SELECTING THE BEST NEIGHBORHOOD TO LIVE IN NEW YORK

By Thahseen AG

SELECTING THE BEST NEIGHBORHOOD TO LIVE IN NEW YORK

- 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.

- The selection is done by:

    - Categorizing Neighborhoods according to the crime reported.
    - Scoring Neighborhoods based on the amenities they have.

Data Sources

- The crime reported in each borough obtained from NYPD Open Data which includes details like offence description,Victim Age and sex, Borough name and latitude and laitude of where the crime has been reported.
- A Borough to Neighborhood mapping data set along with the Latitude and Longitude of each Neighborhood, from https://cocl.us/new_york_dataset
- Foursquare API to get all the venues nearby given the latitude and longitude.

DATA WRANGLING

Only the below details from NYPD crime reported is used in this project.

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)

NYPD Crime Data Set

The resulting data set has 1000 samples with 5 features to the data.

	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

Borough to Neighborhood Data Set

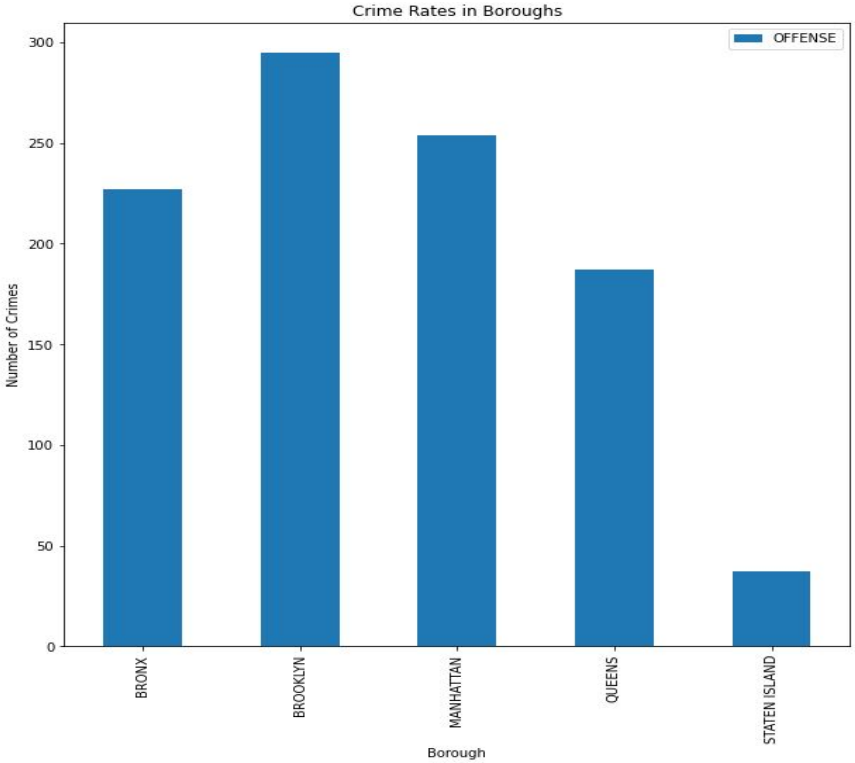
	Borough	Neighbourhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

FourSquare API for Venues in Neighborhoods

	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

New York Map with Crime Clusters

Borough Crime Rates



Neighborhood Crime Rate

- The Data obtained from the NYPD has to be combined with the Borough to Neighborhood to know which neighborhood the crime has occured.
- This achieved using the crime latitude and longitude and then finding the nearest neighborhood to it based on the neighborhoods latitude and Longitude.
- This is achieved using Haversine Formula.

Haversine Formula

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,

φ1, φ2 are the latitude of point 1 and latitude of point 2 (in radians),

λ1, λ2 are the longitude of point 1 and longitude of point 2 (in radians).

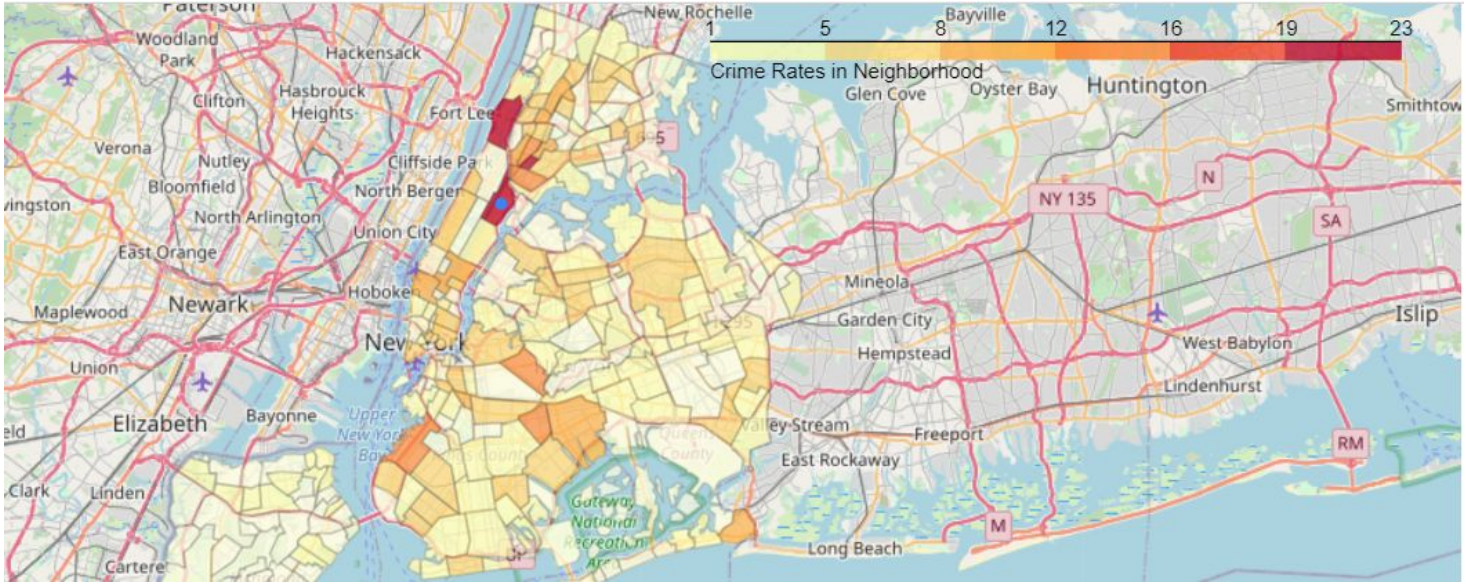
d is the distance between the two points along a great circle of the sphere (see spherical distance),

r is the radius of the sphere.

Neighborhood Crime Data

	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

Neighborhood Crime Data



Selecting The Desired Amenities

- 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.
- To ‘score’ each location based on the amenities/venues nearby a scoring mechanism is used,Weighted Average.

Weighted Average

Calculation that takes into account the varying degrees of importance of the numbers in a data set.

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

Amenities Score,

Amenities	Score
Grocery Store	3
Restaurant	3
Gym	2
Pharmacy	2

Neighborhoods Venues

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

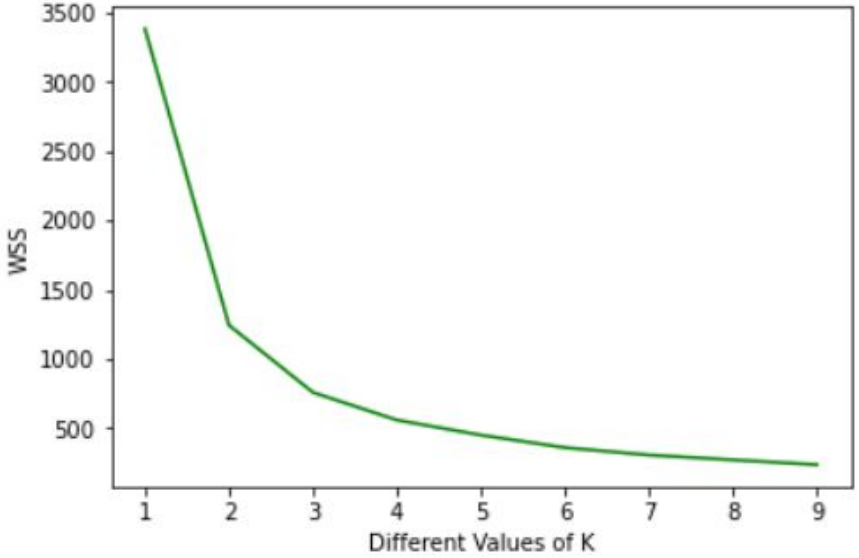
	Neighborhood	Gym	Grocery	Pharmacy	Restaurant	Score
0	Allerton	0.0	0.0	0.0	1.0	0.3
1	Annadale	0.0	0.0	0.0	2.0	0.6
2	Arden Heights	0.0	0.0	1.0	0.0	0.2
3	Arrochar	0.0	0.0	0.0	2.0	0.6
4	Astoria	1.0	0.0	0.0	2.0	0.8

KMeans Cluster

- 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

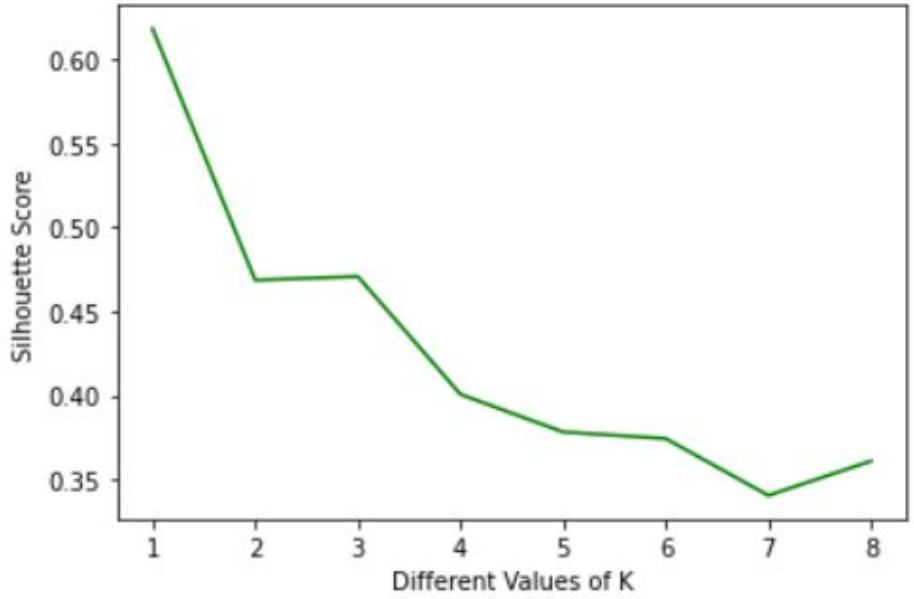
Selecting Best K-Elbow Method/WSS

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.



Selecting Best Value for K-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.



KMeans Model

- From both the method it is clear that the best value for k is 3.
- Fitting the data with k = ‘3’ gives us 3 clusters having 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

Cluster 0

High Number of Crimes.

	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

Cluster 1

Less 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

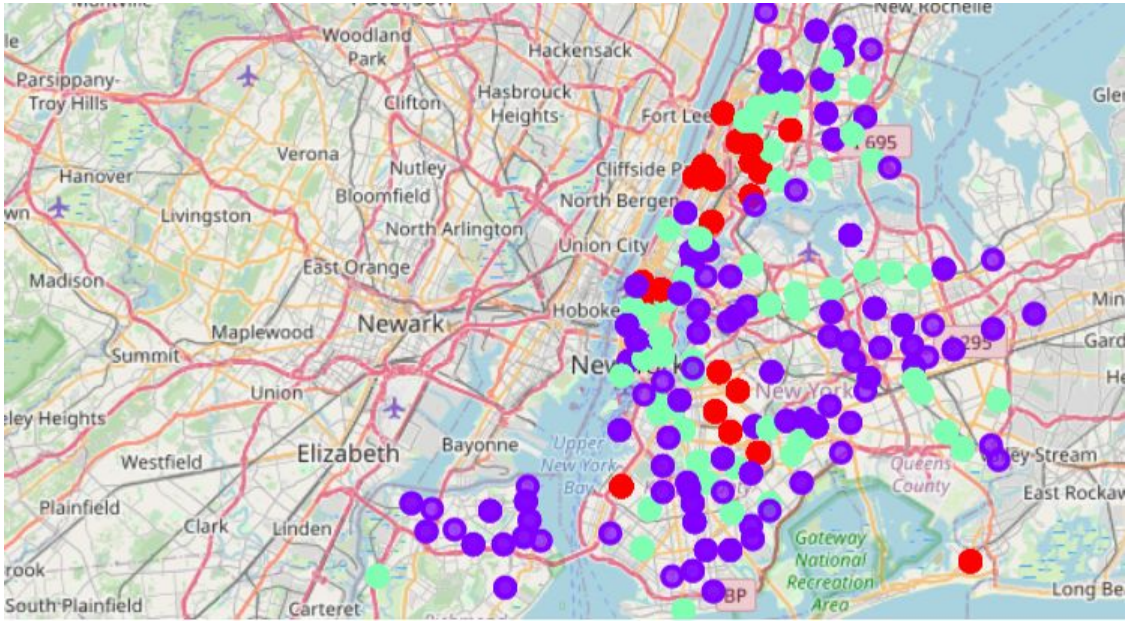
Cluster 2

Moderate 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

Cluster Map

Red-Cluster 0, Blue-Cluster 1, Green-Cluster 2



RESULT AND DISCUSSION

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
...	...	...	...	...	...	...	...	...

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

Sunnyside Neighborhood