capstone project - Battle of the neighbourhoods

- Richard Pears

Introduction

This project aims to determine the safest and cost effective neighbourhood in the city of New York using multiple datasets. We will determine the best neighbourhood with the data available and then identify the most suitable neighbourhoods in the city with the help of the foursquare API which provides us with the venues that are provided in the explore end points. Once suitable neighbourhoods of a borough are determined we will then perform a suitable statistical modelling on the data at hand to gain certain useful insights that will help make a decision in purchasing a home in the city of New York.

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.

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 choosing 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.

Data acquisition and cleaning

Data sources

- The data utilized in this project comes from three main sources,
- Borough data taken from Wikipedia - link
- Crime dataset taken from NY data - 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

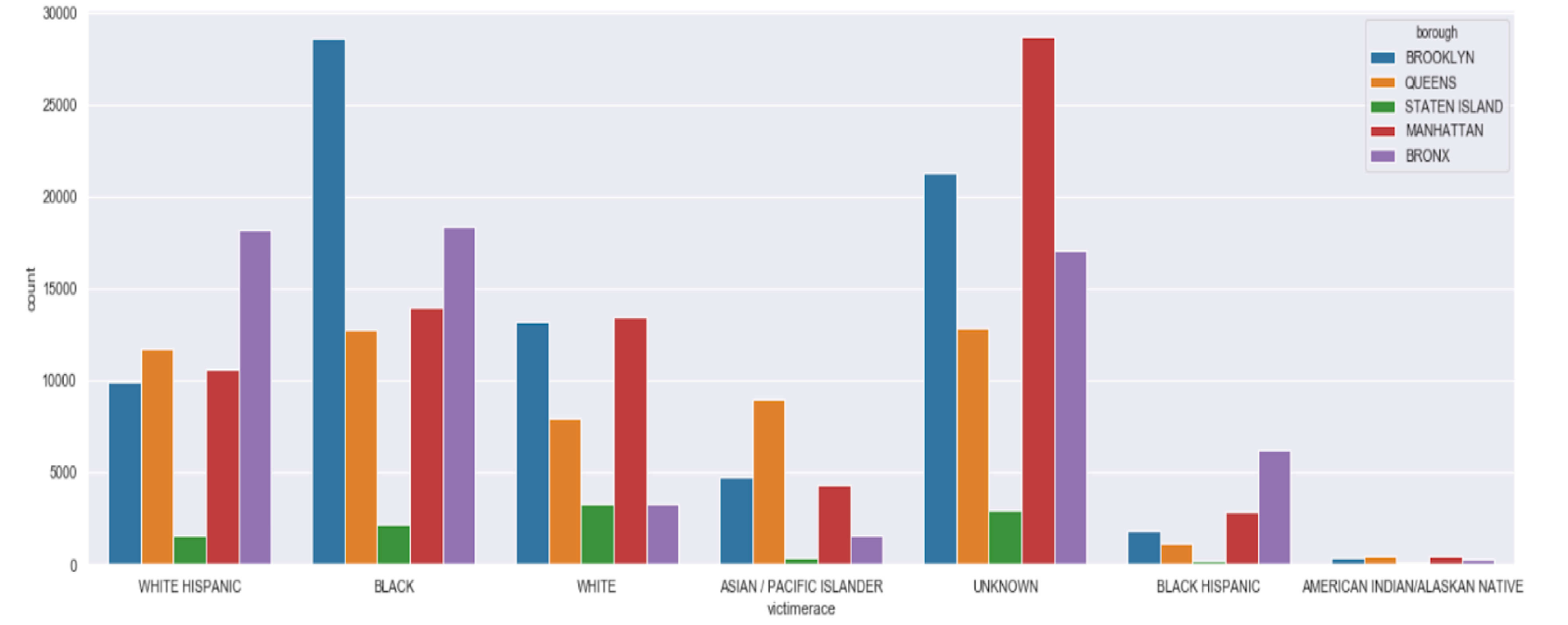
Data cleaning

- The data table taken from Wikipedia is scrapped with beautiful soup and transformed into a data frame, then the community suffix are removed from the neighbourhoods and they are parsed into separate rows.
- The crime dataset is taken from NYC data and they are cleaned for nans and formatted and transformed as required, outliers are identified and removed from the dataset
- The crime data frame is also filtered to showcase real work scenarios as the dataset contains multiple outliers and hence they are removed.
- The sales dataset price ranges are set as per the ideal costs of an apartment within a moderate range of cost and the remaining datapoints are removed, the same is done with the land square feet of each housing unit.
- The sales data is taken from Kaggle and filtered for our price range and unwanted fields are dropped, the data is cleaned for nans and we take the final data frame for our final modelling by merging it with our neighbourhood table for our clustering analysis with k-means

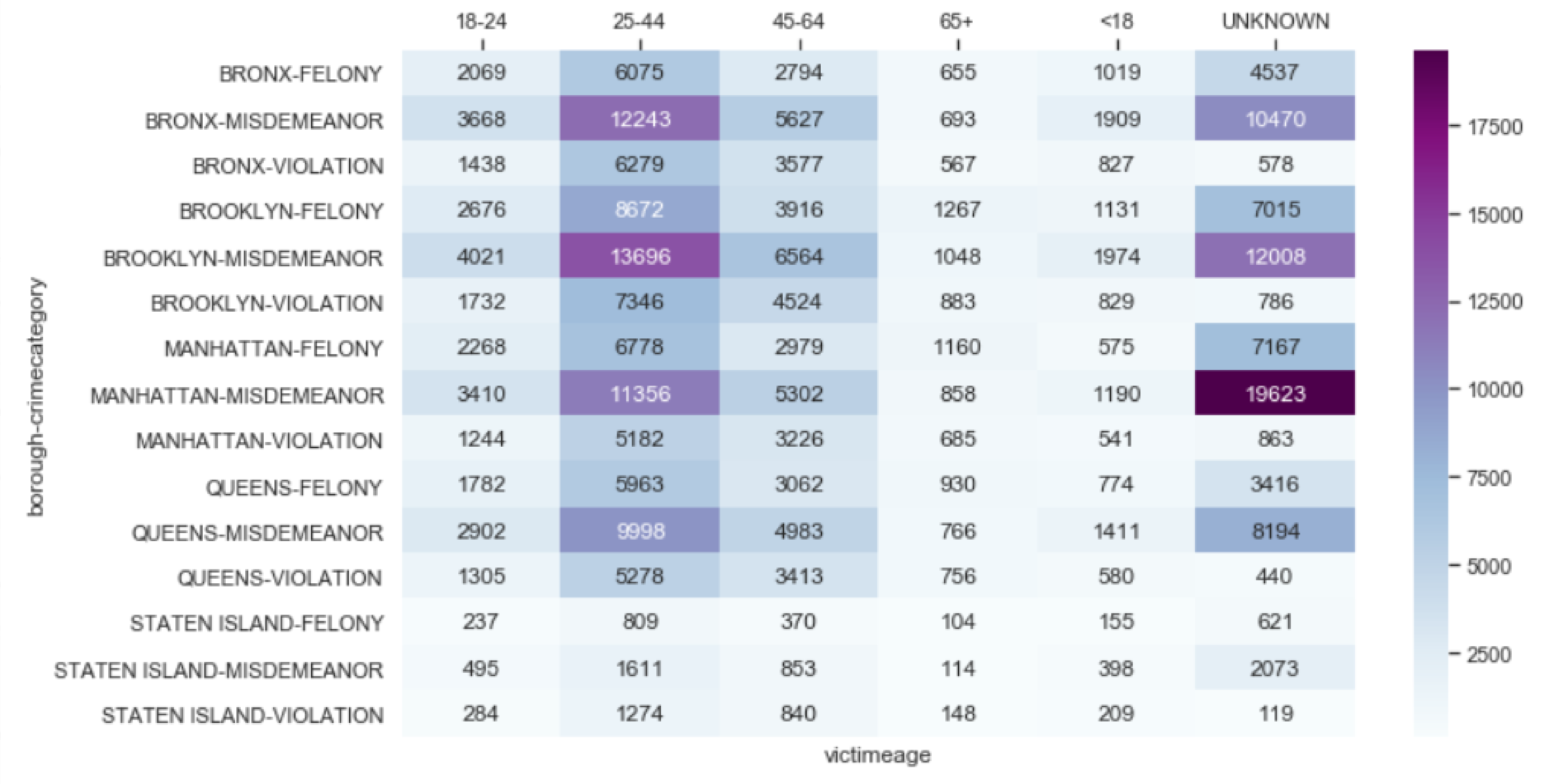
Methodology

Exploratory data analysis

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.

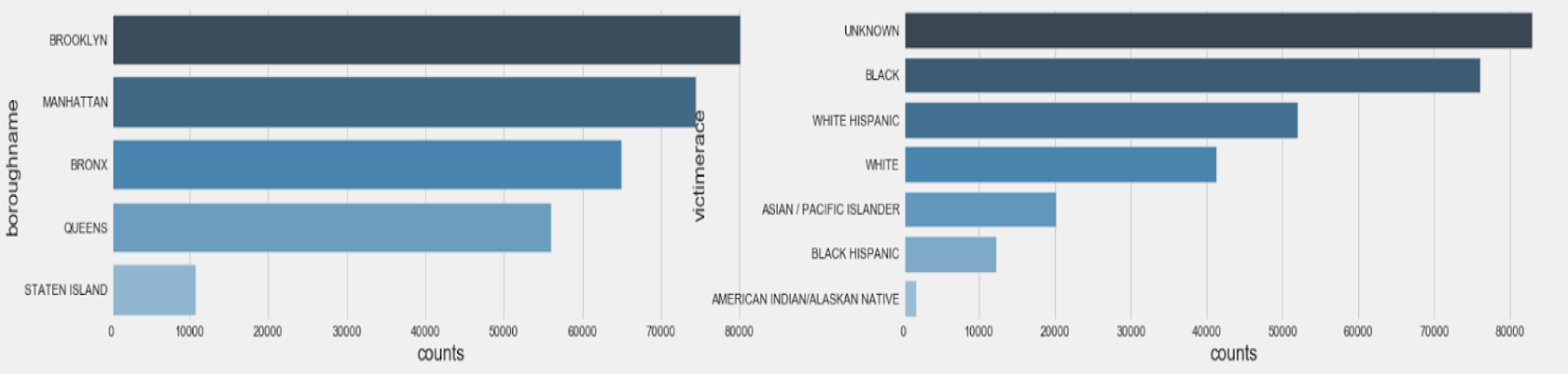


The heatmap for the counts of crimes are shown with a pivot table made with the type of crime and the borough that it had occurred in the location. The plot shows us that the crimes that occurred in Manhattan and Brooklyn are considerably high especially the misdemeanours that occurred in Manhattan.

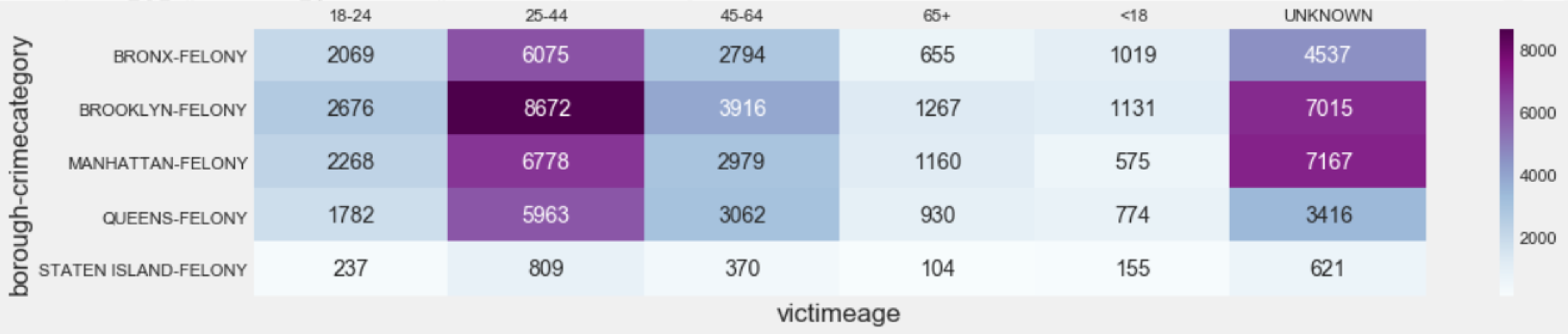


borough-crimecategory	18-24	25-44	45-64	65+	<18	UNKNOWN
BRONX-FELONY	2069	6075	2794	655	1019	4537
BRONX-MISDEMEANOR	3668	12243	5627	693	1909	10470
BRONX-VIOLATION	1438	6279	3577	567	827	578
BROOKLYN-FELONY	2676	8672	3916	1267	1131	7015
BROOKLYN-MISDEMEANOR	4021	13696	6564	1048	1974	12008
BROOKLYN-VIOLATION	1732	7346	4524	883	829	786
MANHATTAN-FELONY	2268	6778	2979	1160	575	7167
MANHATTAN-MISDEMEANOR	3410	11356	5302	858	1190	19623
MANHATTAN-VIOLATION	1244	5182	3226	685	541	863
QUEENS-FELONY	1782	5963	3062	930	774	3416
QUEENS-MISDEMEANOR	2902	9998	4983	766	1411	8194
QUEENS-VIOLATION	1305	5278	3413	756	580	440
STATEN ISLAND-FELONY	237	809	370	104	155	621
STATEN ISLAND-MISDEMEANOR	495	1611	853	114	398	2073
STATEN ISLAND-VIOLATION	284	1274	840	148	209	119

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

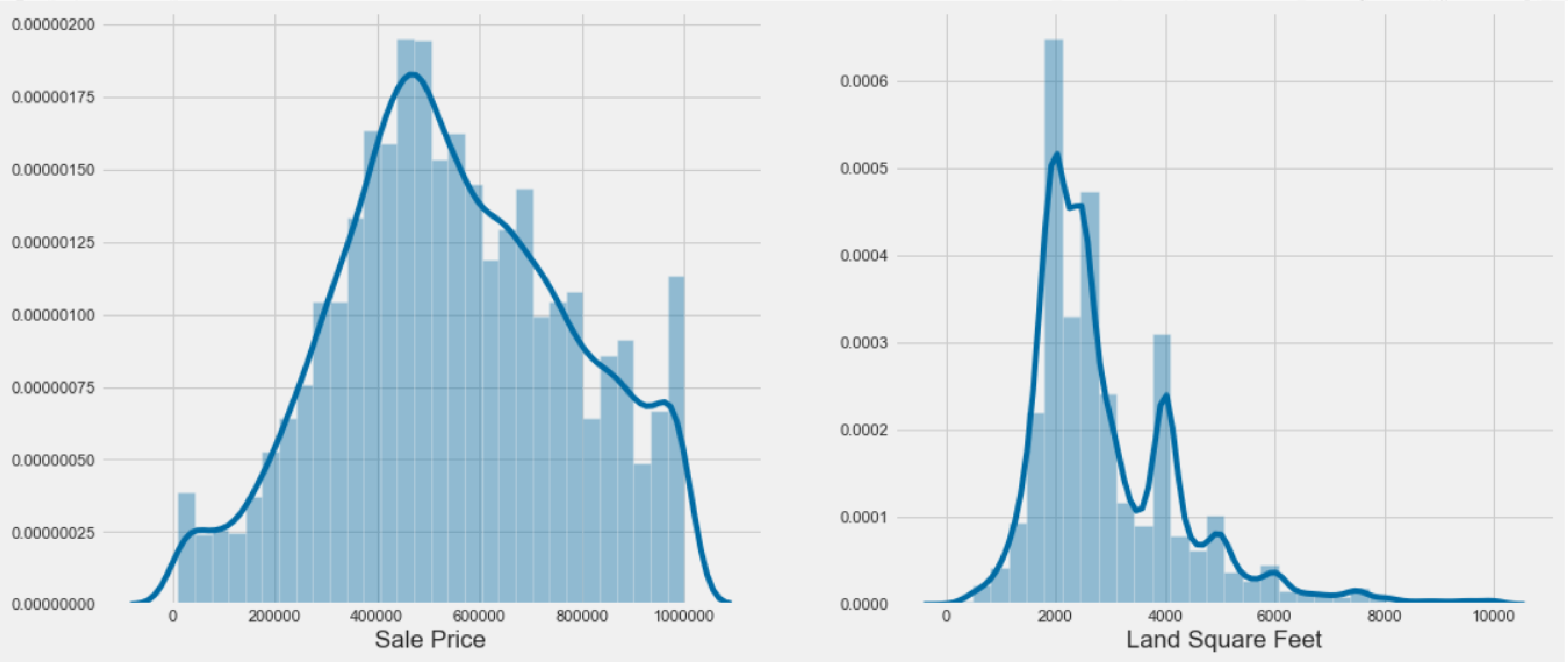


There are three main categories in the crime types, violations, misdemeanours and felonies. The federal bureau categories felonies as a much more grave crime, filtering only for felonies we get, the same heatmap for the age and boroughs

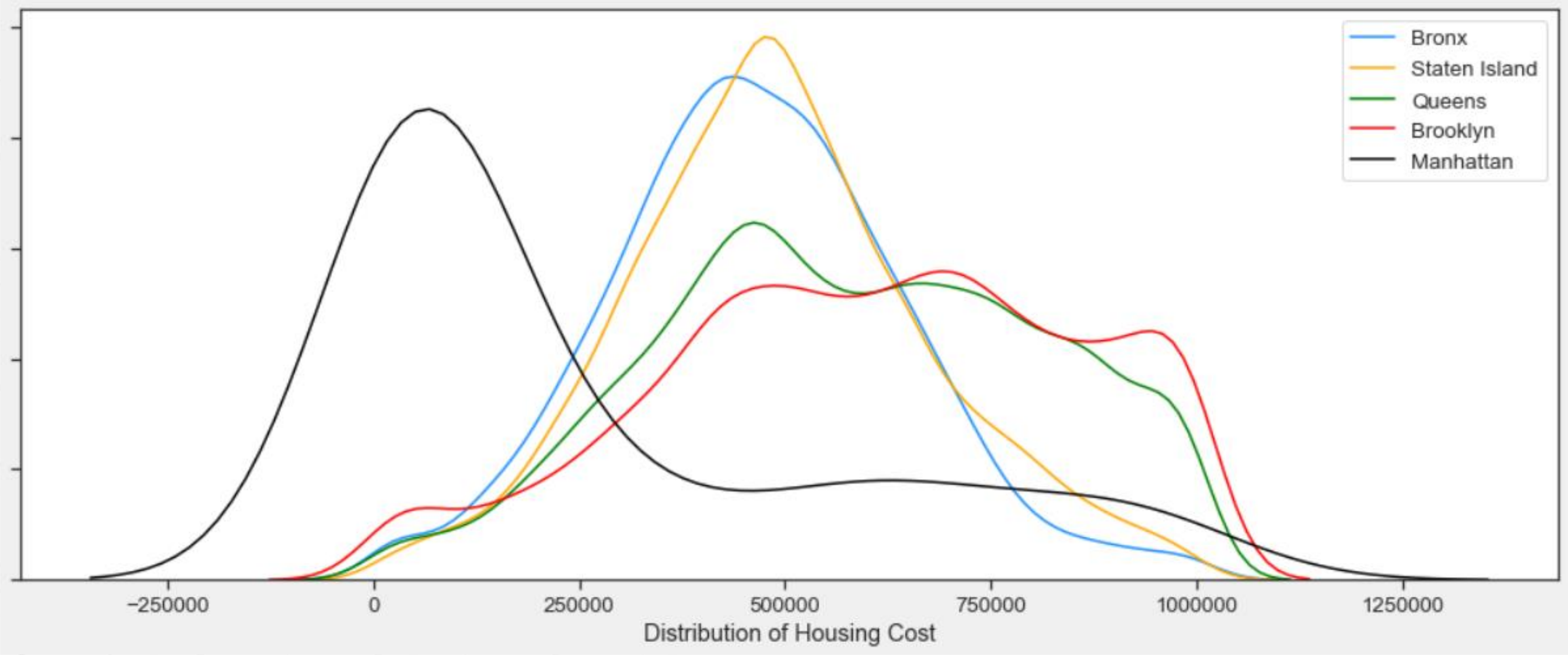


	18-24	25-44	45-64	65+	<18	UNKNOWN
BRONX-FELONY	2069	6075	2794	655	1019	4537
BROOKLYN-FELONY	2676	8672	3916	1267	1131	7015
MANHATTAN-FELONY	2268	6778	2979	1160	575	7167
QUEENS-FELONY	1782	5963	3062	930	774	3416
STATEN ISLAND-FELONY	237	809	370	104	155	621

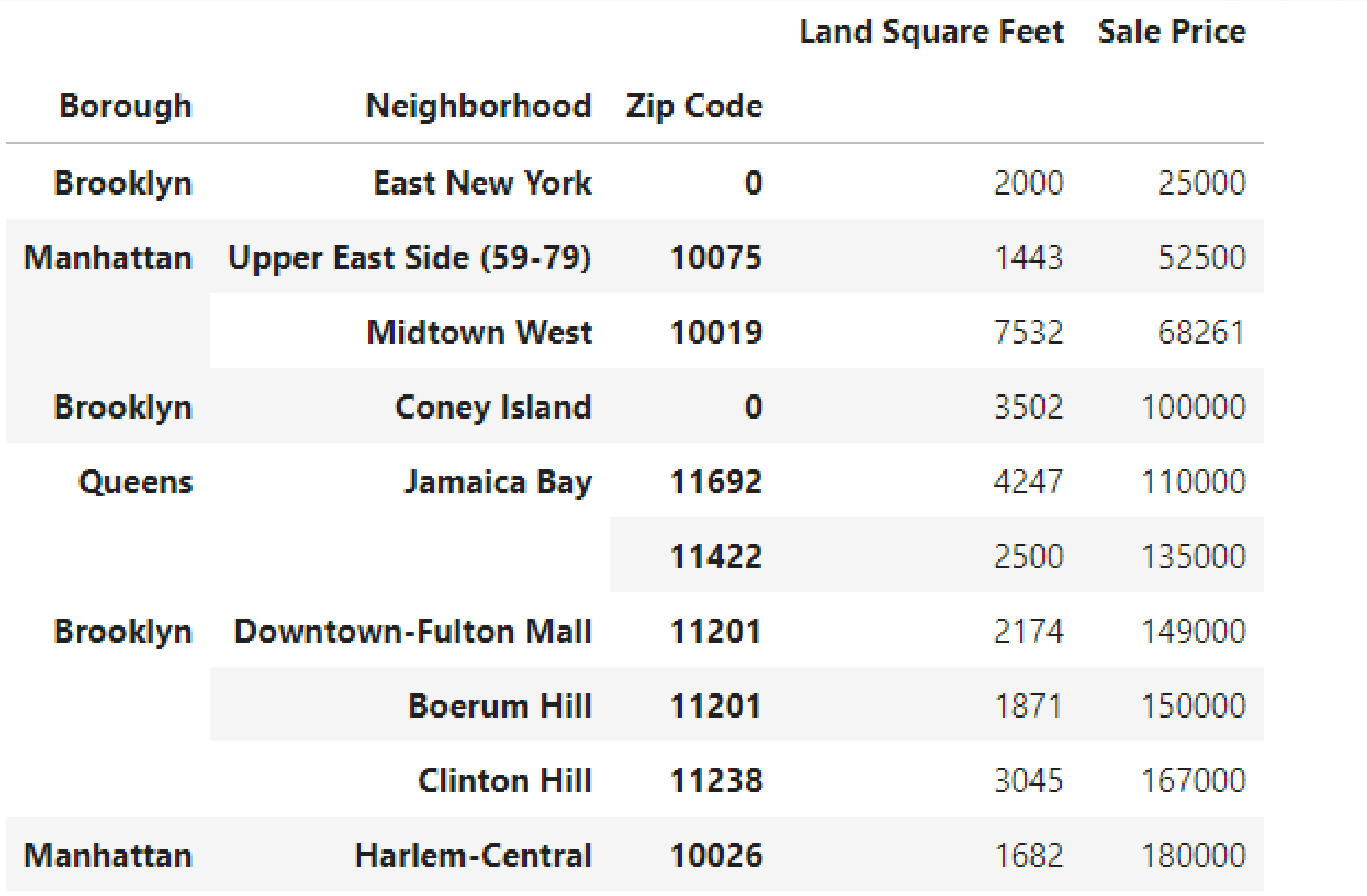
We now take a look at the housing data, and try to interpret the costs of housing across various boroughs. The histogram plot shows us the distribution of the sales prices in the city as well as the land square feet of each individual house.



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 distribution gives a clear picture of the cost of housing in NYC.

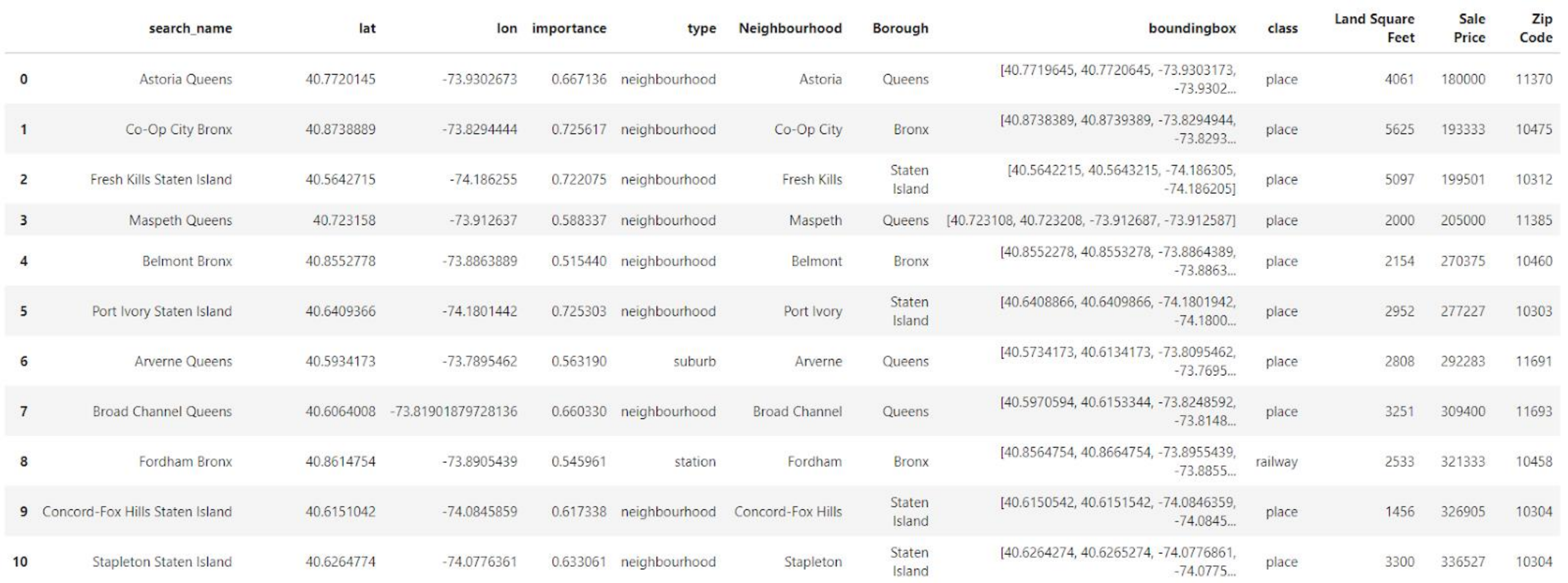


We will use this table which has our neighbourhoods with relatively low cost housing and we will merge this table with the neighbourhood tables to get our final data frame what we will be using for our clustering with k-means clustering algorithm



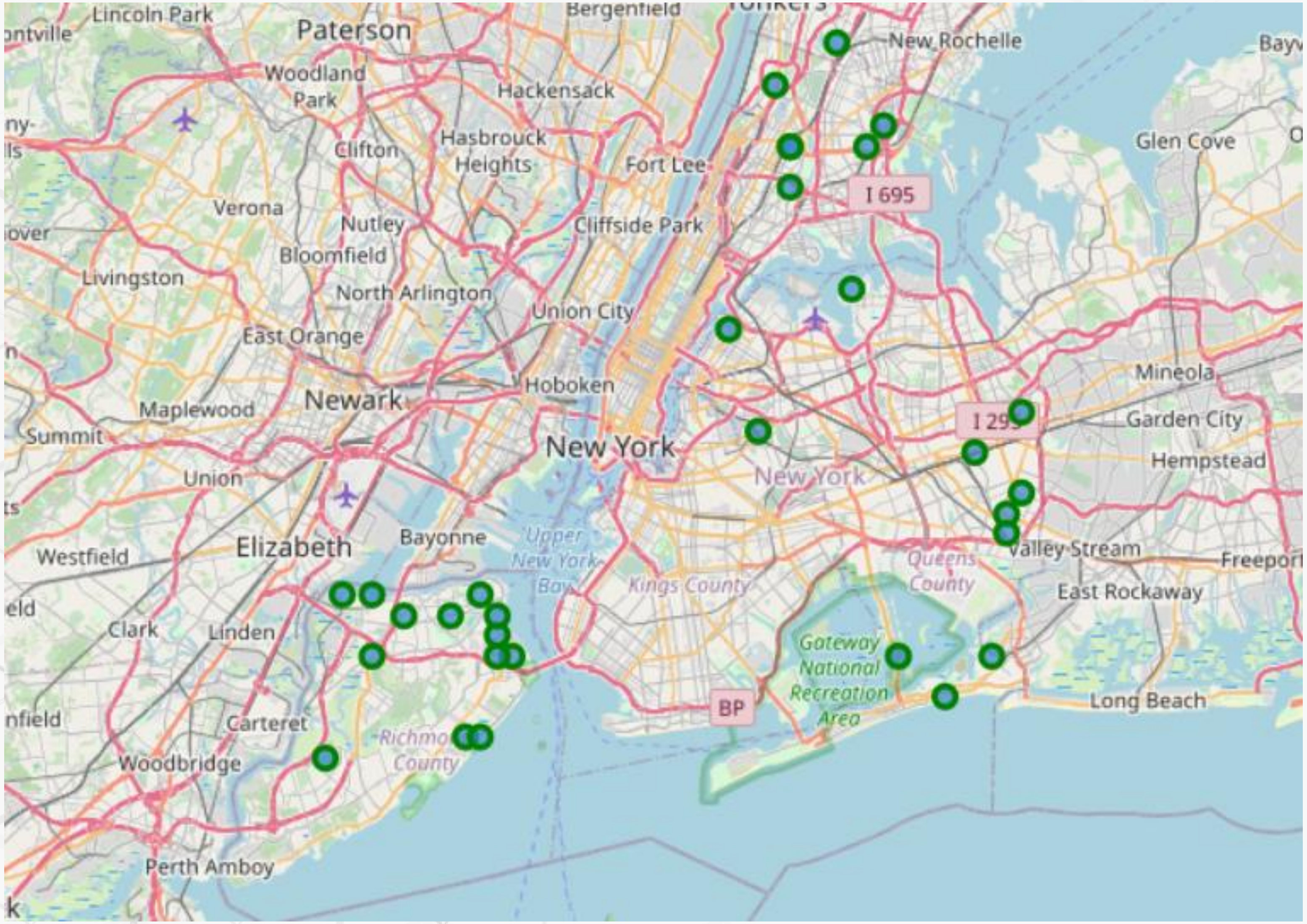
			Land Square Feet	Sale Price
Borough	Neighborhood	Zip Code		
Brooklyn	East New York	0	2000	25000
Manhattan	Upper East Side (59-79)	10075	1443	52500
	Midtown West	10019	7532	68261
Brooklyn	Coney Island	0	3502	100000
Queens	Jamaica Bay	11692	4247	110000
		11422	2500	135000
Brooklyn	Downtown-Fulton Mall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

Final data frame that we get from merging our housing cost data frame with the neighbourhood data frame. This table is used for our modelling 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

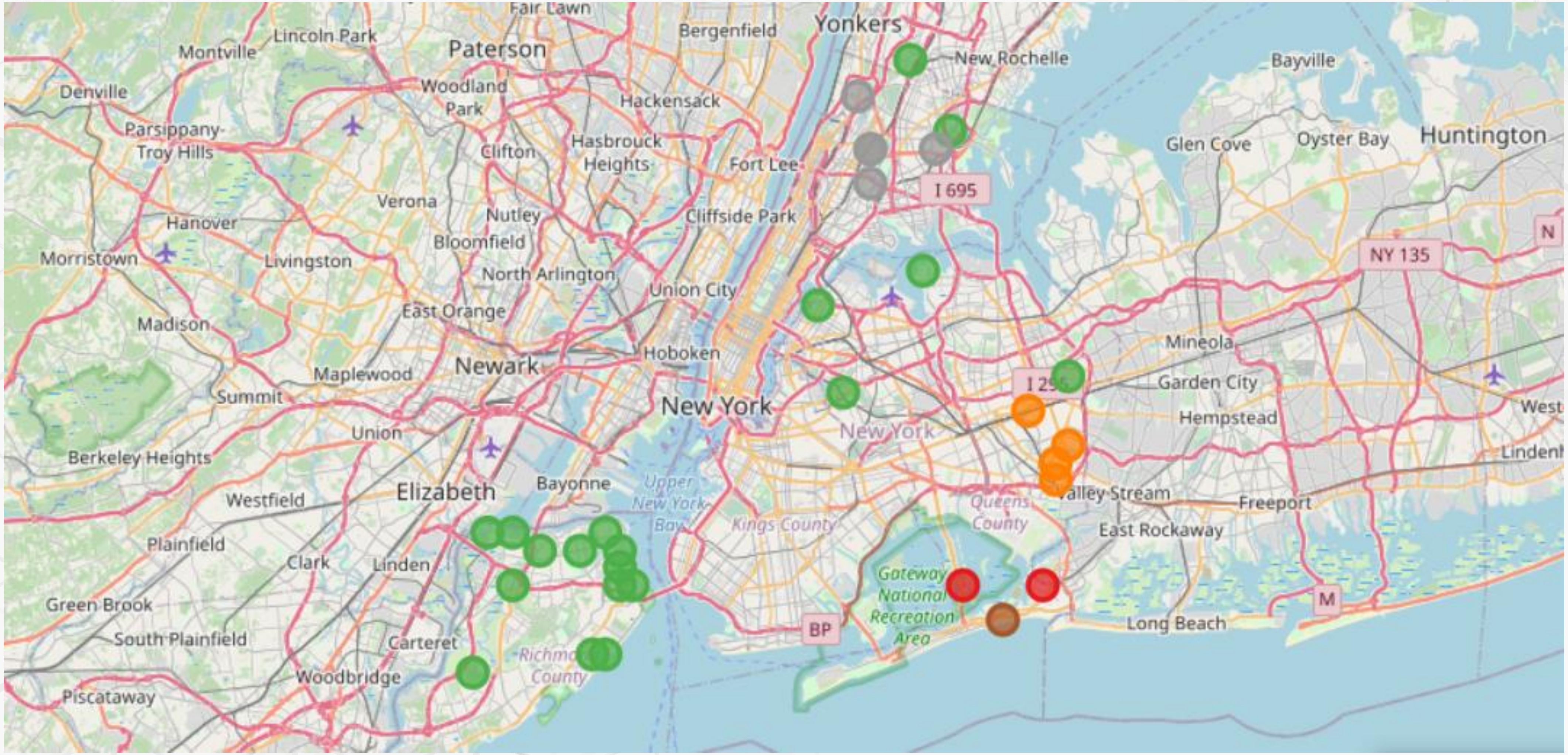
Visualizing our neighbourhoods



Modeling - sklearn kmeans

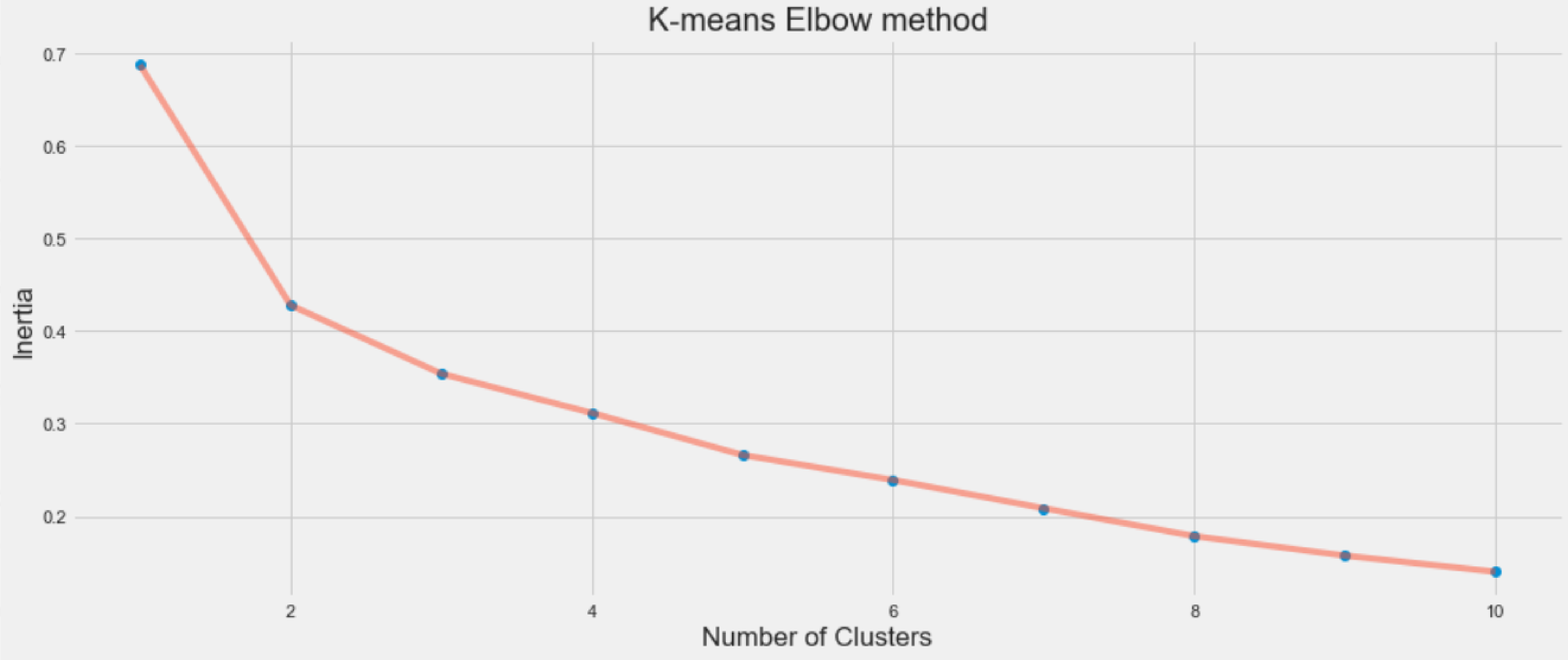
- Analysis of Clusters K=5
- There is a group that has only 1 neighbourhood and shows venues that exist along coastlines, with other minor subtle differences with groups four and two
- There is a group that is the largest cluster group of the lot and pizza and Italian places frequent a lot on this group, since this is the largest cluster we can see than there's a wide array of venues in them
- There is a group that features a Caribbean restaurant, also features venues that are common across coastlines.
- There is a group that shows us that there's a similarity in terms of Italian Restaurants, Parks and Deli / Bodega places
- There is a group that is relatively small due to it being close to the sea line and this proximity gives us unique venues such as surf spots and coffee shops, that aren't found in other cluster groups

Visualization of the five clusters, k-means where k=5

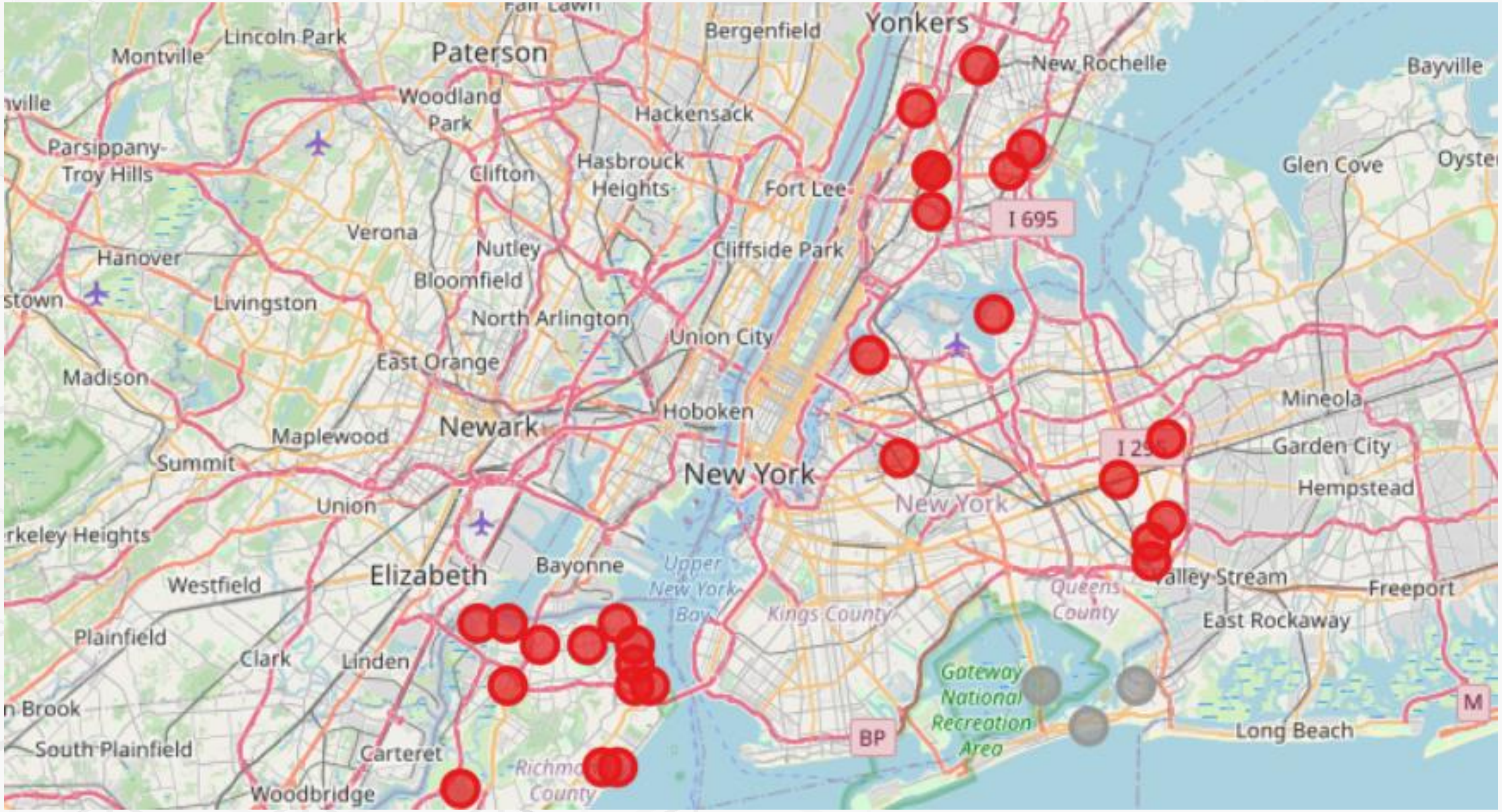


The clusters are then viewed to identify the similarities in the rank of venues, we then apply the elbow method to determine the right number of clusters using the mean squared errors on the dataset.

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



Visualization of the two clusters K-means, where k=2



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 New York City

Key Findings

- Crimes in the city of New York are classified into three main categories- Felonies, misdemeanours 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 misdemeanours, 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 New York.

- 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.