Coursera Capstone Project: Applied Data Science

Jack Niu

15th May

I. Introduction

1.1 Background

Toronto, city, capital of the province of Ontario, southeastern Canada. It is the most populous city in Canada, a multicultural city, and the country's financial and commercial center. Also, with many high-level universities located in GTA (Great Toronto Area), Toronto is also a technology center in North America. Most of the world's largest companies have branches in Toronto. With the fast development of Information Technology, more and more new graduated students are planning to settle in Toronto.

The strength and vitality of the many neighbourhoods that make up Toronto, Ontario, Canada has earned the city its unofficial nickname of "the city of neighbourhoods". There are 140 neighbourhoods officially recognized by the City of Toronto and upwards of 240 official and unofficial neighbourhoods within the city's boundaries. Before 1998, Toronto was a much smaller municipality and formed part of Metropolitan Toronto. When the city amalgamated that year, Toronto grew to encompass the former municipalities of York, East York, North York, Etobicoke, and Scarborough. Each of these former municipalities still maintains, to a certain degree, its own distinct identity, and the names of these municipalities are still used by their residents, sometimes for disambiguation purposes as amalgamation resulted in duplicated street names. The area known as Toronto before the amalgamation is sometimes called the "old" City of Toronto, the Central District or simply "Downtown"

The "former" City of Toronto is, by far, the most populous and densest part of the city. It is also the business and administrative centre of the city. The uniquely Torontonian bay-and-gable housing style is common throughout the former city.

The "inner ring" suburbs of York and East York are older, predominantly middle-income areas, and ethnically diverse. Much of the housing stock in these areas consists of pre-World War II single-family houses and post-war high-rises. Many of the neighbourhoods in these areas were built up as streetcar suburbs and contain many dense and mixed-use streets, some of which are one-way. They share many characteristics with sections of the "old" city, outside the downtown core.

The "outer ring" suburbs of Etobicoke, Scarborough, and North York are much more suburban in nature (although these boroughs are developing urban centres of their own, such as North York City Centre around Mel Lastman Square). The following is a list of some notable neighbourhoods' postal code.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
12	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
18	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
22	M1G	Scarborough	Woburn	43.770992	-79.216917
26	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
32	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
38	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029
44	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577
51	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476
58	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848
65	M1P	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	43.757410	-79.273304
71	M1R	Scarborough	Wexford, Maryvale	43.750072	-79.295849
78	M1S	Scarborough	Agincourt	43.794200	-79.262029
82	M1T	Scarborough	Clarks Corners, Tam O'Shanter, Sullivan	43.781638	-79.304302
85	M1V	Scarborough	Milliken, Agincourt North, Steeles East, L'Amo...	43.815252	-79.284577
90	M1W	Scarborough	Steeles West, L'Amoreaux West	43.799525	-79.318389
95	M1X	Scarborough	Upper Rouge	43.836125	-79.205636
27	M2H	North York	Hillcrest Village	43.803762	-79.363452
33	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556
39	M2K	North York	Bayview Village	43.786947	-79.385975
45	M2L	North York	York Mills, Silver Hills	43.757490	-79.374714
52	M2M	North York	Willowdale, Newtonbrook	43.789053	-79.408493
59	M2N	North York	Willowdale	43.770120	-79.408493
66	M2P	North York	York Mills West	43.752758	-79.400049
72	M2R	North York	Willowdale	43.782736	-79.442259
0	M3A	North York	Parkwoods	43.753259	-79.329656
7	M3B	North York	Don Mills	43.745906	-79.352188
13	M3C	North York	Don Mills	43.725900	-79.340923
28	M3H	North York	Bathurst Manor, Wilson Heights, Downsview North	43.754328	-79.442259

Table 1.1 Some Postal Code of Neighborhoods

1.2 Business Problem

Jack is currently living in North York district in Toronto, with a postal code start from 'M2M'

Now he finds a new job and will be working in the Downtown area. Since the TTC of Toronto is quite slow and unreliable, he doesn't want to waste 2 hours on traveling every day. So he is planning to move to Downtown area.

For the new living space, he hopes it could satisfy the following requirements:

- The new district is similar to the one he is living right now in terms of public services.
- The new district should be convenient to take public transportations

1.3 Data Used

The data for this project has been retrieved and processed from different sources, giving careful consideration to the accuracy of the methods used

The data consists of three parts:

- Borough data with postal code
- Geocoding data
- Venue data with borough details

By joining the first two dataset, we are able to find the centroids of each borough. Then use the data of venue to cluster the boroughs and find the similarity with M2M.

II. Data Processing Method

2.1 Data Importing

First import Borough data with postal code. The data comes from Wikipedia Page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

After importing this data, the initial result shows like the following table:

	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

Table 2.1 Initial Postal Code and Neighborhood Data

Then import the geocoding data, with the latitude and longitude of each neighborhood. The data comes from online source: http://cocl.us/Geospatial_data

After importing this data, the initial result shows like the following table:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Table 2.2 Initial Geocoding Data of Neighborhoods

Getting data from FourSquare's API will be refered in the following sections

2.2 Data Preprocessing

As we see from the initial Postal Code and Neighborhood data, we can easily find there are several NaN value in the Neighborhood column. This is due to the neighborhood classification changes in the last few years, some of the neighborhoods have been merged to other neighborhoods, or just leaving a postal code for future use when the city got expanded. To make the data clean enough for future analyzing, here we should clean out the missing values, and just leave the postal code with at least one real neighborhood inside.

After clean the missing values, we merge the Postal Code data and Geocoding data, with the key on Postal Code. The result shows like the following table:

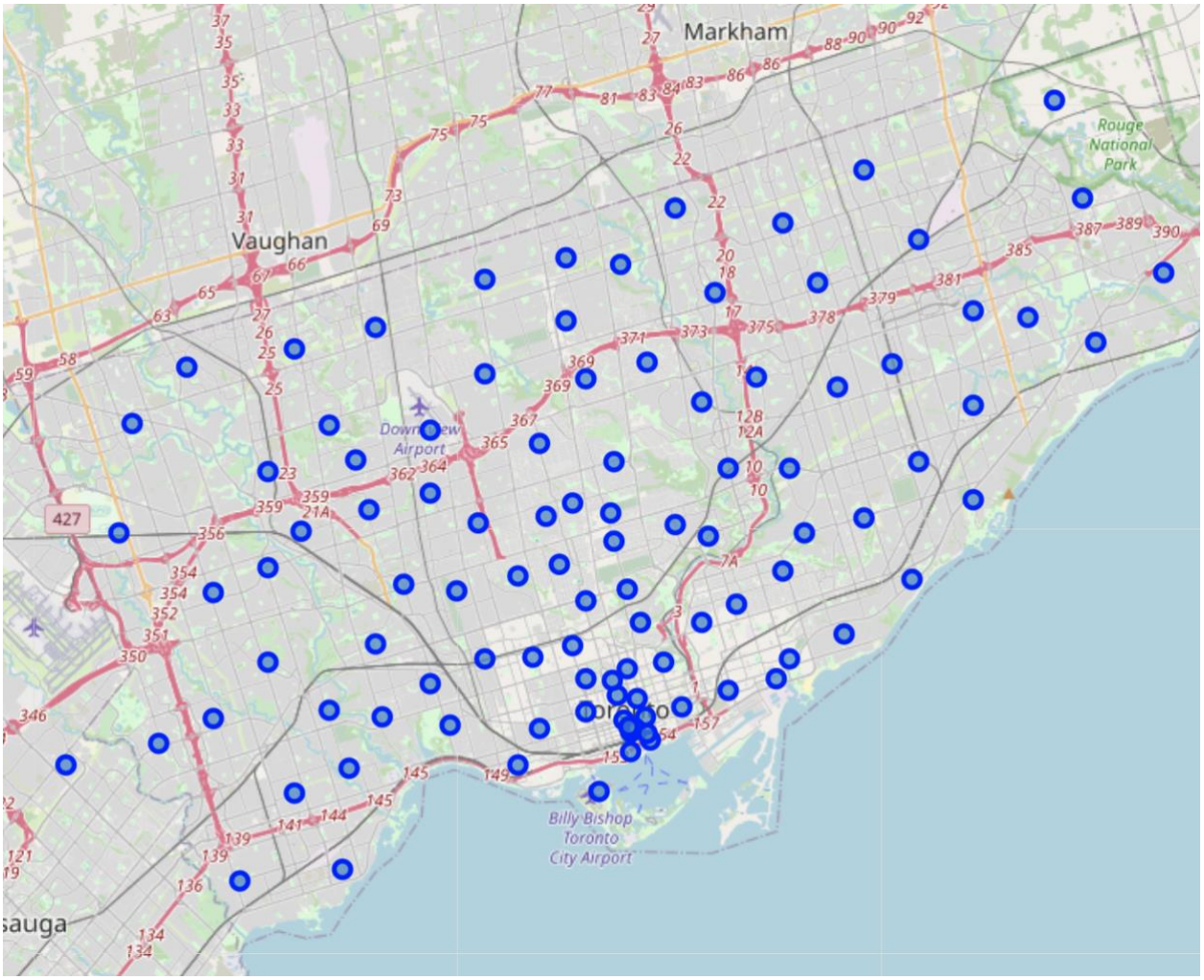
	Postal Code	Borough	Neighborhood	Latitude	Longitude
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
12	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
18	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
22	M1G	Scarborough	Woburn	43.770992	-79.216917
26	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Table 2.3 Cleaned Neighborhood Data

2.3 Data Visualization

Since we want to find a similar neighborhood in downtown compared with the neighborhood with postal code 'M2M', we first need to locate all neighborhoods on map and find the location distribution of them. If all neighborhoods are within the area range we want to learn in this process, then we are good to move forward.

To label all neighborhoods on the map, here we attached the 'folium' package, which is a strong python package used to create map based data visualizations. After loading all data into the map creator, we are able to see the following result:



Graph 2.1 Neighborhoods location visualization

From the above graph, we see all the neighborhoods are in the Toronto area (Not include neighborhoods in nearby cities like Missisauga, Richmond Hill, Pickering etc). We don't need to do any other data filtering on the current data.

III. Get Venues Data

3.1 FourSquare Introduction

FourSquare is a data & intelligence company who provide details location data based on social networks. When a new user reports the location, the location will be stored in the company's dataset. Users could download the information of a location or venues near a location within a certain radius for free, and get some detailed information about the venues like location, owner, rating, categories and so on.

3.2 Venue data downloading

To analyze the similarity of each neighborhood, we need to find the venues data in each neighborhood. For example, how many café shops are there in a neighborhood? How many

categories of venues are there in a neighborhood? How many total venues are there in a neighborhood. To download the data, we use the API call from FourSquare.

Since some neighborhoods in the dataset we have are more likely a living area, which means there are not so many venues in a small radius. If too few venues are in the neighborhoods the system may wrongly cluster it. To avoid such problems, we select a radius of 2500 meters, which is a reasonable distance that people could stand for walking or biking. And because of the venues result limit is 100, we can only return 100 venues for each kind of venues. As we just need to count how many venues of a category in a certain neighborhood, we will only select venue name, venue category, latitude and longitude as the final result. Result shows like the following table:

	name	categories	lat	lng
0	B.B. Cafe	Café	43.791117	-79.418078
1	Starbucks	Coffee Shop	43.796409	-79.419653
2	Eat Bkk Thai Kitchen & Bar	Thai Restaurant	43.795696	-79.419205
3	Daldongnae	Korean Restaurant	43.789729	-79.418104
4	Galleria Supermarket	Supermarket	43.799003	-79.420931

Table 3.1 Sample Venue Data

After getting data for all neighborhoods, join the result of venues data with neighborhood data.

	PostalCode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M1B	43.806686	-79.194353	African Rainforest Pavilion	43.817725	-79.183433	Zoo Exhibit
1	M1B	43.806686	-79.194353	Toronto Pan Am Sports Centre	43.790623	-79.193869	Athletics & Sports
2	M1B	43.806686	-79.194353	Polar Bear Exhibit	43.823372	-79.185145	Zoo
3	M1B	43.806686	-79.194353	Toronto Zoo	43.820582	-79.181551	Zoo
4	M1B	43.806686	-79.194353	Orangutan Exhibit	43.818413	-79.182548	Zoo Exhibit

Table 3.2 Venus of Each Neighborhood

By calculating the total number of venues of each neighborhood, we can get an overview of the business development of each neighborhood.

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
M1B	72	72	72	72	72	72
M1C	58	58	58	58	58	58
M1E	59	59	59	59	59	59
M1G	94	94	94	94	94	94
M1H	100	100	100	100	100	100
M1J	94	94	94	94	94	94
M1K	85	85	85	85	85	85

Table 3.3 Venue Count for Each Neighborhood

IV. Neighborhoods Clustering

4.1 Single Neighborhood Analysis

To understand if one kind of venues exists in a neighborhood, here we use encoder the categories, the result represents which kind of venue it is (in which 0 represents not, 1 represents yes), by calculating the total count of each neighborhood we are able to get an overview of the venue distribution of all neighborhoods.

By calculating the mean value of venues group by neighborhoods, we can find the appearing frequency of venues, based on which we can perform the clustering model to make clusters of the neighborhoods.

	PostalCode	Accessories Store	Afghan Restaurant	Airport	American Restaurant	Amphitheater	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store	...	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wings Joint	Women's Store	Xinjiang Restaurant	Yoga Studio	Zoo	Zoo Exhibit
0	M1B	0.000000	0.00	0.0	0.000000	0.00	0.00	0.00	0.00	0.000000	...	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.027778	0.222222
1	M1C	0.000000	0.00	0.0	0.000000	0.00	0.00	0.00	0.00	0.000000	...	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.000000
2	M1E	0.000000	0.00	0.0	0.000000	0.00	0.00	0.00	0.00	0.000000	...	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.000000	0.000000	0.000000	0.000000
3	M1G	0.000000	0.00	0.0	0.010638	0.00	0.00	0.00	0.00	0.000000	...	0.010638	0.000000	0.00	0.00	0.000000	0.00	0.010638	0.000000	0.000000	0.000000
4	M1H	0.000000	0.00	0.0	0.020000	0.00	0.00	0.00	0.00	0.000000	...	0.010000	0.000000	0.00	0.00	0.020000	0.00	0.010000	0.010000	0.000000	0.000000

Table 4.1 Venues Appearing Frequency Sample

4.2 Clustering

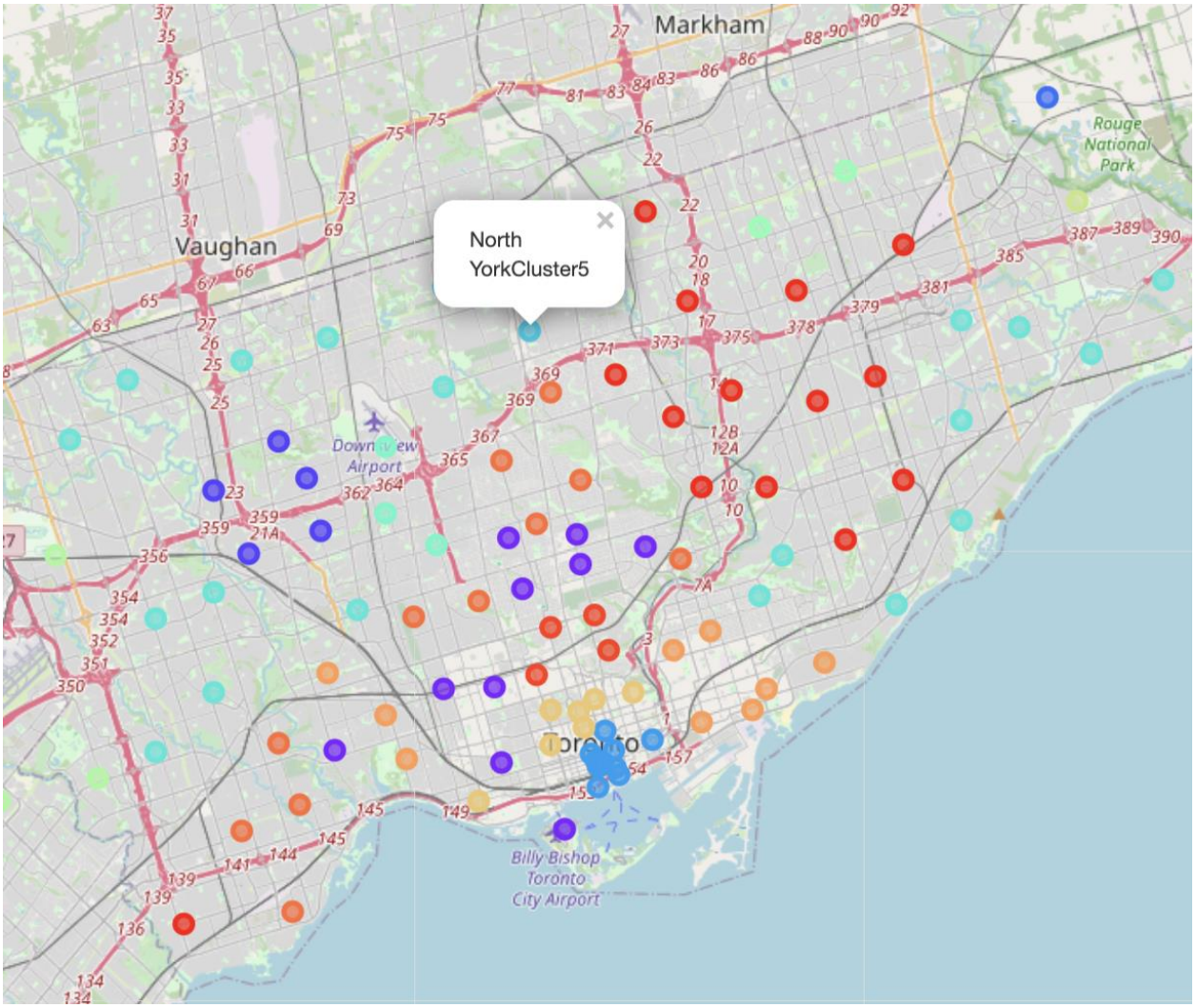
Since we have 103 neighborhoods to cluster, we select 15 as the number of clusters. The method we use is KMeans Clustering. Based on the appearing frequency of each kind of venues, the model could calculate the distance by using the frequency. The frequency calculation process actually already finish the standardscaling process so we can directly perform the model.

After modeling, each neighborhood is assigned a label, which is the cluster it belongs to. Then we can attach the label column into the result table, which shows the neighborhood details, labels and top 10 popular venues.

PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Label	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
M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	10	Zoo Exhibit	Fast Food Restaurant	Park	Pizza Place	Grocery Store	Gas Station	Restaurant	Pharmacy	Zoo	Burger Joint
M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	6	Coffee Shop	Park	Breakfast Spot	Sandwich Place	Pet Store	Grocery Store	Bank	Burger Joint	Liquor Store	Fast Food Restaurant
M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	6	Pizza Place	Coffee Shop	Grocery Store	Hotel	Park	Sandwich Place	Bank	Juice Bar	Supermarket	Discount Store
M1G	Scarborough	Woburn	43.770992	-79.216917	6	Coffee Shop	Fast Food Restaurant	Bank	Pizza Place	Gas Station	Sandwich Place	Chinese Restaurant	Beer Store	Discount Store	Indian Restaurant
M1H	Scarborough	Cedarbrae	43.773136	-79.239476	6	Coffee Shop	Sandwich Place	Gas Station	Pharmacy	Clothing Store	Bank	Gym	Indian Restaurant	Pizza Place	Caribbean Restaurant

Table 4.2 Final Result Table

Since all neighborhoods have their own labels, it is easy for us to view the neighborhood distribution of labels on a map. To mark different labels, we assign different colors to different labels. The result shows like the following graph:



Graph 4.1 Neighborhood with Labels

Now we find the clusters distribution of each neighborhood. The one marked on map is the neighborhood Jack currently living in, and the neighborhoods marked deep blue are the ones he plans to move in.

V. Result and Conclusion

5.1 Result Analysis

From the above graph, we can clearly find that the label of neighborhood ‘M2M’ has no similar one in downtown area. To better evaluate the influence of difference, we find the details of neighborhood ‘M2M’ and cluster in downtown.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Label	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
52	M2M	North York	Willowdale, Newtonbrook	43.789053	-79.408493	5	Korean Restaurant	Coffee Shop	Café	Bubble Tea Shop	Middle Eastern Restaurant	Ramen Restaurant	Japanese Restaurant	Pizza Place	Seafood Restaurant	Dessert Shop

Table 5.1 Details of Neighborhood ‘M2M’

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Label	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
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	4	Coffee Shop	Park	Restaurant	Café	Neighborhood	Plaza	Ice Cream Shop	Mediterranean Restaurant	Gym / Fitness Center	Bakery
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	4	Coffee Shop	Park	Restaurant	Yoga Studio	Café	Plaza	Pizza Place	Japanese Restaurant	Sporting Goods Shop	Sandwich Place
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	4	Coffee Shop	Café	Japanese Restaurant	Park	Plaza	Italian Restaurant	Thai Restaurant	Restaurant	Supermarket	Liquor Store
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	4	Coffee Shop	Café	Plaza	Dessert Shop	Park	Italian Restaurant	Bakery	Liquor Store	Baseball Stadium	Neighborhood
30	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568	4	Coffee Shop	Café	Italian Restaurant	Yoga Studio	Japanese Restaurant	Plaza	Sandwich Place	Park	Pizza Place	Thai Restaurant
36	M5J	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	4	Coffee Shop	Park	Café	Plaza	Hotel	Scenic Lookout	Art Gallery	Theater	Gym	Yoga Studio
42	M5K	Downtown Toronto	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576	4	Coffee Shop	Café	Italian Restaurant	Sandwich Place	Yoga Studio	Park	Plaza	Japanese Restaurant	Baseball Stadium	Hotel
48	M5L	Downtown Toronto	Commerce Court, Victoria Hotel	43.648198	-79.379817	4	Coffee Shop	Café	Park	Sandwich Place	Plaza	Italian Restaurant	Yoga Studio	Cosmetics Shop	Sporting Goods Shop	Japanese Restaurant
92	M5W	Downtown Toronto	Stn A PO Boxes	43.646435	-79.374846	4	Coffee Shop	Café	Plaza	Dessert Shop	Park	Thai Restaurant	Italian Restaurant	Farmers Market	Baseball Stadium	Neighborhood
97	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.382280	4	Coffee Shop	Italian Restaurant	Café	Yoga Studio	Sandwich Place	Plaza	Park	Theater	Restaurant	Japanese Restaurant

Table 5.2 Details of Neighborhoods in Downtown

From the above tables, it is clear to see the overall popular venues of the two different areas are quite similar. The difference is caused by the number of venues, downtown area all have much more venues than it in North York. Which result in the different clusters of neighborhoods. Since in FourSquare it will only return 100 results for all venues, we can not get more detailed data to perform a more accurate model.

The company location that Jack is going to work in is (43.645832, -79.383097). By calculating the distance from the centroid of each neighborhood, we can get an understanding of how far each neighborhood is to the company.

PostalCode	Borough	Neighborhood	Distance
M5K	Downtown Toronto	Toronto Dominion Centre, Design Exchange	0.193271
M5X	Downtown Toronto	First Canadian Place, Underground city	0.296272
M5L	Downtown Toronto	Commerce Court, Victoria Hotel	0.372808
M5H	Downtown Toronto	Richmond, Adelaide, King	0.540263
M5J	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	0.568363
M5W	Downtown Toronto	Stn A PO Boxes	0.667485
M5E	Downtown Toronto	Berczy Park	0.796825
M5C	Downtown Toronto	St. James Town	0.882381
M5B	Downtown Toronto	Garden District, Ryerson	1.303924
M5A	Downtown Toronto	Regent Park, Harbourfront	2.036333

Table 5.2 Distance to Company

By referring the TTC public transportation lines distribution, we find the transportation is way better than it in North York. Since every ride of TTC cost 3 CAD, it could save a lot for Jack if he decides to walk to company. From the table above, we can find the top 5 neighborhoods are not farer than 600 meters to the company, which is an appropriate distance for walking. Jack could select his new room in these neighborhoods: M5K, M5X, M5L, M5H.

5.2 Conclusion

The project has some drawbacks, which make the final result not as detailed as we want. To improve the final result of project, we can amend the project from the following points:

- Increase the venue limits by using premium account, and get more venues that could better cluster the neighborhoods.
- Get public transportation data, like station location, station count in a neighborhood, total traffic volume.
- Get room for rental data, which could reflect how many available rooms are there in a neighborhood. Even though some neighborhoods have both great distance, traffic conditions, and public services, the neighborhood may not be designed for living and don’t have many available rooms.
- Get room price data. Room price is a very important point when people is looking for a room for rental. By analyzing this feature, we are able to find the most suitable neighborhood based on the budget.