Identifying the clusters of Toronto for Real Estate Market

Vardges Bokhyan

May 2020

Contents

1. Introduction 2

1.1 Background 2

1.2 Problem 2

1.3 Stakeholders 2

2. Data 2

2.1 Toronto 2

2.2 Longitude and latitude 3

2.3 Forsquare 3

3. Methodology 3

3.1 Data Preparation 3

3.2 Exploratory Analysis 4

4. Results 9

Cluster 1 9

Cluster 2 10

Cluster 3 10

Cluster 4 10

Cluster 5 10

Cluster 6 10

5. Discussion 11

6. Conclusion 11

1. Introduction

1.1 Background

A startup company has decided to enter into the real estate market in the Canada. The company's CEO acknowledges the challenges and barriers his company is going to face since the sector is very saturated and many agencies already operate in the sector. However, he believes that a modern data-driven approach can help him to identify the gaps in the market. Once the gaps are identified, the company can enter the market by offering the best practice.

1.2 Problem

As possible locations for the initial operations the company has identified the most important city in Canada, Toronto. The city is very famous and attracts younger generation from nearby cities as well as from thousands of foreigners every year. Thus, the CEO of the startup, wants to offer divide the Toronto city based on the preferences of the incomers and younger families willing to buy or rent property in Toronto.

1.3 Stakeholders

This approach is going to be beneficial for both: the client is going to visit the places he/she likes the most, and the company can charge an extra fee for the personalized tour. So the company needs a clear mapping of the city in form of clusters, based on the venues available in different neighborhoods.

2. Data

To address the problem data from several sources should be acquired.

2.1 Toronto

The data from Wikipedia page is going to be used in the analysis. It contains information about postal codes, borough, and Neighborhoods. The postal codes beginning with 'M' are located within city of Toronto in the province of Ontario. The data consist of three columns and 103 rows. Data should be preprocessed before analyzing, in order to avoid any missing data in the table.

Table 1: The postal Codes of City of Toronto

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

2.2 Longitude and latitude

Another data source is going to be used to get the longitude and latitude of geographical regions in order to be able to plot the data.

2.3 Forsquare

Finally, the data should be analyzed by using the Forsquare API services. The data includes information on available restaurants, hotels, historical places, parks, theaters, art galleries, museums, as well as other venues.

3. Methodology

3.1 Data Preparation

The analysis are conducted by using such python libraries as pandas, numpy, matplotlib, folium. The analysis are conducted within Jupiter Notebook from Anaconda distribution.

Before starting the analysis it is crucial to process the data. Some of the steps of the processing are data cleaning, transforming, dealing with missing data, dealing with outliers.

The first step is to obtain the data. The data on Toronto's postal codes is obtained from the respective Wikipedia page by using Python pandas library. Pandas is arguably the most important python library for data science and is widely used to address any kind of problems in the data science field. After obtaining the data, next step is creating a data frame, which is pandas main data structure, and it makes any operation with the data much quicker and easier. Then, I checked for the missing data, and, as you can see from the Table 1 in the Data section, there is indeed missing data in Borough and Neighborhood columns. To address that issue, I filtered out the rows in the Borough column, which contain 'Not assigned'. Next, I grouped the Neighborhoods column by Postal code and Borough, in order to get the Neighborhoods with the same Postal code and Borough in the same raw, thus, getting rid of the duplicate values. This process resulted in a data frame with a shape of (103,3), that is, the table consists of 103 rows and 3 columns.

The next step is to get the corresponding longitude and latitude. This will allow plotting the results and implement the analysis. For this purpose, data called 'Geospatial coordinates' is imported into the Jupiter notebook. Then, the data is merged with the data frame of Toronto Neighborhoods. This enables us to make the first plot with the neighborhoods superimposed on the actual map of Toronto, in order to have a better understanding and overall picture of the region.

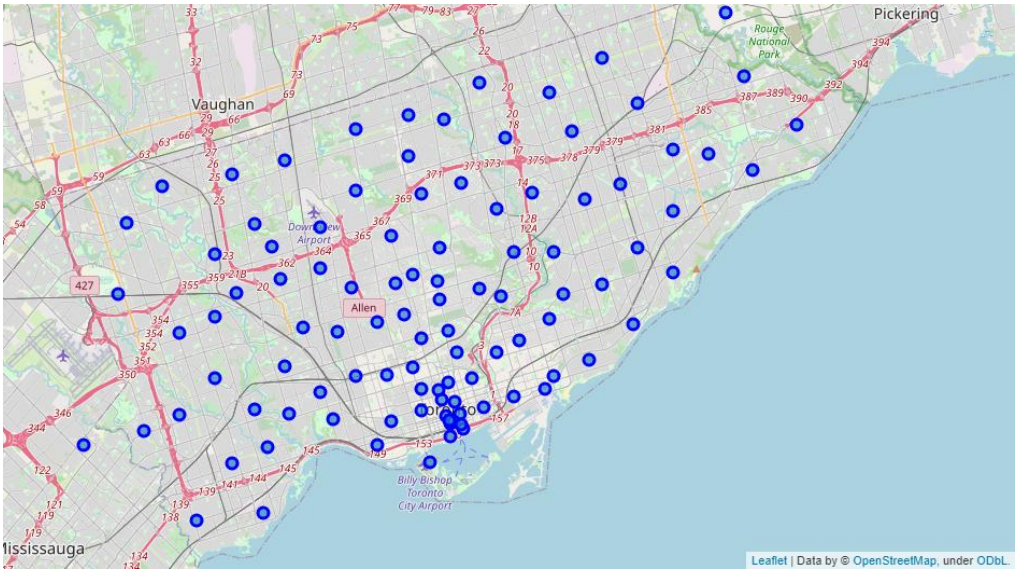

Figure 1: The map of Toronto with superimposed Neighborhoods

3.2 Exploratory Analysis

Since the overall picture is clear, one should dig deeper into the data and understand it better before conducting the actual analysis.

First, we should concentrate only on the neighborhoods within Toronto City, since the startup company has identified this region as the potential target for its business goals.

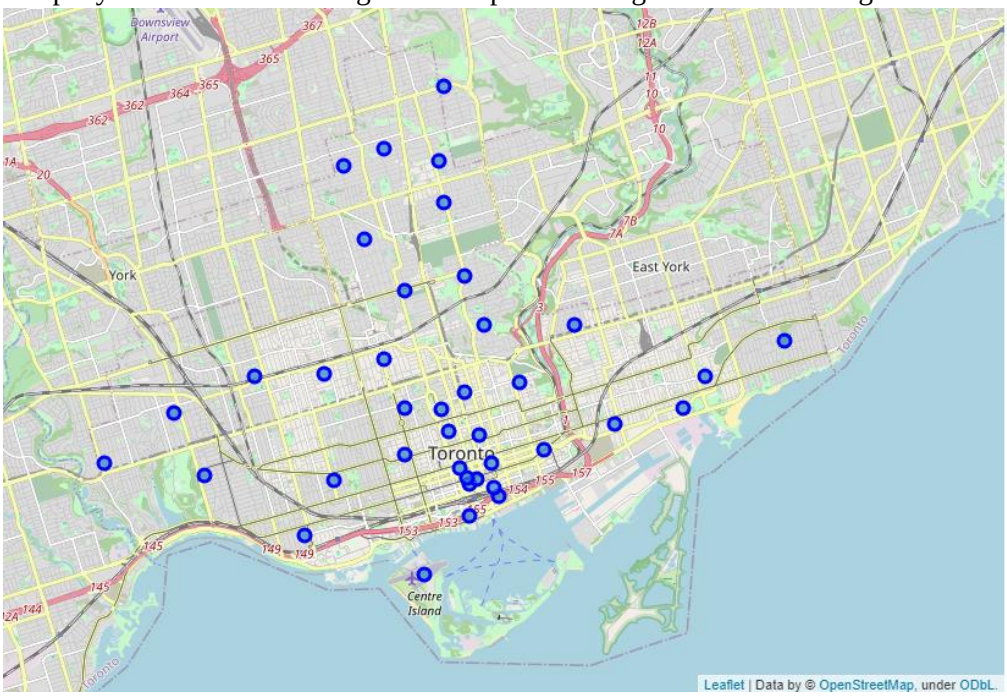

Figure 2: Neighborhoods of Toronto City

The next step is to acquire data on venues. This is done by getting data from Forsquare API, which helps to build location-aware maps. It contains data on available hotels, shops, restaurants, historical and cultural heritage, and generally anything that can be of clients' interest.

Now, when all the data needed is collected one can conduct some exploratory analysis. The table below represents the venues for all the neighborhoods of Toronto City. This table is constructed by extracts the category of the venue, cleaning the json and structuring it into a pandas data frame.

Table 2: Toronto City Neighborhoods and venues: top 5 results

	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

The table 2 is of shape of (1623, 7), that is, it consists of 1,623 rows and 7 columns. It represents the Neighborhoods, respective latitude and longitude of a neighborhood, name of the venue, category of the venue, as well as latitude and longitude of a venue. However, as the shape of the data frame hits us, it is a too large table to easily navigate through , so one alternative way to look at the data is by grouping by neighborhood the number of venues.

Table 3: The Count of Venues by Neighborhood

Neighborhood	*Venue*
Berczy Park	57
Brockton , Parkdale Village , Exhibition Place	23
Business reply mail Processing CentrE	18
CN Tower , King and Spadina , Railway Lands , Harbourfront West , Bathurst Quay , South Niagara , Island airport	17
Central Bay Street	63
Christie	17
Church and Wellesley	78
Commerce Court , Victoria Hotel	100
Davisville	36
Davisville North	8
Dufferin , Dovercourt Village	19
First Canadian Place , Underground city	100
Forest Hill North & West	5
Garden District, Ryerson	100
Harbourfront East , Union Station , Toronto Islands	100
High Park , The Junction South	23
India Bazaar , The Beaches West	20
Kensington Market , Chinatown , Grange Park	55

Lawrence Park	4
Little Portugal , Trinity	41
Moore Park , Summerhill East	1
North Toronto West	20
Parkdale , Roncesvalles	13
Queen's Park , Ontario Provincial Government	38
Regent Park , Harbourfront	48
Richmond , Adelaide , King	94
Rosedale	4
Roselawn	1
Runnymede , Swansea	39
St. James Town	77
St. James Town , Cabbagetown	46
Stn A PO Boxes	95
Studio District	41
Summerhill West , Rathnelly , South Hill , Forest Hill SE , Deer Park	17
The Annex , North Midtown , Yorkville	22
The Beaches	4
The Danforth West , Riverdale	43
Toronto Dominion Centre , Design Exchange	100
University of Toronto , Harbord	36

This table helps to get a quick understanding of the available number of venues within each of the neighborhoods. However, this table is not able to tell what kind of venues are exactly behind those numbers. To answer to that question, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category of venues.

Table 4: The frequency of occurrence of each category of venue by neighborhood

	Neighborhood	Yoga Studio	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Theme Restaurant	Toy / Game Store	Trail	Train Station
0	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00
1	Brockton , Parkdale Village , Exhibition Place	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00
2	Business reply mail Processing CentrE	0.055556	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00
3	CN Tower , King and Spadina , Railway Lands , ...	0.000000	0.058824	0.058824	0.058824	0.117647	0.176471	0.117647	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00
4	Central Bay Street	0.015873	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00
5	Christie	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00
6	Church and Wellesley	0.025641	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.012821	0.000000	...	0.012821	0.000000	0.000000	0.00

With this table, it is possible to figure out which neighborhood has a special type of venue and how many of them are there. This table is much easier to investigate, since it consist of only 39 rows and 231 columns.

It is possible to further simplify the results be print each neighborhood along with the venues. Below presented some of the neighborhoods along with top 5 most common venues.

Table 5: Some Neighborhoods along with the top 5 most common venues

```
----Berczy Park----
                venue  freq
0         Coffee Shop  0.07
1        Cocktail Bar  0.05
2                Café  0.04
3          Restaurant  0.04
4  Seafood Restaurant  0.04

----Harbourfront East , Union Station , Toronto Islands----
                venue  freq
0         Coffee Shop  0.13
1            Aquarium  0.05
2               Hotel  0.04
3                Café  0.04
4  Fried Chicken Joint  0.03

 ----Little Portugal , Trinity----
                          venue  freq
0                           Bar  0.10
1                    Restaurant  0.07
2               Asian Restaurant  0.05
3          Vietnamese Restaurant  0.05
4  Vegetarian / Vegan Restaurant  0.05
```

It is easy to see that the clients who love coffee or/and cocktails or/and seafood restaraunts most probably would be more satisfied by visiting the Berczy Park. While the ones who prefere aquariums and everything closer to the hotel would be better off by visting the Harbourfront East, Union Station, and Toronto Islands. For the real gurmans Little Portugal and Trinity offer a wide variety of restaraunts of different cusines of the world.

Finally, it is more user friendly to present the tables above in one joint table, which will allow the users not only to find the most suitable neigbhorhood for them, but also to comapre different neighborhoods, in order to be able to pick the one which will yield the best practice.

Table 6: Top 10 common venues by Neighborhoods

	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	Berczy Park	Coffee Shop	Cocktail Bar	Beer Bar	Cheese Shop	Restaurant	Café	Bakery	Seafood Restaurant	Bistro	Jazz Club
1	Brockton , Parkdale Village , Exhibition Place	Café	Coffee Shop	Nightclub	Breakfast Spot	Pet Store	Stadium	Intersection	Bakery	Italian Restaurant	Restaurant
2	Business reply mail Processing CentrE	Yoga Studio	Auto Workshop	Park	Comic Shop	Pizza Place	Recording Studio	Burrito Place	Restaurant	Brewery	Light Rail Station
3	CN Tower , King and Spadina , Railway Lands , ...	Airport Service	Airport Lounge	Airport Terminal	Sculpture Garden	Boutique	Rental Car Location	Plane	Boat or Ferry	Harbor / Marina	Bar
4	Central Bay Street	Coffee Shop	Café	Italian Restaurant	Sandwich Place	Ice Cream Shop	Middle Eastern Restaurant	Bar	Thai Restaurant	Burger Joint	Fried Chicken Joint

The table 6 allows to clearly see and get prior knowledge of the most common venues by neighborhoods. However, this table is still too large, and may be inconvenient to go through the whole table in order to come up with a conclusion. Thus, a K-Means clustering technique is applied in order to divide the Toronto City into clusters. This technique allows merging the similar neighborhoods into a single cluster by taking into account their similarities.

To do so, first of all, data on the longitude and latitude should be merged to the table 6. After merging two data frames, it is time to identify the most appropriate number of clusters. To address this problem, the elbow method is employed.

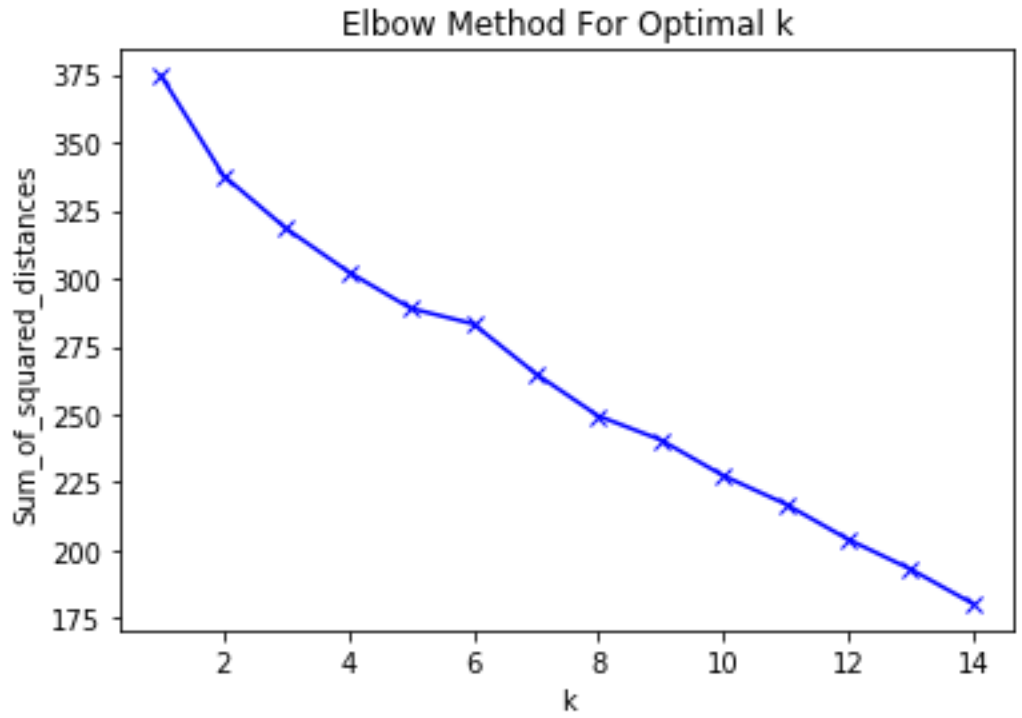

Figure 3: Elbow method for Optimal K

This methods consists of the following steps: first, the number of clusters is set to 1 and the sum squared distance is calculated, then the process is repeated for each following K. In our case the maximum number of K is 14. Then, after calculating the sum squared distances for each K, those numbers are plotted against each other, such that each K corresponds to its sum of square distance. Finally, the optimal number of K's is chosen by identifying the point on the line where the line is bent, in our case this number is equal to 6.

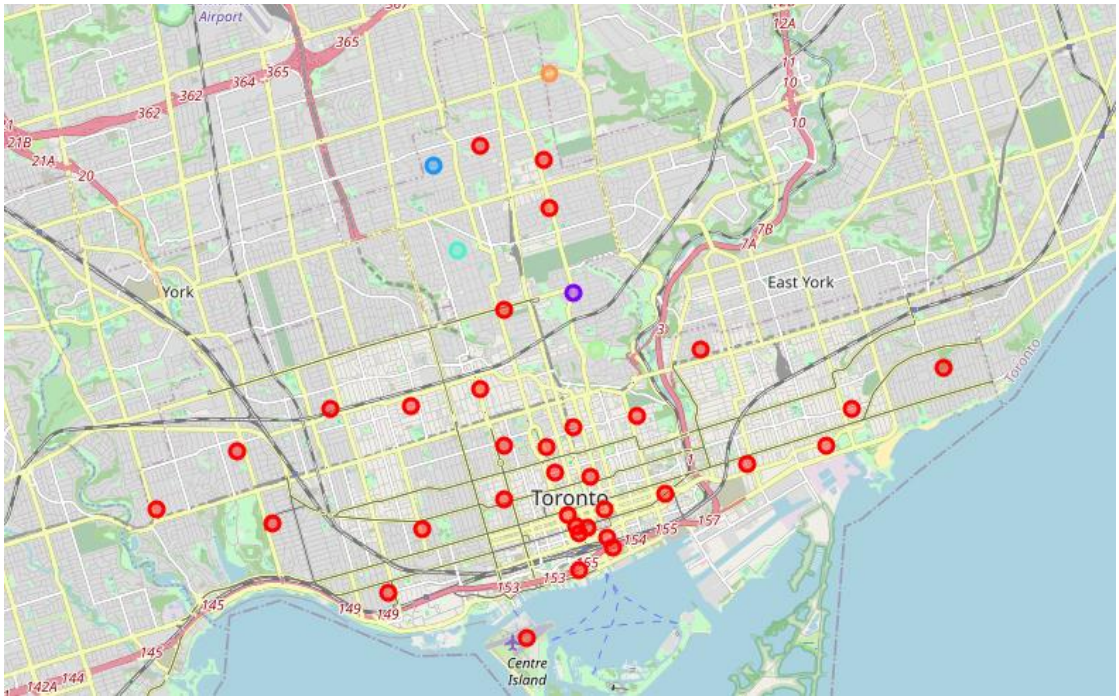

Figure 4: The Toronto City Map and superimposed clusters

4. Results

This section represents the results of the cluster analysis. We'll go one by one and describe the results for each cluster separately.

Cluster 1

This cluster is the largest among all the clusters. A large number of coffee shops, shops , restaurants are the main venues that are common for all the neighborhoods included in this cluster.

Table 7: Cluster 1

	Borough	Cluster Labels	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	Downtown Toronto	0	Coffee Shop	Park	Bakery	Pub	Breakfast Spot	Café	Theater	Mexican Restaurant	Restaurant	Chocolate Shop
1	Downtown Toronto	0	Coffee Shop	Sushi Restaurant	Diner	Yoga Studio	Burrito Place	Beer Bar	Italian Restaurant	Juice Bar	Sandwich Place	Burger Joint
2	Downtown Toronto	0	Clothing Store	Coffee Shop	Café	Cosmetics Shop	Restaurant	Japanese Restaurant	Italian Restaurant	Bubble Tea Shop	Middle Eastern Restaurant	Ramen Restaurant
3	Downtown Toronto	0	Café	Coffee Shop	Cocktail Bar	American Restaurant	Gastropub	Hotel	Restaurant	Gym	Italian Restaurant	Department Store
4	East Toronto	0	Trail	Health Food Store	Pub	Women's Store	Dance Studio	Electronics Store	Eastern European Restaurant	Donut Shop	Doner Restaurant	Dog Run

Cluster 2

The cluster consists of Central Toronto neighborhood. It is rich of parks, women's stores and different kind of restaurants offering Ethiopian and Eastern European cosines. This cluster suits for the people seeking peaceful neighborhoods with parks and dog runs.

Table 8: Cluster 2

	Borough	Cluster Labels	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
29	Central Toronto	1	Park	Women's Store	Deli / Bodega	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Donut Shop	Doner Restaurant	Dog Run	Distribution Center

Cluster 3

This is another cluster for peaceful time. It is very similar to the second cluster, however, here the majority of venues are gardens.

	Borough	Cluster Labels	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
19	Central Toronto	2	Garden	Women's Store	Deli / Bodega	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Donut Shop	Doner Restaurant	Dog Run

Cluster 4

This cluster includes a bus line and trail, may be convenient for people who want a quick access to other parts of the city or country.

	Borough	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
21	Central Toronto	Park	Jewelry Store	Trail	Bus Line	Sushi Restaurant	Deli / Bodega	Electronics Store	Eastern European Restaurant	Donut Shop	Doner Restaurant

Cluster 5

Another cluster suitable for family, with parks and playgrounds as well as Dog runs.

	Borough	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
33	Downtown Toronto	Park	Playground	Trail	Women's Store	Dance Studio	Electronics Store	Eastern European Restaurant	Donut Shop	Doner Restaurant	Dog Run

Cluster 6

This cluster stands out with its swimming schools, parks and bus lines.

	Borough	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
18	Central Toronto	Swim School	Park	Construction & Landscaping	Bus Line	Women's Store	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Donut Shop	Doner Restaurant

5. Discussion

Based on the results discussed above, one can draw conclusions about the potential clustering of the Toronto City. The biggest cluster contains the majority of neighborhoods, which are fool of a restaurants, coffee shops, and other venues. This neighborhoods would be very popular among the younger generation, who are more interested in active style of life. However, as this cluster is large and covers the most of the Toronto it may be divided further into clusters to have a more detailed mapping of the neighborhoods of Toronto City. The other 5 clusters were mostly suitable for elder population with lots of parks, gardens, playgrounds, bus lines, and swimming schools. The latter two may be interesting for younger generation people as well, as they more likely to need those venues for traveling or for learning swimming.

6. Conclusion

The startup company wants to employ a data driven approach to map the Toronto city by identifying and clustering identical neighborhoods into clusters. This will allow the company to offer real estate in the neighborhoods more suiting to the needs and preferences of the clients.

With this purpose in mind, data on Toronto postal codes and neighborhoods, longitude and latitudes, and forsquare venues were collected. After data processing the data of different sources were merged together to form a final data frame, which was going to be analyzed by conducting K-means clustering technique. The analysis showed that the optimal number of clusters should be 6. Hence, the Toronto City neighborhoods were clustered into 6 clusters. The first cluster turned to be the largest one, as it contained the majority of neighborhoods. The similarities of those neighborhoods were in the large number of restaurants, coffee shops, and other entertainment venues. On the other hand, the other 5 clusters were dominantly consisted of gardens, parks, swimming pools and playgrounds, which can be very suitable for elder couples or for any client seeking peace and nature.

This analysis can be further implemented by increasing the number of clusters by breaking down the first cluster, in order to be able to make more personalized offers to the clients.