Battle of Neighbourhoods

Augustine Annan

December 6, 2020

1   Introduction

New York City is a city in the United States of America state of New York. It is the most populous and densely populated city in the United States. As at 2019, the population of Dallas was 8,336,817. It is also one of the world's most populous megacities. New York City is a diverse city, the most linguistically diverse city in the world with over 800 languages spoken. As of 2016, it had the largest foreign-born population of any city in the world. [1]

New York City has an extraordinarily diverse population. It is one of the few cities in the country in which four different racial/ethnic groups each make up at least 10 percent of the population. In 2018, NYC & Company found that New York City welcomed a record 65.2 million visitors, comprising 51.6 million domestic and 13.5 million international visitors, the ninth consecutive year of tourism growth. [2]

We believe it is always very cumbersome task for a traveller or a visitor to make choices from among the many available options since there is dense web information. Information is all fragmented that you need to assemble yourself to make a better decision. This project seeks to present a guide to the popular cuisines in various neighborhoods. Also, we seek to determine the most popular cuisine in New York. There are several cuisines in New York city, from a gold-themed Italian spot to Thai food in the back of a grocery store to an outdoor patio where you can spend a whole day. This project will seek to provide a great start when you're trying to make restaurant decisions.

2   Data Description

In order to help address this concern, we required geographical location data for New York City. As a starting point, we needed the latitudes and longitudes of neighborhoods. We used this data to find restaurants within a certain radius of each neighbourhood.

2.1   New York City

New York City data containing the neighborhoods and boroughs, latitudes, and longitudes was obtained from the data source: https://cocl.us/new york dataset. The data consisted of the following variables:

- *Borough:* Name of Borough

- *Neighborhood*: Name of Neighborhood

- *Latitude:* Latitude of the neighborhood

- *Longitude :* Longitude of the neighborhood

2.2 Foursquare API Data

We needed data about restaurants in all neighborhoods of every borough. In order to gain that information we used "Foursquare" locational information. We connected to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we chose the radius to be 1000 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the neighborhoods. The information obtained per venue are as follows:

- Neighbourhood : Name of Neighbourhood

- Neighbourhood Latitude : Latitude of Neighbourhood

- Neighbourhood Longitude : Longitude of Neighbourhood

- Venue : Venue Name

- ID : ID of the Venue

- Venue Latitude : Latitude of Venue

- Venue Longitude : Longitude of Venue

- Venue Category : Category of Venue

3 Methodology

The main task was to explore each of the neighborhoods individually, plot the map to show the neighbourhoods being considered and then build our model by clustering all of the similar neighbourhoods together and finally plot the new map with the clustered neighbourhoods. We draw insights and then compare and discuss our findings.

Cluster analysis was performed on the retrieved and cleaned dataset. We used the K-means clustering algorithm to ascertain the clusters.

3.1 Data Preparation and Exploration

As part of preparing the data, we selected our important features from https://cocl.us/new york dataset.. Fortunately, our dataset contained the geographical coordinates in the form of latitude and longitude to help us use the Foursquare API. After gathering the required data, it was populated into a pandas Dataframe and then visualized the neighborhoods in a map using Folium package.

Out[10]:
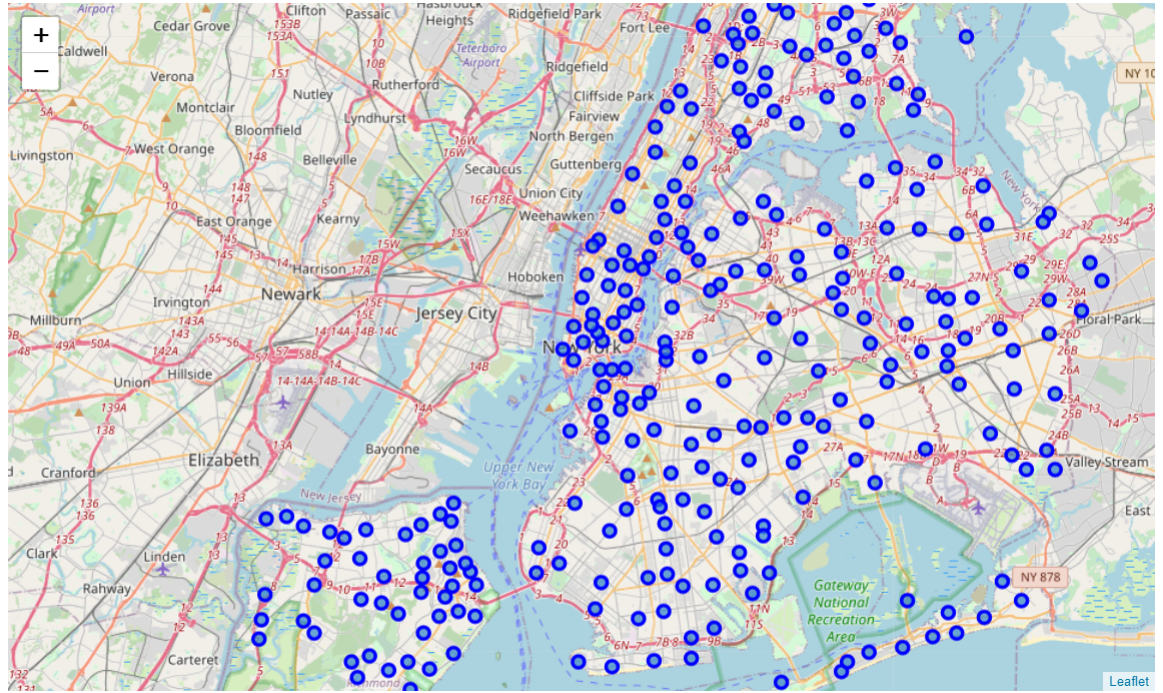
In the next step of exploration, the neighbourhoods were explored in greater detail. We used the Foursquare API to get the top 100 venues that are within 1000 meters from the point of neighborhood coordinates. The following table shows some venues from the first neighbourhood. Since we are interested in venue categories that has the word restaurant, we extracted all venue categories with the name restaurant for our further analysis.

Out[29]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	ID	Venue Latitude	Venue Longitude	Venue Category
1	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	4d375ce799fe8eec99fd2355	40.898152	-73.838875	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Jackie's West Indian Bakery	4c10f6aece57c92804a682d2	40.889283	-73.843310	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Ali's Roti Shop	4c9e50e38afca09379b2ff15	40.894036	-73.856935	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	McDonald's	4be5f0eacf200f47d1fa133c	40.902645	-73.849485	Fast Food Restaurant
5	Wakefield	40.894705	-73.847201	McDonald's	4e4e0e38bd4101d0d7a12e00	40.889435	-73.843369	Fast Food Restaurant
6	Wakefield	40.894705	-73.847201	Hong Kong China King	4edbac4e775bcc53fc0ed5f8	40.889568	-73.842911	Chinese Restaurant
7	Wakefield	40.894705	-73.847201	Mandarin House	4b897beff964a520c43a32e3	40.888139	-73.847119	Chinese Restaurant
8	Wakefield	40.894705	-73.847201	El Mangu Tipico Restaurant	4bec3ac661aca5935cfe8500	40.892891	-73.855825	Spanish Restaurant
9	Wakefield	40.894705	-73.847201	McDonald's	4bc256eb920eb713cd1b1c2c	40.892779	-73.857473	Fast Food Restaurant
10	Co-op City	40.874294	-73.829939	Dumplin Shop	524b6ca6498e92aa03e03ff4	40.876042	-73.835950	Dumpling Restaurant

Next, we check how many restaurants have been collected for each neighbourhood. The following table gives the number of restaurants in just 10 neighborhoods.

3

Number of restaurants	
Neighborhood	
Allerton	11
Annadale	4
Arden Heights	3
Arlington	2
Arrochar	5
Arverne	3
Astoria	31
Astoria Heights	9
Auburndale	35
Bath Beach	33

The next figure shows a bar plot of the neighborhoods with more than 5 restaurants.
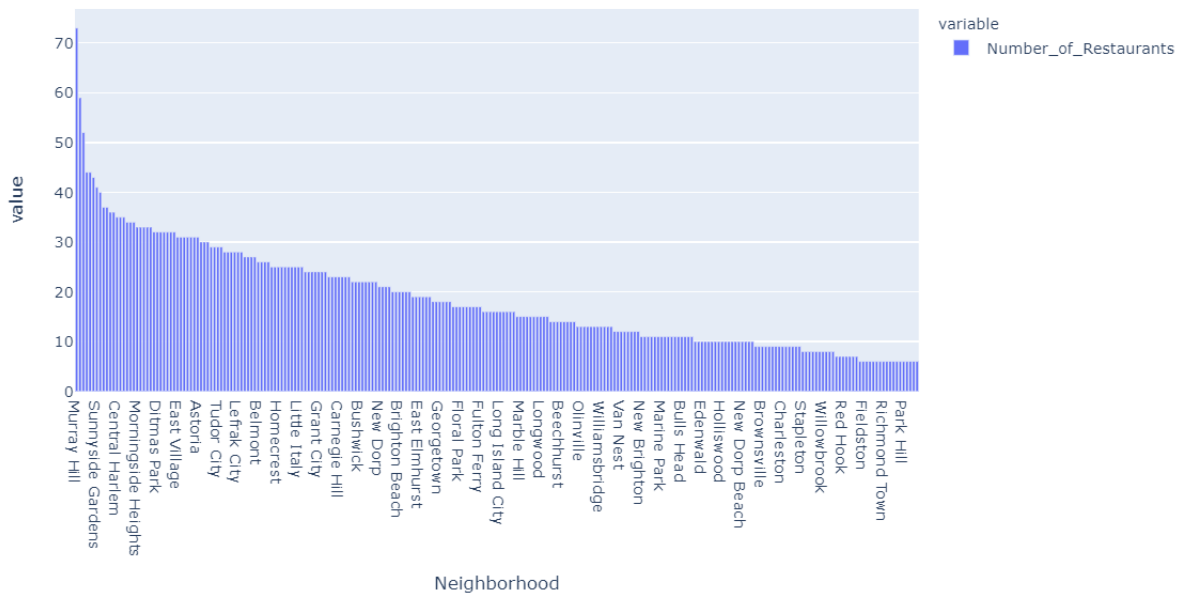
Now, to analyze the neighborhoods further, we focus on the restaurants. We use the one-hot encoding to create dummy variables so that the dataset could be used for machine learning. We then grouped the rows by neighborhood and took the mean of the frequency of occurence of each restaurant category.

After performing manipulations with the dataset, we get the following table (5 neighborhoods) which shows the top ten common restaurants for each neighbourhood.

	Neighborhood	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	Allerton	Fast Food Restaurant	Mexican Restaurant	Caribbean Restaurant	American Restaurant	Chinese Restaurant	Vietnamese Restaurant	Himalayan Restaurant	Filipino Restaurant	French Restaurant	German Restaurant
1	Annadale	Restaurant	American Restaurant	Hawaiian Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	German Restaurant	Gluten-free Restaurant
2	Arden Heights	Italian Restaurant	Mexican Restaurant	Sushi Restaurant	Vietnamese Restaurant	Halal Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant
3	Arlington	Fast Food Restaurant	Vietnamese Restaurant	Hawaiian Restaurant	Ethiopian Restaurant	Falafel Restaurant	Filipino Restaurant	French Restaurant	German Restaurant	Gluten-free Restaurant	Greek Restaurant
4	Arrochar	Italian Restaurant	Mediterranean Restaurant	Polish Restaurant	Middle Eastern Restaurant	Vietnamese Restaurant	Halal Restaurant	Falafel Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant

Finally, we plotted a chart of the first most common cuisine in all the neighborhoods in New York City.
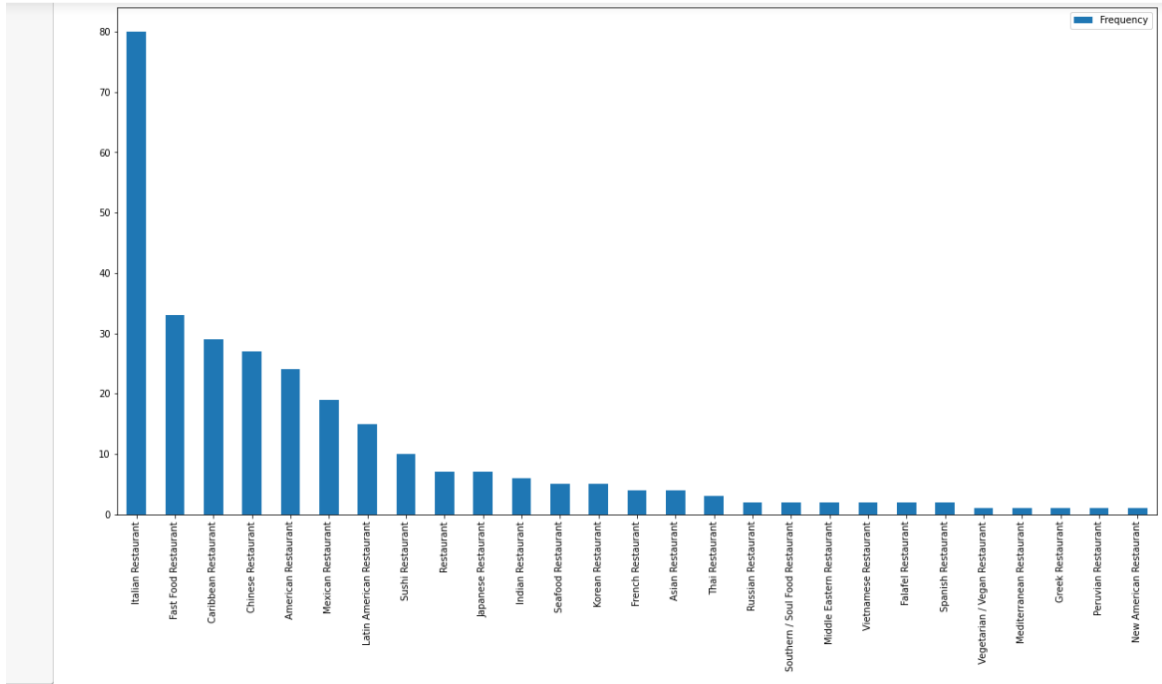
3.2 Cluster Analysis

We performed the clustering analysis on the data by using the K-means clustering algorithm. This unsupervised machine learning technique was prefered for this analysis, since we seek to cluster the neighborhoods based on restaurants. For this algorithm, we need to decide on the number of clusters (K) that we want to use. To avoid trail and error approach, the silhouette scores was used. The following graph shows the silhouette scores for a range of cluster variations.

From the graph, we chose 6 clusters as our optimal K. The algorithm was run for the specified number of clusters and the cluster labels were added to the dataset. Next, the clusters were visualized on the map.

4 Results

In this section, we present the results of our analysis. We present the exploratory analysis results and the main results from the clustering analysis.

4.1 Exploratory Data Analysis

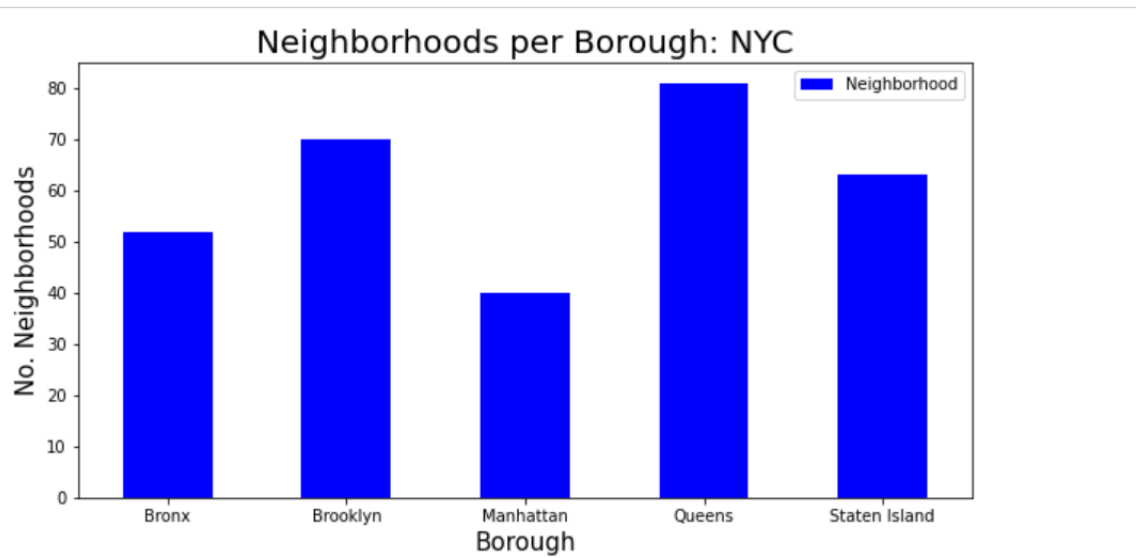Figure 1 shows that Queens Borough has the highest number of neighbourhoods.

Figure 1

Figure 2 shows the silhouette score for a range of cluster variations. We observed that the optimal number of clusters to use is six (6).
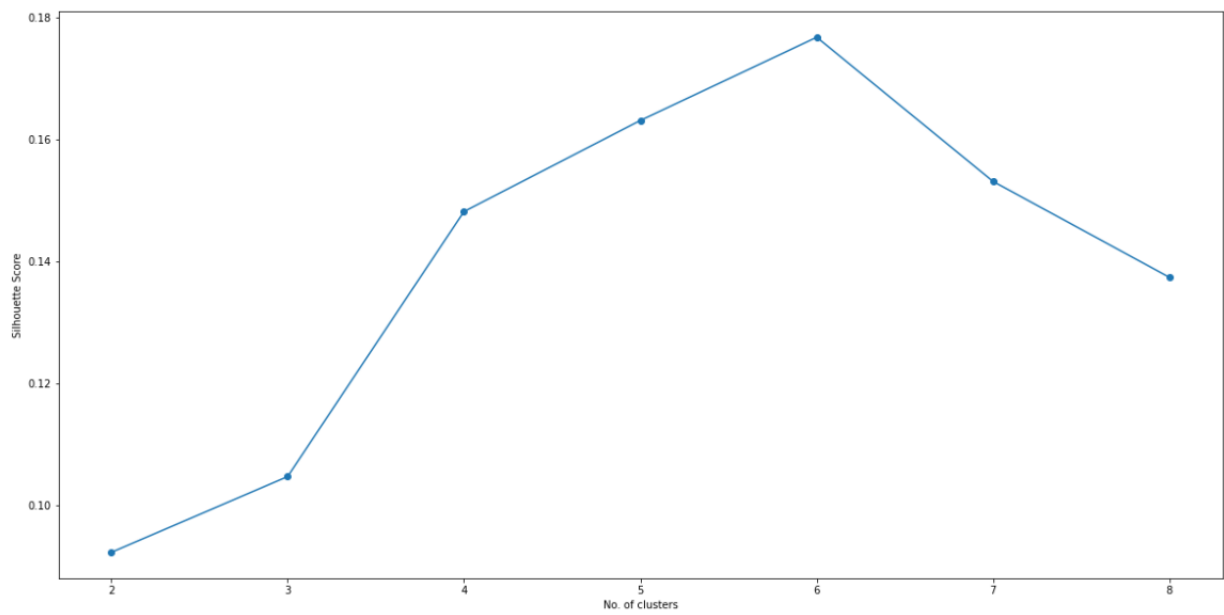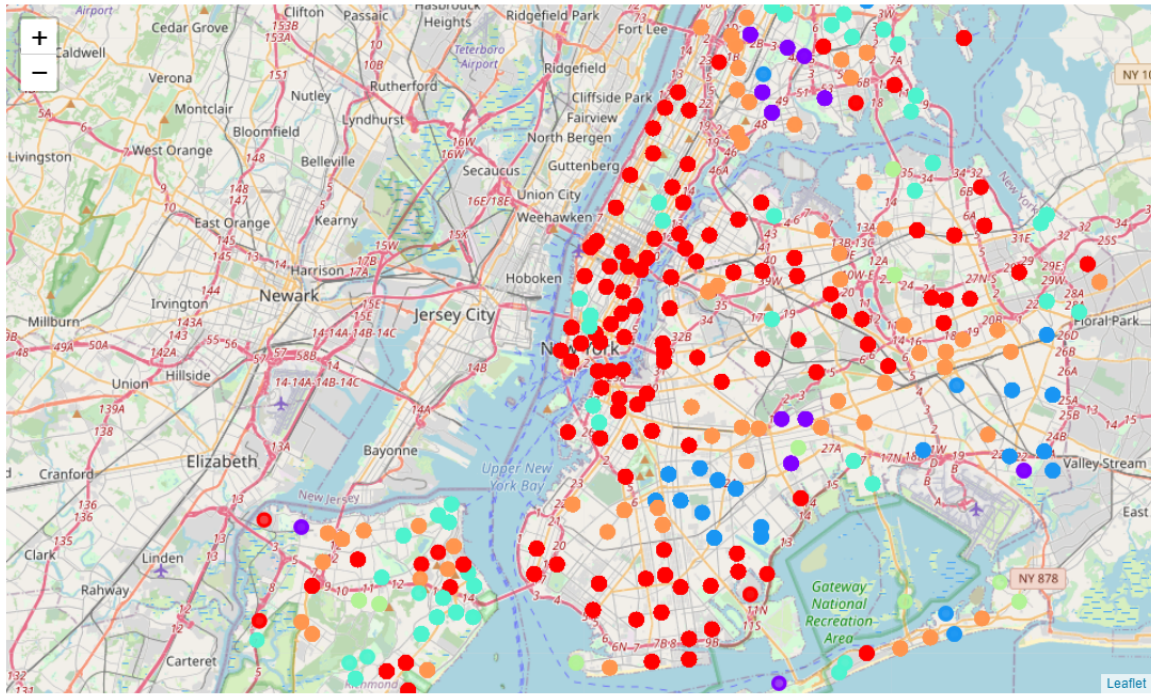

Figure 2: Silhouette Scores for a Range of Cluster Variations

4.2   Cluster Analysis

The clusters were shown the map in the figure below.

4.2.1 Understanding the Clusters

- **Cluster 1**
  The first cluster consist predominantly of Italian restaurants, American restaurants and Mexican restaurants sited mostly among Murray Hill,Elmhurst, Jackson Heights, Woodside, Gramercy and Sunnyside Gardens neighborhoods.

	Frequency
Murray Hill	73
Elmhurst	59
Jackson Heights	52
Woodside	44
Gramercy	43
Sunnyside Gardens	41
Ravenswood	37
Bayside	36
Noho	35
Central Harlem	35

	Frequency
Italian Restaurant	649
American Restaurant	350
Mexican Restaurant	239
Korean Restaurant	188
Sushi Restaurant	173
Chinese Restaurant	159
Japanese Restaurant	138
French Restaurant	87
Indian Restaurant	65
Vietnamese Restaurant	59

(a) Top 10 neighborhoods in Cluster 1   (b) Top 10 Restaurants in Cluster 1

- **Cluster 2**
  Cluster 2 consist of Fast Food restaurants and Latin American restaurants sited mostly among Longwood,Morrisania, Richmond Valley, Cypress Hills, Highland Park and Mount Hope neighborhoods.

	Frequency
Longwood	15
Morrisania	14
Richmond Valley	12
Mount Hope	10
New Lots	10
Cypress Hills	9
Soundview	8
Highland Park	6
East Tremont	6
Brookville	3

	Frequency
Fast Food Restaurant	85
Latin American Restaurant	15

(a) Top 10 neighborhoods in Cluster 2   (b) Restaurants in Cluster 2

- **Cluster 3**
  Cluster 3 consist predominantly of Caribbean restaurants. They are mostly located in the Prospect Park South, Prospect Lefferts Gardens, Rugby and Ramsen Village neighborhoods.

	Frequency
Prospect Park South	32
Prospect Lefferts Gardens	31
Rugby	17
Remsen Village	17
Erasmus	15
Williamsbridge	13
Eastchester	12
Flatlands	11
Wingate	10
Edenwald	10

	Frequency
Caribbean Restaurant	240
Fast Food Restaurant	12
Thai Restaurant	5
Vegetarian / Vegan Restaurant	2

(a) Top 10 neighborhoods in Cluster   (b) Restaurants in Cluster 3

- **Cluster 4**
  The fourth cluster consist predominantly of Greenwich Village, West Village, Lenox Hill, Soho, Belmont and Upper East Side neighborhoods. The most popular and common restaurants among these neighborhoods are Italian restaurants, Sushi restaurants and Chinese restaurants.

	Frequency
Greenwich Village	37
West Village	36
Lenox Hill	29
Soho	27
Belmont	26
Upper East Side	26
Little Neck	25
Fordham	25
Douglaston	24
Cobble Hill	23

	Frequency
Italian Restaurant	684
Sushi Restaurant	11
Chinese Restaurant	10
American Restaurant	2

(a) Top 10 neighborhoods in Cluster 4   (b) Restaurants in Cluster 4

- **Cluster 5**
  The fifth cluster consist predominantly of Chinese restaurants located mostly in Queensboro Hill, Willowsbrook, Malba and Manor Heights neighborhoods.

	Frequency
Queensboro Hill	16
Willowbrook	8
Malba	6
Manor Heights	6
Far Rockaway	5
East New York	4
Sea Gate	2
Prince's Bay	2
Broad Channel	1
Bayswater	1

	Frequency
Chinese Restaurant	49
American Restaurant	2

(a) Top 10 neighborhoods in Cluster 5   (b) Restaurants in Cluster 5

- **Cluster 6**
  The last cluster consist various cuisines including the Latin American, Fast food, Mexican, Chinese, Indian, Middle Eastern and Asian restaurants. These diverse restaurants are mostly located in the Flushing, Sunset Park, Flatbush, Washington Heights and Ditmas Park neighborhoods.

	Frequency
Flushing	44
Sunset Park	40
Flatbush	34
Washington Heights	33
Ditmas Park	32
Inwood	32
Sunnyside	31
Jamaica Center	27
Corona	25
Jamaica Hills	25

	Frequency
Latin American Restaurant	205
Fast Food Restaurant	202
Mexican Restaurant	186
Caribbean Restaurant	160
Chinese Restaurant	153
Indian Restaurant	61
Italian Restaurant	22
Restaurant	19
Middle Eastern Restaurant	12
Asian Restaurant	11

(a) Top 10 neighborhoods in Cluster 6   (b) Top 10 Restaurants in Cluster 6

Figure 8

5   Discussion and Recommendations

The neighbourhoods of New York city are multicultural. There are a lot of different cuisines including Chinese, American, Thai, Italian, and so on. As such, New York has a lot of restaurants to cater for almost every visitor from around the world.
It was obtained that the most common restaurant was the Italian restaurant, which can be found in almost all of the neighborhoods. In general, the least common restaurants among the neighborhoods were the Puerto Rican restaurant and Shanghai restaurant.

Our clustering is completely based on the most common venues obtained from Foursquare data. Hence, we ignored other factors like range of prices of restaurants, distance of closest stations and so on, since we did not have such data available for this project. Our analysis will only help visitors and travellers to get an overview of restaurants distribution in the neighborhoods of New York City. For a more detailed, informed and accurate guidance, the number of restaurants and the radius from neighborhood coordinates could be more expanded. Also more details about the restaurants and neighborhood could be extracted and used.

6   Conclusion

In this project, we tasked ourselves in helping visitors get an overview of restaurants in the neighborhoods of New York City. We performed analysis using the venue data provided by Foursquare as well as neighborhood data scraped from https://cocl.us/new york dataset.295 neighborhoods and 4785 restaurants were involved in this project. The cluster analysis on all restaurant venues gave us insight on the most common restaurants among neighbourhoods in each cluster. Even though this is not an extensive research on restaurants in this major city, this project gives a very good start to visitors and travellers when they are making restaurant decisions.

References

[1] Wikipedia. New york city, Dec 2020.

[2] Andrea Doyle. New york city again sets tourism record as it roars into 2019, Jan 2019.