Coursera Capstone Project - Battle of the Neighbourhoods

Report

1. Introduction

Selecting where to live when moving to a new city is one of the most daunting tasks faced by anyone, especially when one has limited knowledge of the city and its various localities. This project aims to make it easier for a new migrant into a city to quickly understand the different types of neighbourhoods in the city, and help them in making a choice of where to live based on the neighbourhood features and their personal preferences.

The neighbourhoods were clustered into groups based on similarities/ dissimilarities in their profiles. In order to cluster neighbourhoods, the following metrics were selected:

a. Average house rent
b. Crime rate
c. Amenities such as restaurants, grocery stores, shopping centres, etc.

While there are many other features such as connectivity, quality of schools, etc. that may affect a person's choice of residence, the above three factors were ultimately selected because of the ease of gathering the relevant data, and also because of the fact that these factors will be common considerations for migrants across demographics and age groups.

New York City was picked as the city of choice for this project because of its metropolitan nature and due to the large number of people that move to the city every year for work, from all over the world. In addition, it was determined that the data points for each of the metrics described above are more easily available for NYC.

2. Data

The data on the various neighbourhoods and boroughs of New York City was picked from https://geo.nyu.edu/catalog/nyu_2451_34572. This data includes latitudinal and longitudinal coordinates for each neighbourhood in NYC, and was used to visualise the city's neighbourhoods and their clusters.

A snapshot of the dataset for neighbourhoods of NYC is given below:

	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

The data on the three metrics selected for clustering neighbourhoods will be obtained as follows:

a. **Average house rent:** Renthop.com has data on the average house rent for Studio, 1 BHK and 2 BHK apartments by neighbourhood for NYC. The dataset can be found at https://renthop.com/average-rent-in/ney-york-city-ny. However, the dataset is incomplete with rent rates for Studios and 2 BHK apartments in some neighbourhoods missing. Hence, we will be using only the average rent for 1 BHK apartments in each neighbourhood for this metric.

   A snapshot of the resulting dataset for average rent is below:

	Borough	Neighbourhood	Latitude	Longitude	Average_Rent
0	Brooklyn	Bay Ridge	40.625801	-74.030621	1800
1	Brooklyn	Williamsburg	40.707144	-73.958115	3250
2	Brooklyn	Bushwick	40.698116	-73.925258	2391
3	Brooklyn	Carroll Gardens	40.680540	-73.994654	2400
4	Brooklyn	Gowanus	40.673931	-73.994441	3198

b. **Crime rate:** Niche.com (https://www.niche.com) hosts data and ratings on places to live, schools, colleges, and places to work across the US. As part of its ratings, the website features crime ratings for states, cities and localities in the USA, including several New York City neighbourhoods. The crime ratings on the website have been developed through a comprehensive study of open data sources such as the Uniformed Crime Report published by the FBI, as well as local surveys conducted by Niche. Each locality on Niche is given a letter grade for crime and safety based on crime rates for murder, assault, rape, burglary, and other crime statistics, as well as reviews from residents. For the purposes of this project, the crime ratings, where available, for neighbourhoods in NYC were converted from their

letter grades to a numeric scale in the range 1-4, where higher crime rating denotes a safer neighbourhood.

A snapshot of the resulting dataset for the crime rate is below:

	Neighbourhood	Crime Grade	Numeric Crime Grade
0	Bay Ridge	C+	2.0
1	Williamsburg	C+	2.0
2	Bushwick	C	1.0
3	Carroll Gardens	B-	3.0
4	Gowanus	B-	3.0

c. **Amenities:** Foursquare API will be used to explore the venues in each neighbourhood and determine the frequency of occurrence of different categories of venues for each neighbourhood.

A snapshot of the resulting dataset with frequency of occurrence for different categories of venues for each neighbourhood is given below:

	Neighbourhood	Accessories Store	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	...	Video Store	Vietnamese Restaurant	Volleyball Court	W
0	Astoria	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.00	0.000000	0.0	
1	Bay Ridge	0.0	0.035714	0.000000	0.0	0.0	0.000000	0.0	0.011905	0.0	...	0.00	0.011905	0.0	
2	Boerum Hill	0.0	0.011236	0.011236	0.0	0.0	0.011236	0.0	0.022472	0.0	...	0.00	0.000000	0.0	
3	Bushwick	0.0	0.013699	0.000000	0.0	0.0	0.013699	0.0	0.000000	0.0	...	0.00	0.000000	0.0	
4	Carroll Gardens	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.0	0.010000	0.0	...	0.01	0.000000	0.0	

Using the frequency of occurrence of categories of venues, the top 10 venue categories for each neighbourhood were determined. A snapshot of the resulting dataset for top 10 venues in each neighbourhood is given below:

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Astoria	Bar	Middle Eastern Restaurant	Greek Restaurant	Hookah Bar	Seafood Restaurant	Mediterranean Restaurant	Bakery	Food Truck	Salon / Barbershop	Pub
1	Bay Ridge	Italian Restaurant	Spa	Pizza Place	American Restaurant	Bar	Greek Restaurant	Diner	Hookah Bar	Playground	Sushi Restaurant
2	Boerum Hill	Coffee Shop	Dance Studio	Sandwich Place	Bar	French Restaurant	Deli / Bodega	Spa	Furniture / Home Store	Middle Eastern Restaurant	Martial Arts Dojo
3	Bushwick	Bar	Coffee Shop	Mexican Restaurant	Pizza Place	Deli / Bodega	Discount Store	Thrift / Vintage Store	Bakery	Italian Restaurant	Vegetarian / Vegan Restaurant
4	Carroll Gardens	Italian Restaurant	Coffee Shop	Cocktail Bar	Pizza Place	Gym / Fitness Center	Wine Shop	Bar	Bakery	Spa	Thai Restaurant

The final dataset, with data available for each of the above variables (ie: Latitude and Longitude, Average Rent, Numeric Crime Grade, and frequency of occurrence of categories of nearby venues) for each neighbourhood, was then identified. All required data was available for a total of 32 neighbourhoods in NYC. This is the dataset that was used for further analysis and clustering.

A snapshot of this compiled dataset is as given below:

3. Methodology

The compiled dataset given above was then preprocessed for clustering, with any columns not required having been dropped. The MinMaxScaler was used for scaling, since this method would standardise the data without drastically changing its attributes and features, and would not hide any outliers.

The data available for clustering post dropping of columns and preprocessing is given below:

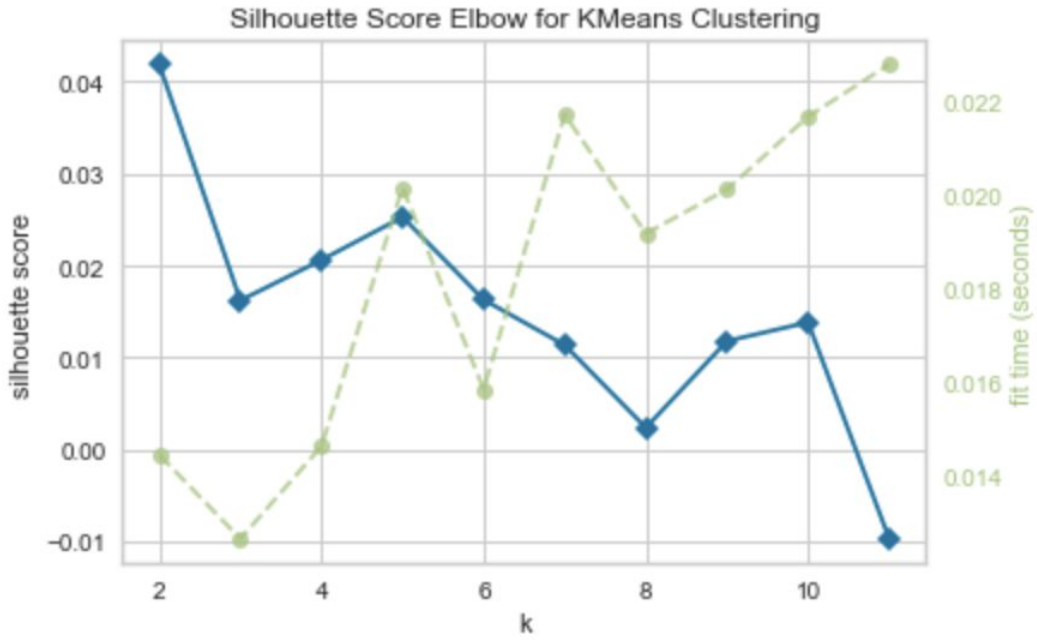
```
array([[0.03296703, 0.33333333, 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.67032967, 0.33333333, 0.        , ..., 0.        , 0.        ,
        0.95959596],
       [0.29274725, 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.2967033 , 0.66666667, 0.        , ..., 1.        , 0.59      ,
        0.        ],
       [0.64747253, 0.66666667, 0.        , ..., 0.        , 0.        ,
        0.4589372 ]])
```

In order to cluster the selected neighbourhoods in NYC, an unsupervised learning algorithm would need to be selected, such that like neighbourhoods are clustered together based on specific similarities. K-means clustering, a popular unsupervised learning algorithm was selected for the purpose of clustering, due to its relatively ease of implementation, computational speed, and guaranteed convergence.

Before using K-means clustering, the optimal value of k (number of clusters) needs to be determined. The elbow method was used for this purpose, with the silhouette value being calculated over a range of values for k. The silhouette value provides a measure of how close

each object is to its own cluster (cohesion) compared to other clusters (separation). The highest silhouette value will correspond to the optimal value of k.

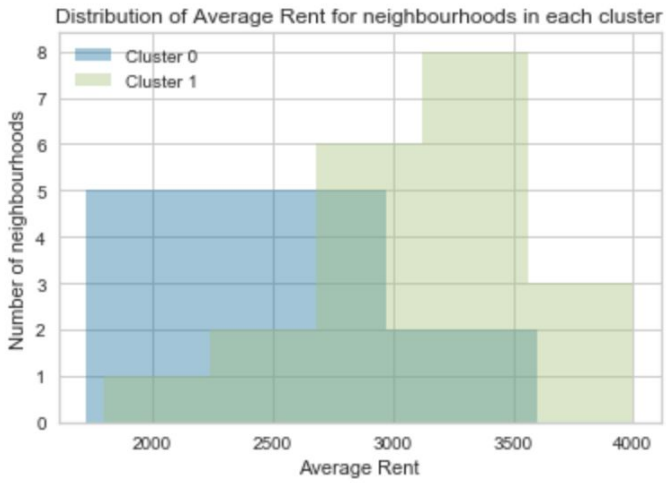
From the below visualization of silhouette values for different values of k, it was determined that the optimal number of clusters for the selected neighbourhoods in NYC is 2.



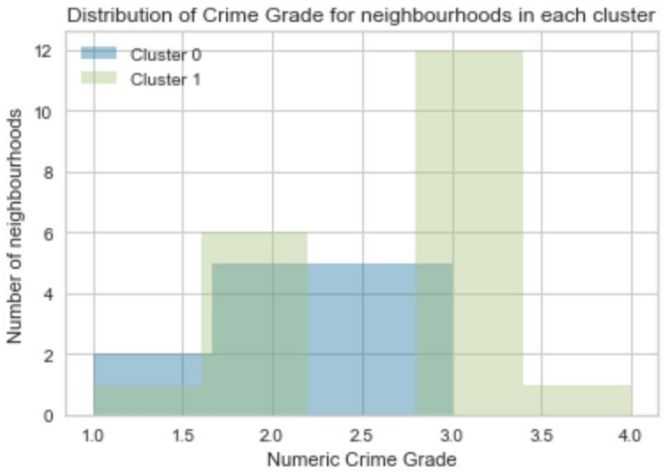
Thus, K-means clustering was performed on the selected NYC neighbourhoods with number of clusters set at 2.

The 2 clusters so formed were then analysed for differences and defining characteristics so as to identify two distinctive groups of neighbourhoods in NYC.

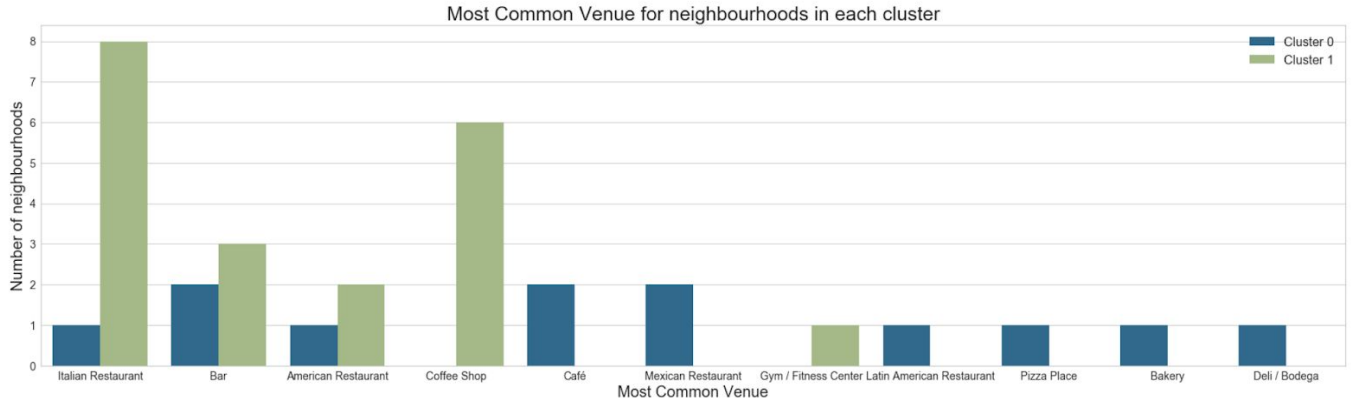
Exploratory analysis of the average rent for each neighbourhood based on the clusters formed revealed that in general, the average rent for neighbourhoods in cluster 0 tends to be lower than that for neighbourhoods in cluster 1, as shown below:



Exploratory analysis of the crime rating for each neighbourhood based on the clusters formed revealed that in general, the level of crime for neighbourhoods in cluster 0 tends to be higher than that for neighbourhoods in cluster 1, as shown below:

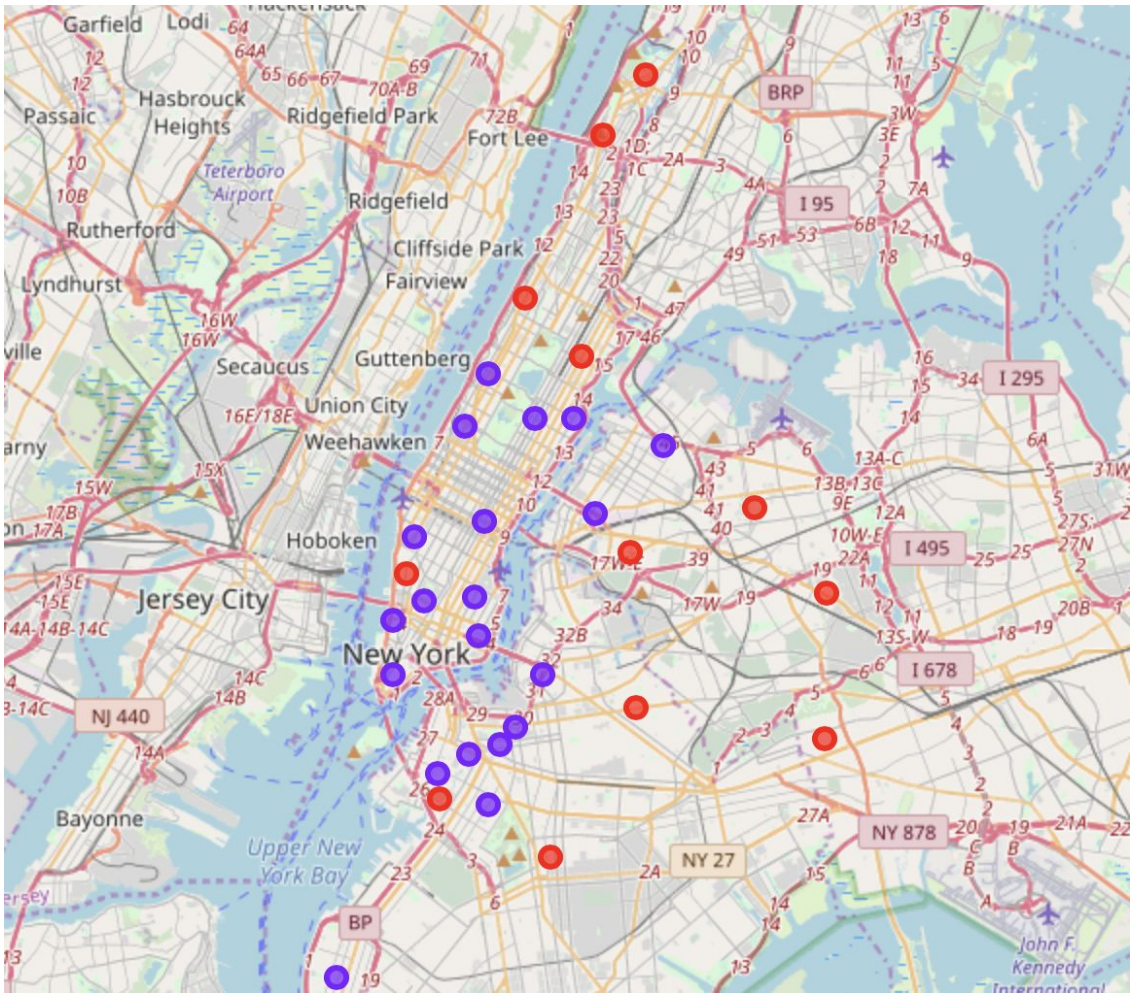


Finally, exploratory analysis of the most common venue categories for each neighbourhood based on the clusters formed revealed that the neighbourhoods in cluster 0 are known for a variety of eateries and cafes, while the neighbourhoods in cluster 1 are hotspots for Italian food and Coffee Shops. The same has been depicted in the below bar graph:



4. Results

The two resulting clusters of NYC neighbourhoods formed through K-means clustering are depicted in the visualization below. While Cluster 0 consists of 12 neighbourhoods, Cluster 1 is comprised of 20 neighbourhoods of the total 32 neighbourhoods analysed.



Further, based on the exploratory analysis performed on each cluster of neighbourhoods based on the average rent, crime grade and most common venues, each cluster can be profiled as follows:

Cluster 0 - Less safe neighbourhoods, with lower rent, and a variety of eateries and cafes

Cluster 1 - More safe neighbourhoods, with higher rent, and hotspots for italian food and coffee shops

5. Discussion

The profiles generated for each cluster as given above now provide a good starting point for anyone looking to move into New York City. Based on preferences and priorities for rent, crime level and nearby venues, a user can clearly identify appropriate neighbourhoods that can then be explored further.

For example, someone who is looking to move into the safest available neighborhoods and doesn't mind spending a little extra on rent, would be better suited to move into neighborhoods in cluster 1. On the other hand, a person who would like to have a variety of restaurant options nearby may prefer neighbourhoods in Cluster 0, despite their lower level of safety.

In general, based on user preferences, the clustering performed enables us to quickly recommend a smaller, more manageable set of neighbourhoods to migrants coming into NYC.

6. Conclusion

This project makes it easier for a new migrant into NYC to quickly understand the different types of neighbourhoods in the city, and help them in making a choice of where to live based on the neighbourhood features and their personal preferences.

While this project only makes use of 32 NYC neighbourhoods for demonstration purposes, the same can be scaled up to include all the neighbourhoods in NYC, provided the required data is available. Further, the analysis can also be enhanced to include other data points, and it can be easily replicated to provide similar analyses for other cities and countries.