SELECTING THE BEST NEIGHBORHOOD TO LIVE IN NEW YORK

Author: Thahseen AG

INTRODUCTION

The Bronx, Brooklyn, Manhattan, Queens, and Staten Island are the five boroughs of New York, which are located where the Hudson River meets the Atlantic Ocean. For each of the five fundamental constituent parts of the newly consolidated city, the word borough was adopted to define a specific form of governmental administration. Manhattan, a densely populated borough that is part of the country, is at its heart.A densely populated borough that is home to some of the world's most significant commercial, environmental, and cultural hubs The Empire State Building and Central Park are two of the city's most prominent landmarks.A place where people live and communicate with each other is a neighborhood. Neighborhoods tend to have their own reputation based on the people who live there and the surrounding locations, or "feel". In terms of big cities, neighborhoods are typically listed, but suburban or rural areas often have neighborhoods.

BUSINESS PROBLEM

The objective is to select a neighborhood in New York with a low crime rate and having the amenities like Grocery Store,Restaurant,Gym,Pharmacy nearby.

DATA

Data that might contribute to selecting a neighborhood include the overall crime that has occured in the neighborhood for the past one year and an amenity score which determines desired amenities near to the neighborhood.

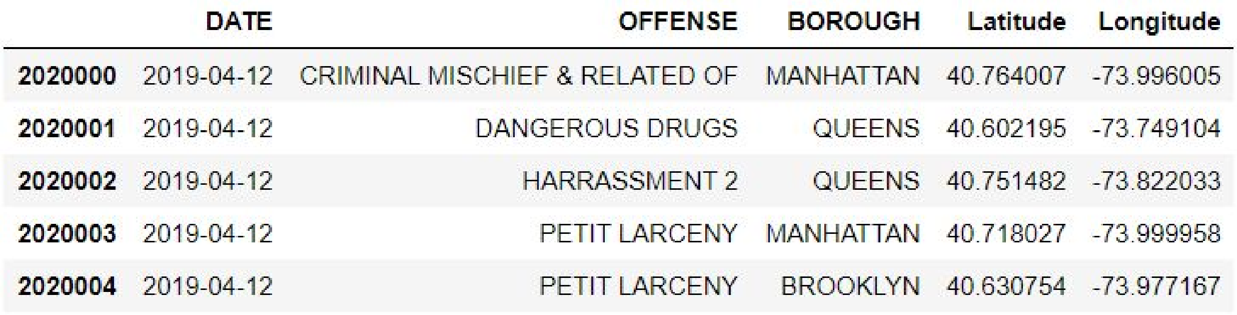
Data Sources and Data Wrangling

The crime reported in each borough will be taken from the NYPD Open Data available.NYPD OPen Data provides all the crimes that has been reported along with various details like offence description,Victim Age and sex, Borough name and most importantly the latitude and laitude of where the crime has been reported. To complement the above data, we will be using a Borough to Neighborhood mapping data set along with the Latitude and Longitude of each Neighborhood, available in https://cocl.us/new_york_dataset. The above datas is combined with the Foursquare data to 'score' each neighborhood on the basis of the availability of desired amenities. Foursquare data provides all the venues nearby given the latitude and longitude. Each venue type is listed under Venue Category.

The Data downloaded from NYPD OPen Data had many undesired columns which were not required for further processing. Only the below features are used for crime processing and the columns are renamed to a more readable name format for easy understandability.The resulting data set has 1000 samples with 5 features to the data.

Column Name	**Renamed Column Name**	**Description**
BORO_NM	Borough	The name of the borough in which the incident occurred
CMPLNT_FR_DT	Date	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
OFNS_DESC	Offence	Description of offense corresponding with key code
Latitude	Latitude	Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude	Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

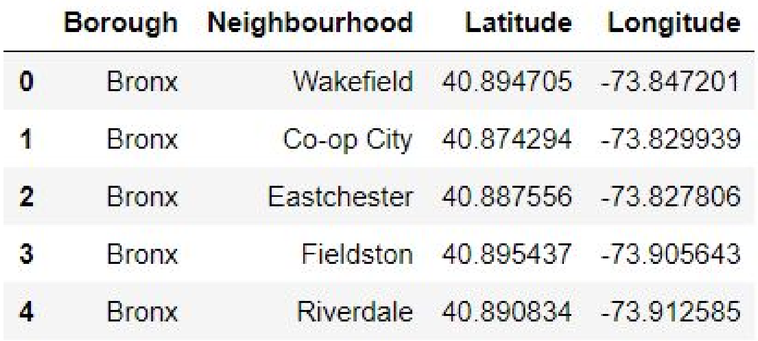
Table 1. New York Crime Data Schema



	DATE	OFFENSE	BOROUGH	Latitude	Longitude
2020000	2019-04-12	CRIMINAL MISCHIEF & RELATED OF	MANHATTAN	40.764007	-73.996005
2020001	2019-04-12	DANGEROUS DRUGS	QUEENS	40.602195	-73.749104
2020002	2019-04-12	HARRASSMENT 2	QUEENS	40.751482	-73.822033
2020003	2019-04-12	PETIT LARCENY	MANHATTAN	40.718027	-73.999958
2020004	2019-04-12	PETIT LARCENY	BROOKLYN	40.630754	-73.977167

Fig 1. New York Crime Data Set

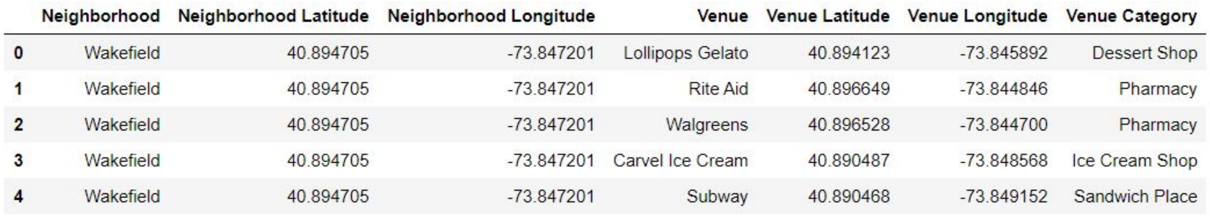
The data from the New York Data set include each borough with its neighborhoods and Latitude and Longitude.



	Borough	Neighbourhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Fig 2. Neighborhood Latitude and Longitude Data Set

Use the foursquare API to obtain the venues near the neighborhood.Each neighbourhood is listed with Neighborhood Latitude and Longitude,Venue Name,Category and Venue Latitude and Longitude.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield	40.894705	-73.847201	Subway	40.890468	-73.849152	Sandwich Place

Fig 3. Venue Details for Neighborhood

METHODOLOGY

The data obtained from the NYPD is combined with the New York Data set to determine which neighborhood the crime occured.

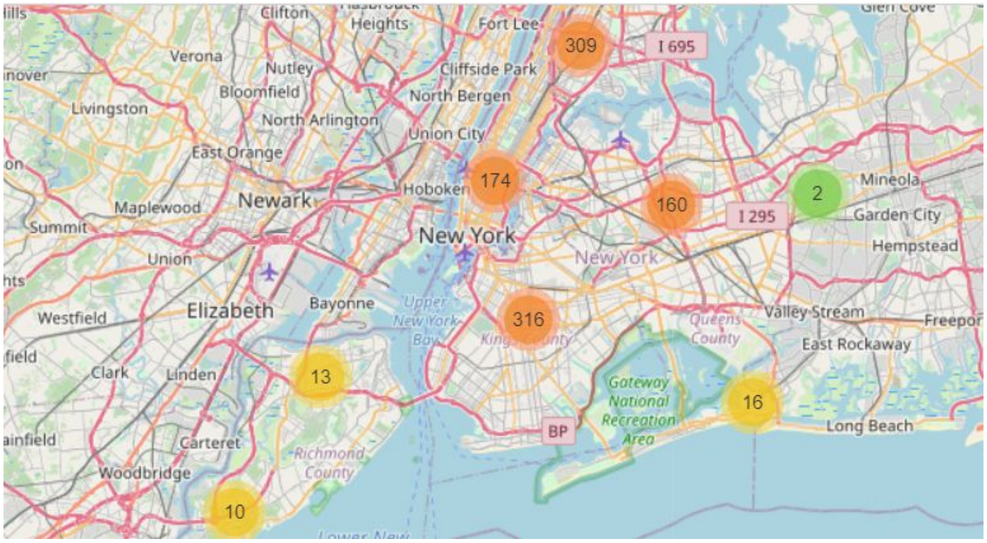


Fig 4. New York Map visualised with Crime Latitude and Longitude

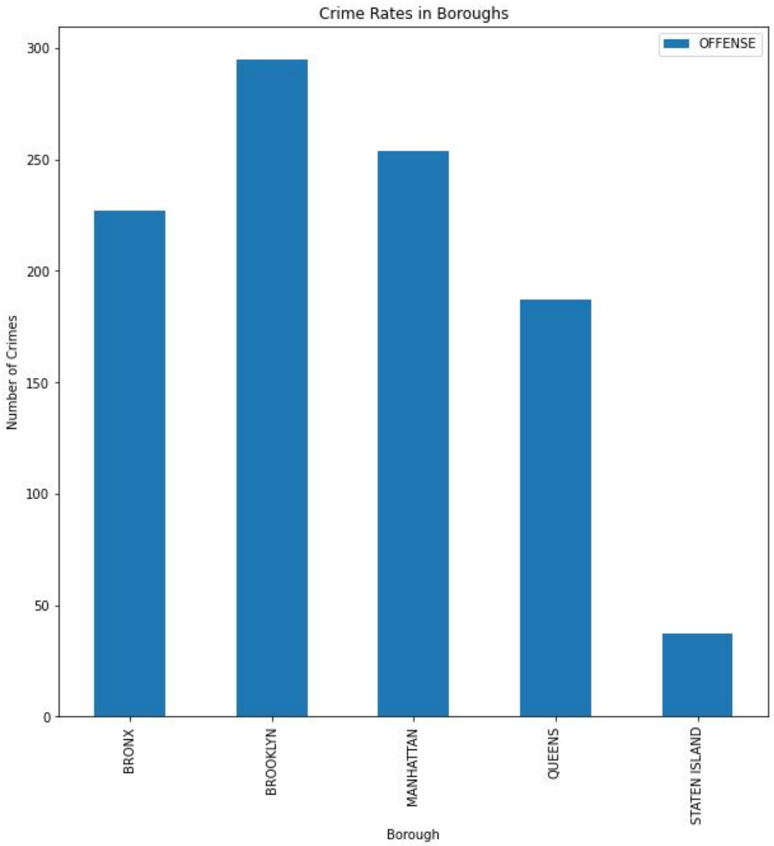


Fig 5. Borough Crime Rates

This is done with the Haversine formula, which determines the great-circle distance between two points on a sphere given their longitudes and latitudes.

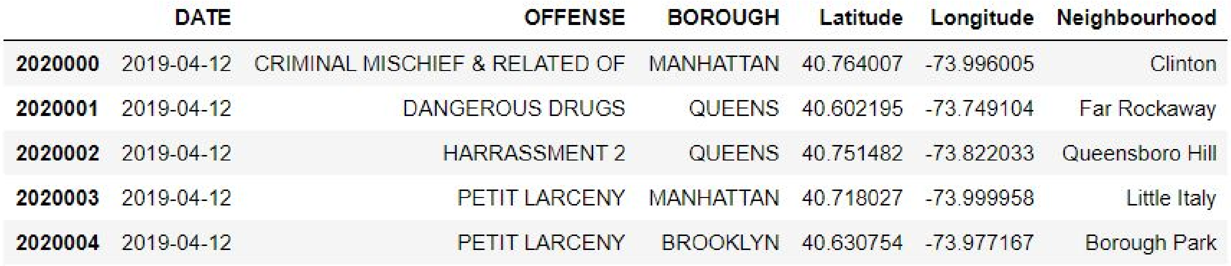
$$d = 2r\ arcsin(\sqrt{sin^2(\frac{\varphi_2-\varphi_1}{2}) \ + cos(\varphi_1)cos(\varphi_2)sin^2(\ \frac{\lambda_2-\lambda_1}{2})}\ )$$

where,
φ_1, φ_2 are the latitude of point 1 and latitude of point 2 (in radians),
λ_1, λ_2 are the longitude of point 1 and longitude of point 2 (in radians).

d is the distance between the two points along a great circle of the sphere (see spherical distance),

r is the radius of the sphere.

So given the Neighborhood Latitude and Longitude and Crime Latitude and Longitude you can use the Haversine formula to find the distance between the 2 points. This can be used to find the nearest neighborhood in which the crime occured. Once the nearest neighborhood to each crime is obtained, the crime data set is merged with the neighborhood details for each crime.



	DATE	OFFENSE	BOROUGH	Latitude	Longitude	Neighbourhood
2020000	2019-04-12	CRIMINAL MISCHIEF & RELATED OF	MANHATTAN	40.764007	-73.996005	Clinton
2020001	2019-04-12	DANGEROUS DRUGS	QUEENS	40.602195	-73.749104	Far Rockaway
2020002	2019-04-12	HARRASSMENT 2	QUEENS	40.751482	-73.822033	Queensboro Hill
2020003	2019-04-12	PETIT LARCENY	MANHATTAN	40.718027	-73.999958	Little Italy
2020004	2019-04-12	PETIT LARCENY	BROOKLYN	40.630754	-73.977167	Borough Park

Fig 6. Neighborhood Crime Details Data Set

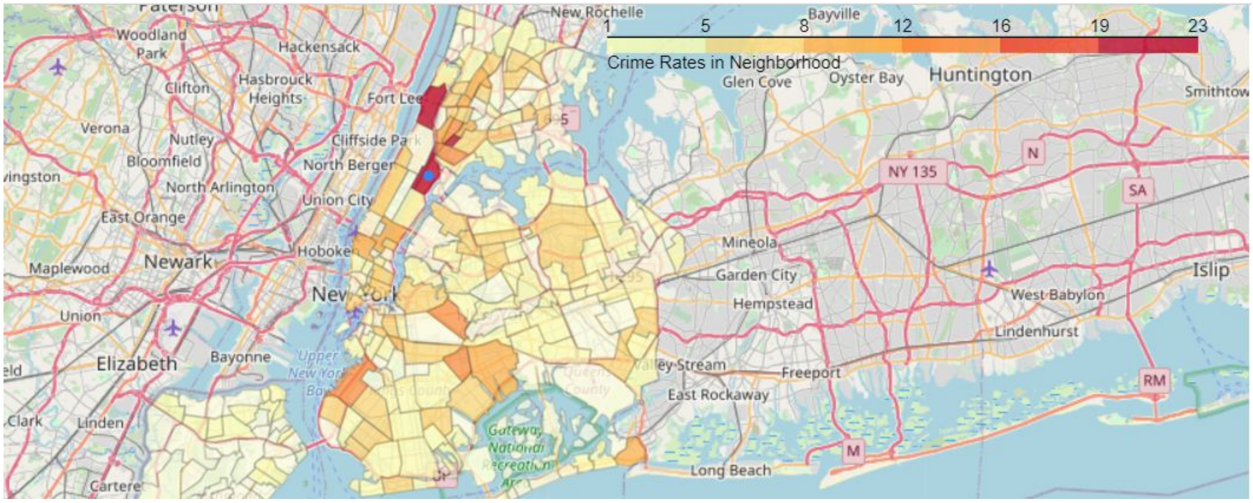


Fig 7. Neighborhood Crime Rates

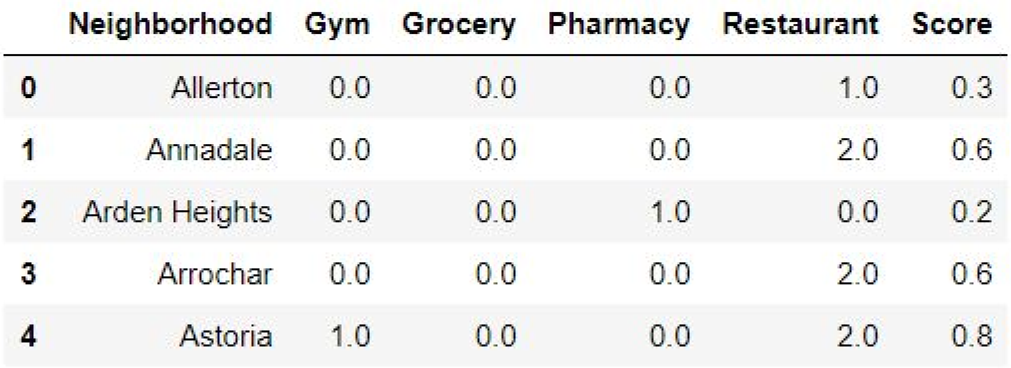
Using Foursquare api all the venues in each neighborhood are collected. For better categorization the Venue Category given as eg: Asian Restaurant or Indian Restaurant are renamed to just 'Restaurants'. The desired neighborhood is selected on the basis of different 'amenities' or required venues nearby. The user/client requires 4 main venues nearby in their selected neighborhoods ie. Grocery Store, Gym, Restaurant, Pharmacy. To 'score' each location based on the amenities/venues nearby a scoring mechanism is used. The 'score' for desired amenities is calculated by the weighted average. Weighted average is a calculation that takes into account the varying degrees of importance of the numbers in a data set. In calculating a weighted average, each number in the data set is multiplied by a predetermined weight before the final calculation is made.

$$Weighted\ Average = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Amenties	Score
Grocery Store	3
Restaurant	3
Gym	2
Pharmacy	2

Table 2. Amenities Score

Each of the categories is then grouped with only the desired amenities and is then scored with a weighted average score to determine desirability of the neighborhood.



	Neighborhood	Gym	Grocery	Pharmacy	Restaurant	Score
0	Allerton	0.0	0.0	0.0	1.0	0.3
1	Annadale	0.0	0.0	0.0	2.0	0.6
2	Arden Heights	0.0	0.0	1.0	0.0	0.2
3	Arrochar	0.0	0.0	0.0	2.0	0.6
4	Astoria	1.0	0.0	0.0	2.0	0.8

Fig 8. Amenities Score using Average Weight

This Amenities score data set is then combined with the crime data set so that we can find the best desirable neighborhood. This is done using the Machine KMeans Clustering Method.

KMEANS CLUSTERING

KMeans takes data points as inputs and groups them into k clusters, which is non overlapping. The less variation we have within clusters, the more homogeneous/similar the data points are in the cluster. We have used the KMeans clustering algorithm to find a neighborhood with the least amount of crimes and desired amenities/venues.

Selecting the best K

In this project we have used 2 methods to find the best value of K.

Within Cluster Sum of Squared Errors(WSS) or Elbow Method:

The Squared Error for each point is the square of the distance of the point from its predicted cluster center.WSS is the sum of all the Square Error for each cluster. We choose the k for which the WSS first starts to diminish.

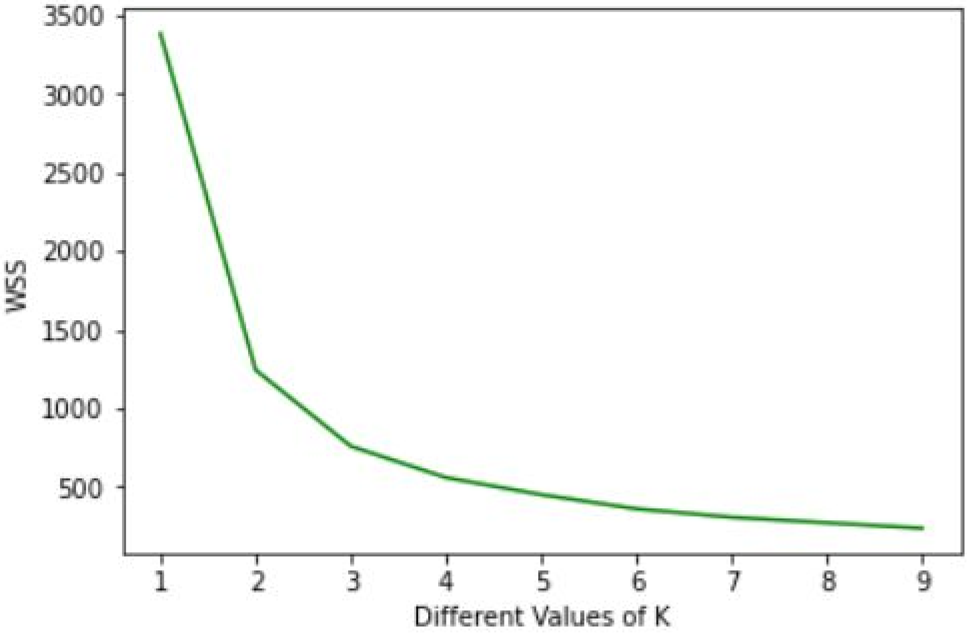


Fig 9. Plotting WSS vs K

From the above figure we can see that the plot looks like an arm with a clear elbow at k=3.

The Silhouette Method:

The silhouette value measures how similar a point is to its own cluster compared to other clusters. The value ranges between +1 and -1. A high value is desirable and indicates that the point is in the correct cluster.

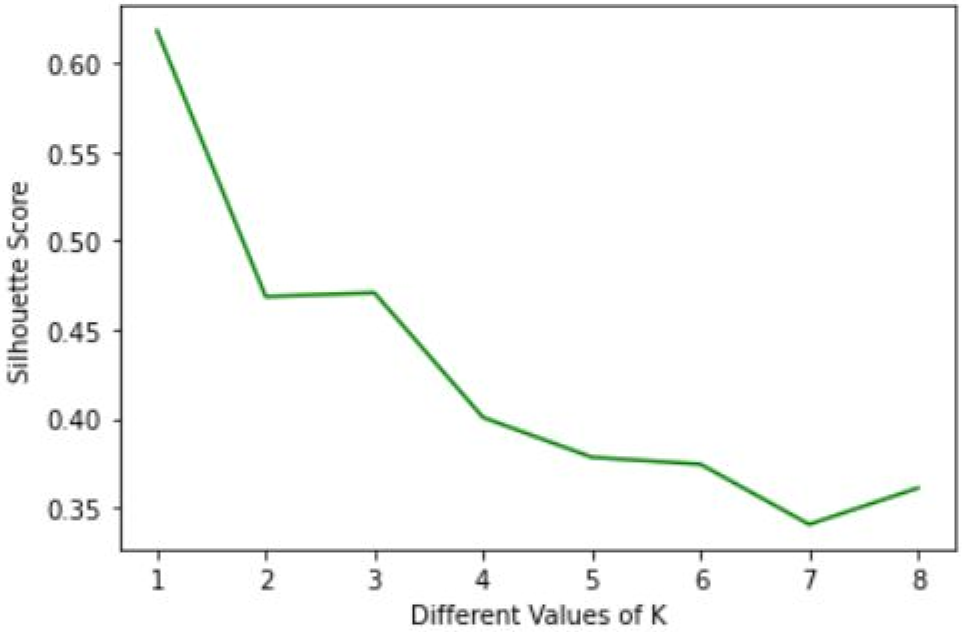
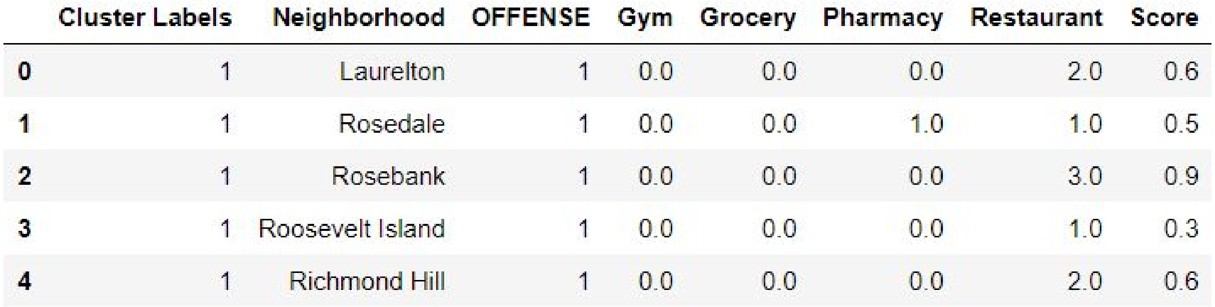


Fig 10. Silhouette score vs K

The silhouette reaches its global maximum at the optimal k ie, k=3 as there is a peak in 3. From both the methods it is determined that building the cluster with value k is going to be most useful.

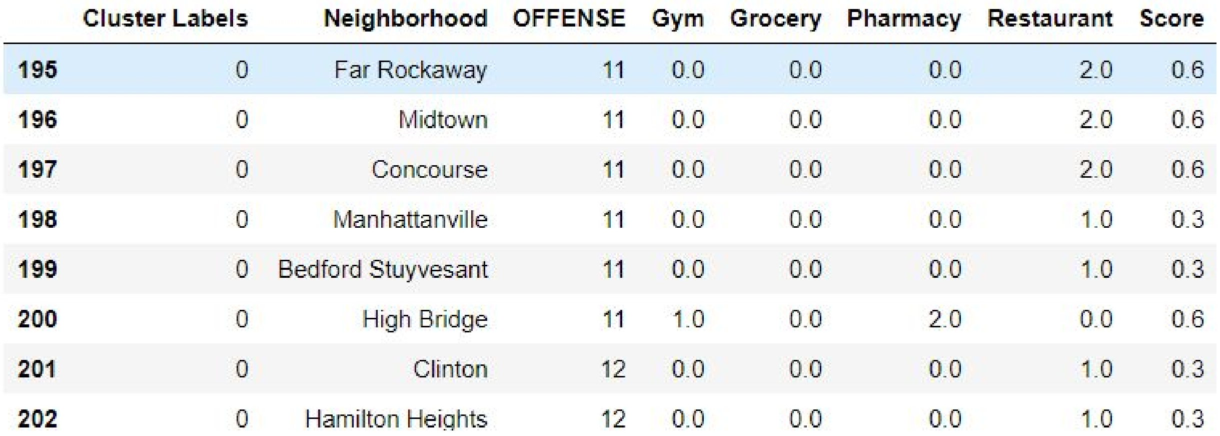
The KMeans algorithm is used with the k value equals 3. This gives us a cluster of 3 neighborhoods with similar characteristics.



	Cluster Labels	Neighborhood	OFFENSE	Gym	Grocery	Pharmacy	Restaurant	Score
0	1	Laurelton	1	0.0	0.0	0.0	2.0	0.6
1	1	Rosedale	1	0.0	0.0	1.0	1.0	0.5
2	1	Rosebank	1	0.0	0.0	0.0	3.0	0.9
3	1	Roosevelt Island	1	0.0	0.0	0.0	1.0	0.3
4	1	Richmond Hill	1	0.0	0.0	0.0	2.0	0.6

Fig 11. Cluster Labels of Each Neighborhood.

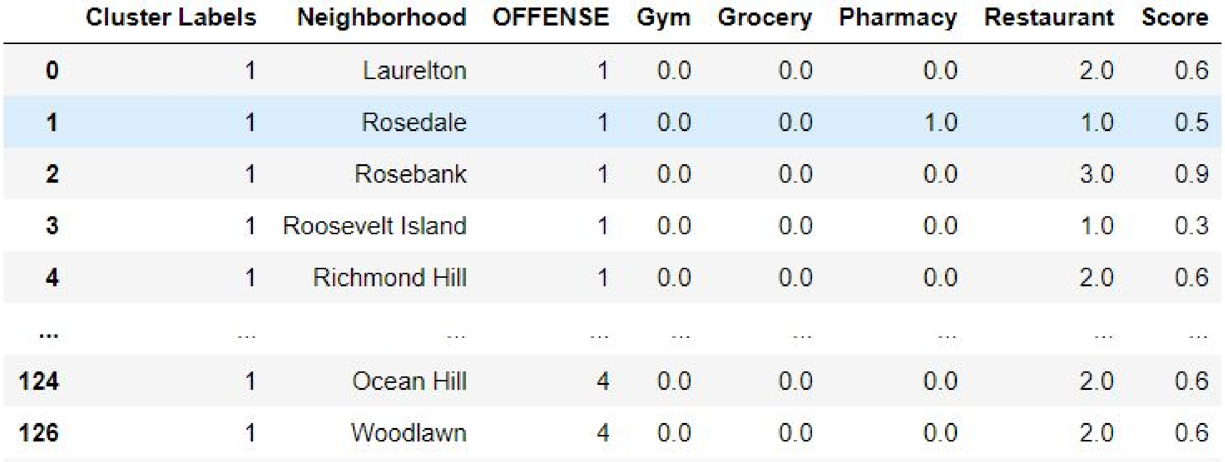
This information is combined with the neighborhood Latitude and Longitude to visualize the newly formed clusters.



	Cluster Labels	Neighborhood	OFFENSE	Gym	Grocery	Pharmacy	Restaurant	Score
195	0	Far Rockaway	11	0.0	0.0	0.0	2.0	0.6
196	0	Midtown	11	0.0	0.0	0.0	2.0	0.6
197	0	Concourse	11	0.0	0.0	0.0	2.0	0.6
198	0	Manhattanville	11	0.0	0.0	0.0	1.0	0.3
199	0	Bedford Stuyvesant	11	0.0	0.0	0.0	1.0	0.3
200	0	High Bridge	11	1.0	0.0	2.0	0.0	0.6
201	0	Clinton	12	0.0	0.0	0.0	1.0	0.3
202	0	Hamilton Heights	12	0.0	0.0	0.0	1.0	0.3

Fig 12. Cluster 0

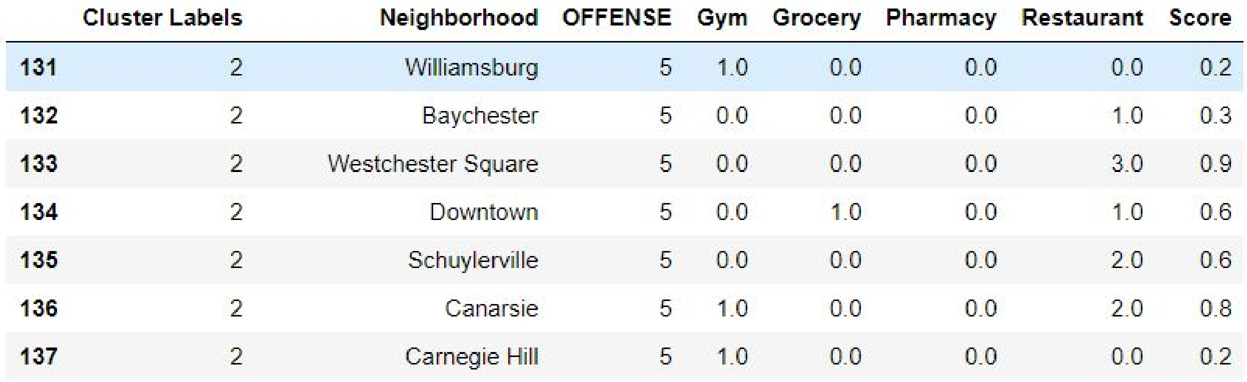
Cluster 0 has neighborhoods with high number of crimes.



	Cluster Labels	Neighborhood	OFFENSE	Gym	Grocery	Pharmacy	Restaurant	Score
0	1	Laurelton	1	0.0	0.0	0.0	2.0	0.6
1	1	Rosedale	1	0.0	0.0	1.0	1.0	0.5
2	1	Rosebank	1	0.0	0.0	0.0	3.0	0.9
3	1	Roosevelt Island	1	0.0	0.0	0.0	1.0	0.3
4	1	Richmond Hill	1	0.0	0.0	0.0	2.0	0.6
...	...	...	...	...	...	...	...	...
124	1	Ocean Hill	4	0.0	0.0	0.0	2.0	0.6
126	1	Woodlawn	4	0.0	0.0	0.0	2.0	0.6

Fig 13. Cluster 1

Cluster 1 has the neighborhoods with less number of crimes.



	Cluster Labels	Neighborhood	OFFENSE	Gym	Grocery	Pharmacy	Restaurant	Score
131	2	Williamsburg	5	1.0	0.0	0.0	0.0	0.2
132	2	Baychester	5	0.0	0.0	0.0	1.0	0.3
133	2	Westchester Square	5	0.0	0.0	0.0	3.0	0.9
134	2	Downtown	5	0.0	1.0	0.0	1.0	0.6
135	2	Schuylerville	5	0.0	0.0	0.0	2.0	0.6
136	2	Canarsie	5	1.0	0.0	0.0	2.0	0.8
137	2	Carnegie Hill	5	1.0	0.0	0.0	0.0	0.2

Fig 14. Cluster 2

Cluster 2 has neighborhoods with the moderate amount of crimes.

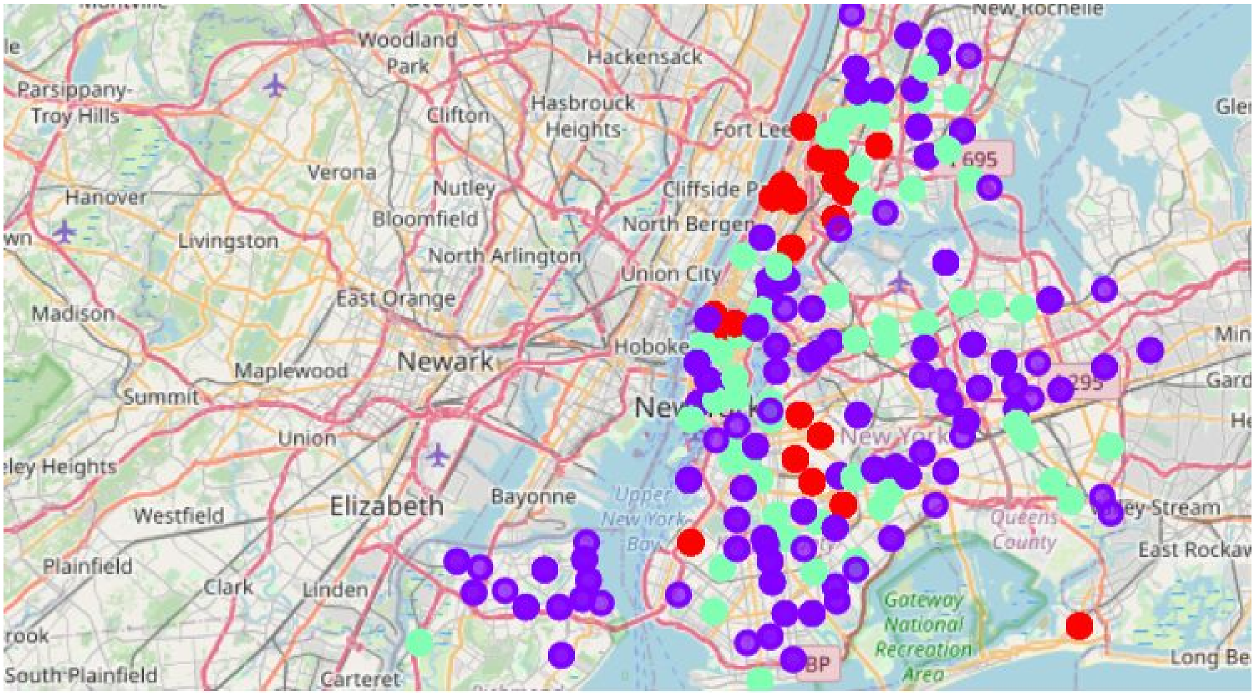
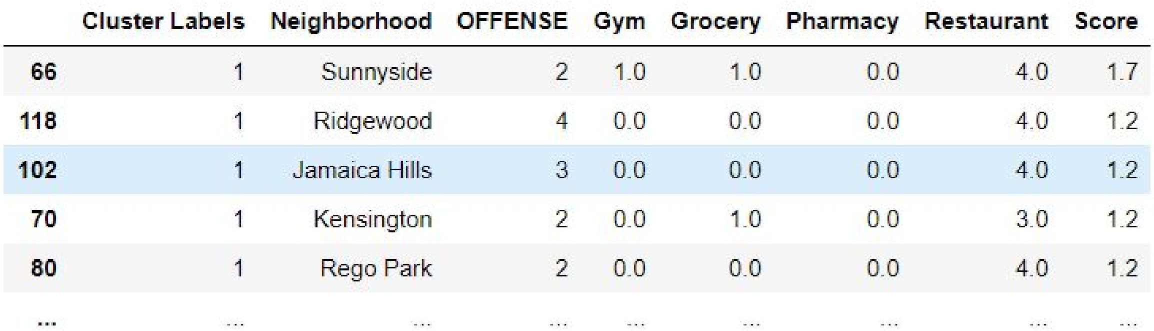


Fig 15. Neighborhood Clusters

Green indicates cluster 2, Blue is cluster 1 and Red is cluster 0.

RESULTS AND DISCUSSION

Cluster 0 has neighborhoods with high Crime Rate and Cluster 2 has neighborhoods with moderate crime rate. The careful examination of all the 3 clusters indicates that the Cluster 1 has neighborhoods with best amenities score and less amount of crime rate.

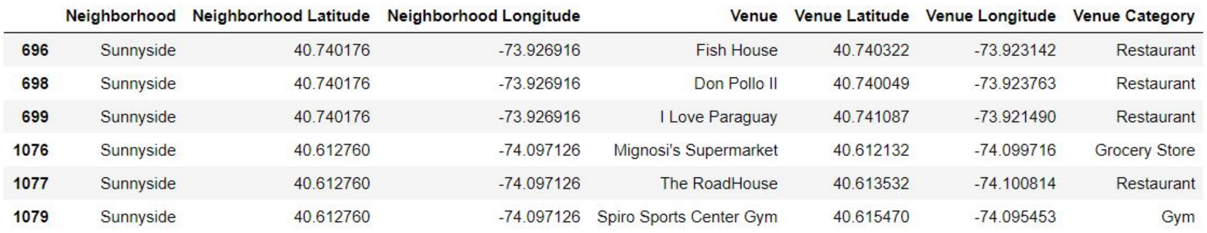


	Cluster Labels	Neighborhood	OFFENSE	Gym	Grocery	Pharmacy	Restaurant	Score
66	1	Sunnyside	2	1.0	1.0	0.0	4.0	1.7
118	1	Ridgewood	4	0.0	0.0	0.0	4.0	1.2
102	1	Jamaica Hills	3	0.0	0.0	0.0	4.0	1.2
70	1	Kensington	2	0.0	1.0	0.0	3.0	1.2
80	1	Rego Park	2	0.0	0.0	0.0	4.0	1.2
...	...	...	...	...	...	...	...	...

Fig 16. Cluster 1 with highest amenities score

CONCLUSION

Based on the above I conclude that Sunnyside is the best neighborhood to live in with having only 2 offences reported in past one year and having a amenities score of 1.7.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
696	Sunnyside	40.740176	-73.926916	Fish House	40.740322	-73.923142	Restaurant
698	Sunnyside	40.740176	-73.926916	Don Pollo II	40.740049	-73.923763	Restaurant
699	Sunnyside	40.740176	-73.926916	I Love Paraguay	40.741087	-73.921490	Restaurant
1076	Sunnyside	40.612760	-74.097126	Mignosi's Supermarket	40.612132	-74.099716	Grocery Store
1077	Sunnyside	40.612760	-74.097126	The RoadHouse	40.613532	-74.100814	Restaurant
1079	Sunnyside	40.612760	-74.097126	Spiro Sports Center Gym	40.615470	-74.095453	Gym

Fig 17. Sunnyside Neighborhood.

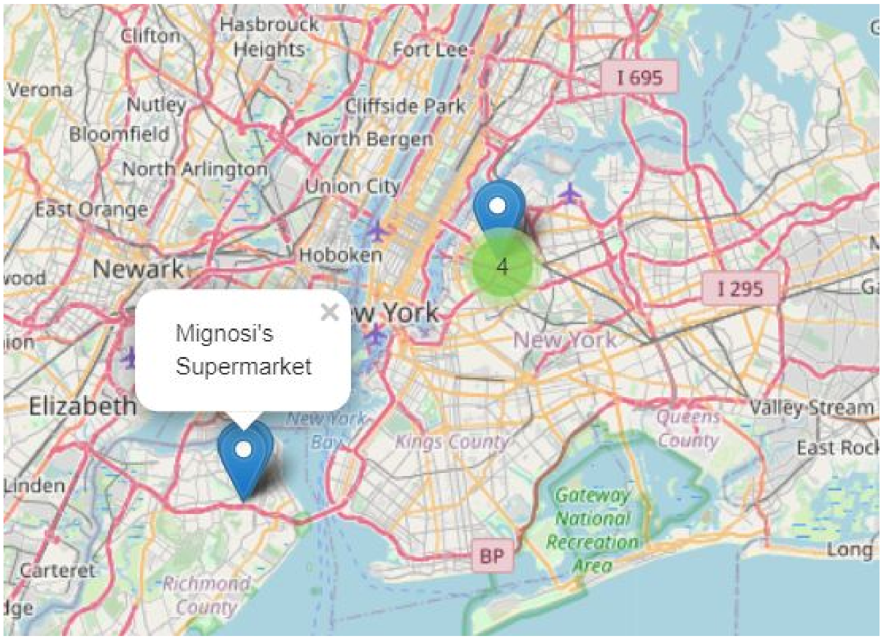


Fig 17. Sunnyside Neighborhood.