
G3	Population aged 25+ with a bachelor's degree o...
H1	Unemployment rate
H2	Households with children under 18 years old
H3	Borough
I1	Year

Method	Optimization	Best Model	Final Score
Backward Selection Linear Regression with a stopping rule	Predictors are eliminated starting with the highest p-value. Model selection stops when all p-values < 0.5 Adjusted R2 for 10-fold CV	B1+C2+C3+E2+G3+H2+I1+H3_BROOKLYN+H3_QUEENS	0.699
Backward Stepwise Selection Linear Regression	R-squared of a full model is calculated. For number of predictors k = p, p -1,..,1 the model with highest R-squared is searched. Finally among the best models for each k, the one with the highest adjusted R-squared is select as the best model Adjusted R2 for 10-fold CV	A1+A3+B1+C2+E2+F2+G3+H2+I1+H3_BROOKLYN+H3_QUEENS	**0.718**
Lasso Regression	Grid Search CV for the best alpha (0.01,0.1,1) Adjusted R2 for 10-fold CV	alpha = 0.01 A1+A2+A3+B1+B2+C1+C2+C3+D1+D2+D3	0.682
Ridge Regression	Grid Search CV for the best alpha (0.01,0.1,1) Adjusted R2 for 10-fold CV	alpha = 1 A2+'B2+C2+C3+D1+H1+H2+I1+H3_BROOKLYN+H3_MANHATTAN+H3_QUEENS	0.650
Random Forest Regression	Grid Search CV for depth, n_estimators Out of Bag Error	all the predictors n_estimators = 100, d = 5	0.713

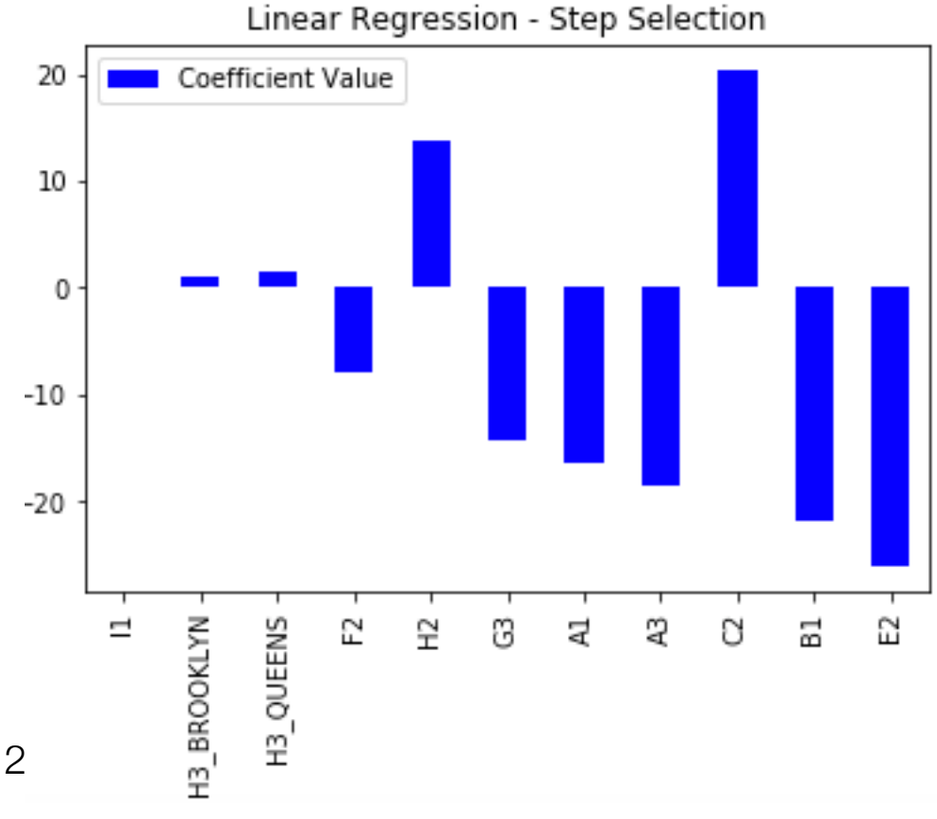
Based on backward stepwise selection model, 11 demographics factors were found that either negatively or positively correlate with severe crime rates







A1	Labor force participation rate
A3	Population aged 65+
B1	**Foreign-born population**
C2	**Single-person households**
E2	**Born in New York State**
F2	Population aged 25+ without a high school diploma
G3	Population aged 25+ with a bachelor's degree o...
H2	Households with children under 18 years old
H3	Borough (Queens, Brooklyn or other)
I1	Year



Unsupervised Learning

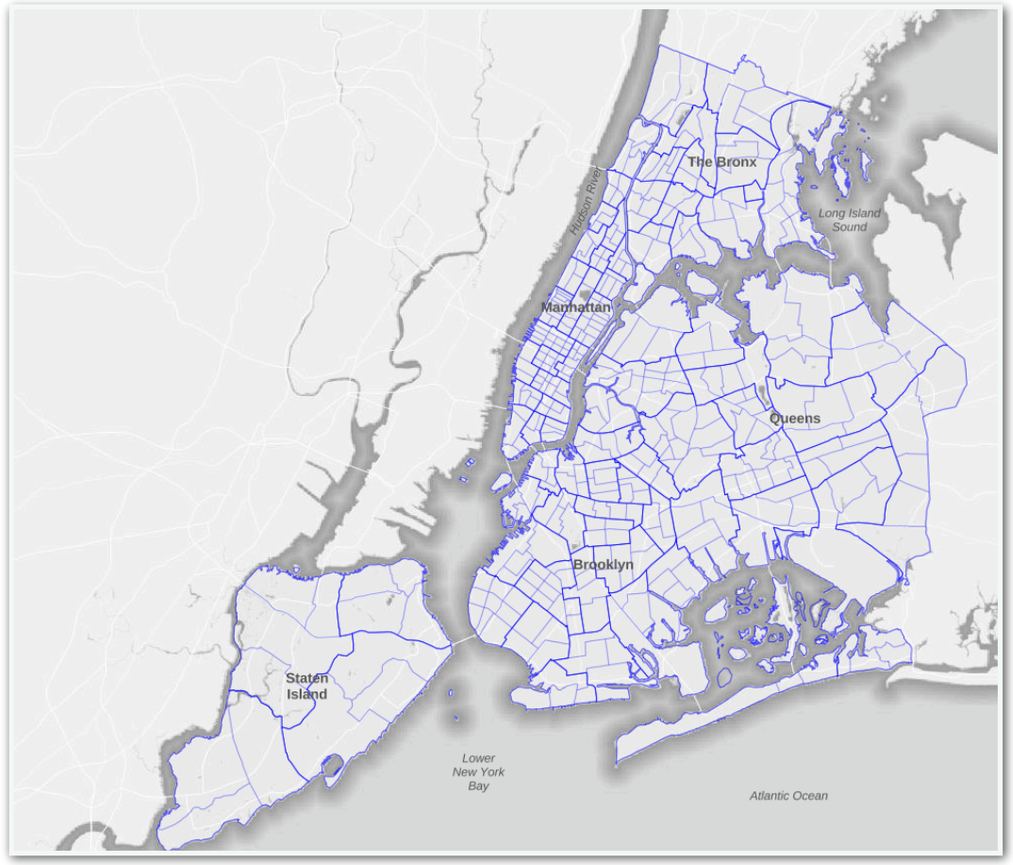
The goal of unsupervised learning is to check how strong are the similarities between the 77 NYC precincts and if there are in any homogenous groupings among them

Features: Offence Codes

Offence Code	101	102	103	104	105	106	107	109	110	111	..
0	0.000558	0.000105	0.000000	0.003695	0.037196	0.036429	0.070731	0.467960	0.018755	0.005822	..
1	0.000937	0.000117	0.000000	0.004861	0.079656	0.107477	0.082116	0.396112	0.015521	0.005798	..
2	0.000554	0.000000	0.000000	0.004985	0.077550	0.058674	0.092889	0.570801	0.021987	0.009928	..
3	0.001295	0.000000	0.000072	0.008780	0.127600	0.111407	0.073983	0.308672	0.036848	0.005614	..
4	0.001031	0.000000	0.000000	0.007256	0.085836	0.077384	0.095442	0.411328	0.027458	0.005648	..
5	0.000933	0.000124	0.000						8	0.011192	..

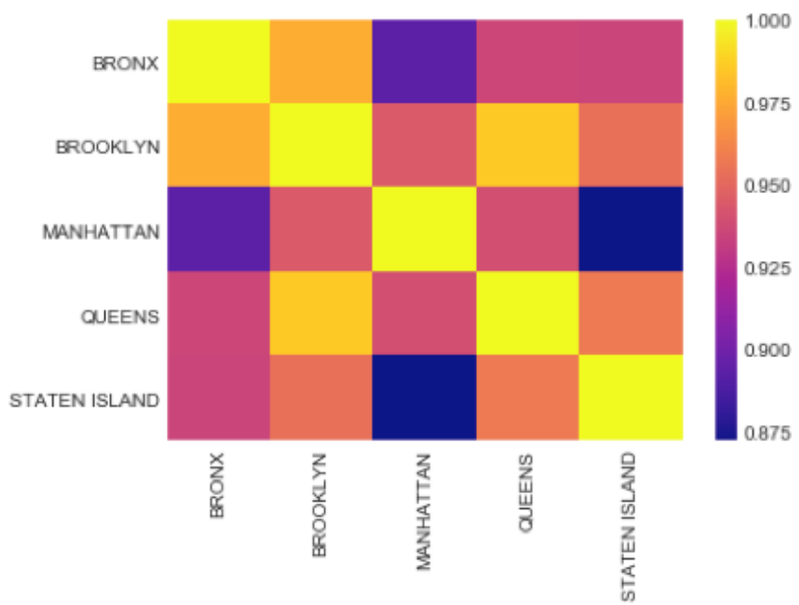
Offence frequency vectors

Observations: 77 Precincts



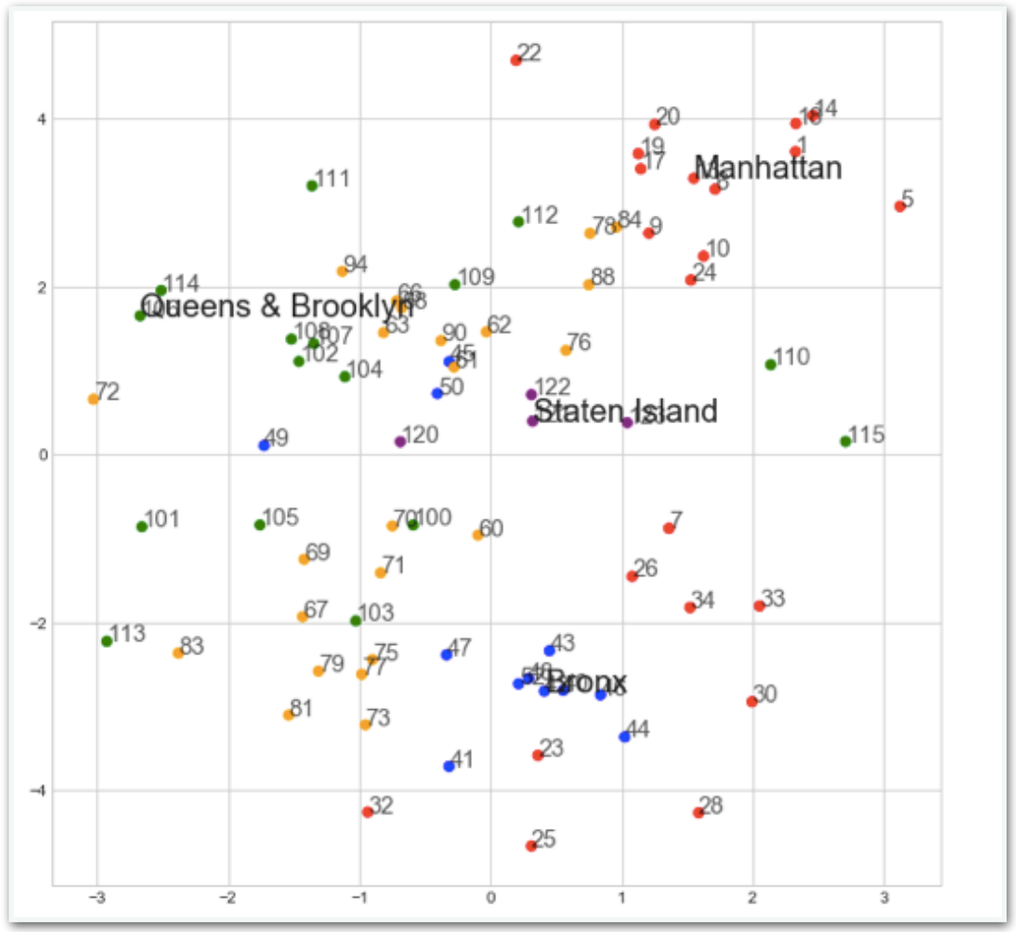
Boroughs generally are very similar to each other crime-pattern wise with Manhattan being the most different from all other boroughs and Brooklyn and Queens showing the closest resemblance

	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

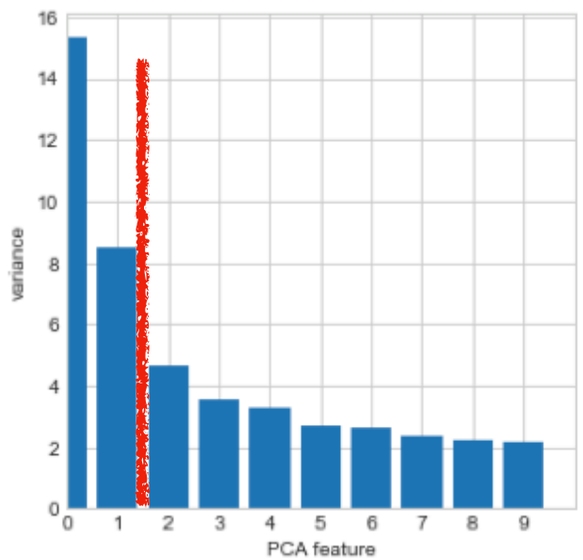


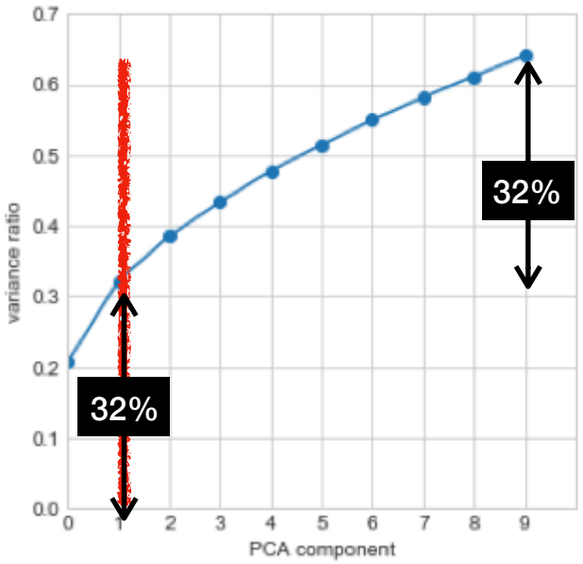
T-SNE decomposition used to transform offence code frequencies into two dimensional space, showed some visual clustering of the precincts

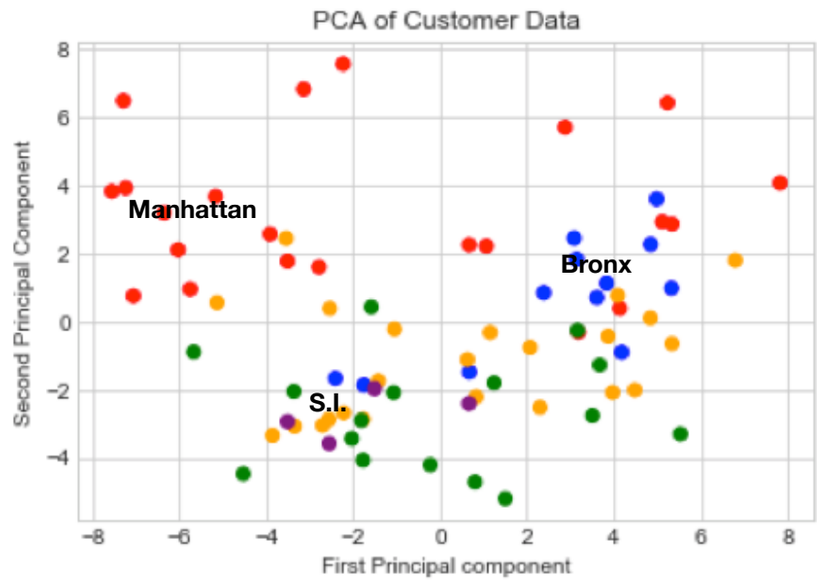
t-SNE preserves nearness of the samples



PCA analysis of variance showed that the dataset can be decomposed into two intrinsic dimensions, that revealed similar clustering to t-SNE method

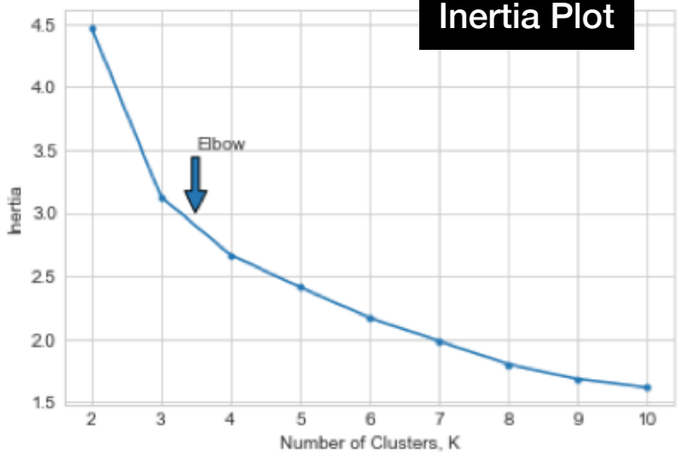




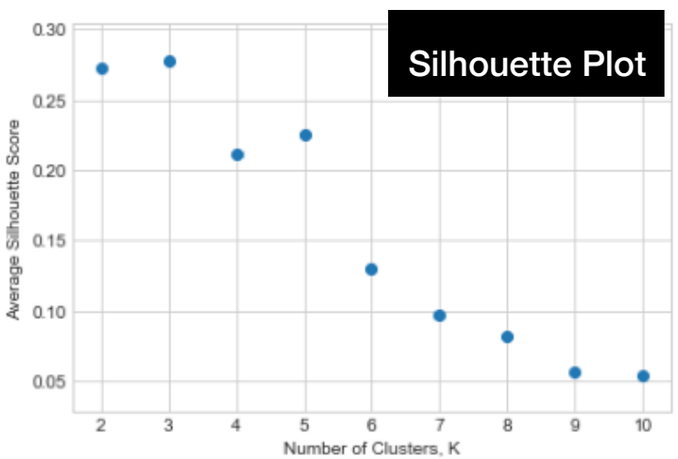


Three methods that were used to specify the number of clusters for K-Means clustering method, give similar results of best K around 2 - 3 clusters

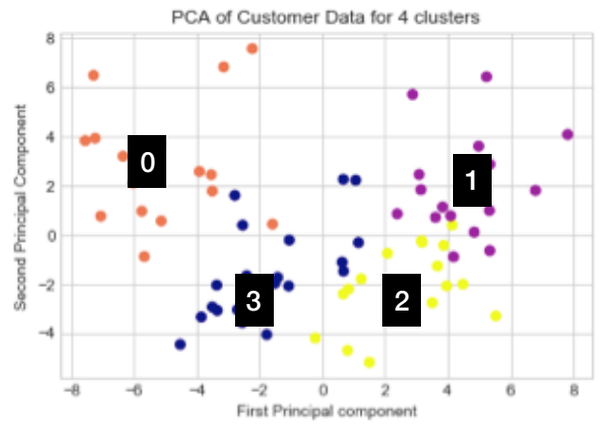
Inertia Plot



Silhouette Plot



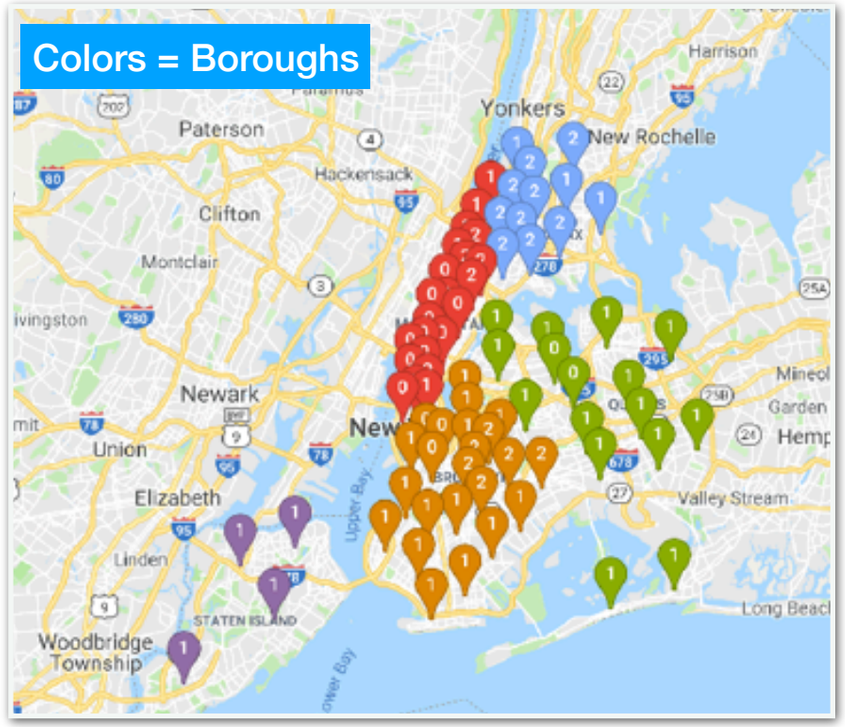
PCA Plots



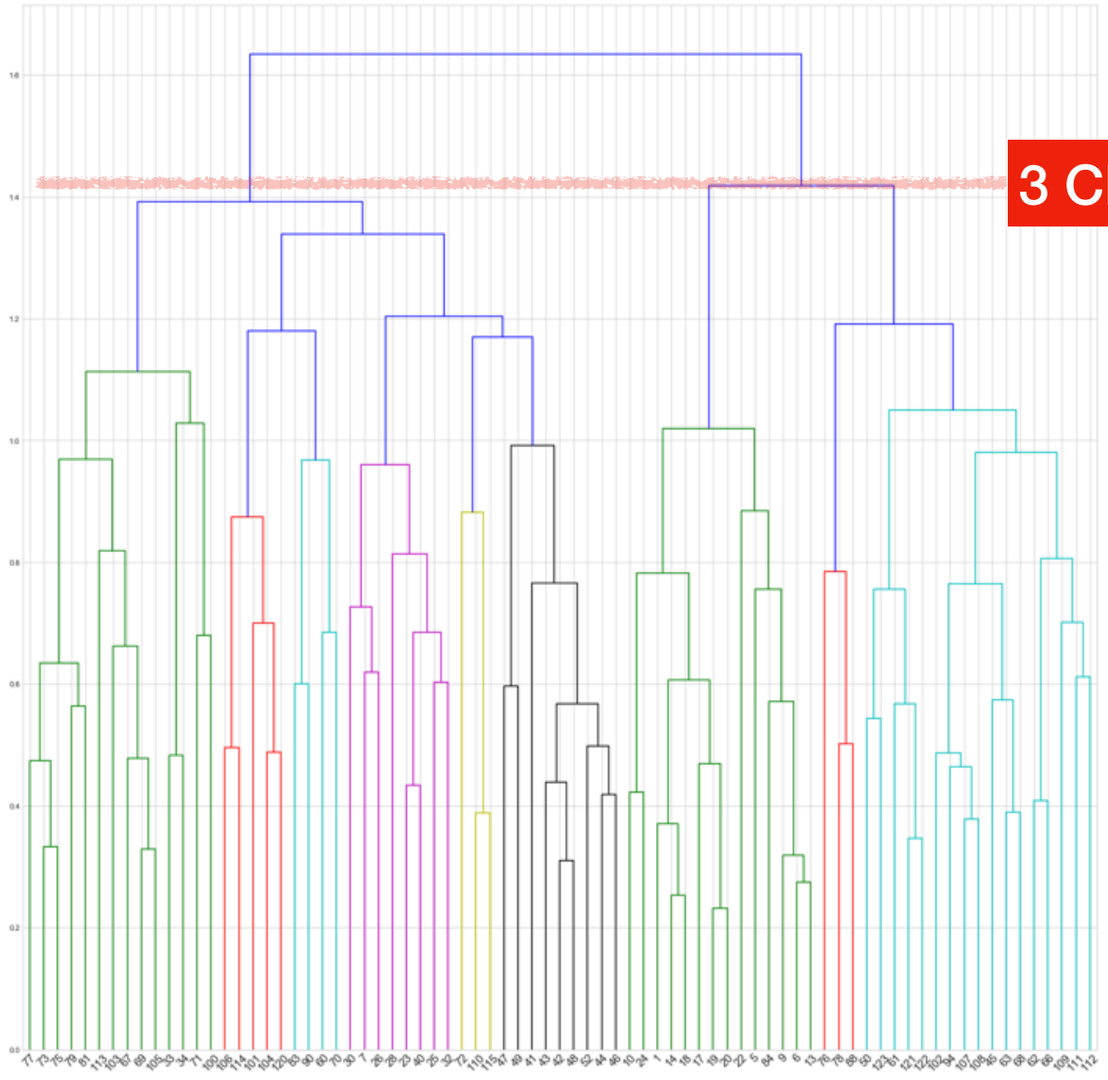
K-Means method for K = 3 produced clusters that often group neighboring precincts. All clusters show similar most frequent offence codes

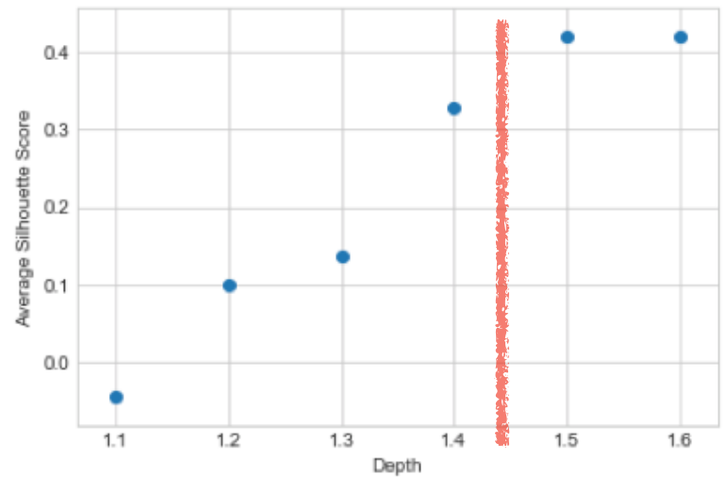
LABEL	BRONX	BROOKLYN	MANHATTAN	QUEENS	STATEN ISLAND
0	0	0	8	1	0
1	5	6	5	8	0
2	0	7	0	2	1

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



Hierarchical clustering was the second clustering method used that showed better silhouette scores for greater depths i.e. smaller number of clusters. Three clusters were chosen for the further study to compare the results with K-Means



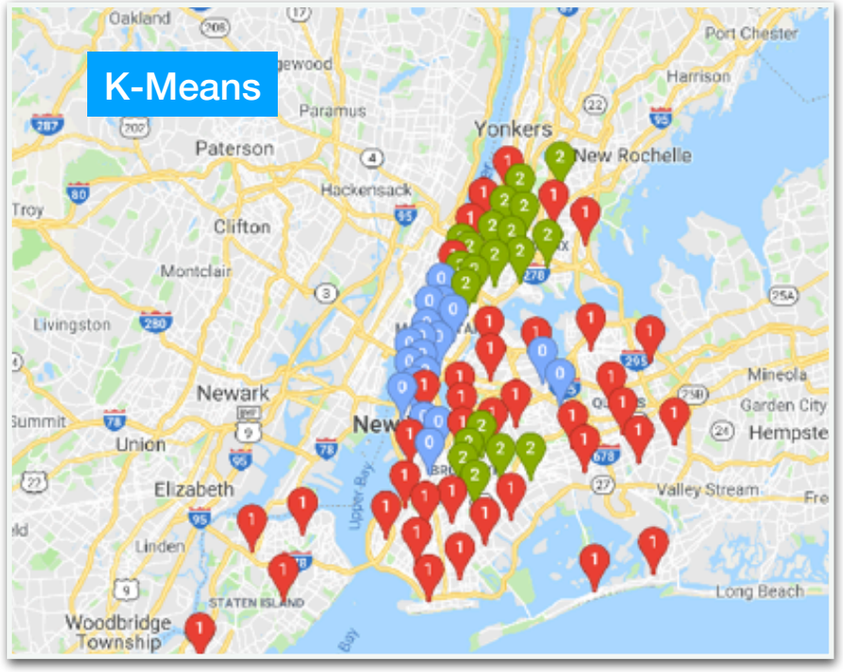


Correlation - based distance was used as dissimilarity measure

Hierarchical clustering produced similar clusters to those in K-Means method once codes

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

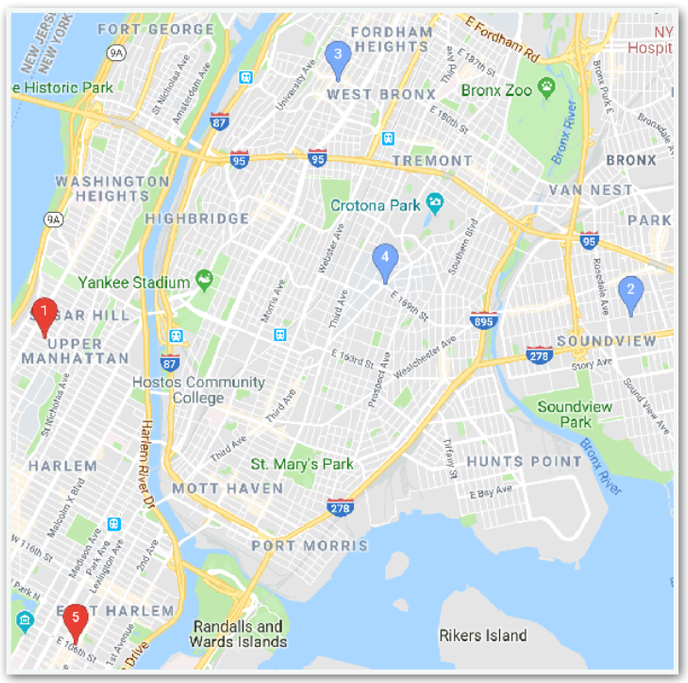
Offence Code	Description	Offence Level	Label
341	PETIT LARCENY	MISDEMEANOR	1
578	HARRASSMENT 2	VIOLATION	1
344	ASSAULT 3 & RELATED	MISDEMEANOR	1
351	CRIMINAL MISCHIEF &	MISDEMEANOR	1
235	DANGEROUS DRUGS	MISDEMEANOR	1
341	PETIT LARCENY	MISDEMEANOR	2
109	GRAND	FELONY	2
578	HARRASSMENT 2	VIOLATION	2
344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



A tool to study cosine similarities among precincts was created based on NMF that shows on the map the location of the most similar precincts and the offence codes that contribute the most to those similarities

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

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



Results Summary

- Number of offenses per year decreases on the city level as well as in each borough
- Crime rates show much bigger differences if studied per area or population of a given borough

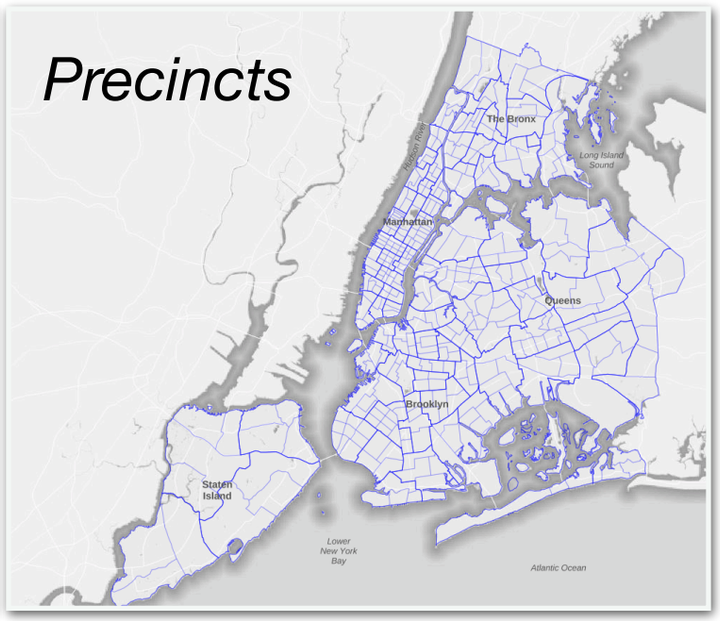
- A linear regression model was developed to understand relation between serious crime rates and some demographic factors. Generally, crimes rates mostly decrease with the increase of native New Yorkers or foreign-born citizens, and increase with the number of single households

- The study of the housing market showed influence of the Great Recession, therefore no significant correlations between crime rates in respective boroughs and housing market indicators were found

- 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 developed to compare similar precincts and their location

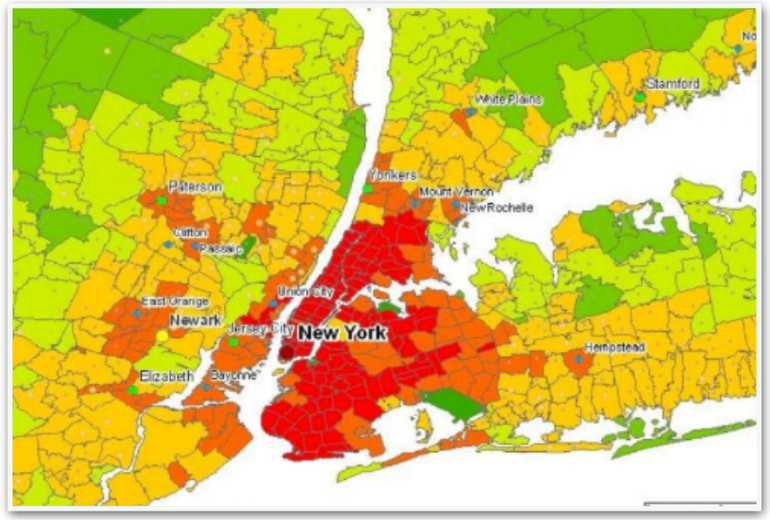
Assumptions and Limitations

- Two datasets used operate on different granularity level - precincts and neighborhoods often overlap but still differ in size and number
- Homogenous population density was assumed but it varies among precincts
- No information on number of police resources deployed in each borough or complaints dataset limitations



Different sub-borough granularities

Neighborhood ≠ Precinct



Different densities

Different polices resources ?

Recommendations

- The continuation of this analysis requires more information on the data collection conditions
- Neighborhoods are more natural divisions than precincts = more useful information
- Not every type of crime is indicator of (un)safeness - selection of severe crime codes can have more practical application

Verification of NYPD database

Neighborhood Level Analysis

Analysis per crime type
What crimes decrease safety?