Applied Data Science Capstone Project

The Battle of the Neighborhoods

Finding the most suitable residential neighborhood for families with children in Toronto, Canada

Siew Cheng Aw

20 July 2020

Table of Contents

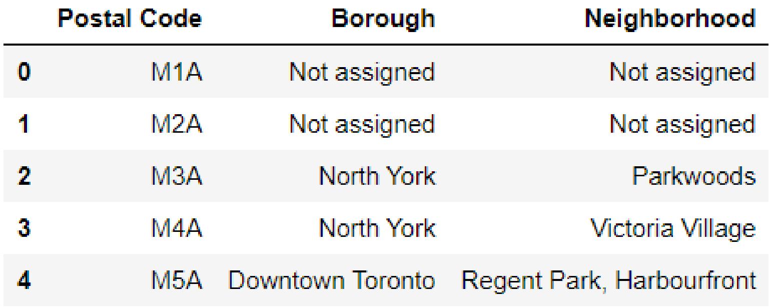
Introduction

Deciding where to stay can be challenging and especially so if you have children. Not only do you need to consider about your needs and wants, you also need to think about your children's education. Besides finding a location which has access to grocery stores and eateries, you would want to look for a neighborhood with schools which offer the best education for your children. This report aims to find the most suitable neighborhood for families with children and help them make better informed decision on where to live and in particular, in Toronto, Canada.

Data

This section lists the data which will be used for analysis in this report.

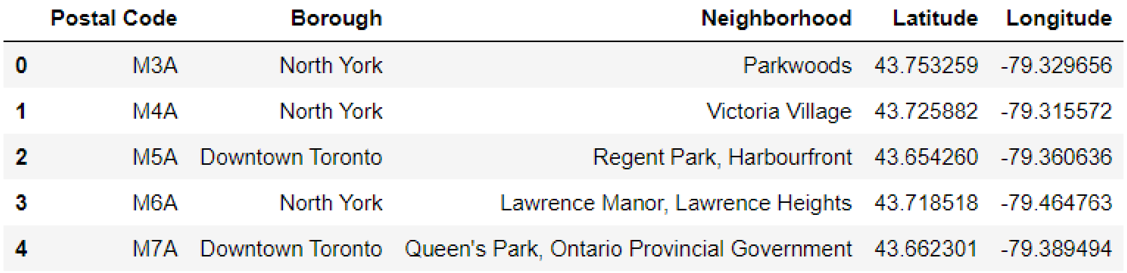
- A list of neighborhoods in Toronto, Canada which is taken from this website https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. It is stored in a dataset as follows:



	Postal Code	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

Figure 1: A sample of neighborhoods in Toronto, Canada with postal codes and boroughs

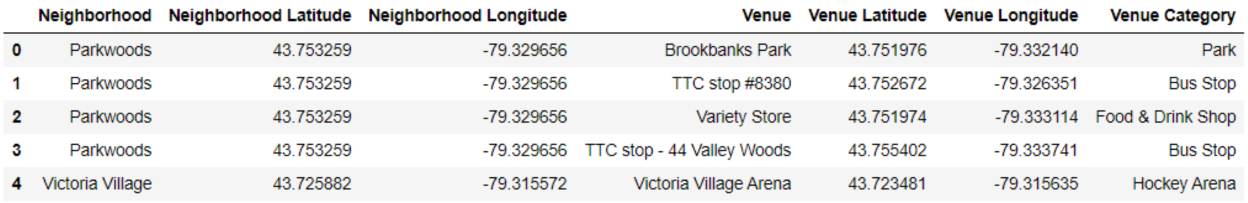
- Using GeoPy, a Python client, I get the latitude and longitude coordinates for each neighborhood and include them as columns to the dataset.



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

Figure 2: A sample of neighborhoods with latitude and longitude values

- I use the Foursquare API to get information of venues in each of the neighborhood and store them in a dataset.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	TTC stop #8380	43.752672	-79.326351	Bus Stop
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Parkwoods	43.753259	-79.329656	TTC stop - 44 Valley Woods	43.755402	-79.333741	Bus Stop
4	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena

Figure 3: A sample of neighborhoods with venue information

Methodology

Firstly, I needed to clean the dataset of all neighborhoods in Canada (Figure 1) scraped from Wikipedia, making sure that the postal codes are unique and boroughs and neighborhoods have valid values. Next, I added the latitude and longitude coordinates to the dataset with location values returned from GeoPy (Figure 2). With the latitude and longitude values for each neighborhood, I retrieved the information of a maximum of 100 venues within a radius of 500m from the coordinates for each neighborhood (Figure 3) with the Foursquare API and compiled them into a dataset.

With the derived dataset from above, I set out to explore the neighborhoods. To visualize the neighborhoods in Toronto on a map, I created a map of Toronto with the Folium library in Python and superimposed the locations of the neighborhoods in the dataset on the map. The locations are represented by the blue circles in the following map.

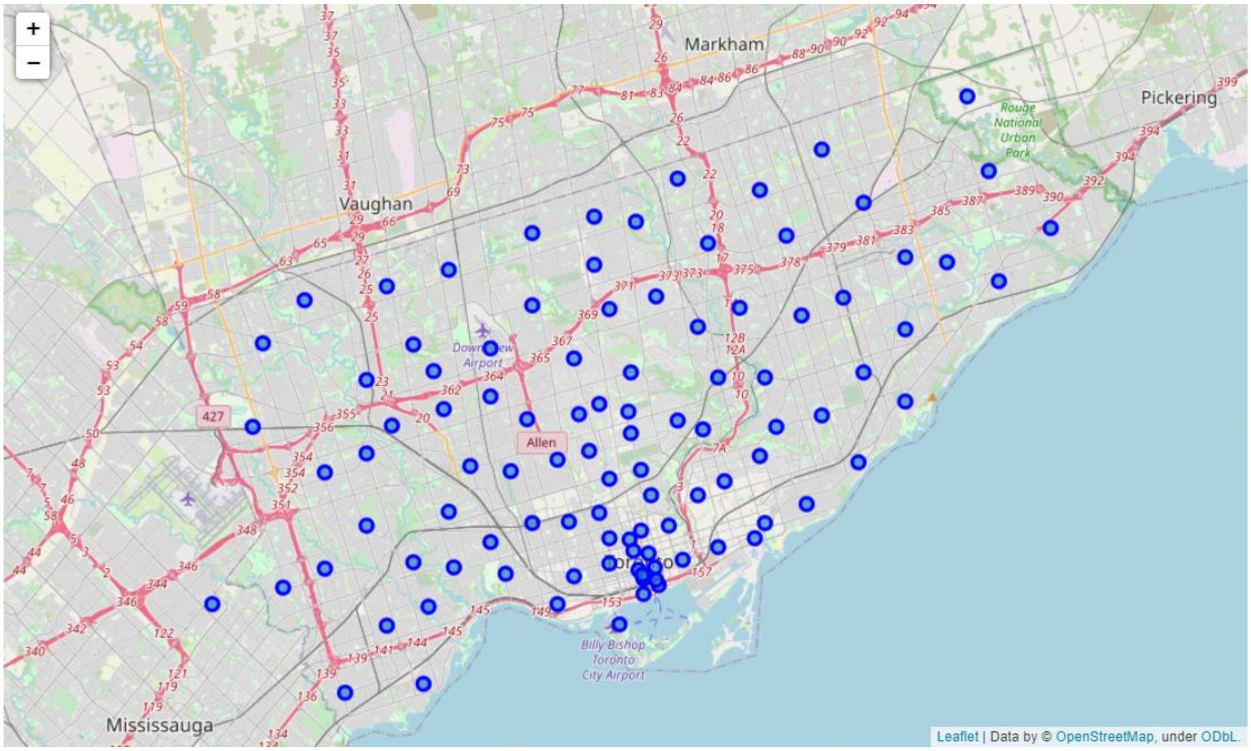

+
−
Markham
Pickering
Rouge National Urban Park
Vaughan
Downsview Airport
Allen
427
Toronto
Billy Bishop Toronto City Airport
Mississauga
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 4: A map of locations of neighborhoods in Toronto generated via Folium

I have charted the total number of venues and unique venue categories in each Toronto neighborhood in the bar chart below. Note that the maximum number of venues retrieved from Foursquare is limited at 100 per neighborhood and the venues are within 500m radius of the coordinates of the neighborhood. There are 10 boroughs and 99 neighborhoods in Toronto.

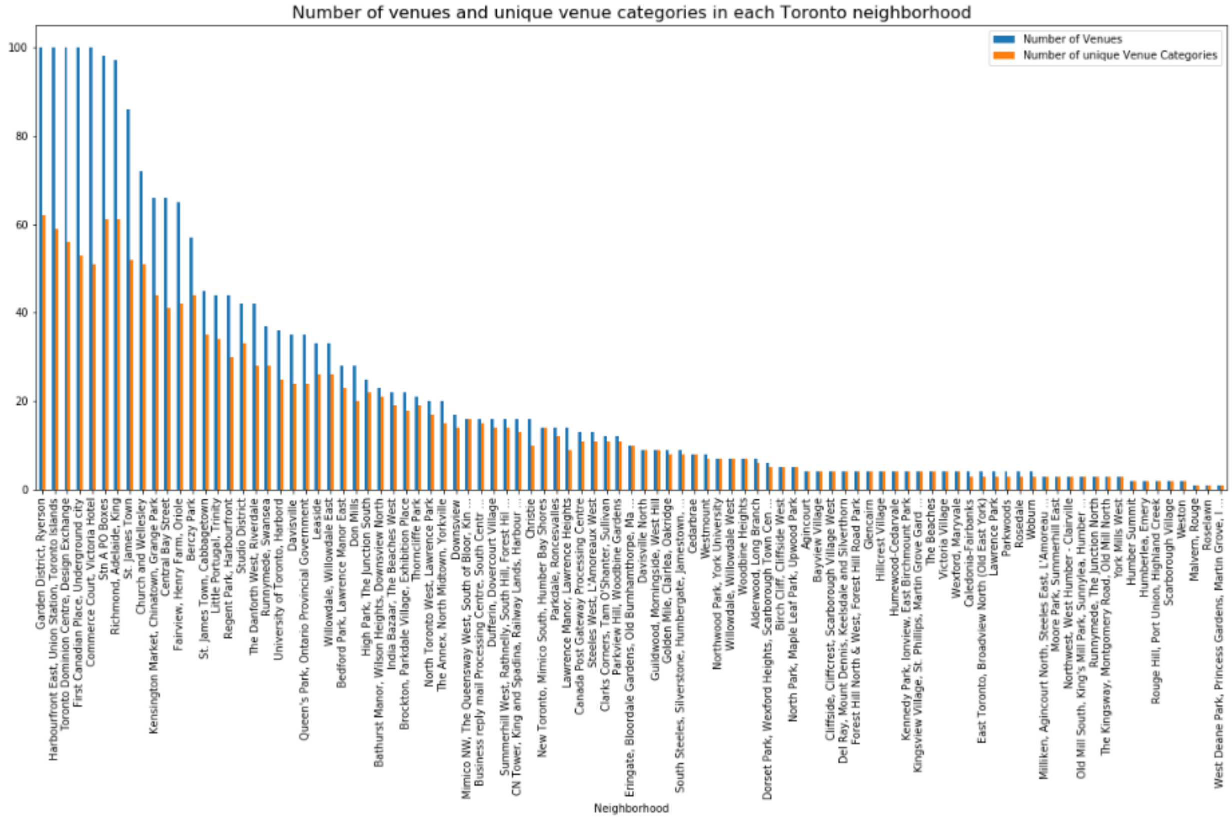
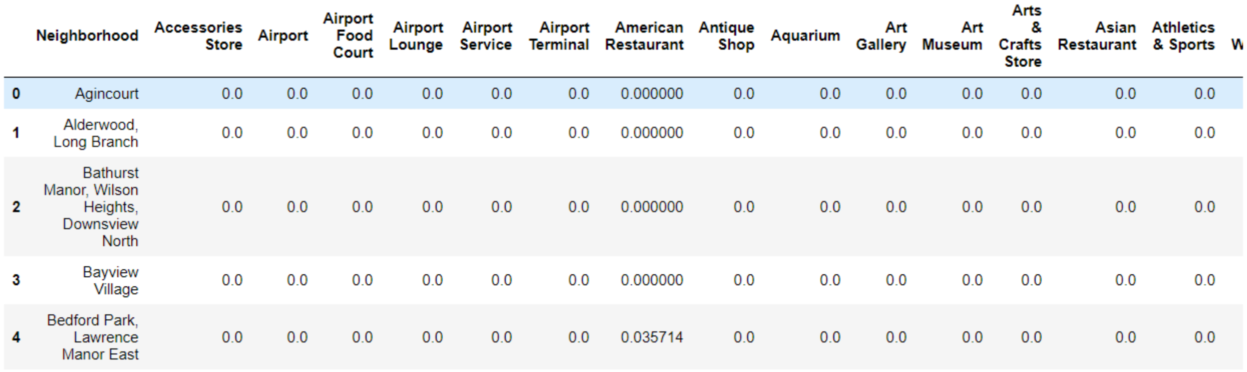


Figure 5: A bar chart showing the number of venues and unique venue categories in each neighborhood

There is a total of 274 unique venue categories in Toronto. I have removed neighborhoods which have less than 4 venue categories from the dataset as I found that the machine learning algorithm which I would be using in this report had better results when they are removed.

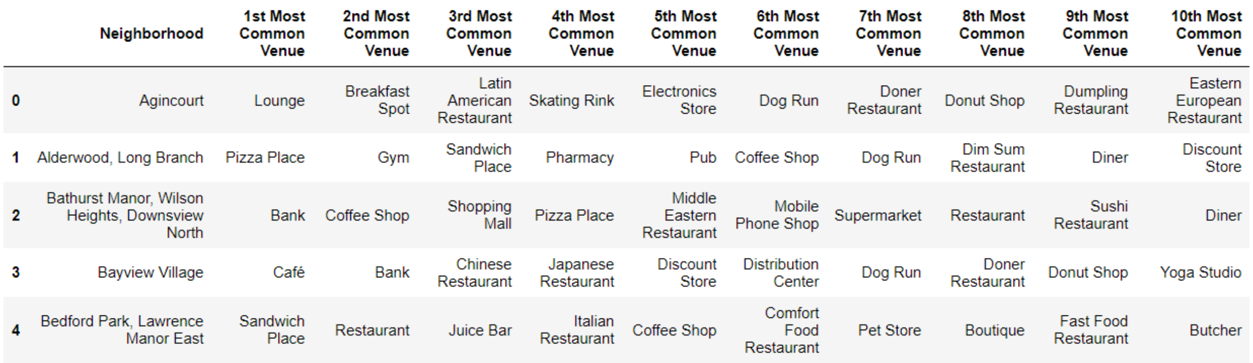
To find out if there is any natural grouping of neighborhoods based on the venues, I decided to perform cluster analysis using the K-Means clustering which is one of the popular unsupervised machine learning algorithms. To prepare the dataset for analysis, I used one hot encoding to split the categorical values in “Venue Category” in the dataset into multiple columns where each column represented a unique categorial value. I then calculated the mean of the frequency of occurrence of each venue category for each neighborhood.



	Neighborhood	Accessories Store	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	W
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.035714	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Figure 6: A sample of the dataset after one hot encoding and calculation of mean of frequency of occurrence of each venue category for each neighborhood

To have a better idea of what the above dataset represents for each neighborhood, I created another dataset using the dataset above and populated it with the top 10 venue categories for each neighborhood.



	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	Agincourt	Lounge	Breakfast Spot	Latin American Restaurant	Skating Rink	Electronics Store	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant
1	Alderwood, Long Branch	Pizza Place	Gym	Sandwich Place	Pharmacy	Pub	Coffee Shop	Dog Run	Dim Sum Restaurant	Diner	Discount Store
2	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Shopping Mall	Pizza Place	Middle Eastern Restaurant	Mobile Phone Shop	Supermarket	Restaurant	Sushi Restaurant	Diner
3	Bayview Village	Café	Bank	Chinese Restaurant	Japanese Restaurant	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Yoga Studio
4	Bedford Park, Lawrence Manor East	Sandwich Place	Restaurant	Juice Bar	Italian Restaurant	Coffee Shop	Comfort Food Restaurant	Pet Store	Boutique	Fast Food Restaurant	Butcher

Figure 7: A sample of the dataset which shows the top 10 venue categories for each neighborhood

Next, I proceeded to find the optimal number of clusters, K, of the K-Means using the Elbow method. With the Yellowbrick library in Python, the optimal K is found to be at 5.

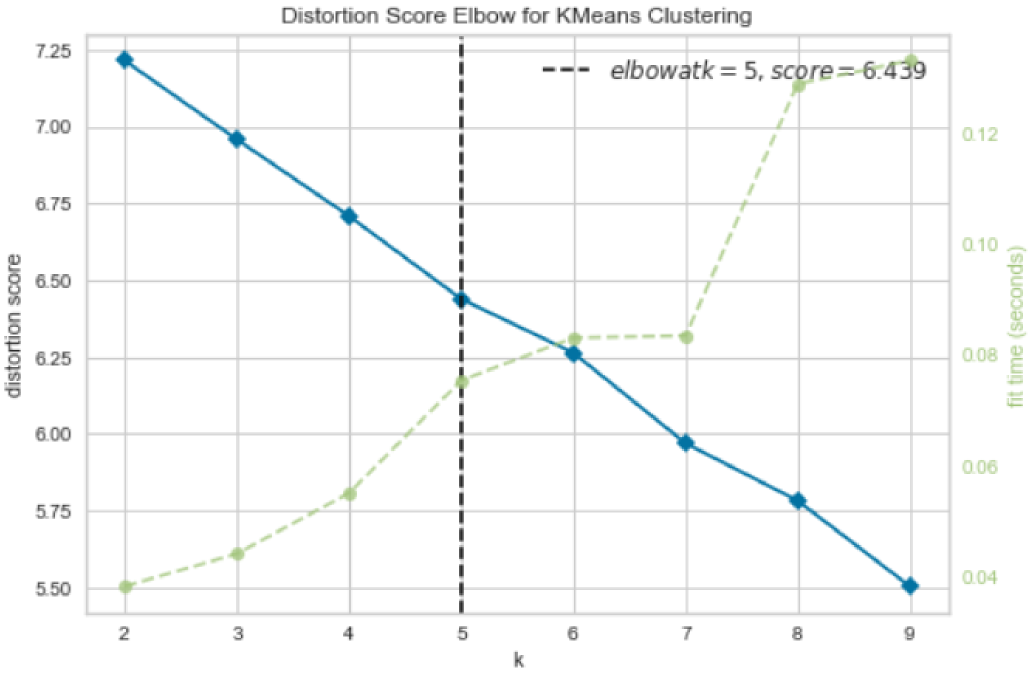


Figure 8: The Elbow method for finding the optimal k using the Yellowbrick library

Results

I ran the K-Means algorithm with K = 5 and added the cluster labels to the dataset in Figure 7 and plotted the clusters on the map with Folium as follows:

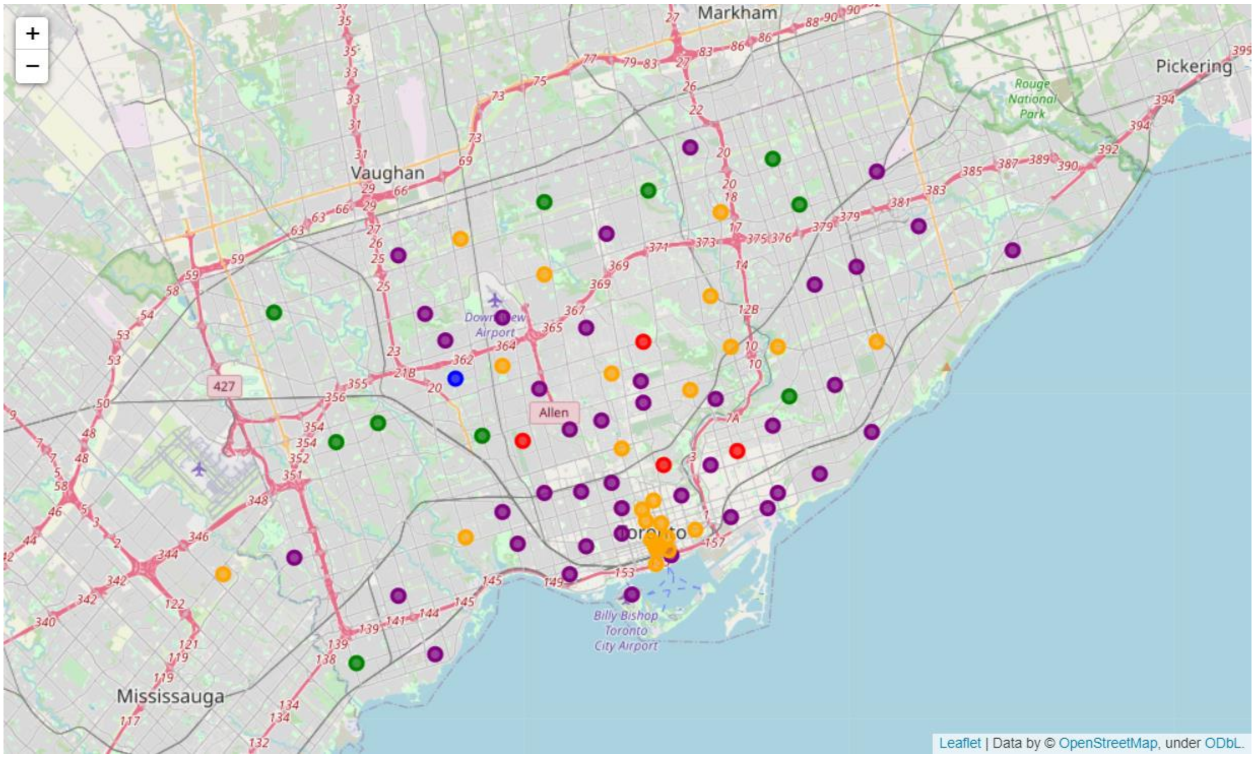


Figure 9: Visualization of clusters of neighborhoods

I explored the clusters to determine what defined the clusters.

- Cluster 0 – Represented by yellow circles in the above map (Figure 9)

There are 25 neighborhoods in Cluster 0.

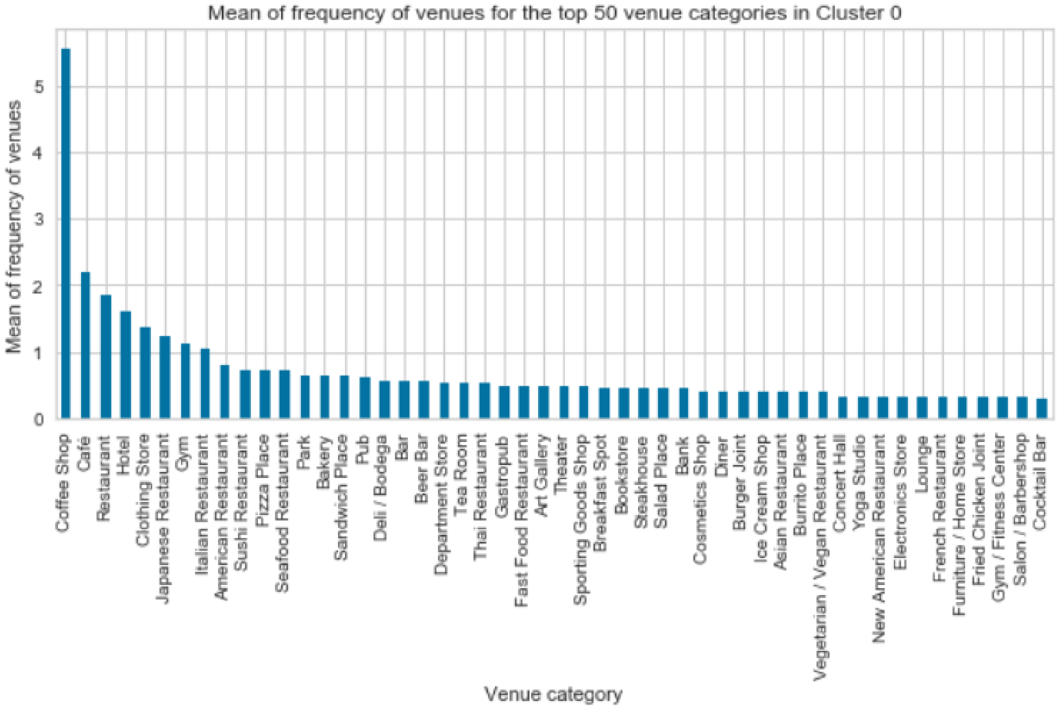


Figure 10: Mean of Frequency of venues for the top 50 venue categories in Cluster 0

Most of the Downtown Toronto neighborhoods fall in this cluster. The neighborhoods in this cluster have good access to coffee shops, cafes and restaurants which serve diverse types of cuisine. There are also hotels and retail stores. This cluster may be the shopping and dining area of Toronto. There could be residential areas in these neighborhoods though their property prices may be on the higher side.

- Cluster 1 – Represented by red circles in the above map (Figure 9)

  There are 4 neighborhoods in Cluster 1.

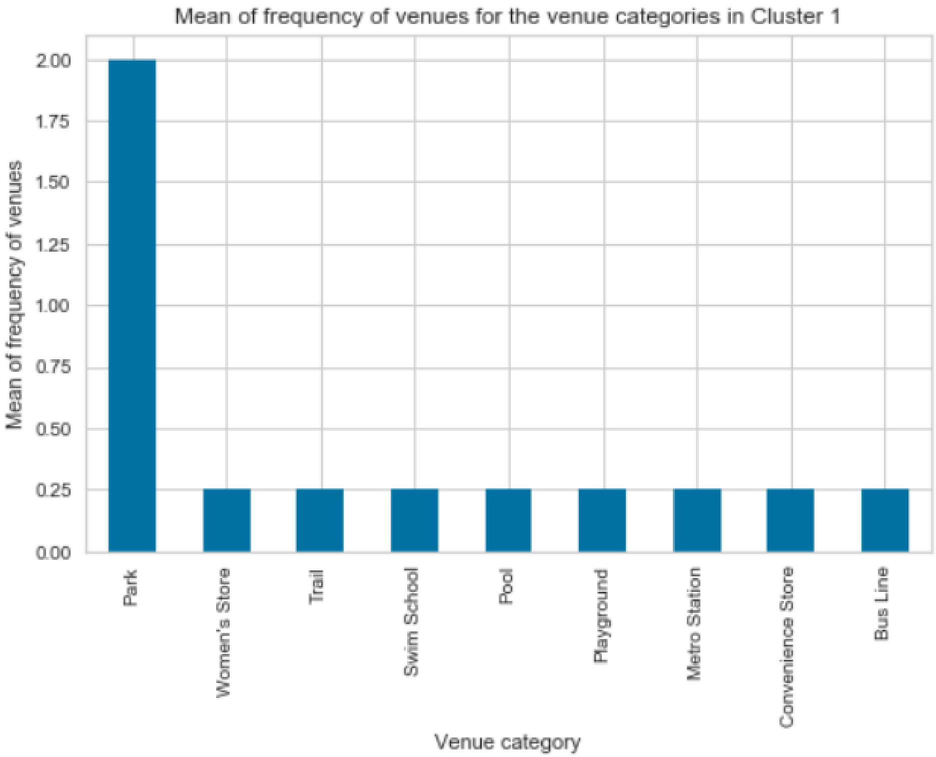


Figure 11: Mean of frequency of venues for the venue categories in Cluster 1

The neighborhoods in this cluster seem to be the more relaxed, green and quiet parts of Toronto with more parks, trails, playgrounds, pools available in this cluster compared to other clusters.

- Cluster 2 – Represented by blue circles in the above map (Figure 9)

  There is only 1 neighborhood in Cluster 2.

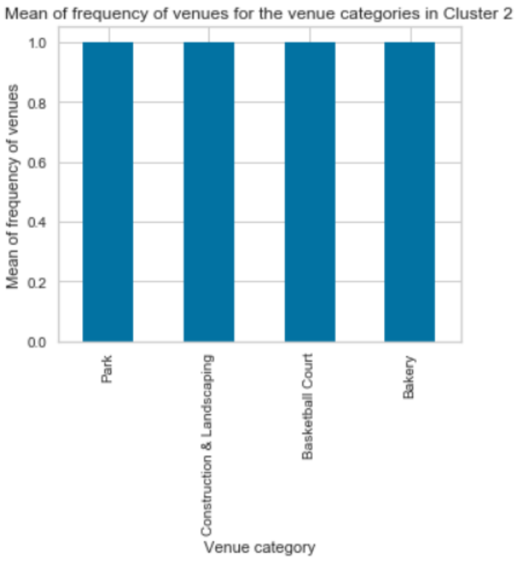


Figure 12: Mean of Frequency of venues for the venue categories in Cluster 2

This neighborhood contains a park, a bakery, a basketball court and a company.

- Cluster 3 – Represented by green circles in the above map (Figure 9)

  There are 10 neighborhoods in Cluster 3.

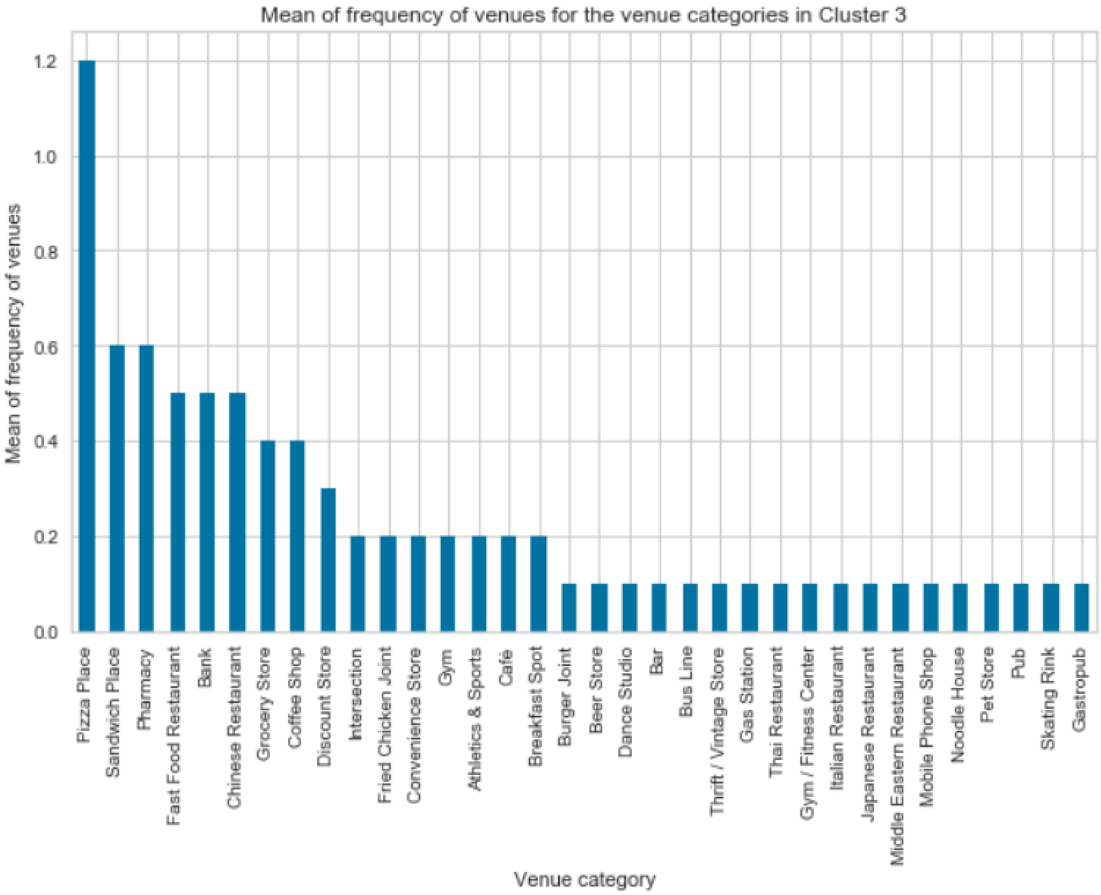


Figure 13: Mean of frequency of venues for the venue categories in Cluster 3

The neighborhoods in this cluster has good access to eateries, pharmacies, banks and grocery stores. This looks like residential areas in Toronto and as compared to Cluster 0, the neighborhoods are probably less dense due to the lower means of frequencies of the venue categories and the distances of the neighborhoods from Downtown Toronto.

- Cluster 4 – Represented by purple circles in the above map (Figure 9)

There are 41 neighborhoods in Cluster 4.

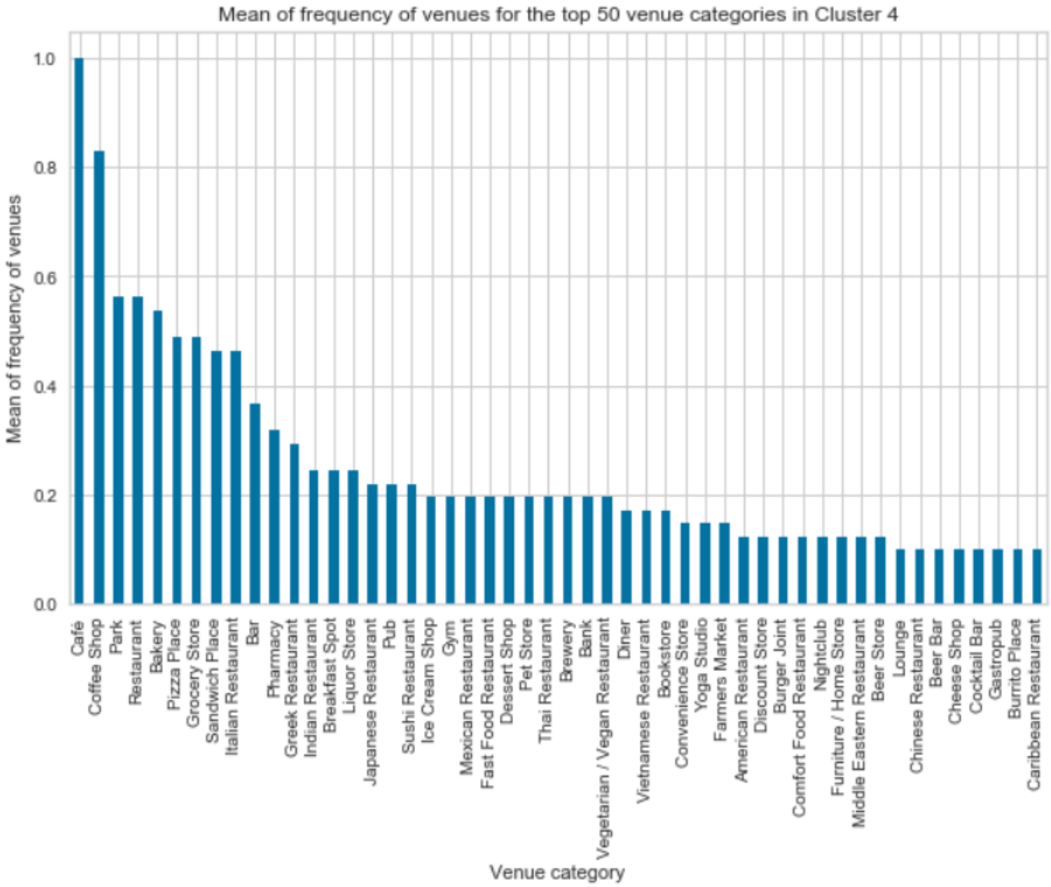


Figure 14: Mean frequency of venues in the top 50 venue categories in Cluster 4

The neighborhoods in this clusters have quite a number of cafes, coffee shops, parks and restaurants and other eateries. This cluster appears to contain neighborhoods where people meet and eat and could potentially be a vibrant residential cluster due to access to a variety of eateries.

Discussions

Based on the results from the clustering algorithm from the previous section, clusters 0, 3 or 4 may have the suitable neighborhood for families with children due to the access to eateries, stores and parks.

As the availability of good schools is important, I have compiled a list of the top public schools in Toronto based on this website https://justo.ca/blog/the-best-public-schools-in-the-city-of-toronto/ and superimposed the schools' locations on the map in Figure 9.

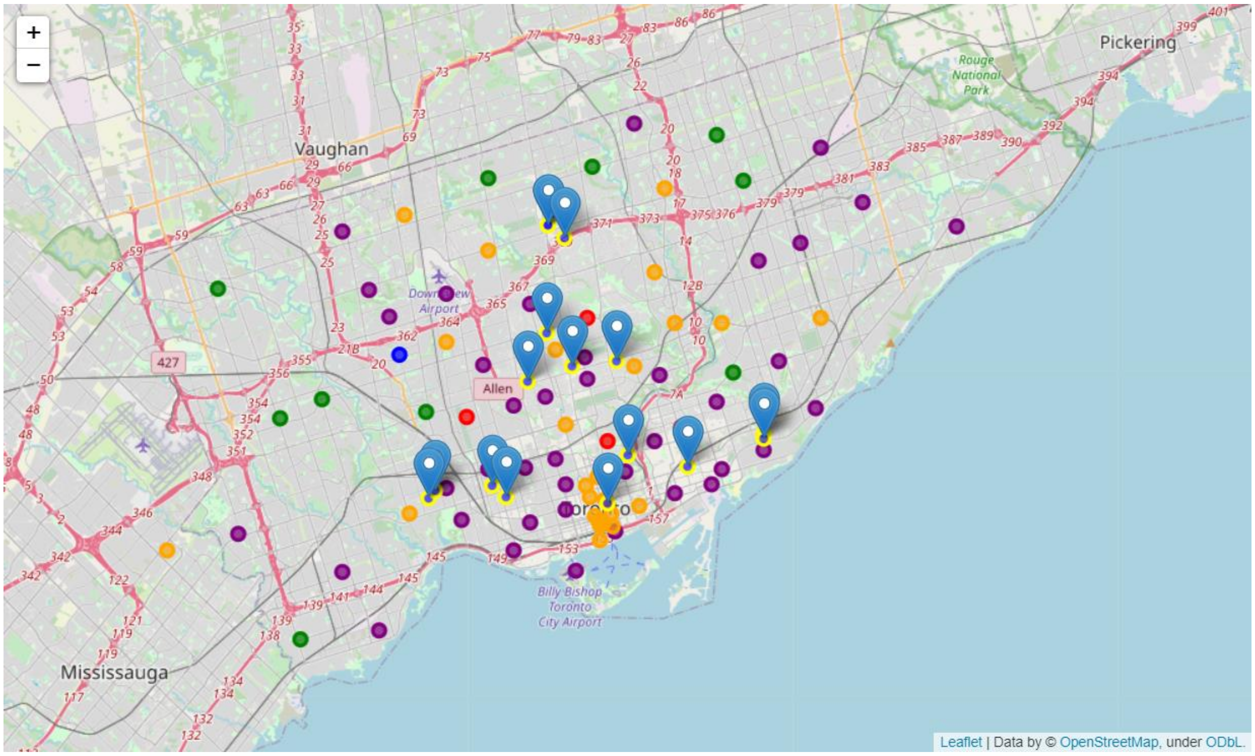


Figure 15: Locations of top schools in Toronto

The top public schools seem to be mostly near neighborhoods in cluster 0 and 4.

Conclusion

Based on the results and observations gathered so far, it seems that neighborhoods in Clusters 0 and 4 are potentially suitable neighborhoods to stay for families with children. However, there are a few other things to take note:

- The venues in the neighborhoods are based on Foursquare API, which means that if a venue exists only if a Foursquare account user recommends it. There are other venues which exist but not recommended in Foursquare. Hence, the information on the venues may not be conclusive.
- I have only explored the top public schools. Families may prefer private school or other public schools for their children.
- Some families may prefer neighborhoods that are quiet and closer to parks, trails and playgrounds. Hence, they may prefer neighborhoods in Cluster 1 to Clusters 0 and 4.

Some suggestions for future work to extend the analysis would be:

- Besides access to eateries, stores and schools, safety is also an important factor in deciding where to stay. Find data on crime rates in Toronto.
- It would be useful to find out the property prices in each neighborhood. Find data on property prices in Toronto.
- Population density of the neighborhood would give the readers of the report a better idea of the neighborhoods.