Finding Similar Neighborhoods between Toronto, ON and Queens, New York

Collins Opoku-Baah

1. Introduction

1.1 Background

The world has become one giant global village and thus, each and every minute, people travel from one place to another. The purpose for traveling could be temporary that is for vacation, business, visits etc. or could be permanent e.g. school, work etc. When people live in a particular region for a long time, they tend to embrace the cultures (e.g. food, clothing, language etc.) of that region, making it very difficult to transition into other neighborhoods. For example, a person who loves seafood and attend yoga classes will want to move to a new place with such venues in order to continue having their pleasant life experience.

1.2 Problem

Finding a place that share similar venue as your current neighborhood can be burdensome considering how developed most cities in the world are currently. Hence, this project aims to find neighborhoods between two big cities namely Toronto, ON and Queens, New York that are similar with respect to venues. To do this, I will employ machine learning approaches and other techniques to segment and cluster neighborhoods in these two cities.

2. Data Acquisition and Cleaning

2.1 Data Sources

The data for this project comprised the venue locations that are within a radius of 750 meters around the neighborhoods in the two big cities, which are Toronto, ON, Canada and Queens, New York, USA. First, I obtained the various neighborhoods for the Toronto and Queens, New York from the web sources. The neighborhoods for Toronto, ON were obtained by scraping the website, 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M' and that for Queens, New York were obtained by downloading json file from https://cocl.us/new_york_dataset.

2.2 Data Cleaning

Both datasets contained more information than needed for this project. While the Toronto data contained information about all the Boroughs and Neighborhoods in Canada, the Queens data contained all the Boroughs and Neighborhoods in New York. With regards to the data for Toronto, I created a dataframe containing only the neighborhoods under Boroughs named Toronto. Likewise, I created another dataframe containing only the neighborhoods in Queens Borough. Both dataframes were then combined into a single data frame containing the neighborhoods in both cities.

2.3 Feature Selection

The features for this project were constructed from the distinct categories of venue locations in the neighborhoods of the two cities. However, in order to get these venues, the latitude and longitude

coordinates for each of the neighborhoods were obtained using geocoder API. While the dataset for Queens already came with these coordinates, that of Toronto didn't. The geocoder library was used to obtain the coordinates for each of the neighborhoods in our combined datasets.

After successfully obtaining these coordinates, the Foursquare API was used to obtain location venues such as yoga places, restaurants, etc. that are within a 750 meters' radius around these neighborhoods. The number of venues for each neighborhood was limited to 100. One-hot encoding was employed to create features based on the distinct venue categories. The resulting dataframe was grouped by neighborhoods to obtain a dataframe with neighborhoods in the rows and distinct venue categories in the columns. During the grouping, the number of venue under each category for a particular neighborhood was averaged by the total number of venues for that neighborhood. This yielded a fair representation of the proportion of each venue category for that neighborhood. See figure 1 for an example of the resulting dataframe.

	Neighborhood	Borough	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	...	Vietnamese Restaurant	Warehouse Store	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop
0	Adelaide	Downtown Toronto	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.010000	0.000000
1	Arverne	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.062500
2	Astoria	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.010000
3	Astoria Heights	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
4	Auburndale	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
5	Bathurst Quay	Downtown Toronto	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.038462	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
6	Bay Terrace	Queens	0.02381	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.02381	0.0	0.000000	0.000000
7	Bayside	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.014925	0.000000
8	Bayswater	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
9	Beechhurst	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
10	Bellaire	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
11	Belle Harbor	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
12	Bellerose	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.043478

Figure 1. An example of the data for the project.

3. Methodology

3.1 Exploring Locations of Neighborhoods using Folium Maps

First, I explored the locations of the neighborhoods in the two cities, that is Toronto, Ontario, CA and Queens, New York, USA using folium maps. In all, there were 154 neighborhoods with Toronto having 73 and Queens having 81.

Figure 2a. Folium map showing the neighborhoods of Queens, New York.

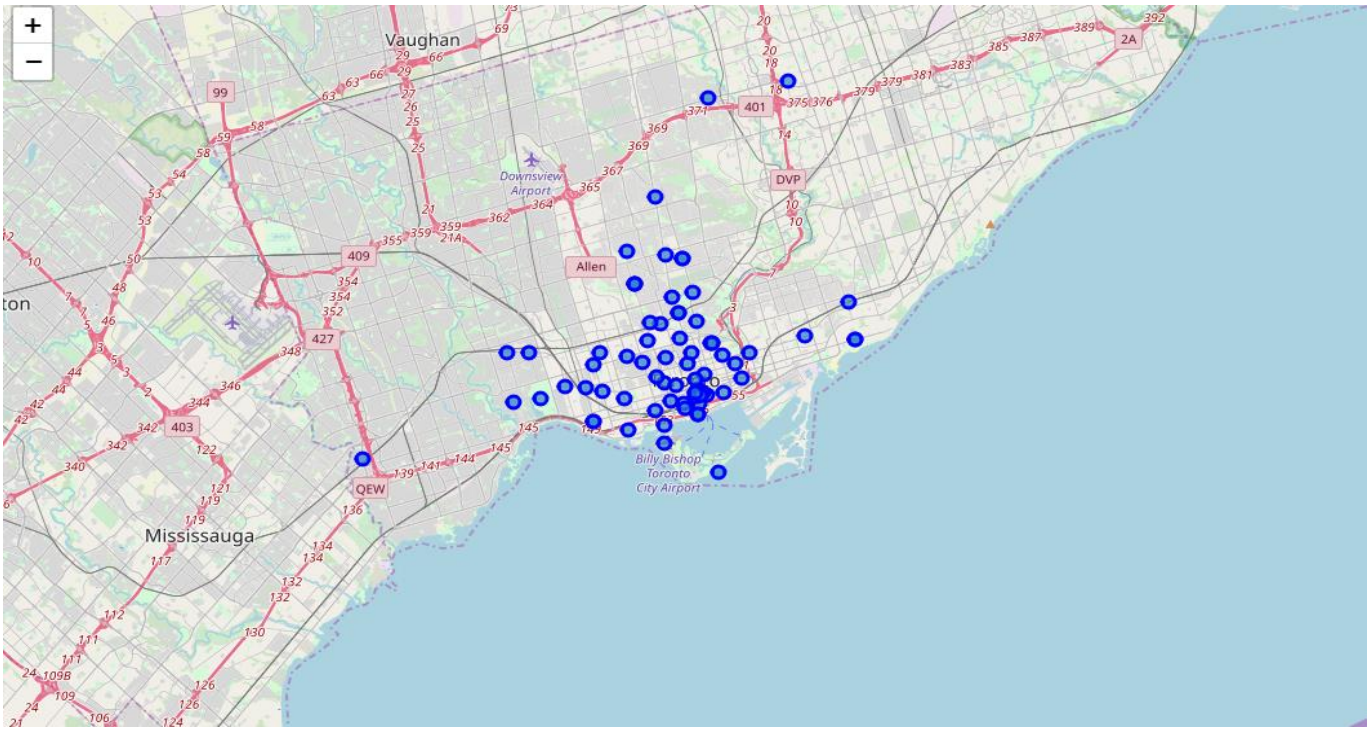


Figure 2b. Folium map showing the neighborhoods of Toronto, Ontario.

3.2 Exploring the Foursquare API Data

After running the Foursquare API for all the neighborhoods of the two cities, it was observed that neighborhoods in Toronto had a higher mean number of venues than that of Queens.

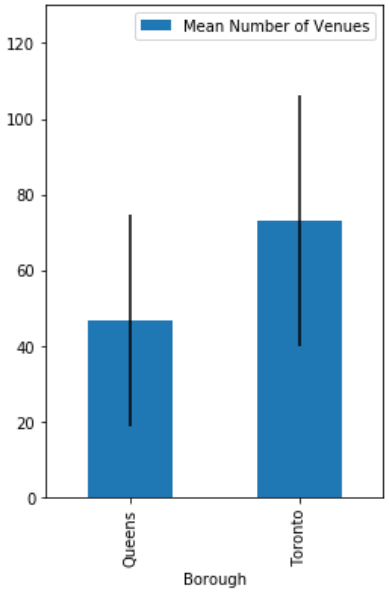


Figure 3. Mean Number of Venues for Toronto, Ontario and Queens, New York.

After applying one-hot encoding the create features (See the feature selection for more details), I created a figure that illustrated the first 25 most common venues for each city combining data across all the neighborhoods. While the first two most common venues for Toronto was Coffee shop and café probably due to colder weather conditions, that for Queens New York were Pizza place and Deli/Bodega which are both fast food places.

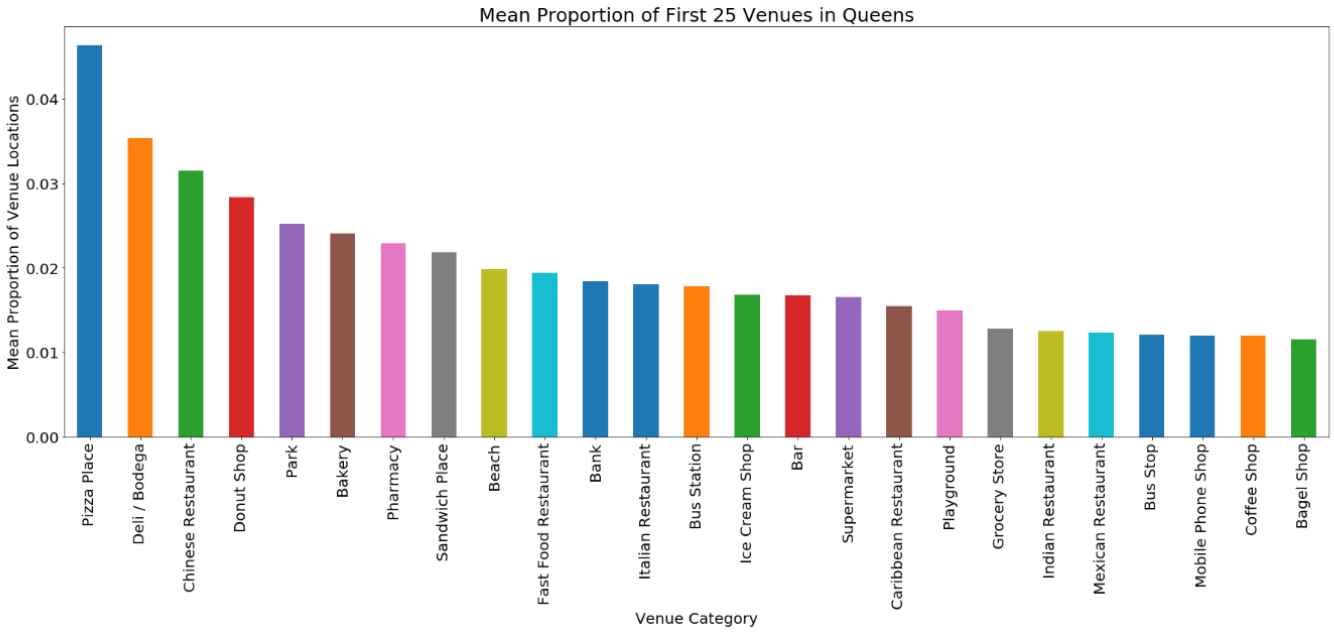


Figure 4a. Mean Proportion of First 25 Venues in Queens, New York.

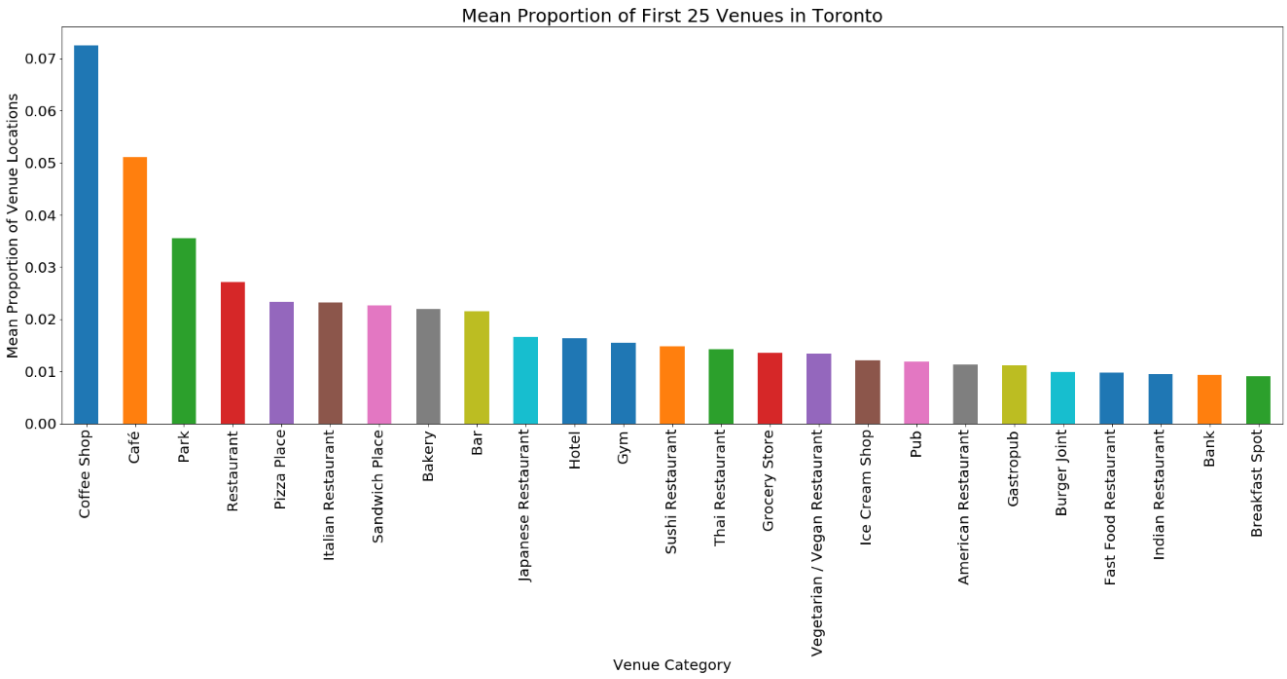


Figure 4b. Mean Proportion of First 25 Venues in Toronto, Ontario.

4. Results and Discussion

4.1 Clustering and Segmentation of Neighborhoods

In order to determine similar neighborhoods between the two cities, K Means clustering approach was used to segment and cluster these neighborhoods. The K Means model was run using number of clusters ranging from 2 to 9. The inertia value of each cluster which also measures how close the individual points are within the clusters were compared across the 8 models. I was discovered that the model with nine clusters had the lowest inertia. This also meant that increasing the number of clusters could further decrease the inertia, however, the upper limit was set at 9 in order not to facilitate individual exploration of clusters.

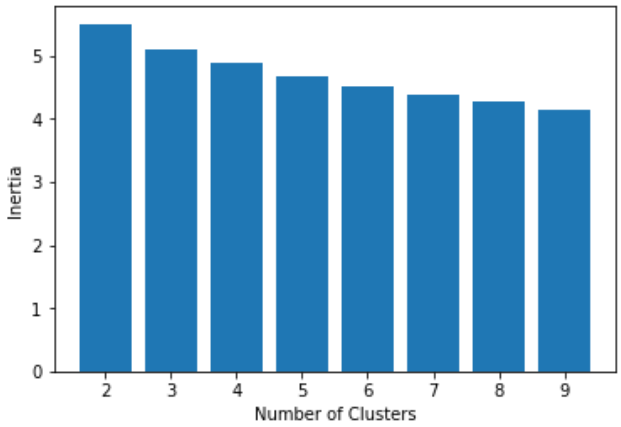


Figure 5. Inertia Values for the Cluster Models.

Upon analyzing the model with nine clusters, it was found that 6 out of the 9 clusters had neighborhoods in both Toronto and Queens while the remaining three had only neighborhood in Queens. Moreover, most of the neighborhoods in each city were similar. However, these neighborhoods in for Queens formed a different cluster from those in Toronto.

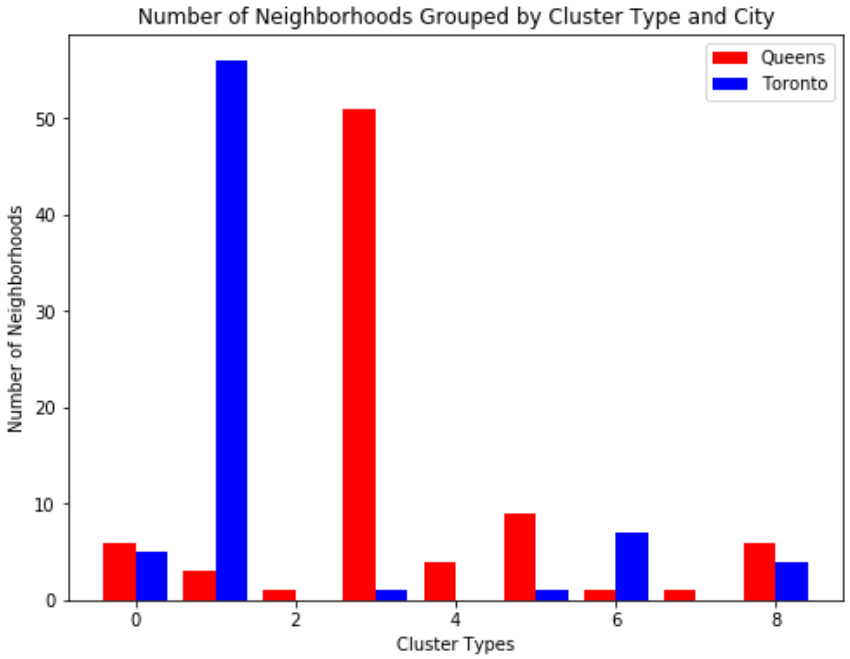
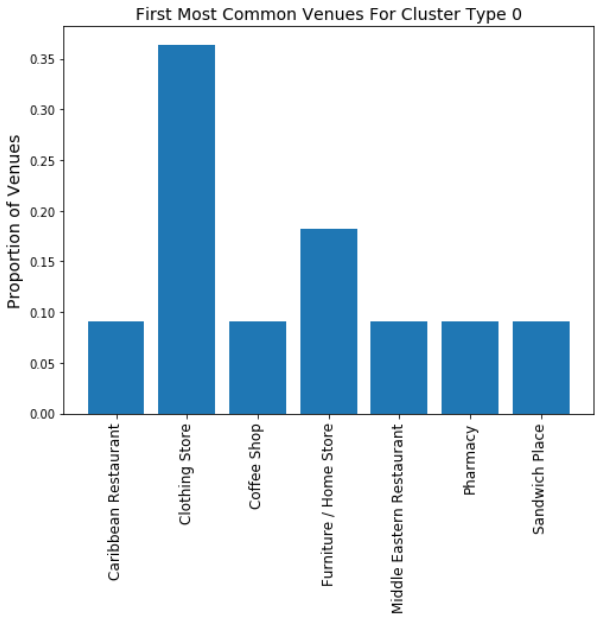
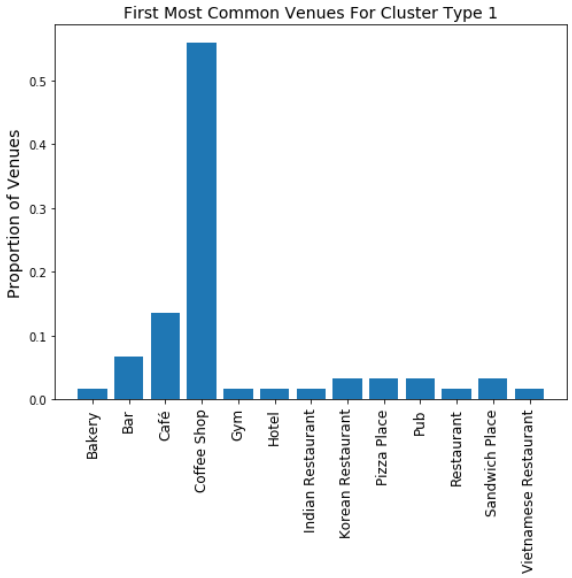


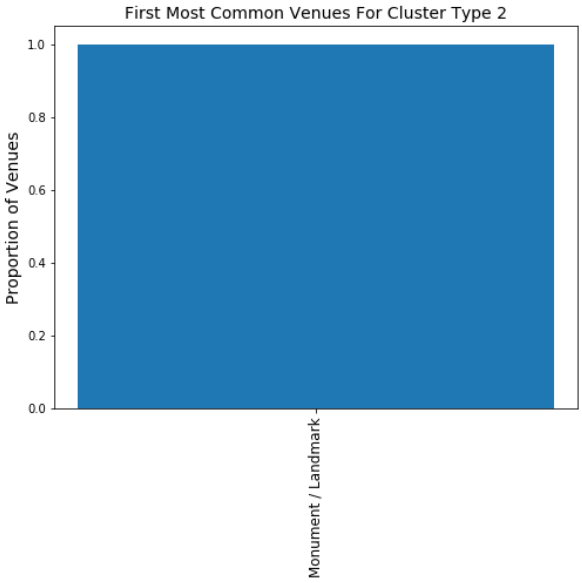
Figure 6. Number of Neighborhoods Grouped by Cluster Type and City.

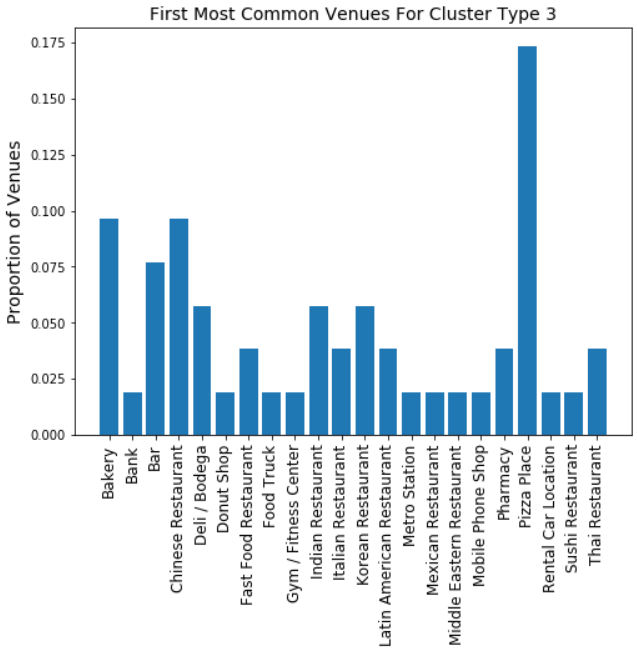
Comparing each cluster type revealed the difference in the pattern of venues amongst them. For Cluster Type 0, the most common venue was Clothing Store which signifies that this place may be surrounded by several fashion stores and will be good place for people who may want to go on shopping. Cluster Type 1 had Coffee Shop and Café as the most common venues, quite similar to the overall pattern in Toronto. This explains why most of the neighborhoods in Toronto were in Cluster Type 1.



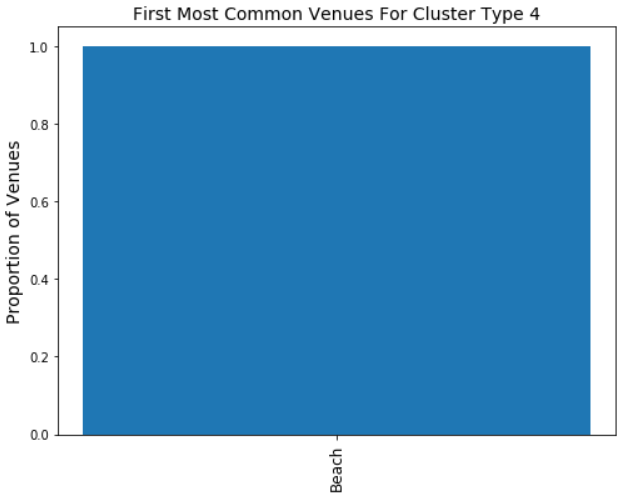


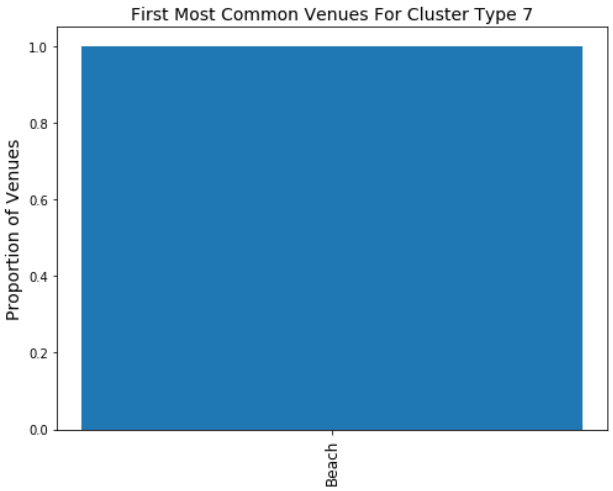
Cluster Type 2 was one of the cluster types with only a neighborhood in Queens. This neighborhood is characterized by Landmark/Monuments signifying a place for tourist interest. Like Cluster Type 1, Cluster Type 3 was a representative of the whole borough of Queens and most of the venues were fast food places and restaurants.



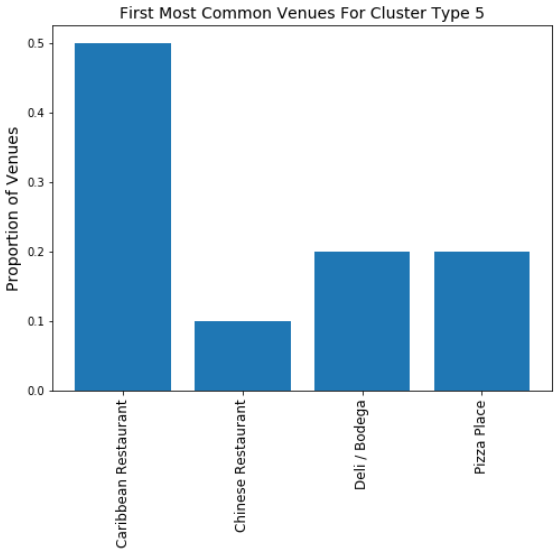


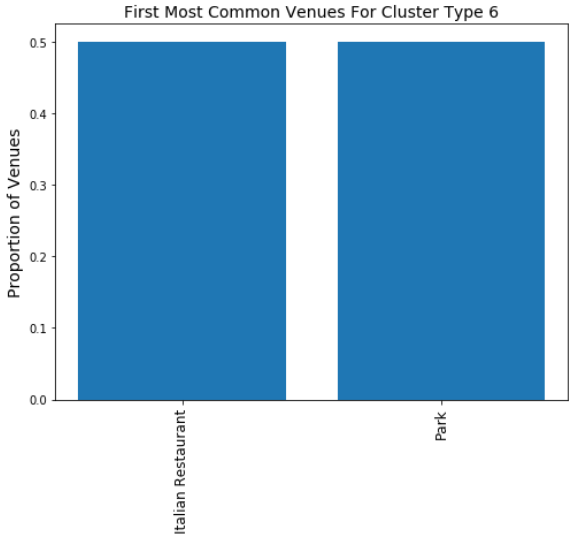
It appeared that Cluster Types 4 and 7 were similar because the most common place for both was beach. Again both of them were comprised neighborhoods that were found only in Queens and would be a nice place for vacation and recreation purposes. However, these clusters were differently grouped because they had other venues that were not similar.



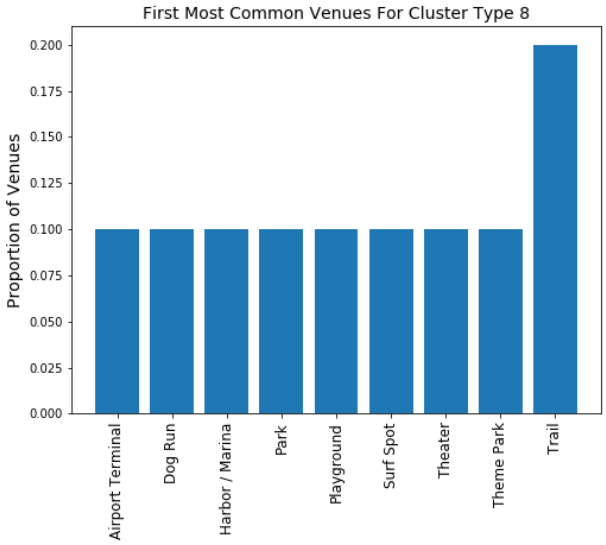


Although different, Cluster Types 5 and 6 comprised neighborhoods with lot of restaurants and fast food joints.





Lastly, Cluster Type 8 has venues for sports and recreational activities.



5. Conclusion

In this project, I compared neighborhoods in Toronto and Queens, New York to determine which neighborhoods were similar in terms of the venues they had. I identified that the mean number of neighborhoods in Toronto was higher than that of New York. Also, while the main venues for Toronto were coffee shops and cafes, those for Queens, New York were Pizza place and Deli representing fast food places. Segmenting the neighborhoods into nine clusters, I discovered that Toronto and Queens both had neighborhoods that shared similar neighborhood structures in 6 out of nine clusters. This project showed that clustering models and location data can be used to determine places that share similar venue categories. This will be useful for people who are moving from one place to another and wanting either something similar to or different from their previous locations.

6. Future Direction

In the future, it will be interesting to include other information on these neighborhoods such as climate, socioeconomic status, crime rates, prices of housing etc. This will enrich the model and discover interesting similarities between the neighborhoods.