Capston Project

The Battle of the Neighborhoods

Written by Pegah Aziziyanesfahani

1. Introduction/Business Problem

Toronto, New York and City of London are one of the most visited cities in the world. They have diversity in many fields such as technology, art and tourism. We want to search about their similarity and dissimilarity in their neighborhoods. Each city has its own diversity by neighborhood. Nowadays, geo information is very common which helps people to extract different types of information regarding location of a place. Moreover, many applications give you various types of information about different places all around the world. All above mentioned cities are popular in the tourisms industry with their different neighborhoods and places to visit. Now we can compare their neighborhoods of these cities and determine how are similarity and dissimilarity between them.

For example, one question can be as follows: **Is New York City more like Toronto or City of London?** The goal of this project is to figure out similarity or dissimilarity between these three cities according to information of their neighborhoods with data analysis.

2. Data and Problem Solution

For this problem, we use Foursquare API and some datasets to explore the data of these cities in terms of their neighborhoods. The Last version of dataset also includes the information about the places around each neighborhood like restaurants, hotels, coffee shops, parks, theaters, art galleries, museums and many more. Data resources are as follows:

- City of London data were obtained from Doogal website which covers UK postcodes and map tools.

- New-york city information is extracted from json dataset

- Canada post code lists were extracted from Wikipedia list and geo information of each neighborhoods were merged

- Using Foursquare API to get the venues records for all three cities

3. Data Cleaning

For each city, data downloaded or scraped from multiple sources were combined into
one table. Each city has a dataset in form of dataframe with information about neighborhoods with the latitude and longitude coordinates of each neighborhood. For each city, we get following information:

- New York: New York is the main point in this project and its similarity to other cities will be compared. All necessary and useful information will be extracted from the downloaded json data for creating a dataframe with Borough,Neighborhood, Latitude and Longitude Columns.
- Toronto: We extracted the related information for all neighborhoods in Toronto from Wikipedia. Then, we merged the geospatial data from another data source based on their same postal code column. In this project, the created dataframe was sliced to create a new dataframe of Toronto data with only borough's name

with Toronto. There are 39 neighborhoods that their borough's name contain Toronto.

- The City of London (also known simply as "the City") is considered as another city for comparing. The City is the historic core of metropolis of Greater London. The city of London is not considered as a London borough because it is governed by the City of London Corporation separately. A comma separated file will be used for this project which consists of geospatial data of all neighborhoods within city of London. This dataset has updated information of all postal codes for neighborhoods. Only in-used postal codes separated from dataset for each neighborhoods. In the last step, our dataframe consists of neighborhoods with their latitude and longitude values.

The following picture (Figure 1) shows the number of neighborhoods and boroughs between these three cities. According to our dataset, New York has 5 boroughs and 306 neighborhoods. Toronto has 10 Borough and 10 Neighborhoods and there are 25 neighborhoods in City of London.

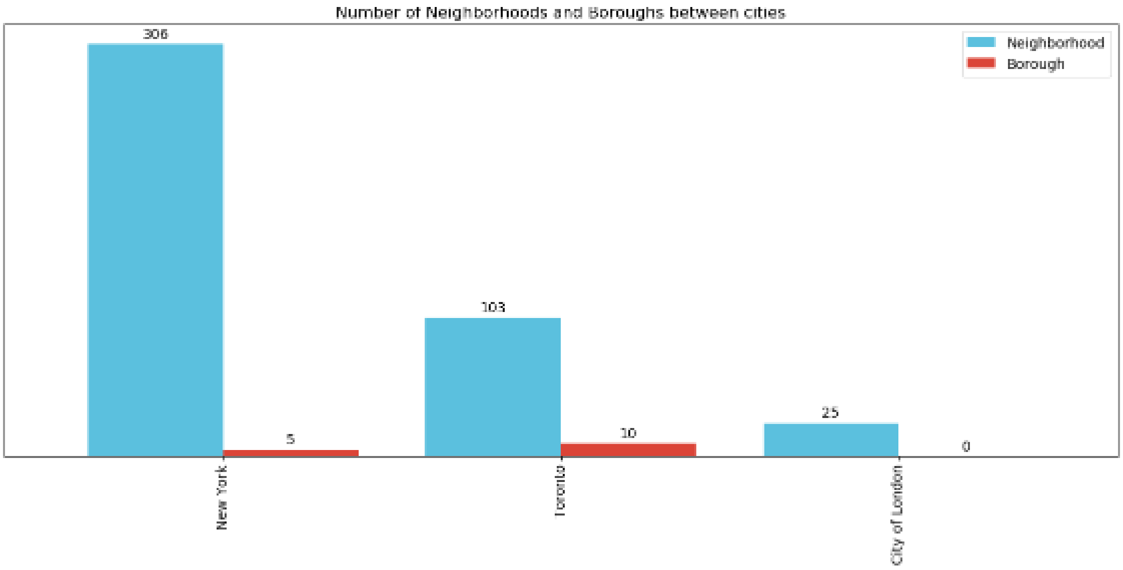


Figure 1: Number of Neighborhoods and Boroughs between Cities

4. Visualization of Neighborhoods in each Cities

In this section, we visualize the neighborhoods for each city. First of all, we use geopy library to get the latitude and longitude values of City. Then, we use Folium library for visualization. Folium is a great visualization library. We can easily zoom into the above map, and click on each circle mark to reveal the name of the neighborhood and its respective borough.

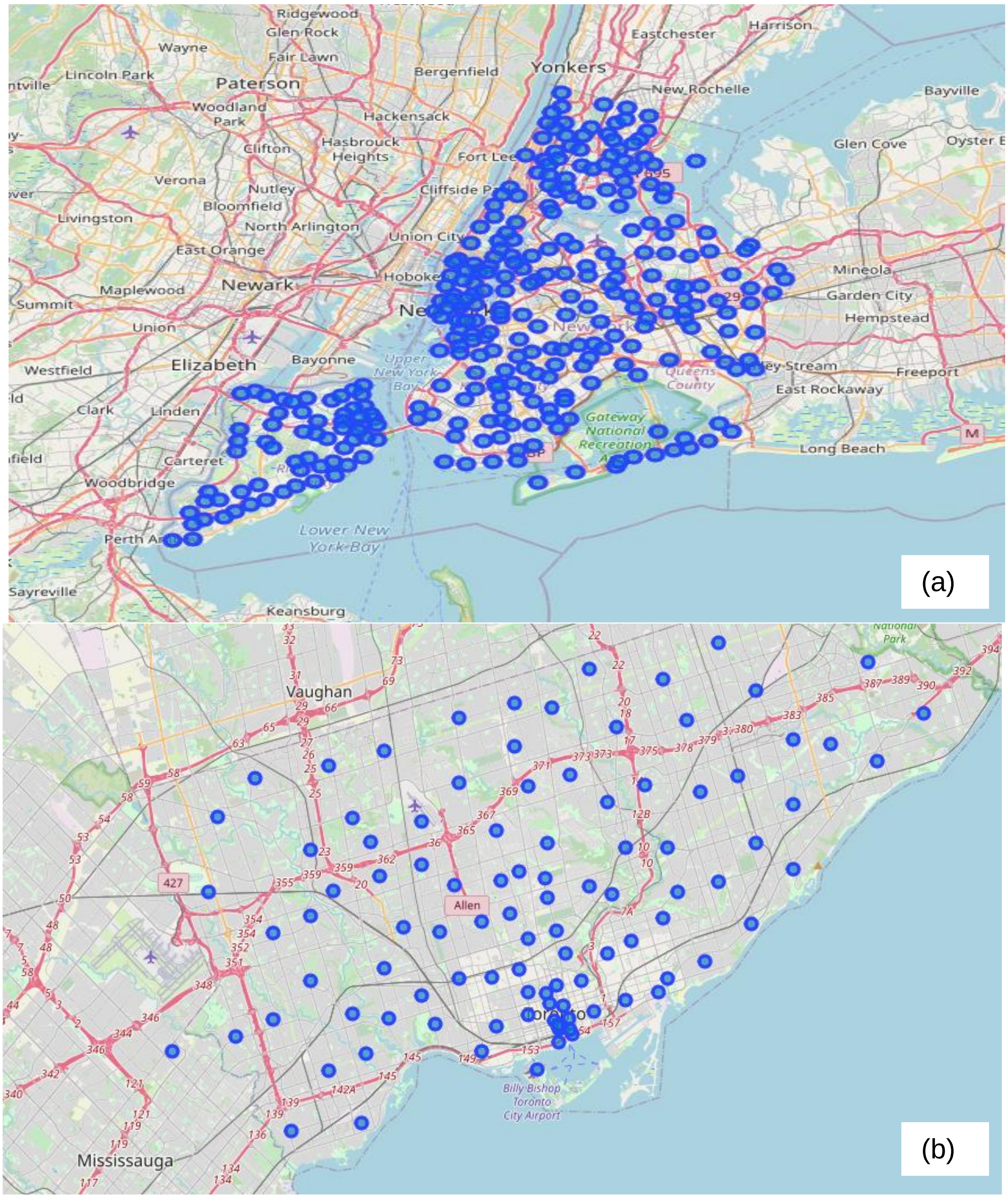


Figure 2: Visualization of Neighborhoods on Map – (a) New York Neighborhoods, (b) Toronto Neighborhoods

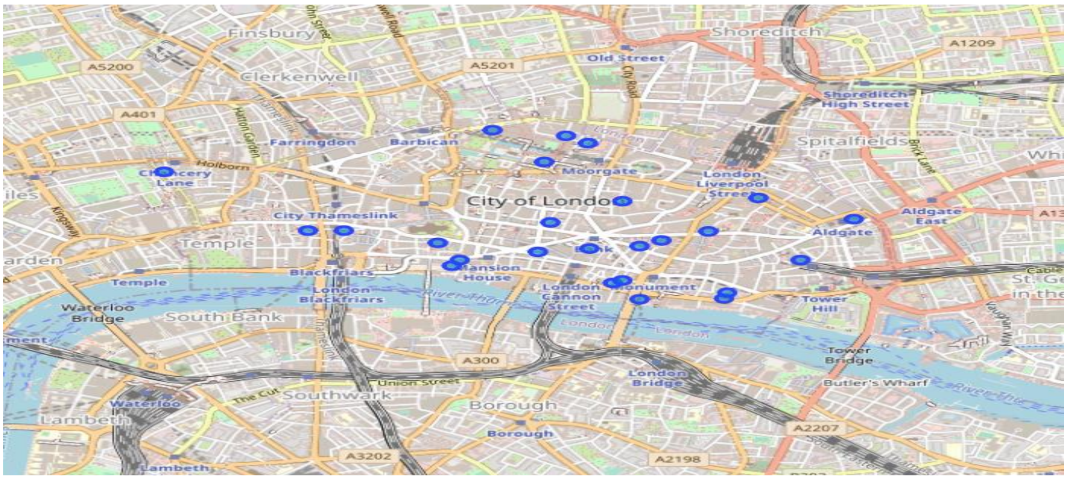


Figure 3: Visualization of City of London Neighborhoods on Map

We can also group the neighborhoods on map to see them as clusters in each part of the cities. This helps for showing the huge number of neighborhoods on map. For example, Figures 4 and 5 show neighborhoods clusters in all these cities:

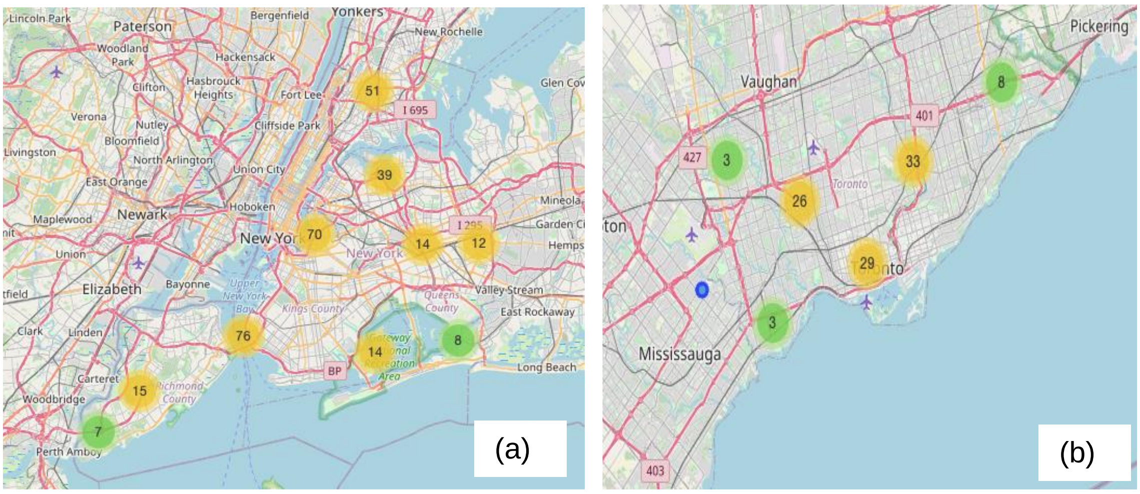


Figure 4: Neighborhood Clusters- (a) New York, (b) Toronto

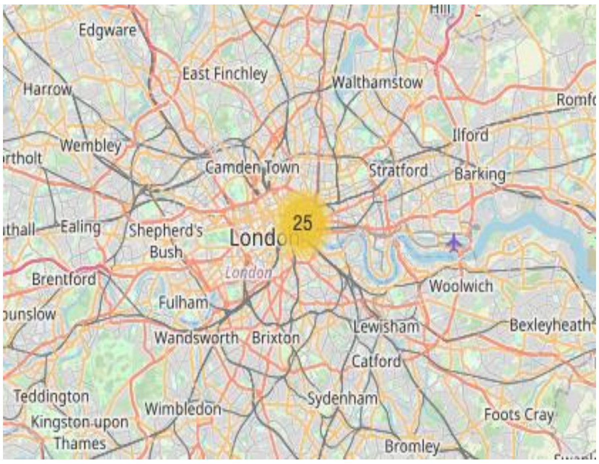


Figure 5: Neighborhood Clusters- City of London

4. Methodology

Machine learning technique, “Clustering” is used to segment the neighborhoods with similar objects on the basis of each neighborhood data. This will help to explore similarity or dissimilarity between these cities. We can understand how similar New York is by comparing the neighborhoods in other cities.

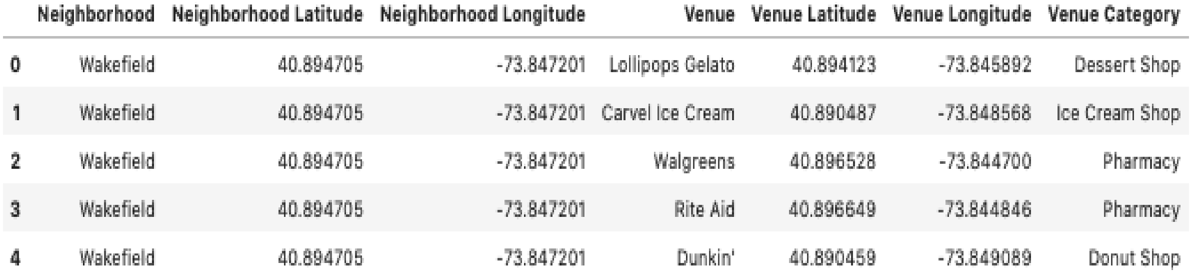
5. Data Analysis

5.1 Venue categories Data

In this section, the datasets for each city are being used to describe the venues of their neighborhoods. Venues information are retrieved from Foursquare which is a popular source of location or venue data. API service will be utilized to access and download venues data.

5.1.1 Venues Data for New York City

Figure 6 shows the retrieved data about venues in New York city neighborhoods. For each venue, venue name, category, latitude and longitude were retrieved. There are 427 uniques categories in all the returned venues of New York city.

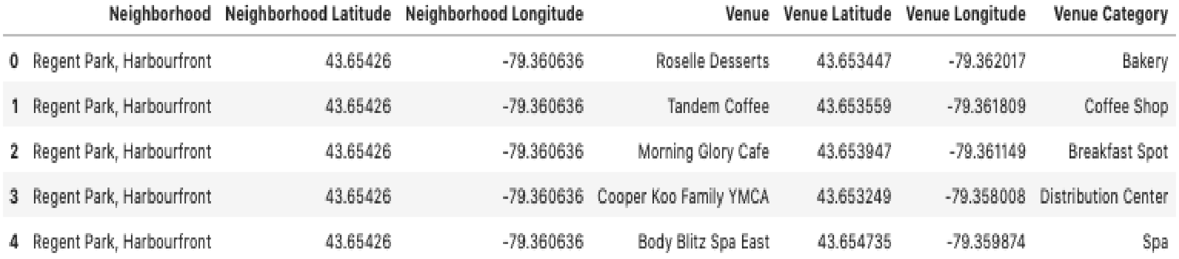


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Figure 6: Venues of New York City

5.1.2 Venues Data for Toronto

Similar to what has been done for NYC, a dataframe that describes the venues of Toronto neighborhoods was created. Figure 7 shows the dataframe for more than 1600 venues in Toronto. Each venue belongs to one of 236 uniques categories.

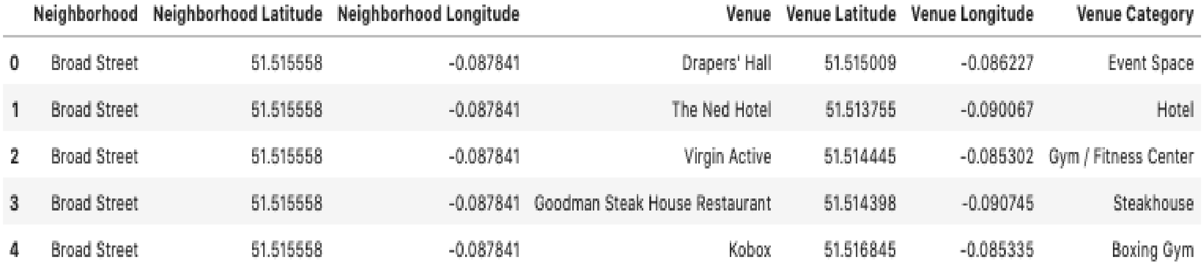


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
3	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
4	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

Figure 7: Venues of Toronto City

5.1.3 Venues Data for City of London

Just like above mentioned cities, dataframe for City of London was created. It can be seen in Figure 8 the venues for City of London and it has 151 unique categories in this Dataframe.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Broad Street	51.515558	-0.087841	Drapers' Hall	51.515009	-0.086227	Event Space
1	Broad Street	51.515558	-0.087841	The Ned Hotel	51.513755	-0.090067	Hotel
2	Broad Street	51.515558	-0.087841	Virgin Active	51.514445	-0.085302	Gym / Fitness Center
3	Broad Street	51.515558	-0.087841	Goodman Steak House Restaurant	51.514398	-0.090745	Steakhouse
4	Broad Street	51.515558	-0.087841	Kobox	51.516845	-0.085335	Boxing Gym

Figure 8: Venues of City of London

5.2 Most common Venue Categories

In this section, some visual analytics were done to explore and understand the venues data of neighborhoods in each city. The most common venue categories showed for all three cities. Figure 9 shows a bar pot of the most common venues in New York City. The most common category is Pizza Place with circa 450 venues in NYC. In the second rank, Coffee Shop is the most common category. In third rank comes Italian restaurant with almost 300 venues in this city.

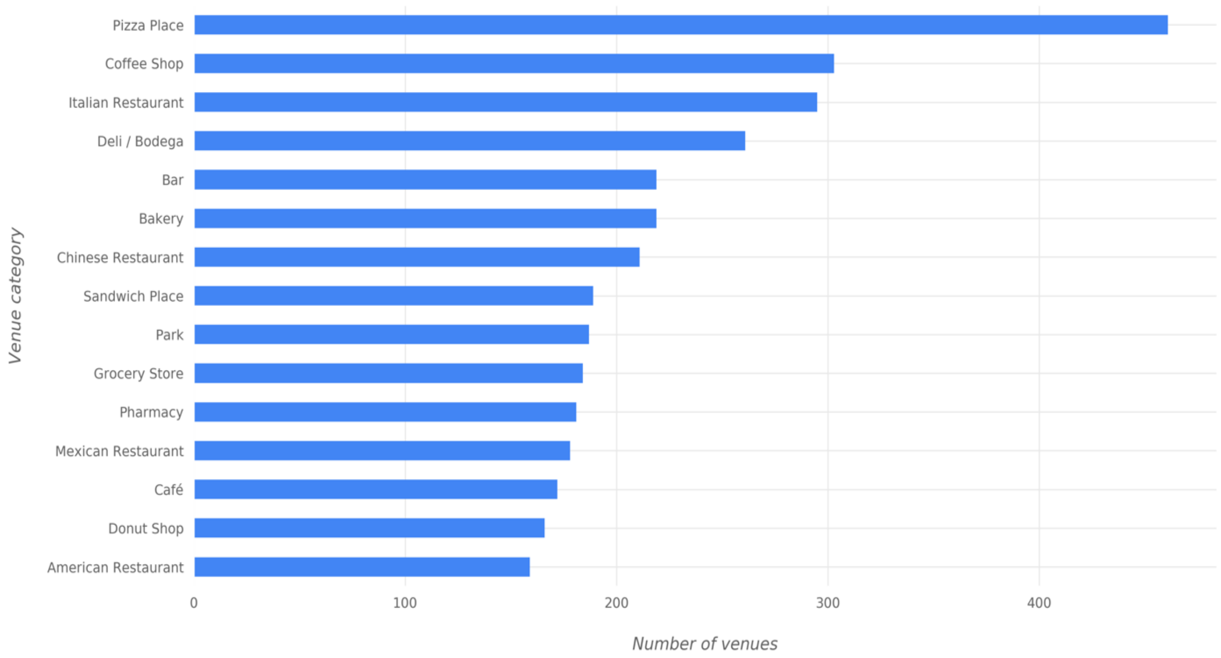


Figure 9: The most common venue Categories in New York City

Figure 10 describes the most common venues in Toronto city. Coffee shop is the most common category in Toronto with around 145 venues. Cafe comes in the second rank with almost 90 venues and Restaurant shows in the third place with around 55 venues.

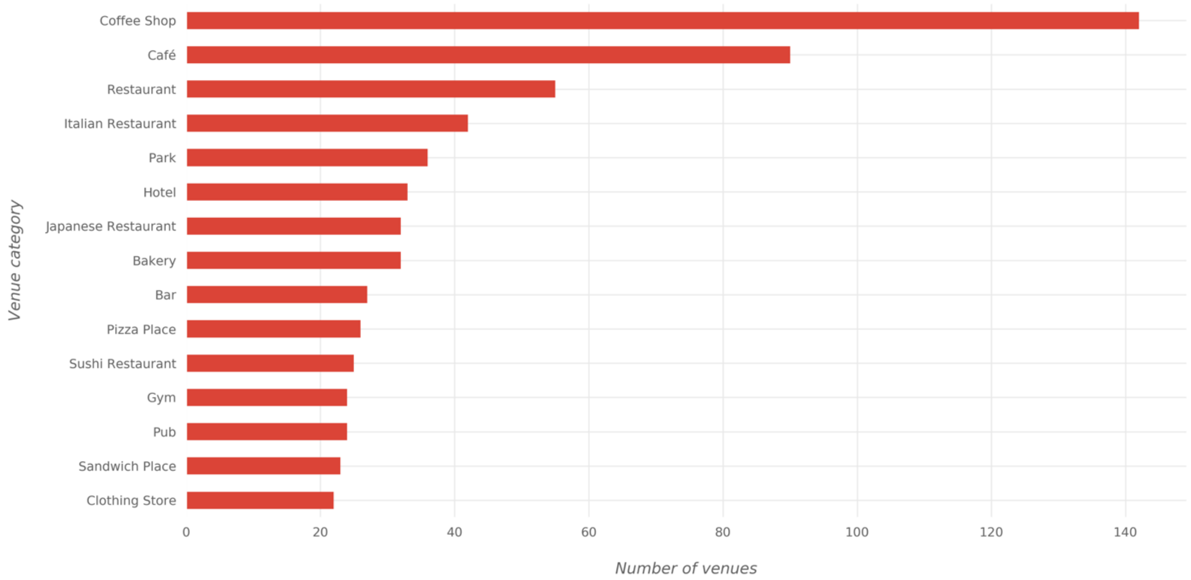


Figure 10: The most Common Categories in Toronto

As can be seen in Figure 9 and Figure 10, there are similarities between the most common categories in New York City data and in Toronto data. Figure 11 shows the most common venues category for City of London. In the first rank appears Coffee Shop with almost 145 venues. The Second place between venues categories is Gym/Fitness Center for City of London. Hotel shows as the third rank between venues categories. There are also some similar categories between New York city and City of London as well.

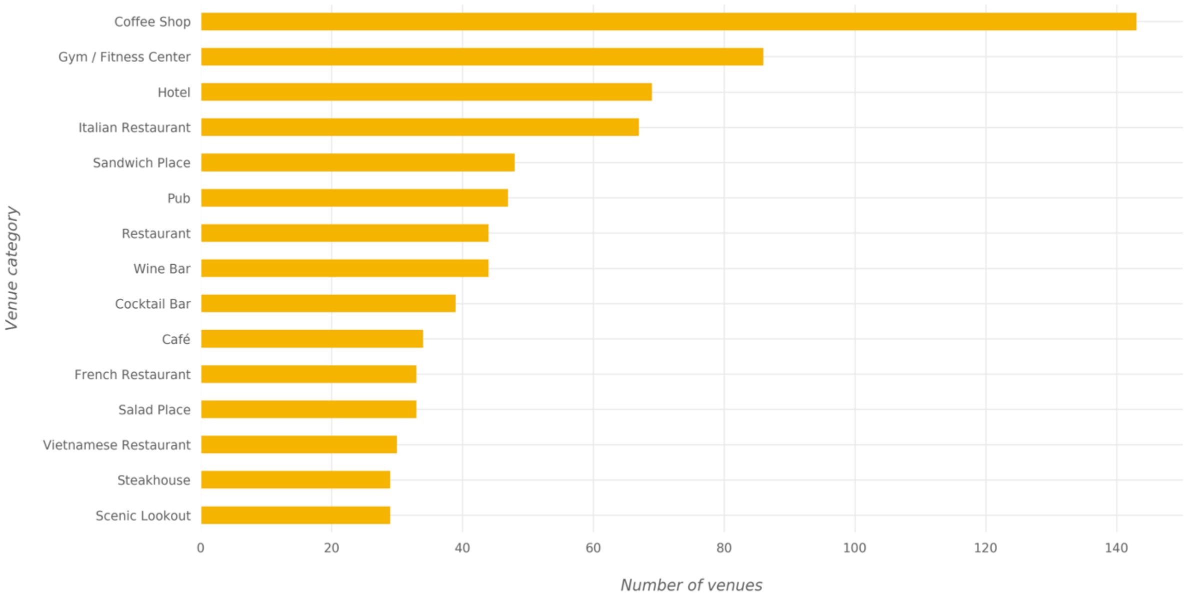


Figure 11: The most Common Categories in City of London

5.3 Venue Category in Neighborhoods

The venues categories that exist in more neighborhoods in each city are showed in the following figures. There are 306 neighborhoods in New York city. As is shown in Figure 12, Pizza Place is the most widespread category that exists in almost 200neighborhoods. Deli/ Bodega category is in almost 165 neighborhoods in New York City. The third widespread venue in New York City is Chinese Restaurant.

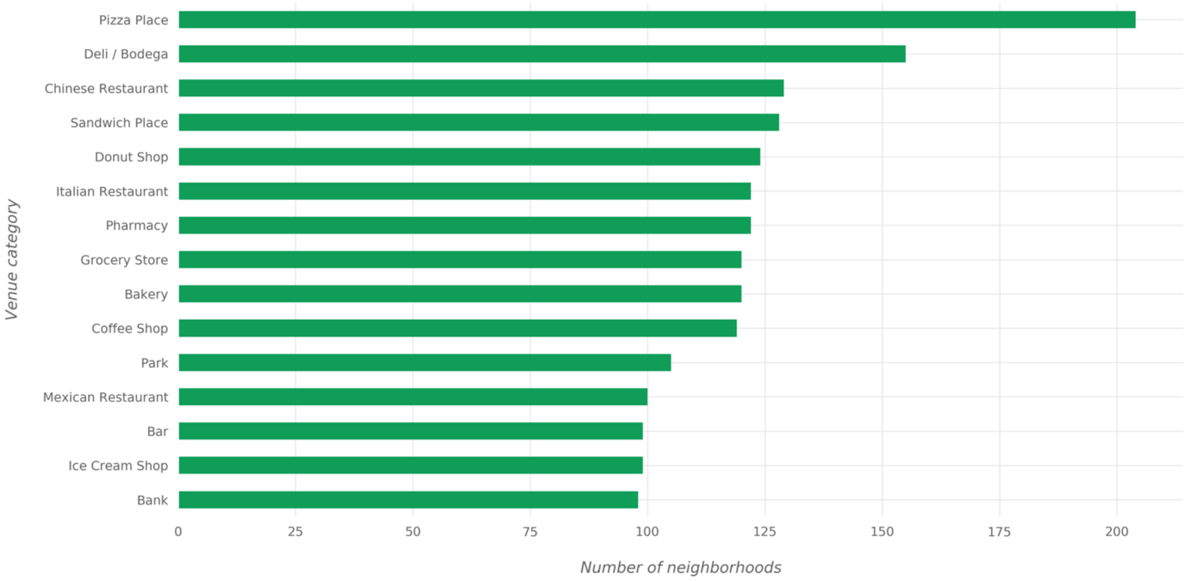


Figure 12: Venue Category in New York Neighborhoods

There are 103 neighborhoods in Toronto and only 39 neighborhoods are considered for this project analysis. Restaurant and Café come in the first and second places respectively and they spread in almost 28 neighborhoods in Toronto. Park appears in the third place as widespread venue category.

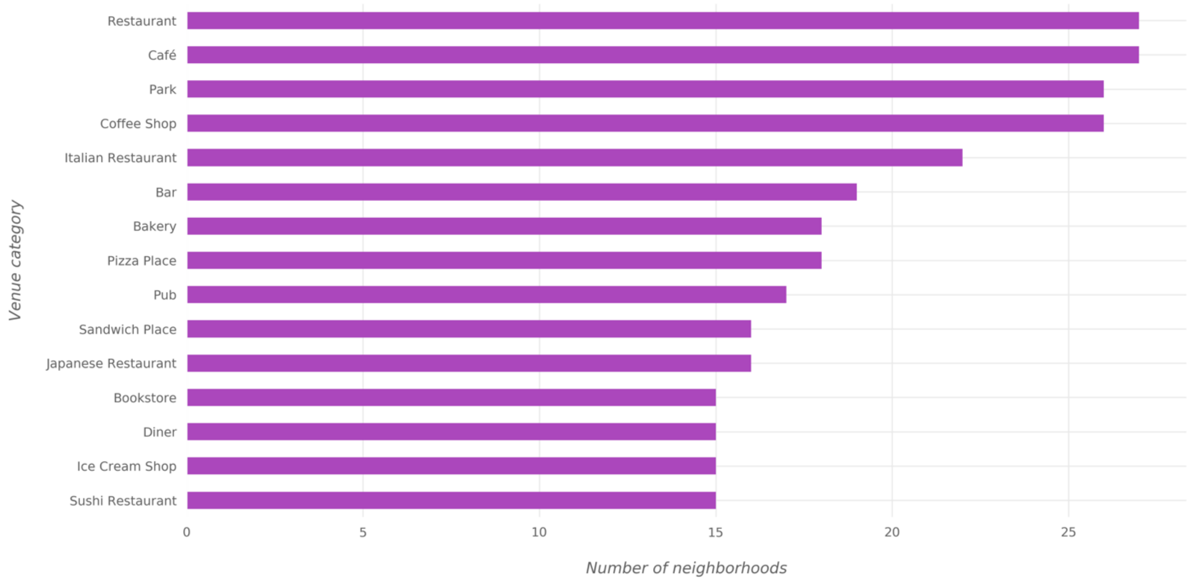


Figure 13: Venue Category in Toronto Neighborhoods

Figure 14 shows the most widespread venue categories in Toronto. There is the difference between the most widespread categories in Toronto and the most common categories. Coffee Shop and Gym Center come in the first and second places respectively with venues in 25 neighborhoods. Hotel Category appears in the third place with venues in circa 23 neighborhoods.

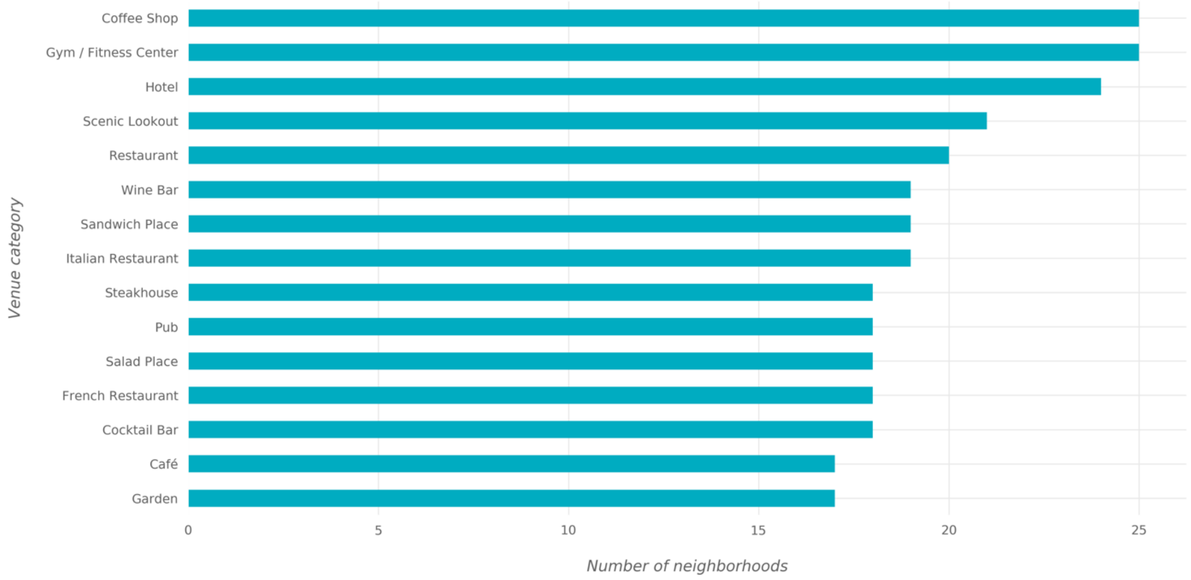


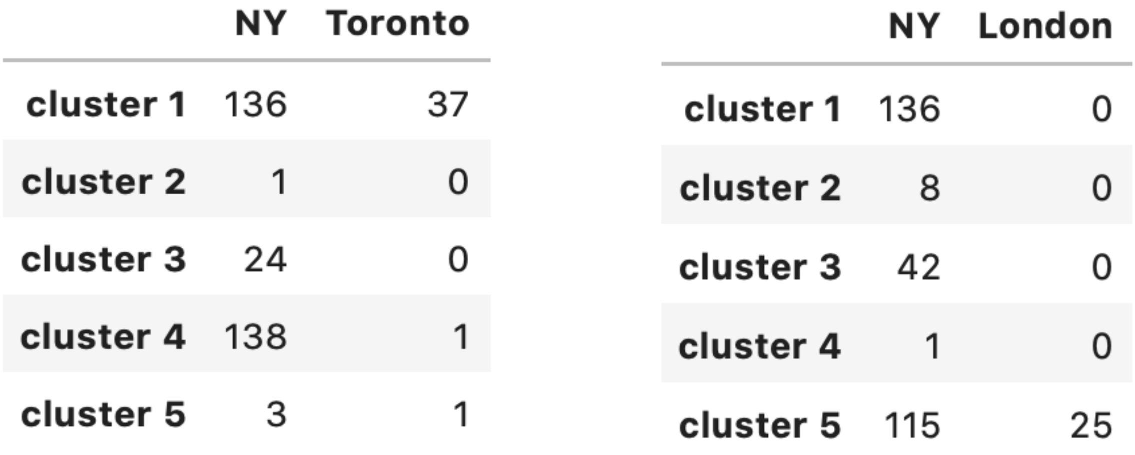
Figure 14: Venue Category in City of London Neighborhoods

5.4 Clustering of Neighborhoods

We did clustering the neighborhoods once in each city separately. Then, we did clustering firstly on NYC and Toronto neighborhoods and secondly on NYC and City of London neighborhoods. It helps to see if there is more similarity between NYC and Toronto or City of London. “Neighborhood” and “ Venue Category” are selected as features for clustering process. It helps to cluster neighborhoods based on similarity of venue categories. Firstly, one-hot coding will be applied on data for the analysis. Secondly, the data of two comparing cities will be combined. The next step is the aggregation of values for each neighborhood. Rows will be grouped by neighborhoods and by taking the mean of frequency of occurrence of each venue category. For clustering neighborhoods in New York and Toronto or City of London, the aggregated dataframes will be combined. For recognizing the neighborhoods, a text string is added to each of neighborhood name before merging. Finally, the merged dataframe is used to specify the most common categories for each neighborhood in both cities that are selected for comparing. K-means Clustering produced cluster-labels for each neighborhood within data. It labels each neighborhood in the city that belongs to the similar cluster. The output is 5 different clusters that each cluster is expected to contain similar neighborhood based on venue categories.

6. Results and Disscussion

Figure 15 shows the number of neighborhoods that are in each cluster. Figure 15.a is between New York City and Toronto and Figure 15.b shows the neighborhoods within each cluster between New York City and City of London.



	NY	Toronto
cluster 1	136	37
cluster 2	1	0
cluster 3	24	0
cluster 4	138	1
cluster 5	3	1

	NY	London
cluster 1	136	0
cluster 2	8	0
cluster 3	42	0
cluster 4	1	0
cluster 5	115	25

Figure 15: Number of neighborhoods in each cluster- (a): NYC and Toronto (b): NYC and City of London

Figure 16 shows the number of New York City neighborhoods and the number of Toronto neighborhoods in five resulting clusters. For example on cluster 1 there are 37 similar neighborhood in Toronto like New York city neighborhoods. Toronto has one similar neighborhood like New York city in cluster 4 and cluster 5 respectively. This similarity is based on venues category between two cities.

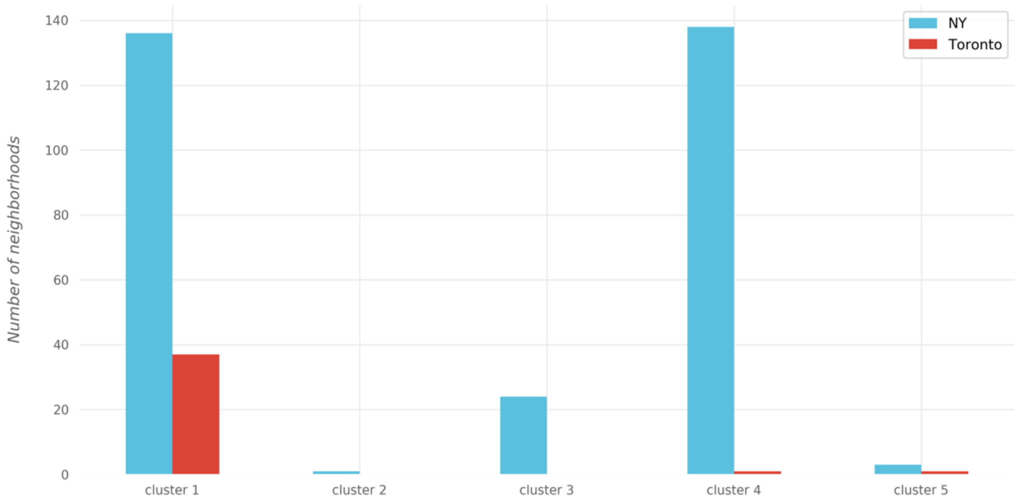


Figure 16: Number of Neighborhoods within Cluster between NYC and Toronto

Figure 17 shows the most common venues categories according to New York and Toronto top venues data. According to results, 45% of venues categories in both cities is Pizza place. 30% of the venues categories is Coffee shop which appears on second place. Italian Restaurant can be found in almost 9% of venues.

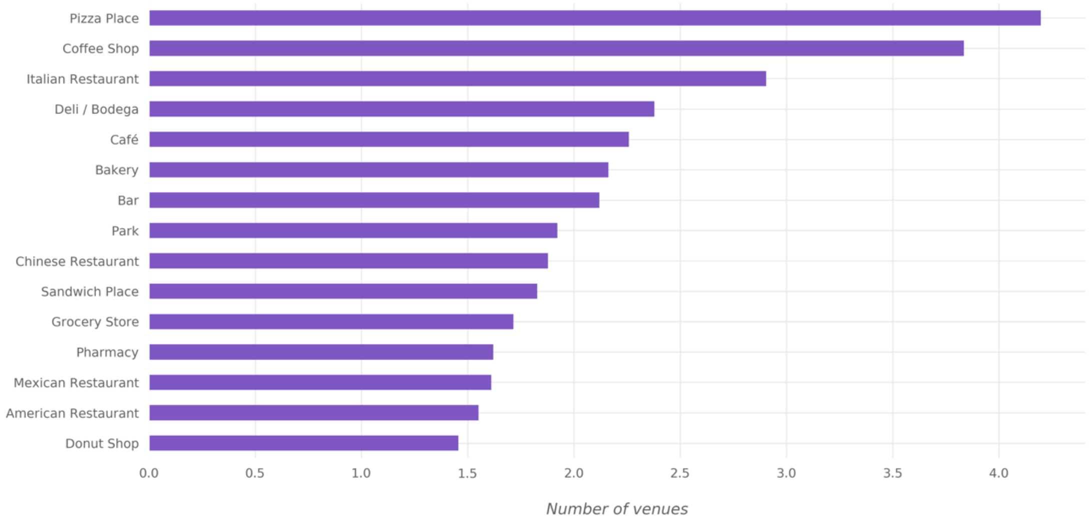


Figure 17: The most Common Venues Categories in New York and Toronto

Figure 17 shows that the all 25 neighborhoods in City of London are similar to 115 neighborhoods in New York City. They were all grouped in cluster 5. There are similar venue categories just like Toronto city that are similar in neighborhoods of both New York and City of London. Almost 42% of venues belong to Pizza Place. Around 38% of venues are categorized as Coffee Shop. Italian restaurant is in almost 32% of venues between City of London and New York city (Figure 19).

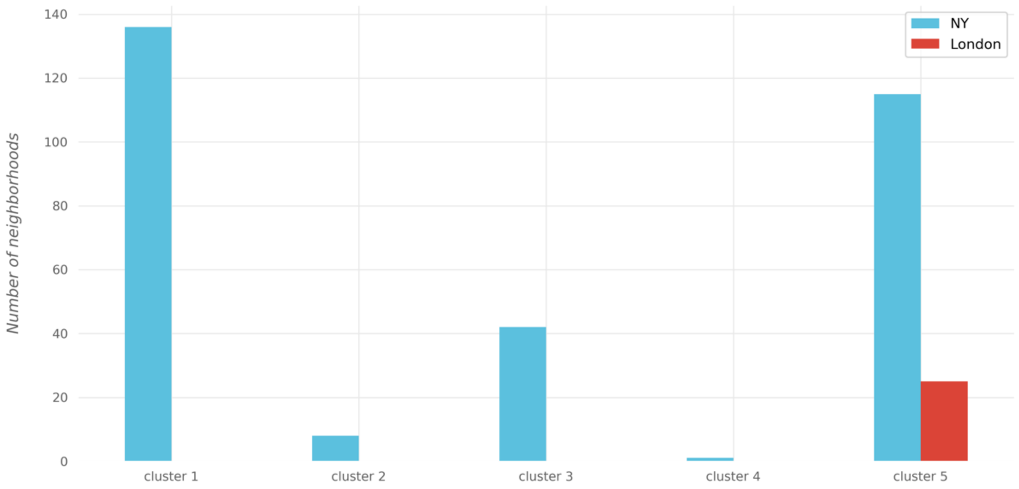


Figure 18: Number of Neighborhoods within Cluster between NYC and City of London

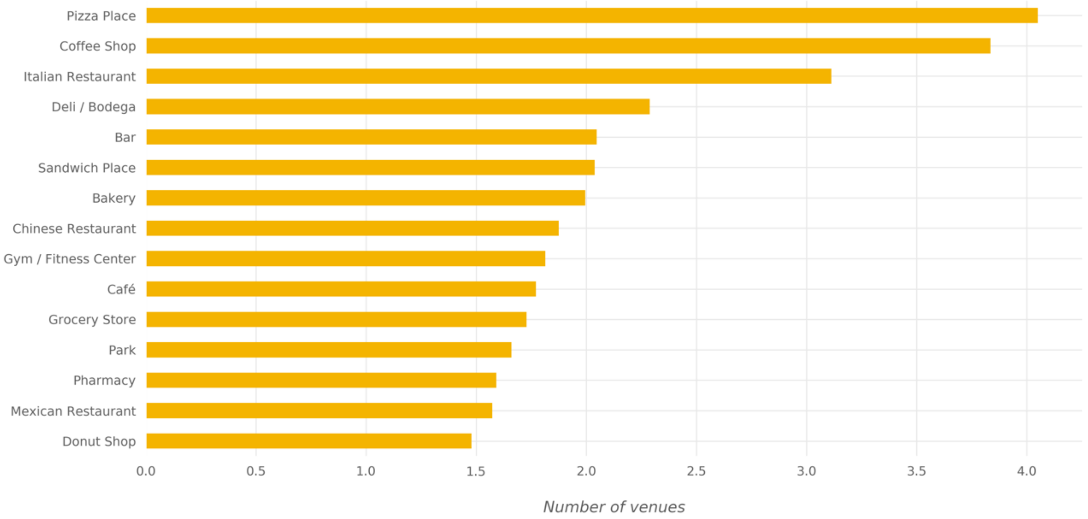


Figure 19: The most Common Venues Categories in New York and City of London

7. Conclusion

In this project, the neighborhoods of New York City , Toronto and City of London were clustered into various groups based on the venues categories in their neighborhoods. The results shows that there are similar venue categories that are more common in some clusters than others. There are more common venues in neighborhoods of New York City and Toronto. According this analysis, we cannot say that there is dissimilarity between New York City and City of London according to their neighborhoods. As it showed above, all 25 neighborhoods within City of London were clustered with 115 neighborhoods in New York City. We could see that there are common venue categories between these three cities and they are all same.