Introduction

In this age of convenient and efficient transportation means, travelling around the world becomes easier and more common. In fact, people relocate around the world more often too with the globalization trend.

When people move from one city to another, most simply pick an area/neighborhood near work/school first, then move to a different area/neighbor if desired after settling down and getting to know the surrounding areas more. Moving, sometimes several times, isn't unusual in this scenario. It is obviously not an optimal process since we are limited by the scope of information we have access to, often via word of mouth or physically checking out a few nearby areas. Moving, also brings anxiety for the common fear of unknown. It would be helpful if we are able to compare the new city to the city we currently live in to identify areas that might be a good living location for us in the new city, and be better prepared by understanding the difference in advance. It would potentially minimize the need or frequency of relocations, which is a big hassle for those relocating with a family, and ease the mental burden of relocation.

However, efforts to research a new city often only offer the city-level information, for example, population, histories, economic condition, climates, etc. It gives the overall picture of the city, which is more suitable for tourists but not for selecting an area for living. For the latter purpose, it is the neighborhood-level details, such as what kind of shops, entertaining facilities, athletic centers and schools nearby, that are the focus of considerations.

Fortunately, with the advancement in technology, there are many location data platforms like **Foursquare** that provide detailed information on all kinds of venues around any geographical locations of interest. In this project, we intend to marry the rich location data provided by **Foursquare** and the power of machine learning to undertake comparisons of neighborhoods in two (or more) cities to fill this void of comparative information at the neighborhood level. We hope to help making relocation an easier and better experience with this additional dimension of information (packaged in a tool, if turn into a future App). Neighborhood-level comparisons among cities could also be valuable for people exploring and searching for their next stop (city) in life. For this project, we will focus on comparing the neighborhoods in New York City and Toronto as an illustrative example. The objective will be to provide a summary of how different/similar the two cities are based on their neighborhoods, as well as to offer recommendations for relocating between the two cities.

Data

In order to use the **Foursquare** platform to gather neighborhood venue information, we need data that contains the neighborhoods exist in each city as well as the latitude and longitude coordinates of each neighborhood.

We will first install packages and load the necessary libraries. The codes in the cell below was run twice. First time included the installation of packages geopy and folium, which took a long time and generated a lot of distracting outputs. The second time were run with the two installation lines commented out to hide the outputs from installation.

1. Load and extract neighborhood data for New York City

For New York city, this data had been compiled and exists in one file at https://cocl.us/new_york_dataset for this IBM course. The original source of this data is

from https://geo.nyu.edu/catalog/nyu_2451_34572. There are 306 neighborhoods in New York City. Below is an example of five records in this data.

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

2. Load and create neighborhood data for Toronto

For Toronto, the list of neighborhood and corresponding postal code will be scraped from this Wikipedia page https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. Even though we could use geocoder Python package to retrieve the geographical location data based on postal codes, this package is unreliable (could get stuck in the process for unreasonably long time if using a while loop to ensure getting a result for each postal code). So, we will use the csv file containing the geographical location data for each of the postal code in Toronto that is provided for this IBM course at http://cocl.us/Geospatial_data. There are 103 neighborhoods (rows) in the final data, with an example below showing 5 records.

	Postal Code	Borough	Neighborhood	Latitude	Longitude	City
0	M3A	North York	Parkwoods	43.753259	-79.329656	Toronto
1	M4A	North York	Victoria Village	43.725882	-79.315572	Toronto
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	Toronto
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	Toronto
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Toronto

3. Using Foursquare API to retrieve the venues data for neighborhoods

Once we have the neighborhood data with the appropriate geographical data for both cities, we will then use the **Foursquare API** to retrieve the venues information within a certain range of the radius (say, 500 or 1000 meters, we use 500 meters in this project) for each neighborhood. Service and activity venues, nearby within a neighborhood, are characteristics of a neighborhood and reflect the convenience and life style of people living in the area. Hence, quantifying these venues into categories and the associated venue counts are meaningful features to use for classifying neighborhoods into clusters/groups. Because our purpose is to compare the two cities, we will compile a combined data set for clustering analysis based on neighborhood venue features, and then examine the distribution of the clusters/groups between the two cities.

When exploring the data extracted from **Foursquare API**, we realized that some neighborhood names are not unique in the combined neighborhood data from the previous two steps. Some names are used by two neighborhoods from different Boroughs, or some neighborhoods are associated with more than one pair of latitude and longitude coding. Because our **Foursquare API** requests are based on each pair of (latitude, longitude) in the combined neighborhood data, for later analysis, we will include the latitude and longitude information as grouping factors for rolling up data to neighborhood level, being aware that Don Mills and Downsview, from Toronto, are represented by 2 and 4, respectively, sub-locations in the data. Below is an example showing 5 rows of the extracted venues data from **Foursquare API**.

	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New York City	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	New York City	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	New York City	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	New York City	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	New York City	Wakefield	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station

Methodology

Having data containing venues retrieved from **Foursquare API**, one natural question to ask is how many venues each neighborhood has. The total number of venues could be used as an indicator for how busy a neighborhood is. In addition, the data also assign venues to a category. Hence, we could also count the unique venue categories in each neighborhood. This measure reflects the diversity of services/activities provided in the area and could be used as a surrogate for convenience of life in the neighborhood. We are going to roll the per-venue-per-row data up to per neighborhood level with these two measures. We'll use visualizations to compare the two cities with regard to these two overall neighborhood measures, with super-imposed histograms and scatter plot.

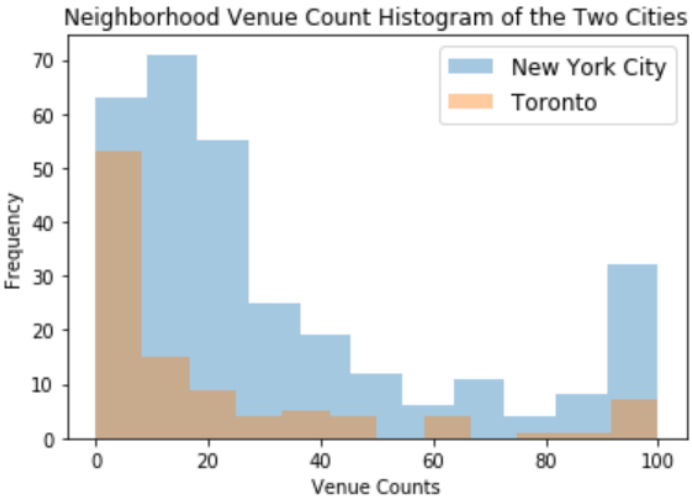
The number of venues in a category is a feature that provides a clue about a neighborhood. We decide to use the venue counts in all of the 458 unique categories as features for clustering analysis to segment the neighborhoods from the two cities.

1. We'll use one-hot coding to turn the venue categories into column features.
2. Roll up the data to get the venue counts in each category at the neighborhood level.
3. Use the elbow-method, Silhouette-method and Davies-Bouldin score to help determine the number of clusters for k-means clustering analysis.
4. With selected cluster number, run k-means clustering to assign neighborhoods into clusters.

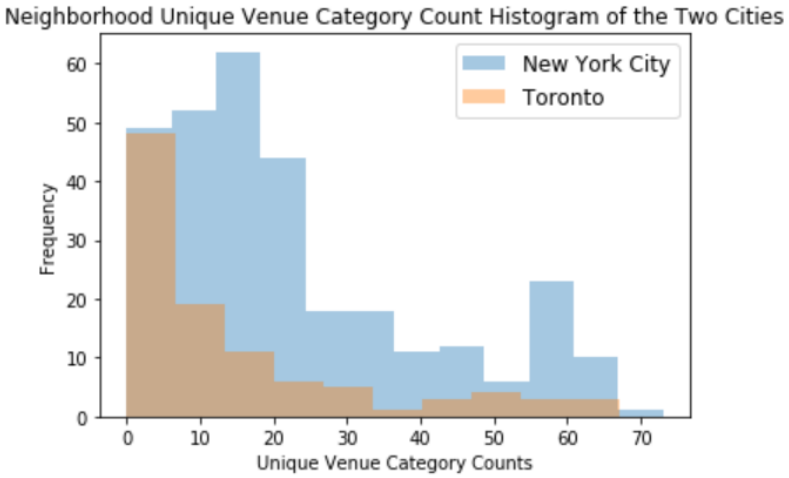
We will then examine the clusters based on the total number of venues and the unique venue category counts in a neighborhood to gain some understanding of each cluster. The cluster distribution between the two cities will be compared using cross tabulation with chi-squared test to study whether there is a difference. A side-by-side bar plot will also be provided to assist visual comparison. Finally, the maps of the two cities will be presented using folium map with neighborhoods superimposed as circle marker colored according to their designated clusters.

Results

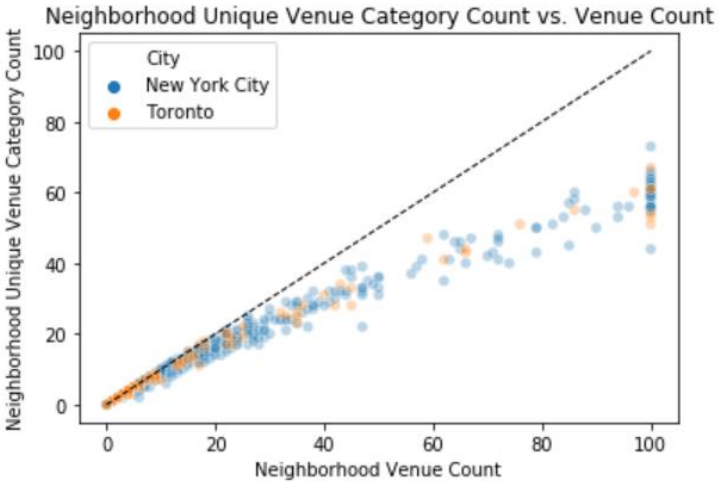
It is no surprise that New York City has more neighborhoods than Toronto across the whole spectrum of venue counts or unique venue category counts, since it has about 3 times more neighborhoods (Figure 1A-B). However, the histogram patterns follow a somewhat similar lopsided U-shape for both cities, with the number of neighborhoods highest for low venue counts/unique venue category counts. The number of neighborhood drops with increasing venue counts/unique venue category counts then turns upward a little at very high venue counts/unique venue category counts. The scatter plot between the venue count and unique venue category count shows that these two measures are quite close at lower venue counts but the gap between them (vertical distance from the diagonal line) grows larger with increasing venue count, indicating more duplicated venues from some venue categories. The neighborhoods from the two cities are relatively well mixed across the whole spectrum of these two measures in the scatter plot. Hence, the two cities are not too different in the sense that we are likely to find neighborhoods of certain characteristics from both cities most of the time.



A.



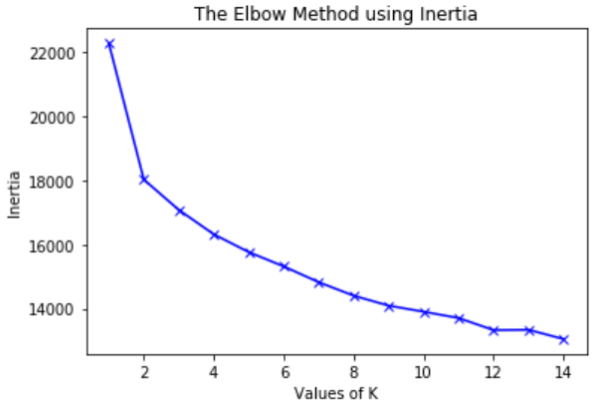
B.



C.

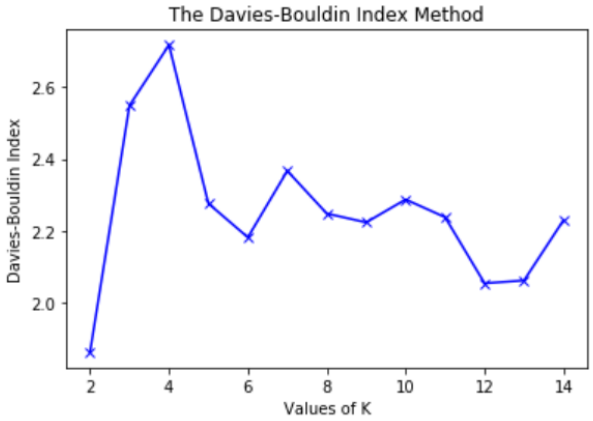
Figure 1. Histogram of venue count (A) and unique venue category count (B) across the neighborhood by city, as well as a scatter plot of these two measures (C). Dash line in the scatter plot is the diagonal line where the two measures equal to each other.

There isn't a clear "elbow" when using the elbow-method to help determine the optimal cluster number to use for k-means cluster analysis (Figure 2). It seems like 3-6 clusters could be used. While Silhouette-method suggests 3 clusters, the Davies-Bouldin score suggests 4 clusters. When using 3 clusters, it results in a very large cluster (including over 300 neighborhoods) and 2 rather small clusters. When using 5 or 6 clusters, there is a cluster of only 2 neighborhoods. With 4 clusters, we have better sized clusters that are not super large (>300) nor super small (<5). Therefore, we decide to use 4 clusters.



A.

B.

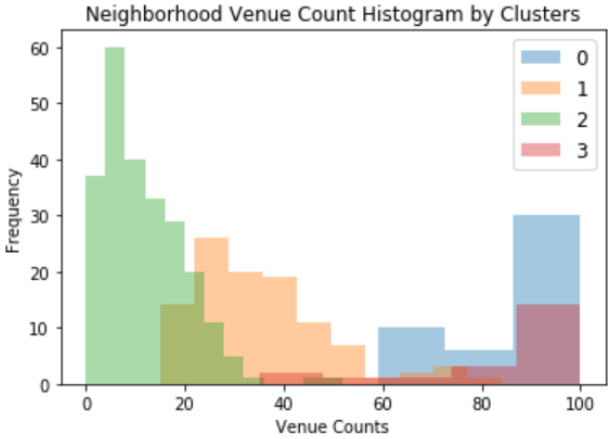


C.

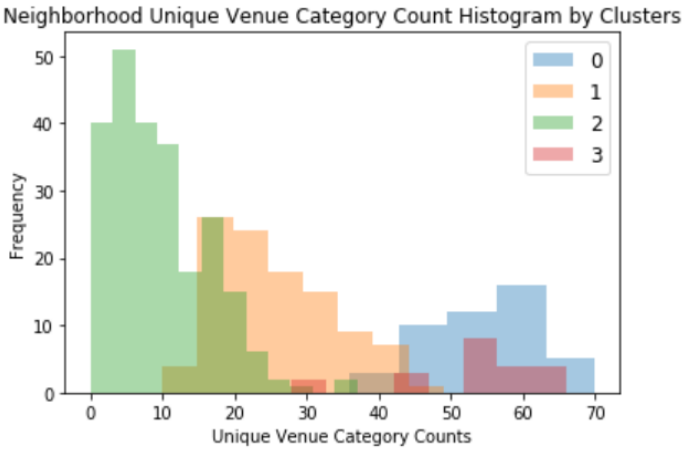
Figure 2. Elbow-method (A), Silhouette-method (B), and Davies-Bouldin score (C) to show the inertia of clustering with different number of clusters (values of K) for k-means clustering analysis.

This results in 46,104, 238, and 21 neighborhoods in clusters 0, 1, 2, and 3. Further examining the clusters and their venue count (Figure 3A) and unique venue category count distribution (Figure 3B), we could describe the clusters as below. At this level of details, we are not sure what is the difference between the two high venue counts/high unique venue category counts clusters (cluster 0 and 3).

- Cluster 0: Most with high venue counts (>60) and high unique venue category counts (>40)
- Cluster 1: Most with medium venue counts (20-60) and medium unique venue category counts (20-40)
- Cluster 2: Most with low venue counts (<20) and low unique venue category counts (<20)
- Cluster 3: Most with high venue counts (>60) and high unique venue category counts (>40)



A.



B.

Figure 3. Histogram of neighborhood venue count and unique venue category count by neighborhood cluster.

Table 1 is the cross tabulation of the clusters against the cities. Chi-squared test based on this table indicates there is difference in the neighborhood cluster distribution between the two cities ($p<0.001$). From Figure 4, Toronto has higher proportion of neighborhoods in cluster 2 (the low venue counts/low unique venue category count cluster) than New York City, while has lower proportion of neighborhoods in cluster 1 (the medium venue counts/medium unique venue category counts cluster). For the two clusters with high venue counts/high unique venue category counts, i.e. clusters 0 and 3, we noticed the difference between the two clusters could maybe be explained by the city. Cluster 0 could be called “unique NY style high venue counts/high unique venue category

counts neighborhood", since neighborhoods in cluster 0 are almost exclusively from New York. While cluster 3 contains more general high venue counts/high unique venue category counts neighborhoods. However, both cities have at least one neighborhood in each cluster, making it possible to find a target neighborhood group transitioning from one city to the other.

Table 1. Cross tabulation of neighborhood cluster distribution between New York City and Toronto.

City	New York City	Toronto
Cluster Labels		
0	45	1
1	93	11
2	160	78
3	8	13

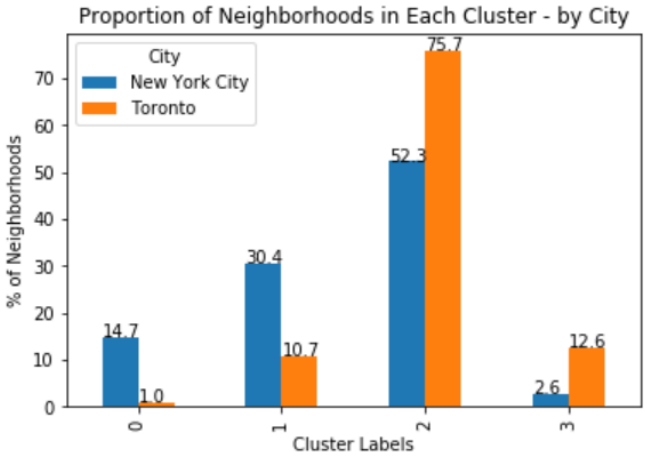
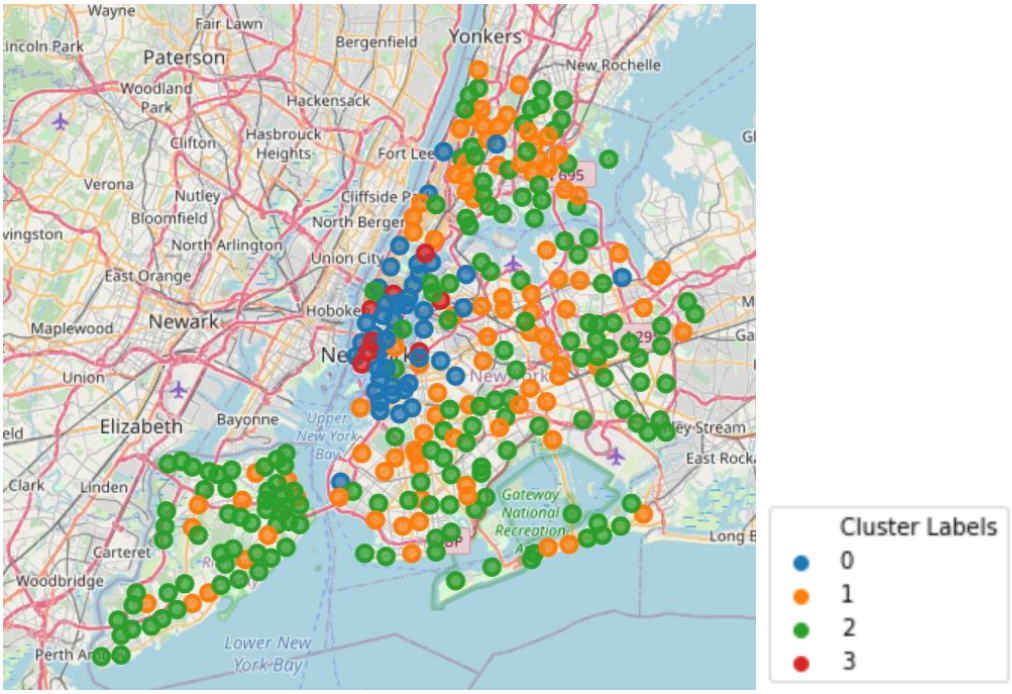
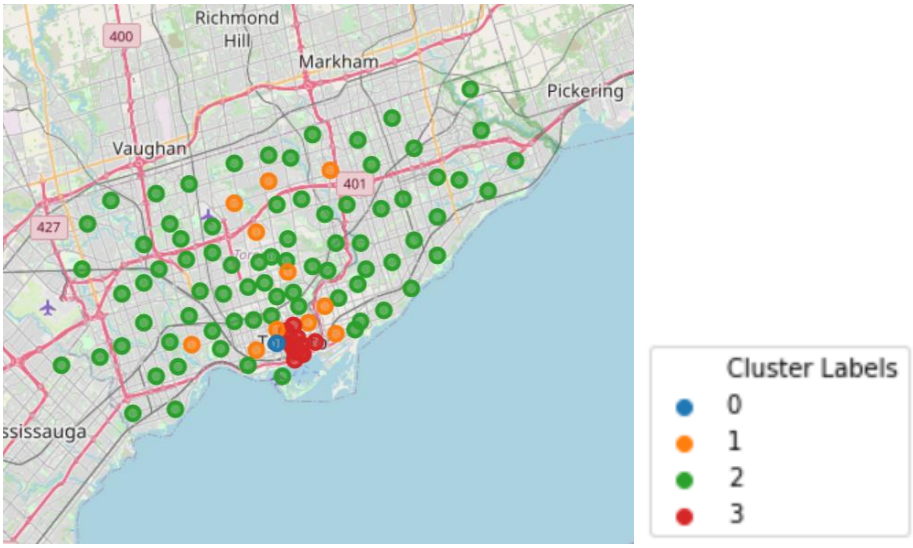


Figure 4. Proportion of neighborhoods in each cluster by city.

The two city maps with neighborhoods superimposed as markers color-coded by the cluster they belong to could help geographically visualize the clusters (Figure 5). We could see another similarity between the two cities from the two maps that high (blue and red markers) venue counts/high unique venue category counts neighborhoods tend to aggregate around the city center (Manhattan borough for New York City and Downtown Toronto borough for Toronto), and medium (orange markers) and low (green markers) venue counts/unique venue category counts neighborhoods spread out and are farther away from the city center. For people who are relocating between the two cities, they could easily tell what options they have if they know the ideal neighborhoods in their current city.



A.



B.

Figure 5. City maps for New York City (A) and Toronto (B) with neighborhoods superimposed as markers and colored according to their clusters.

Discussions

There are room for improvement in the approaches to compare the neighborhoods in the two cities.

When interacting with **Foursquare API**, we could change the radius of the search. For people who have cars and don't mind driving a lot, they could expand the search radius. But, for people who are environmentally sensitive and avoid driving as much as possible, maybe a narrower search radius will be more appropriate. We also see that some neighborhoods maxed up the 100 venue retrieve

limit, and could use a larger number to more completely capture the true number of venues in those high venue count neighborhoods.

In addition, we could use the venue data differently, say, only clustering the neighborhoods based on the top 100 most retrieved venue categories, since those are most likely related to issues/concerns of day-to-day life. The other approach will be clustering the data based on certain focused venue categories, such as work-out facilities, education related venues, healthcare focused venues or a combination of some focused categories.

As we see from the cross tabulation (Table 1), there is only one cluster 0 neighborhood in Toronto. If that is the target cluster, it is very limited. One could expand the search into cluster 3, knowing they are both high venue count/high unique venue category count type of neighborhoods. This will change the recommended number of neighborhoods from 1 to 14. We could further narrow it down to a number more manageable (between 5-10) by imposing some restrictions such as only recommend those neighborhoods with parks/playgrounds if the relocating family has young children; or more than one gym/yoga venues if the relocating person is very active and likes to sample before settling down with a workout studio.

There are, however, a large number of neighborhoods in some clusters for a city. Options to narrow it down could be both imposing restrictions or repeating the clustering analysis on the specific cluster subset across both cities. The process could continue until reaching a reasonable number of recommended neighborhoods.

Last, other un-supervised machine learning techniques other than k-means clustering could be used for this purpose and extra data with other types of details (such as population density, race composition, education resources) could be incorporated.

Conclusions

Indeed, New York City and Toronto are different with regard to their neighborhoods. New York City has more neighborhoods and its neighborhoods are more densely packed. Using the venue information extracted via **Foursquare API** for each neighborhood, we cluster the neighborhoods into 4 clusters. The distribution of the neighborhoods in the 4 clusters are different between the two cities as well, with Toronto having larger proportion of neighborhoods in the low venue count/low unique venue category count cluster. Also, among high venue count/high unique venue category count neighborhoods, most of the New York City neighborhoods segregate into a cluster that only contain one neighborhood from Toronto, demonstrating a unique neighborhood cluster that seems to be New York City specific.

However, we also observe that the distribution of neighborhood venue count and unique venue category count follow similar patterns for both cities. The distributions of the clusters are similar for both cities too, showing high venue count/high unique venue category count neighborhoods aggregating around city center while the others spreading out and further away from the center. And, most importantly, there are neighborhoods in each cluster for both cities, meaning that one can always find a few target neighborhoods if they know their ideal neighborhood in their current city. Even though it might need a bit more further restrictions or subsequent clustering analysis to come up with a reasonable number of recommended neighborhoods.

In summary, New York City and Toronto are different in some ways but there are similarities between the two with regard to the day-to-day living in their neighborhoods. With this analysis, we demonstrate that one can find welcome and comfort relocating between the two cities with location data from platforms like **Foursquare** and the machine learning techniques to compare cities.