Capstone Project: The Battle of Neighbourhoods

IBM Data Science Professional Certificate

Mohsen Zargoush

December 2019

1 Introduction

1.1 Background

Toronto, with 2,731,571 population in 2016, is the capital of Ontario Province, the most populous city in Canada, and the fourth largest populous city in North America. Toronto is one of the most multicultural cities in the world and recognized as an international center of arts, business, finance, and culture. As Toronto is a significant destination for Canada's immigrants, its population is very diverse. There are more than 200 distinct ethnic origins among people who live in this city. While the primary language is English, over 160 languages are spoken in the city. Each year over 43 million tourists visit Toronto.

Toronto contains a geographical area previously administered by different municipalities. These municipalities have each created a unique history and identity over the years, and their names still are used by people. Former municipalities include East York, Swansea, Forest Hill, Mimico, North York, Scarborough, Etobicoke, Parkdale, Weston and York. Toronto consists of hundreds of small and large neighborhoods.

1.2 Problem Definition

In this project, we want to explore the neighborhoods in Toronto and group them into similar and dissimilar clusters. There can be many factors to consider some regions similar, including the facilities, events, restaurants, parks, schools, etc. in each neighborhood.

1.3 Interest

This study can be interesting for those who want to live temporary or for a long period in Toronto including new residents, tourists, and people who want to change their neighborhood. Imagine that someone wants to live a new neighborhood (whether they are tourists or Toronto residents), it is important for them to know their new neighborhood and compare it to their previous or desired districts. Hence, this project will help them to know every area in Toronto and choose their new and favorite neighborhood.

2 Data Acquisition and Cleaning

2.1 Data Sources

The following Wikipedia page is used to get information about neighborhoods in Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. This defines the scope of this project, which is the city of Toronto in Canada.

Also, we use the following CSV file to extract the geographical coordinates of different postal codes (neighborhoods): http://cocl.us/Geospatial_data. This is required to get the venue data and plot the map.

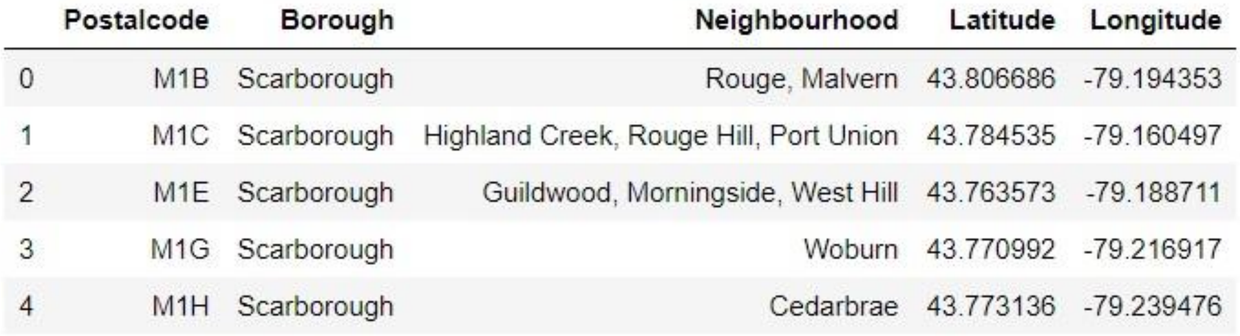
Finally, we request the venue data for each neighborhood from the Foursquare API. This data is used to execute clustering on the neighborhoods.

2.2 Data Cleaning

We combine the data downloaded from multiple sources into one table. After transforming the data into the Pandas data frame, we ignore the rows with ‘Not assigned’ label in the Borough column. Then we merge the neighborhoods with the same postal code. Finally, if a neighborhood has ‘Not assigned’ name, we consider the name of their borough as their neighborhood’s name.

2.3 Feature Selection

After all the merging and cleaning data that we mentioned above, we consider postal code, borough, neighborhood’s name, latitude, and longitude of each neighborhood as shown in the following table (there are 103 rows and five columns). Note that in the methodology section, we will discuss how to consider and insert different events for each neighborhood as a new data frame.

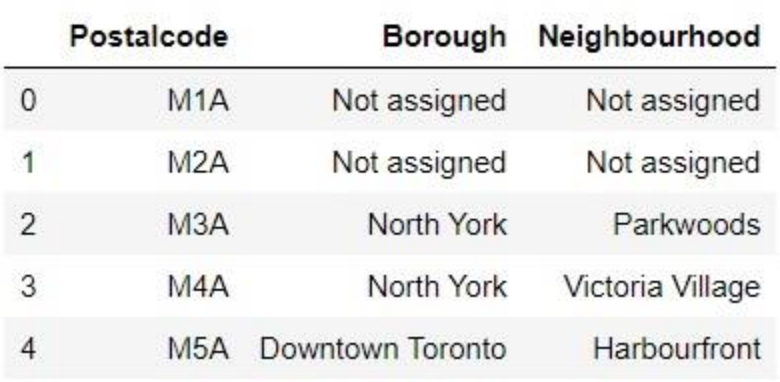


	Postalcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

3 Methodology

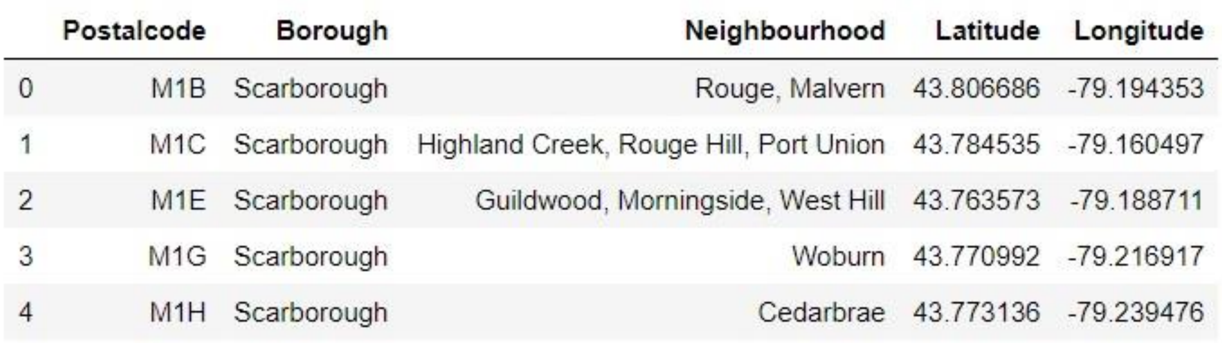
3.1 Preparing the Primary Data

First, we use the BeautifoulSoup package to read the data about Toronto neighborhoods on the Wikipedia page, and then we transform it into the Pandas data frame as below.



	Postalcode	Borough	Neighbourhood
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	Harbourfront

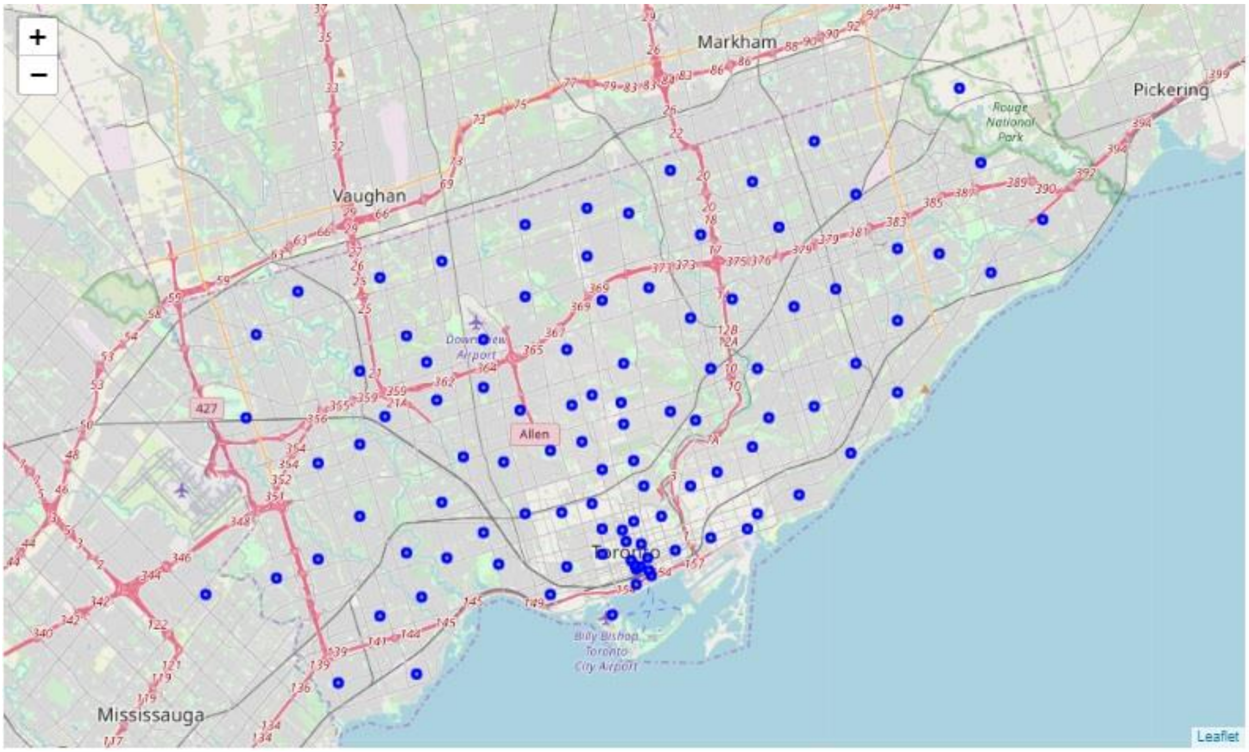
Second, we use the CSV file to extract the geographical coordinates of different neighborhoods. Finally, after doing some data cleaning mentioned in section 2.2, we combine the data as follows.



	Postalcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

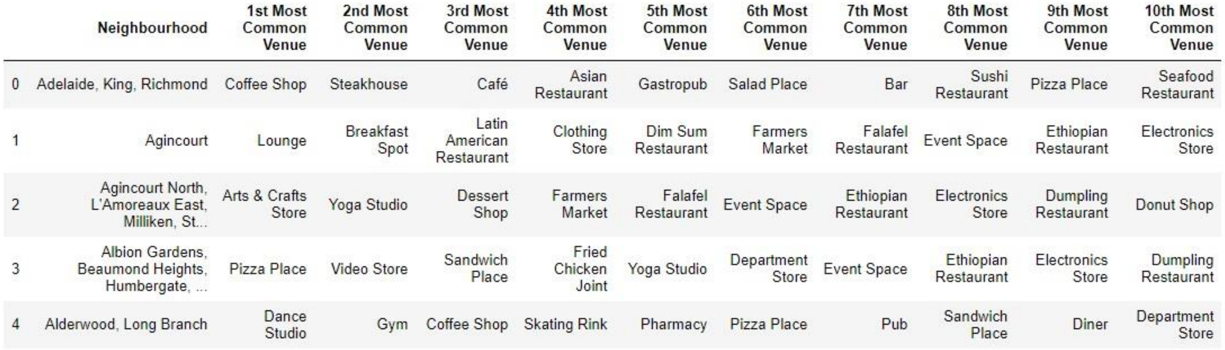
3.2 Showing the Toronto Neighborhoods on the Map

By using the latitude and longitude of each neighborhood, and the Python folium library, we generate the following map to visualize the data (Toronto neighborhoods).



3.3 Using Foursquare API to Explore each Neighborhood

By using the Foursquare API, we explore the neighborhoods to find out what venues exist in each neighborhood. We get the top 50 venues of each neighborhood within the radius of 600 meters of their geographical coordinates. Eventually, we create a new data frame as follows to display the ten most common venues of each neighborhood.



	Neighbourhood	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	Adelaide, King, Richmond	Coffee Shop	Steakhouse	Café	Asian Restaurant	Gastropub	Salad Place	Bar	Sushi Restaurant	Pizza Place	Seafood Restaurant
1	Agincourt	Lounge	Breakfast Spot	Latin American Restaurant	Clothing Store	Dim Sum Restaurant	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store
2	Agincourt North, L'Amoreaux East, Milliken, St...	Arts & Crafts Store	Yoga Studio	Dessert Shop	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Dumpling Restaurant	Donut Shop
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Pizza Place	Video Store	Sandwich Place	Fried Chicken Joint	Yoga Studio	Department Store	Event Space	Ethiopian Restaurant	Electronics Store	Dumpling Restaurant
4	Alderwood, Long Branch	Dance Studio	Gym	Coffee Shop	Skating Rink	Pharmacy	Pizza Place	Pub	Sandwich Place	Diner	Department Store

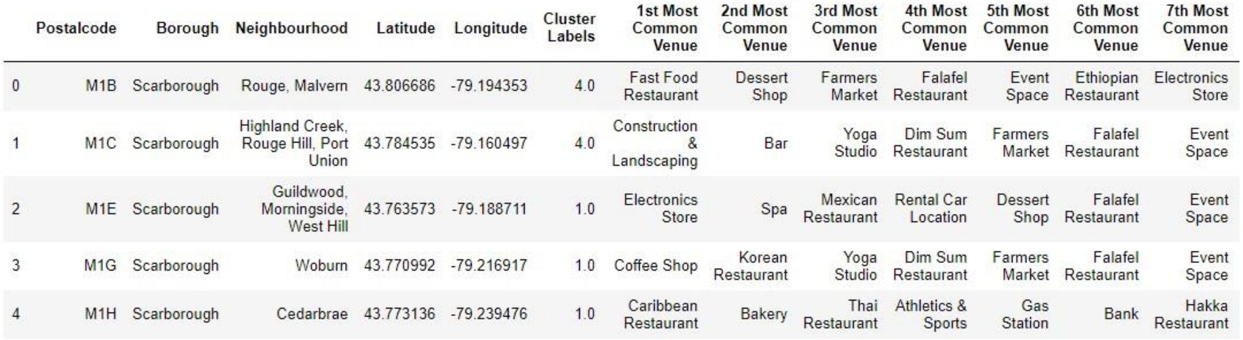
3.4 K-Means Clustering Algorithm

There are some common venues among neighborhoods. So, the K-means algorithm is a suitable way to group neighborhoods into different categories in which each category shows similar neighborhoods.

K-means algorithm is a popular unsupervised machine learning algorithm, which is used for clustering data. Note that we categorize neighborhoods into 6 clusters.

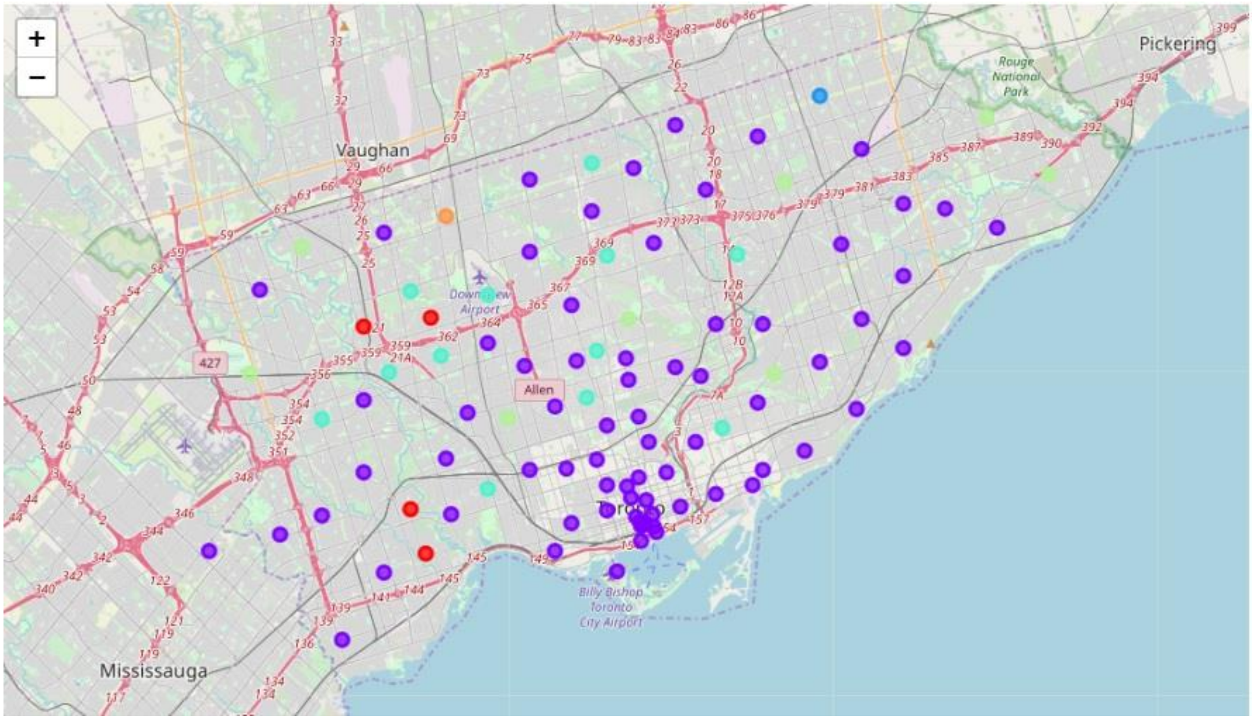
4 Results

The cluster label of each neighborhood is shown below.



	Postalcode	Borough	Neighbourhood	Latitude	Longitude	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
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	4.0	Fast Food Restaurant	Dessert Shop	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	4.0	Construction & Landscaping	Bar	Yoga Studio	Dim Sum Restaurant	Farmers Market	Falafel Restaurant	Event Space
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	1.0	Electronics Store	Spa	Mexican Restaurant	Rental Car Location	Dessert Shop	Falafel Restaurant	Event Space
3	M1G	Scarborough	Woburn	43.770992	-79.216917	1.0	Coffee Shop	Korean Restaurant	Yoga Studio	Dim Sum Restaurant	Farmers Market	Falafel Restaurant	Event Space
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	1.0	Caribbean Restaurant	Bakery	Thai Restaurant	Athletics & Sports	Gas Station	Bank	Hakka Restaurant

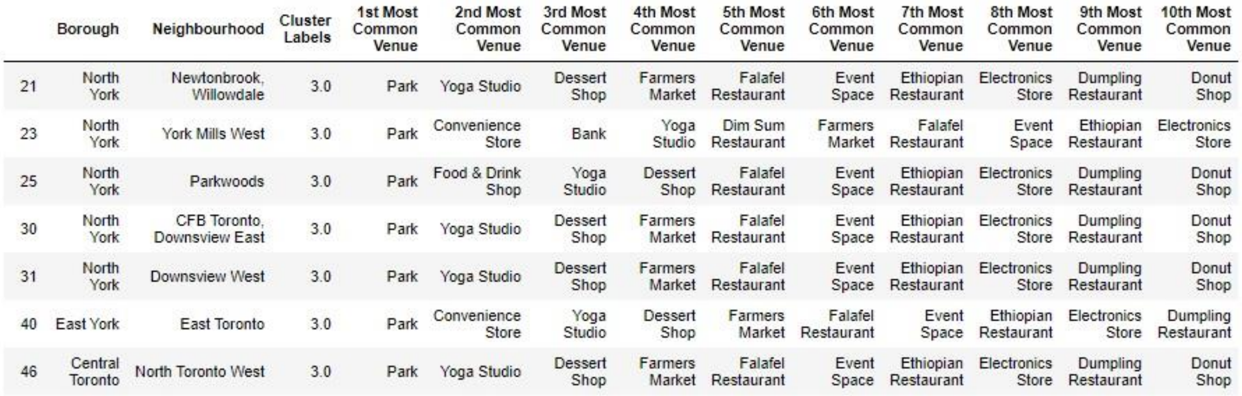
The following map visualizes the cluster of each neighborhood in Toronto by using the Folium package and Matplotlib library.



5 Discussion

Now, we know about the most common venues in each neighborhood. Also, we categorized similar neighborhoods into 6 clusters. This helps those who want to live in a new place to choose the neighborhood which is similar to their previous or desired neighborhood.

We can analyze the clusters and see similar neighborhoods in each cluster. For example, the below table shows the neighborhoods in cluster 4.



	Borough	Neighbourhood	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
21	North York	Newtonbrook, Willowdale	3.0	Park	Yoga Studio	Dessert Shop	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Dumpling Restaurant	Donut Shop
23	North York	York Mills West	3.0	Park	Convenience Store	Bank	Yoga Studio	Dim Sum Restaurant	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store
25	North York	Parkwoods	3.0	Park	Food & Drink Shop	Yoga Studio	Dessert Shop	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Dumpling Restaurant	Donut Shop
30	North York	CFB Toronto, Downsview East	3.0	Park	Yoga Studio	Dessert Shop	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Dumpling Restaurant	Donut Shop
31	North York	Downsview West	3.0	Park	Yoga Studio	Dessert Shop	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Dumpling Restaurant	Donut Shop
40	East York	East Toronto	3.0	Park	Convenience Store	Yoga Studio	Dessert Shop	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Dumpling Restaurant
46	Central Toronto	North Toronto West	3.0	Park	Yoga Studio	Dessert Shop	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Dumpling Restaurant	Donut Shop

Imagine someone wants to change his neighborhood from North York borough to a similar neighborhood in East York. According to the above table, East Toronto is a proper choice. And if they want to move to Central Toronto, North Toronto West is a suitable option. These neighborhoods are considered similar, as there are some common venues among them including Park, Yoga Studio, Farmers Market.

6 Conclusion and Future Directions

In this project, we explored the neighborhoods in Toronto through preparing data, categorize neighborhoods into six groups by performing K-means clustering algorithm (which is an unsupervised machine learning algorithm). Lastly, we developed recommendations to the people who want to live temporary or for a long period in Toronto including new residents, tourists, and people who want to change their neighborhood.

As new research, some can consider other algorithms to cluster neighborhoods and compare the results of different algorithms. Also, we can find a way to determine the optimal number of clusters (k) before performing the K-means algorithm.

7 References

- Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- CSV file for geographical data: http://cocl.us/Geospatial_data
- Foursquare API