IBM Foursquare - Battle of the Neighbourhoods

Clustering Neighbourhoods in New York City

Richard Pears

May, 2020

1. Introduction

1.1 Background

When most Individuals relocate to a new city in search of a better job or a better life, they seldom think about the location of where to live in a city,generally the choice is left to immediate availability and convenience but Its not the best approach in selecting a neighbourhood to live in.Everyone chooses a location based on proximity to their workplace or they chose a place that is more suited to their budget but it's not easy to identify that without having prior knowledge regarding a place. To minimize the risk and errors we will be looking into proper research before making a decision on where to start searching for good housing, as making a decision like this can't be changed that easily and it involves a lot of work, time and money hence its important that the right decision is made in the first attempt in picking a house in the ideal neighbourhood based on one's requirement

1.2 Problem

The main issues that will be tackled in this project are the cost of housing and the safety of its neighbourhood,both these are determined using the housing sales data from Kaggle and NYC crime data from the cities database hosted online. The goal is to identify a borough and pick the best neighbourhood from that borough that has relatively low housing cost and is a safe place for a family, and implement k-means clustering to group them using the Foursquare data for venues in respective neighbourhoods.

1.3 Interests

Individuals who are trying to determine the cost and safety of the neighbourhoods in NY and the proximity of venues can utilize this report to make an informed decision before chosing a home in the city and it also provides information on the ideal cost of a neighbourhood for any business trying to open a new outlet in a particular neighbourhood.

2. Data acquisition and cleaning

2.1 Data sources

The data utilized in this project comes from three main sources,

- Borough data taken from wikipedia - link
- Crime dataset taken from nydata - link
- NYC sales data taken from kaggle - link
- Foursquare API - Venues for locations

The datasets are collect and imported from the data folder for our exploratory data analysis and for understanding our objectives

2.1A Borough data

The borough data provides details in regards to the neighbourhood with corresponding boroughs, although the population data that's in the table isn't accurate to current population numbers we dont require those metrics hence it is ignored.

2.1B Crime data

The details of all crime reports in the city allows us to determine the locations where crimes are more frequent, the boroughs that report a high number of crimes are identified with this data and the data fields that are irrelevant are dropped and renamed.

- The crime data is over the time period of 11 months from nov-18' to sept-19'
- The unwanted columns are removed from the dataframe as they add no value to the objective at hand as they mainly comprise of codes of reference
- The columns are renamed to make them more readable to the viewer and descriptions are given below

2.1C NYC housing sales

Here we have a dataset taken from kaggle that gives us property sales from the years 2016 and 2017, we will utilize this to determine the neighbourhoods that support moderate housing costs, the data is considerably large and there are multiple missing values. After removing the rows with multiple Nan's we also get a filter on the housing cost and the square feet area. We will utilize this to determine the neighbourhoods that support moderate housing costs, and help us identify neighbourhoods that are skyhigh when it comes to rentals and property values.

2.2 Data Cleaning

The common practices for data cleaning are utilized and nan values from important fields are addressed, along with the outliers that are identified and removed.

Community Boards in New York City

Community Board (CB)	Area km^2	Pop. Census 2010	Pop./ km^2	Neighborhoods
Bronx CB 1	7.17	91,497	12,761	Melrose, Mott Haven, Port Morris
Bronx CB 2	5.54	52,246	9,792	Hunts Point, Longwood
Bronx CB 3	4.07	79,762	19,598	Claremont, Concourse Village, Crotona Park, Morrisania
Bronx CB 4	5.28	146,441	27,735	Concourse, Highbridge
Bronx CB 5	3.55	128,200	36,145	Fordham, Morris Heights, Mount Hope, University Heights
Bronx CB 6	4.01	83,268	20,765	Bathgate, Belmont, East Tremont, West Farms
Bronx CB 7	4.84	139,286	28,778	Bedford Park, Norwood, University Heights
Bronx CB 8	8.83	101,731	11,521	Fieldston, Kingsbridge, Kingsbridge Heights, Marble Hill, Riverdale, Spuyten Duyvil, Van Cortlandt Village
Bronx CB 9	12.41	172,298	13,884	Bronx River, Bruckner, Castle Hill, Clason Point, Harding Park, Parkchester, Soundview, Unionport
Bronx CB 10	16.76	120,392	7,183	City Island, Co-op City, Locust Point, Pelham Bay, Silver Beach, Throgs Neck, Westchester Square
Bronx CB 11	9.32	113,232	12,149	Allerton, Bronxdale, Indian Village, Laconia, Morris Park, Pelham Gardens, Pelham Parkway, Van Nest
Bronx CB 12	14.56	152,344	10,463	Baychester, Edenwald, Eastchester, Fish Bay, Olinville, Wakefield, Williamsbridge, Woodlawn
Brooklyn CB 1	12.82	160,338	12,507	Greenpoint, Williamsburg, Williamsburg Houses
Brooklyn CB 2	7.72	98,620	12,775	Boerum Hill, Brooklyn Heights, Brooklyn Navy Yard, Clinton Hill, Dumbo, Fort Greene, Fulton Ferry, Fulton Mall, Vinegar Hill
Brooklyn CB 3	7.67	143,867	18,757	Bedford-Stuyvesant, Ocean Hill, Stuyvesant Heights
Brooklyn CB 4	5.31	104,358	19,653	Bushwick
Brooklyn CB 5	14.61	173,198	11,855	City Line, Cypress Hills, East New York, Highland Park, New Lots, Starrett City
Brooklyn CB 6	9.01	104,054	11,549	Carroll Gardens, Cobble Hill, Gowanus, Park Slope, Red Hook
Brooklyn CB 7	10.96	120,063	10,955	Greenwood Heights, Sunset Park, Windsor Terrace
Brooklyn CB 8	4.25	96,076	22,606	Crown Heights, Prospect Heights, Weeksville
Brooklyn CB 9	4.07	104,014	25,556	Crown Heights, Prospect Lefferts Gardens, Wingate
Brooklyn CB 10	10.57	122,542	11,593	Bay Ridge, Dyker Heights, Fort Hamilton

The data taken from wikipedia that contains a table containing information regarding the neighbourhoods and boroughs in new york city and the the other details aren't accurate and are removed

The data is formatted into a pandas dataframe and the community board column is transformed into another field only containing the borough names.The neighbourhoods are separated into separate values and iterated to each borough, The other columns are not required and are dropped.

The table is then parsed to get relevant details with the geopy library to get location data and add that into the dataframe for each corresponding neighbourhood and borough.

	search_name	lat	lon	importance	type	Neighbourhood	Borough	boundingbox	class
0	Melrose Bronx	40.8256703	-73.9152416	0.536855	station	Melrose	Bronx	[40.8206703, 40.8306703, -73.9202416, -73.9102...	railway
1	Mott Haven Bronx	40.8089897	-73.9229147	0.623710	neighbourhood	Mott Haven	Bronx	[40.8089397, 40.8090397, -73.9229647, -73.9228...	place
2	Port Morris Bronx	40.8015147	-73.9095811	0.606728	neighbourhood	Port Morris	Bronx	[40.8014647, 40.8015647, -73.9096311, -73.9095...	place
3	Hunts Point Bronx	40.8126008	-73.8840247	0.673005	neighbourhood	Hunts Point	Bronx	[40.8125508, 40.8126508, -73.8840747, -73.8839...	place
4	Longwood Bronx	40.8162916	-73.8962205	0.532702	station	Longwood	Bronx	[40.8112916, 40.8212916, -73.9012205, -73.8912...	railway
5	Claremont Bronx	40.8341667	-73.9102778	0.350000	park	Claremont	Bronx	[40.8341167, 40.8342167, -73.9103278, -73.9102...	leisure
6	Concourse Village Bronx	40.823868149999996	-73.92118091218828	0.500000	residential	Concourse Village	Bronx	[40.8227742, 40.8249622, -73.9229217, -73.9194...	landuse
7	Crotona Park Bronx	40.838901899999996	-73.89386451155042	0.644356	park	Crotona Park	Bronx	[40.8344186, 40.8434525, -73.9011324, -73.886757]	leisure
8	Morrisania Bronx	40.8292672	-73.9065253	0.540718	neighbourhood	Morrisania	Bronx	[40.8292172, 40.8293172, -73.9065753, -73.9064...	place
9	Concourse Bronx	40.8185618	-73.927303	0.589675	station	Concourse	Bronx	[40.8135618, 40.8235618, -73.932303, -73.922303]	railway
10	Highbridge Bronx	40.83653245	-73.92959563928645	0.400000	residential	Highbridge	Bronx	[40.8351109, 40.8387844, -73.9306484, -73.9280...	landuse
11	Fordham Bronx	40.8614754	-73.8905439	0.545961	station	Fordham	Bronx	[40.8564754, 40.8664754, -73.8955439, -73.8855...	railway
12	Morris Heights Bronx	40.8498223	-73.919859	0.631291	neighbourhood	Morris Heights	Bronx	[40.8497723, 40.8498723, -73.919909, -73.919809]	place

Crime data

The cleaning is done by removing a lot of data points that aren't required, the data also only stores information regarding the boroughs that correspond to each crime that was committed and the total data points are more than eighty thousand, however there are multiple Nan values and the location data although does contain latitudinal and longitudinal fields lacks over more than 80 percent of its data, hence it cannot be utilized to obtain the neighbourhoods that correspond to the crime incidents.

We will for the sake of interpretation be removing multiple fields that are deemed unnecessary for the scope of our project as they aren't relevant in determining our ultimate goal which is to determine the safety of locations in the city of new york.

CMPLNT_NUM - removed
CMPLNT_FR_DT - complaintdate
CMPLNT_FR_TM - complainttime
CMPLNT_TO_DT - complaintlastdate
CMPLNT_TO_TM - complaintlasttime
ADDR_PCT_CD - removed
RPT_DT - reporteddate
KY_CD - removed

OFNS_DESC - offencedescription
PD_CD - removed
PD_DESC - removed
CRM_ATPT_CPTD_CD - status
LAW_CAT_CD - crimecategory
BORO_NM - borough
LOC_OF_OCCUR_DESC - crimelocation
PREM_TYP_DESC - premisestype
JURIS_DESC - removed
JURISDICTION_CODE - removed
PARKS_NM - removed
HADEVELOPT - removed
HOUSING_PSA - removed
X_COORD_CD - removed
Y_COORD_CD - removed
SUSP_AGE_GROUP - suspectage
SUSP_RACE - suspectrace
SUSP_SEX - suspectsex
TRANSIT_DISTRICT - removed
Latitude - removed
Longitude - removed
Lat_Lon - location
PATROL_BORO - removed
STATION_NAME - removed
VIC_AGE_GROUP - victimeage
VIC_RACE - victimerace
VIC_SEX - victimesex

The data frame is cleaned and the significant outliers are removed.

	complaintdate	complainttime	complaintlastdate	complaintlasttime	reporteddate	offencedescription	status	crimecategory	borough	crimelocation	premisestype	suspectage	suspectrace	suspectsex	
0	11/30/2018	22:00:00	01/07/2019	17:00:00	01/09/2019	PETIT LARCENY	COMPLETED	MISDEMEANOR	BROOKLYN	INSIDE	RESIDENCE - APT. HOUSE	UNKNOWN	UNKNOWN	F	(40.6577578 -73.9517740
1	11/30/2018	06:50:00	NaN	NaN	01/06/2019	HARRASSMENT 2	COMPLETED	VIOLATION	QUEENS	INSIDE	RESIDENCE-HOUSE	25-44	BLACK	M	(40.7082189 -73.7360338
2	11/30/2018	13:15:00	01/10/2019	16:00:00	01/10/2019	PETIT LARCENY	COMPLETED	MISDEMEANOR	QUEENS	INSIDE	STREET	25-44	BLACK HISPANIC	M	(40.6995935 -73.8940697
3	11/30/2018	11:00:00	12/05/2018	09:00:00	01/06/2019	THEFT-FRAUD	COMPLETED	FELONY	BROOKLYN	INSIDE	RESIDENCE - APT. HOUSE	UNKNOWN	UNKNOWN	U	(40.6086413 -73.9904862
4	11/30/2018	00:01:00	11/30/2018	23:59:00	01/08/2019	HARRASSMENT 2	COMPLETED	VIOLATION	STATEN ISLAND	FRONT OF	RESIDENCE-HOUSE	25-44	BLACK	F	(40.6413541 -74.0906960

.

The data does contain a significant number of missing values but they are not in the important fields that we need to use for our objective.

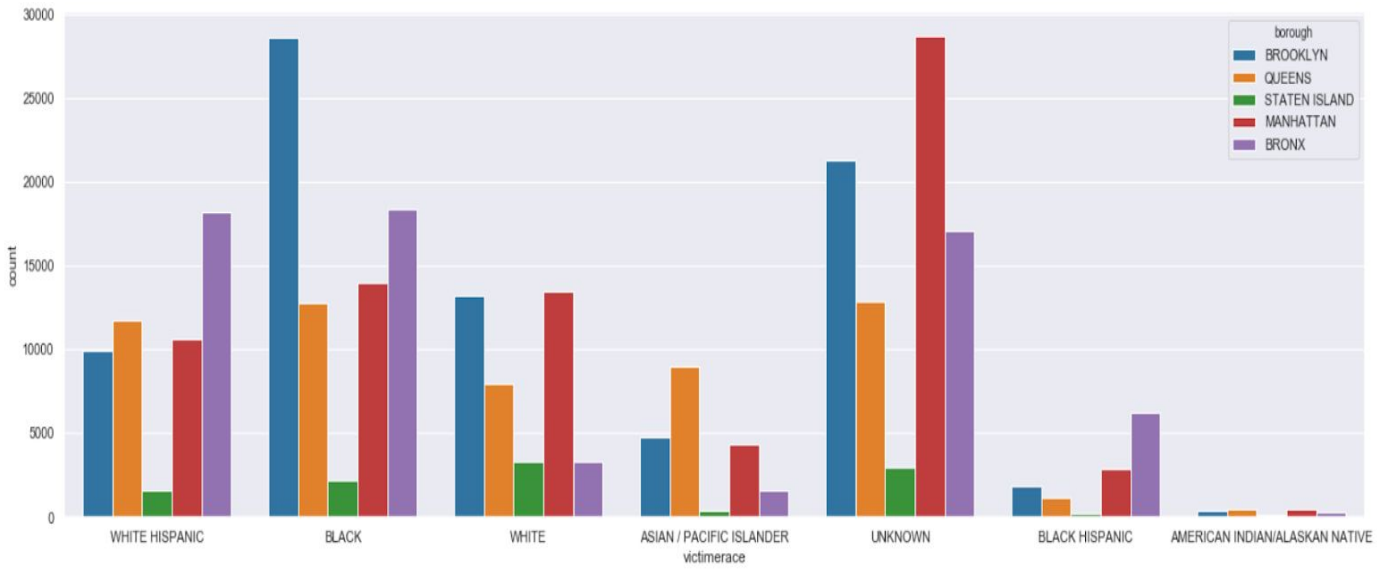
Sales data

The boroughs are marked as binned values, using the information provided they are renamed to their names.The sale prices have a few outliers that need to be removed to have the sales price ranges more practical and within the scope of what we need to be searching for hence its set from 10000 to 1000000 dollars.The same is applied to the land square feet field that sets the range filter to a realistic housing area from 100 to10000 square feet.

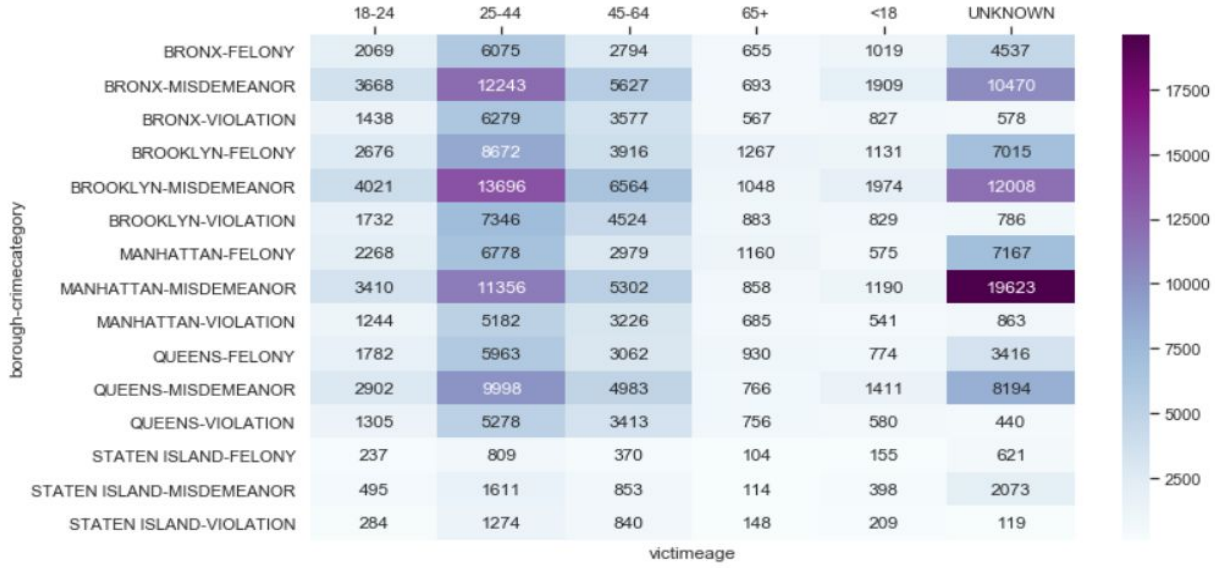
	Borough	Neighborhood	Building Class Category	Tax Class At Present	Block	Lot	Building Class At Present	Address	Apartment Number	Zip Code	Residential Units	Commercial Units	Total Units	Land Square Feet	Gross Square Feet	Year Built	Tax Class At Time Of Sale	Building Class At Time Of Sale	Sale Price	Sale Date
0	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2A	392	6	C2	153 AVENUE B		10009	5	0	5	1633	6440	1900	2	C2	6625000	2017-07-19 00:00:00
1	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2	399	26	C7	234 EAST 4TH STREET		10009	28	3	31	4616	18690	1900	2	C7	-	2016-12-14 00:00:00
2	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2	399	39	C7	197 EAST 3RD STREET		10009	16	1	17	2212	7803	1900	2	C7	-	2016-12-09 00:00:00
3	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2B	402	21	C4	154 EAST 7TH STREET		10009	10	0	10	2272	6794	1913	2	C4	3936272	2016-09-23 00:00:00
4	Manhattan	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2A	404	55	C2	301 EAST 10TH STREET		10009	6	0	6	2369	4615	1900	2	C2	8000000	2016-11-17 00:00:00

3. Exploratory Data Analysis and methodology

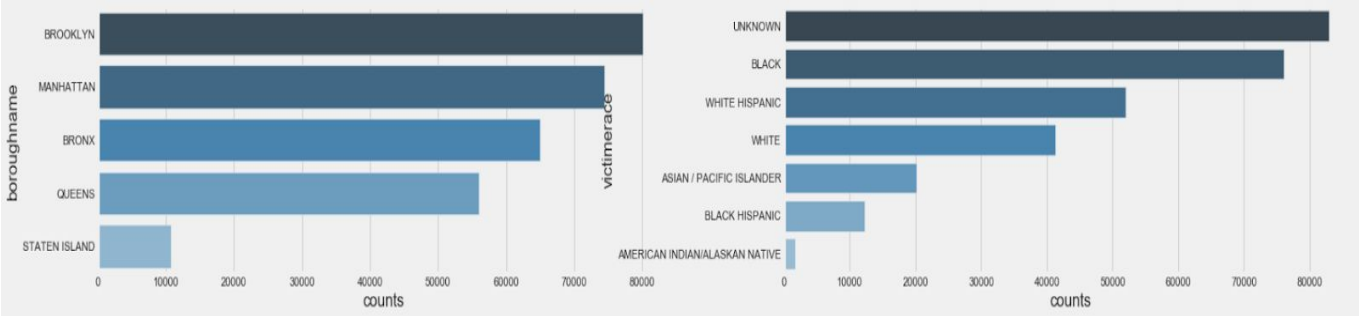
The analysis performed on the crime data shows us significant insights on the rate of crimes and the types that are prevalent in certain boroughs. The first plot on this data shows us the relations between the victims race and the borough that they are prevalent in.



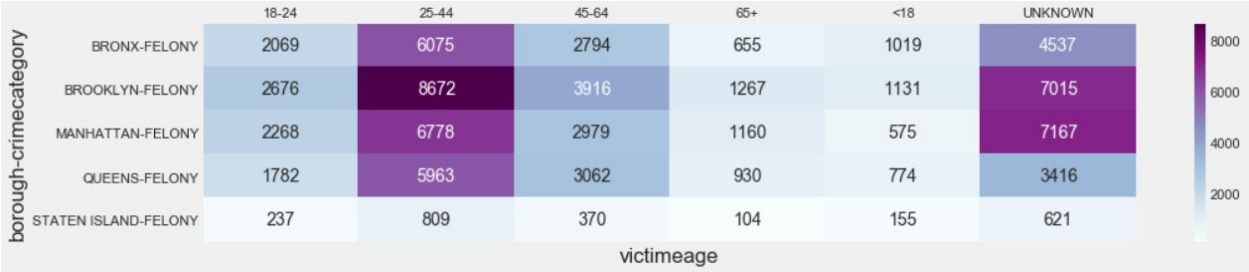
The heatmap for the counts of crimes are shown with a pivot table made with the type of crime and the borough that it had occurred in.The plot shows us that the crimes that occured in manhattan and brooklyn are considerably high especially the misdemeanors that occured in manhattan.



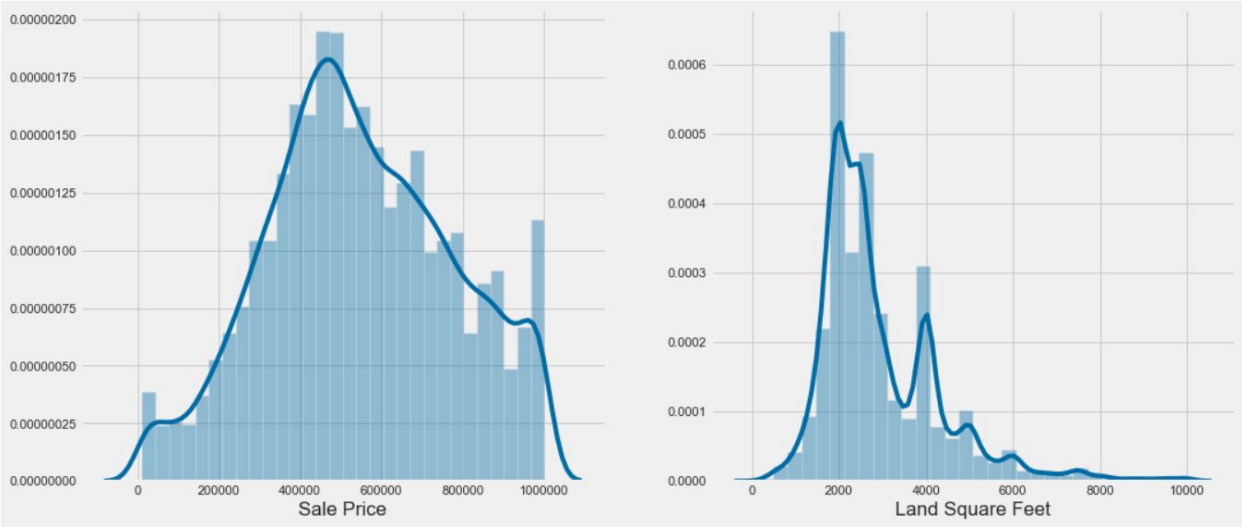
The Borough crime counts and the race counts shows us the relative race victims that are generally targeted in new york.



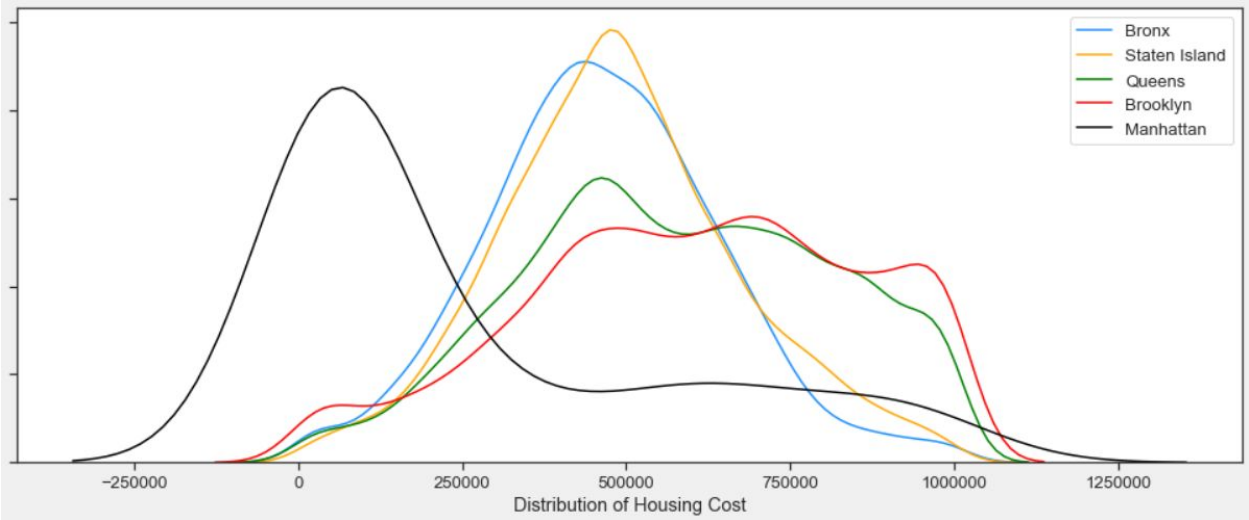
There are three main categories in the crime types, violations, misdemeanors and felonies.The federal bureau categories felonies as a much more grave crime, filtering only for felonies we get, the same heatmap for the age and boroughs.



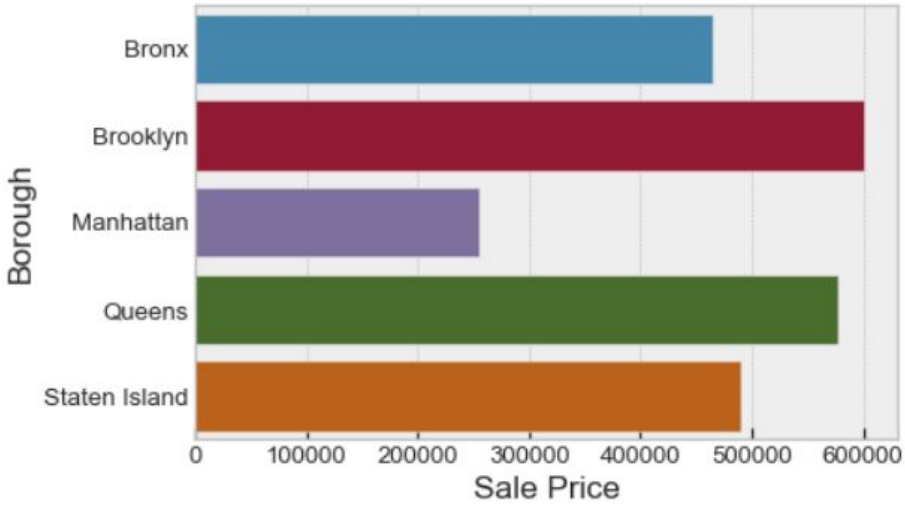
We now take a look at the housing data, and try to interpret the costs of housing across various boroughs. The histogram plot shows us the distribution of the sales prices in the city as well as the land square feet of each individual house.



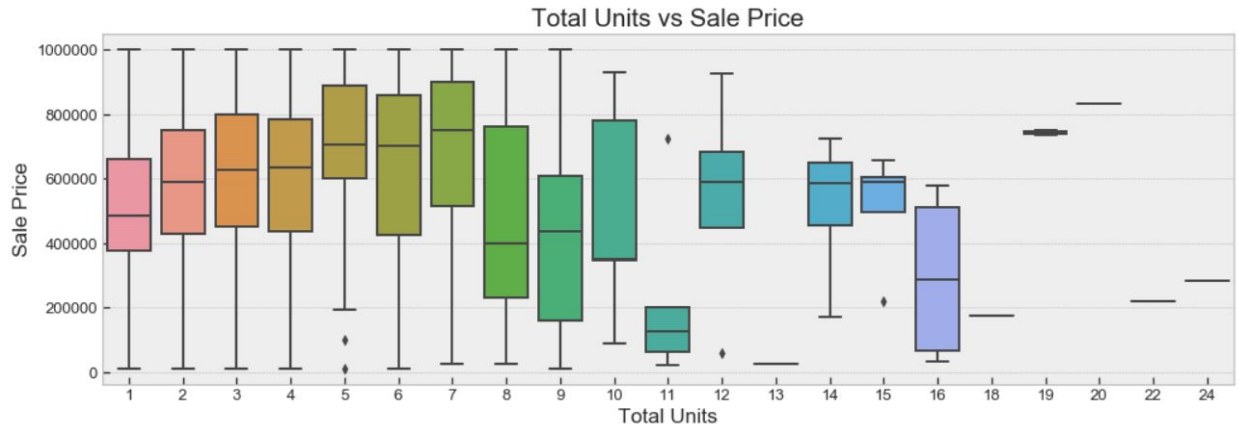
The distribution plot gives us the cost of housing across all the boroughs, here we can notice that the cost of housing in manhattan is relatively low, but however when we cleaned the data there weren't many data points on manhattan.



The bar plot shows us the mean of housing sales in the city with respect to the sale price average we can determine the ideal borough.



The plot shows us the comparison of the sales prices of homes with the total units that are available in each building.This enables us to determine the min max and range of the cost for housing.



Once we analysed the data from the crime and the housing we determined that the less safe boroughs were Brooklyn and Manhattan and we removed those two boroughs from our housing dataset. The data that we have in the housing dataset is pivoted and groped to show the average values in the land square feet and sales prices.

We will be using this table for the rest of our analysis where we plot the neighbourhoods and their corresponding venues using the data we get from the Foursquare api.

We will use the table that we have as shown below to proceed with our analysis.Since the cost of a house is wide and diverse across the neighbourhoods and it's hard to determine the

cheapest borough with the dataset at hand, and not only that the preliminary EDA shows that the boroughs have neighbourhoods that are both in high and low cost.

Borough	Neighborhood	Zip Code	Land Square Feet	Sale Price
Brooklyn	East New York	0	2000	25000
Manhattan	Upper East Side (59-79)	10075	1443	52500
	Midtown West	10019	7532	68261
Brooklyn	Coney Island	0	3502	100000
Queens	Jamaica Bay	11692	4247	110000
		11422	2500	135000
Brooklyn	Downtown-Fulton Mall	11201	2174	149000
	Boerum Hill	11201	1871	150000
	Clinton Hill	11238	3045	167000
Manhattan	Harlem-Central	10026	1682	180000

If we compare this with our crime analysis of newyork it favours Staten island over the other boroughs and this is also due to the fact that Staten Island is actually quite far away and most people prefer to live in the other boroughs.

For this project we need to find the middle ground between the safety of a borough and the cost of housing in its neighbourhood and hence we narrow it down to a mix of boroughs that are relatively cheap, for getting good locations from the foursquare api we will be looking at venues that are relevant to a person who wishing to have all essential amenities that are easily accessible and not at a very far distance from the location of the neighborhood.

We transform the data and get relevant location data using geopy and we finally get the datatable that we will be using for our modeling and visualization of clusters.

	search_name	lat	lon	importance	type	Neighbourhood	Borough	boundingbox	class	Land Square Feet	Sale Price	Zip Code
0	Astoria Queens	40.7720145	-73.9302673	0.667136	neighbourhood	Astoria	Queens	[40.7719645, 40.7720645, -73.9303173, -73.9302...	place	4061	180000	11370
1	Co-Op City Bronx	40.8738889	-73.8294444	0.725617	neighbourhood	Co-Op City	Bronx	[40.8738389, 40.8739389, -73.8294944, -73.8293...	place	5625	193333	10475
2	Fresh Kills Staten Island	40.5642715	-74.186255	0.722075	neighbourhood	Fresh Kills	Staten Island	[40.5642215, 40.5643215, -74.186305, -74.186205]	place	5097	199501	10312
3	Maspeth Queens	40.723158	-73.912637	0.588337	neighbourhood	Maspeth	Queens	[40.723108, 40.723208, -73.912687, -73.912587]	place	2000	205000	11385
4	Belmont Bronx	40.8552778	-73.8863889	0.515440	neighbourhood	Belmont	Bronx	[40.8552278, 40.8553278, -73.8864389, -73.8863...	place	2154	270375	10460
5	Port Ivory Staten Island	40.6409366	-74.1801442	0.725303	neighbourhood	Port Ivory	Staten Island	[40.6408866, 40.6409866, -74.1801942, -74.1800...	place	2952	277227	10303
6	Arverne Queens	40.5934173	-73.7895462	0.563190	suburb	Arverne	Queens	[40.5734173, 40.6134173, -73.8095462, -73.7695...	place	2808	292283	11691
7	Broad Channel Queens	40.6064008	-73.81901879728136	0.660330	neighbourhood	Broad Channel	Queens	[40.5970594, 40.6153344, -73.8248592, -73.8148...	place	3251	309400	11693
8	Fordham Bronx	40.8614754	-73.8905439	0.545961	station	Fordham	Bronx	[40.8564754, 40.8664754, -73.8955439, -73.8855...	railway	2533	321333	10458
9	Concord-Fox Hills Staten Island	40.6151042	-74.0845859	0.617338	neighbourhood	Concord-Fox Hills	Staten Island	[40.6150542, 40.6151542, -74.0846359, -74.0845...	place	1456	326905	10304
10	Stapleton Staten Island	40.6264774	-74.0776361	0.633061	neighbourhood	Stapleton	Staten Island	[40.6264274, 40.6265274, -74.0776861, -74.0775...	place	3300	336527	10304

We visualized the data that we have to show the neighbourhoods that we have in our table to later use for our statistical modeling and clustering analysis.

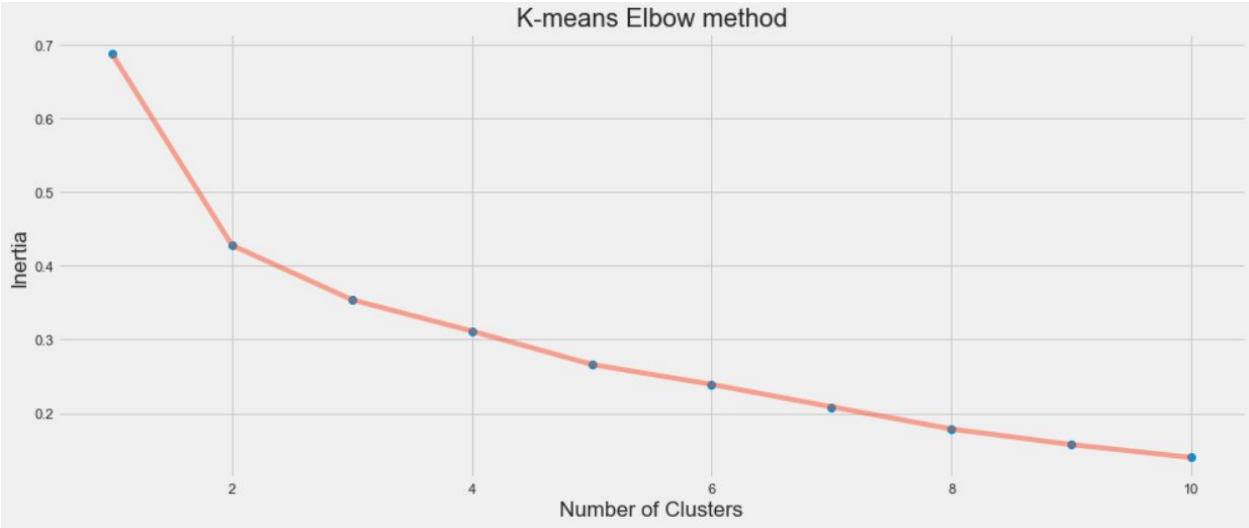
The data is parsed using the foursquare api and corresponding venues are obtained for our clustering analysis.We utilize k-means clustering on our transformed data and the venues that we have to get the ranks of certain venue categories based on the frequency of occurrence of the venue.

We have close to 30 neighbourhoods and we can now proceed to get the clusters and map them using folium to visualize them on the map.They are grouped into 5 clusters and each cluster group is mapped to the dataframe using cluster numbers.

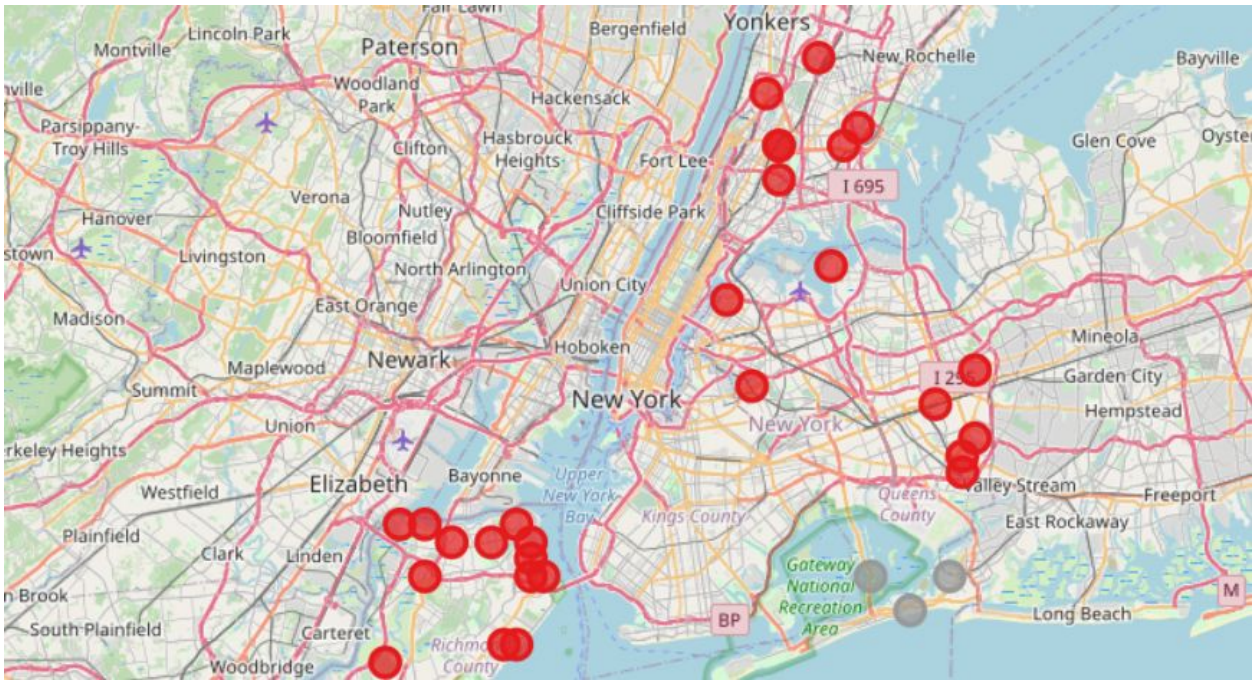
The cluster maps have venues ranked in order of the occurrence of certain venues and they are shown in the dataframe and it is the end inference we need to determine the choice of neighbourhood.

The clusters are then viewed to identify the similarities in the rank of venues, we then apply the elbow method to determine the right number of clusters using the mean squared errors on the dataset.

The elbow method shows us the ideal number of clusters that give us a more clear differentiation on the cluster groups.



Visualizing the clusters when k is two gives us the following map with the cluster groups.



4. Results and Conclusion

This project helps us to understand how to approach a process of selection and elimination based on a set of prerequisites that are deemed essential to choosing a good neighbourhood. The main aspects of the selection are the following three parameters.

- Housing Cost
- Neighbourhood Safety
- Proximity of Essential Venues

The most suitable neighbourhoods are selected across the various boroughs,as all boroughs have certain neighbourhoods that vary to both extreme ends in the cost of housing as well as the safety of the neighbourhood.
The neighbourhoods are selected and with the Foursquare API the venues are collected and they are clustered together for essential venues that are of higher importance than the rest,they are grouped accordingly with the help of k-means and are depicted on the map of the Boroughs in Newyork City

The data has been loaded from various sources and exploratory data analysis has been performed on all of the datasets,all necessary cleaning procedures and mining procedures have been followed to ensure the data is in the proper format.

The findings and interpretations from the visualizations that are gained from the crime and housing datasets, enabled us to narrow it down to our select group of neighbourhoods that we are able to utilize with our foursquare Api to identify ideal locations that are suitable based on

venues preferences.The Average cost of housing and square feet of each housing unit is also made available for making a calculated decision on purchase of a house.The findings are made to match with real world knowledge of the city in NewYork to make the correct decision in the EDA process.

5. Key Findings

- Crimes in the city of New York are classified into three main categories- Felonies, misdemeanors and violations. The US federal system deems Felonies as a bigger crime in comparison to the other two.
- Staten Island is much further away from the city and it takes time for one to travel to the other Boroughs in the city, hence it also can impact the decision making process for choosing a suitable home in the aforementioned borough.
- Crimes in Brooklyn and Manhattan show not only an average high on all categories but especially in the form of misdemeanors, our crime data only is Borough specific and does not contain neighbourhood data hence we remove these two boroughs, not only on the aspect of safety but also on the average cost of housing.
- Housing costs in Manhattan are found to be the lowest, this could also be due to the fact that after stripping the data for ranges between housing costs and land square feet into realistic numbers and removing data points that have Null values we don't have much data points from Manhattan.
- We can presume that Bronx and Staten Island have decent housing in the states, as over the past few years crime rates have dropped drastically in the Bronx which was once a crime infested borough and Staten Island has seen a steady increase in its housing and development from the State of Newyork.
- Our final dataset contains boroughs from Bronx, Staten Island and Queens.It is hard to correlate a whole Borough to a particular parameter like cost or safety hence we look at individual boroughs for determining the cheapest locations.
- The main reason we mix and choose from 3 different boroughs is because it is hard to choose one borough as the ideal as most boroughs have neighbourhoods that are at extreme ends in terms of safety and cost of housing.
- The venues are picked from the Foursquare API and they are clustered into groups with k-means, and each group tends to have certain similarities in them along the rankings of their venues and the frequency of a venue's categorical importance and occurrence in that area.

Notes and further analysis

This project can be modified using the selections and even refined even further with more data specific to each neighbourhood to determine the best neighbourhood to live in based on personal preferences. This is also helpful to someone who needs to shortlist areas based on distances from their places of work, their choice of essential venues etc. The end results shown in the clusters only denote one point of view of extremely cheap neighbourhoods, They can be modified with a selective range on the cost and mitigation on the safety to narrow it down to a list of ideal neighbourhoods to live in the city of New York.