Capstone Project

The Battle of Neighbourhoods – City of Toronto

Henry – September 09, 2019

1. Introduction

1.1. Background

In "Applied Data Science Capstone" course, we explored, segmented, and clustered the neighbourhoods in the city of Toronto based on venues in each postcode area.

This capstone project will continue the same analysis with additional features (such as elementary and secondary schools' ratings, 2016 census data from Statistic Canada, and housing data from HouseSigma) to better describe the characteristics of each neighbourhood.

1.2. Problem

There are many factors that could characterize a neighbourhood besides venues or interest points.

Other factors could include, but not limited to, house and rental pricing, availability of good elementary and secondary schools for the children, average family incomes for selection reference.

Due to time limitation, only detached houses will be considered in this project.

This analysis tries to answer the question: Which neighbourhoods have the similar features from where a family with young children can purchase a detached residential property?

1.3. Interest

Families with young children of school age who are looking for a residential property in City of Toronto.

Note: *Residential property or house* referring here could be a detached house, semi-detached house, townhouse, or apartment.

2. Data Acquisition and Cleaning

2.1. FSA Postcodes of City of Toronto

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

HTML scraping to get all Forward Sortation Area (FSA) postcodes, boroughs, and neighbourhoods in City of Toronto.

These data cleaning steps will be applied:

- The dataframe will consist of three columns: Postcode, Borough, and Neighbourhood
- Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.
- More than one neighbourhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighbourhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighbourhoods separated with a comma as shown in row 11 in the above table.
- If a cell has a borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough. So for the 9th cell in the table on the Wikipedia page, the value of the Borough and the Neighbourhood columns will be Queen's Park.

Note: The FSA postcodes are the first three characters of Canadian postal codes. For example, *M5H* is the FSA postcode for the postal code '*M5H* 2N2' (where Toronto City Hall is located).

In addition, the latitude and longitudes coordinates of each FSA postcode will be loaded from http://cocl.us/Geospatial_data/Geospatial_Coordinates.csv file.

2.2. Foursquare Location Data

Foursquare API will be used to find out the following in a postcode area:

- Venues: Limited to maximum 100 venues
- Schools: School ranking average in 5000m (5km)
    - Elementary School: 4f4533804b9074f6e4fb0105
    - High School: 4bf58dd8d48988d13d941735
    - Middle School: 4f4533814b9074f6e4fb0106
    - Private School: 52e81612bcbc57f1066b7a46
- Transportation: How many stations or stops in 800m (10 min walking distance)

- Bus Station: 4bf58dd8d48988d1fe931735
- Bus Stop: 52f2ab2ebcbc57f1066b8b4f
- Train Station: 4bf58dd8d48988d129951735
- Transportation Service: 54541b70498ea6ccd0204bff

Above listed Foursquare venue category ids can be found at https://developer.foursquare.com/docs/resources/categories

2.3. Fraser Institute - School Rating

HTML scraping to find out school ratings:

- Elementary: http://ontario.compareschoolrankings.org/elementary/SchoolsByRankLocationName.aspx?schooltype=elementary
- Secondary: http://ontario.compareschoolrankings.org/secondary/SchoolsByRankLocationName.aspx?schooltype=secondary

School ratings are between 0 and 10. Division by 10 (ten) will be needed to normalize school ratings for clustering.

We will ignore schools without school ratings from this data source and we won't take into account school boundaries.

2.4. Statistics Canada - 2016 Census

HTML scraping to find out demographic and economic data (such as average family income) in a FSA postcode area.

For example, hyperlink for M5H postcode (as Code1 parameter in the URL): https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=FSA&Code1=M5H

3. Data Loading

3.1. Load Neighbourhoods in Toronto with Geospatial Coordinates

Postcodes in City of Toronto have “M” as first letter. These postcodes and neighbourhood names were scrapped from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

Data cleaning was applied according procedures described in previous section.

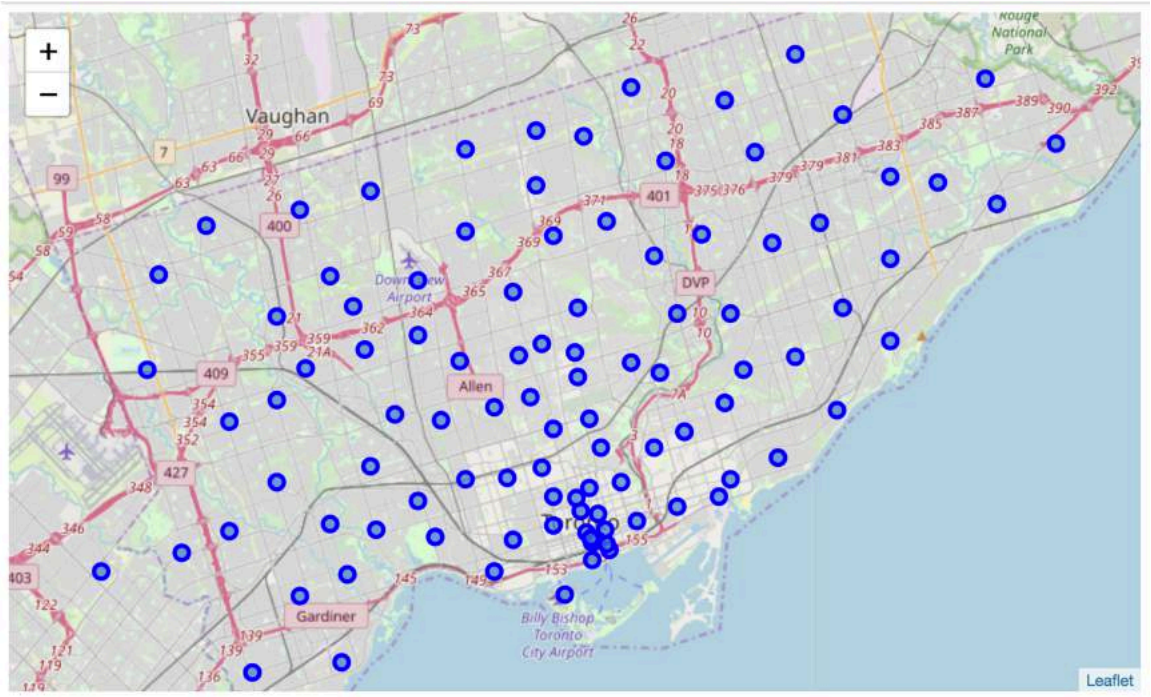
Then, each postcode was cross-referencing to latitudes and longitudes in http://cocl.us/Geospatial_data/Geospatial_Coordinates.csv.

Here is the excerpt of the Borough and Neighbourhoods in City of Toronto with geospatial coordinates (latitudes and longitudes):

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Port Union, Rouge Hill	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
...	...	...	...	...	...
98	M9N	York	Weston	43.706876	-79.518188
99	M9P	Etobicoke	Westmount	43.696319	-79.532242
100	M9R	Etobicoke	Kingsview Village, Martin Grove Gardens, Richv...	43.688905	-79.554724
101	M9V	Etobicoke	Albion Gardens, Beaumond Heights, Humbergate, ...	43.739416	-79.588437
102	M9W	Etobicoke	Northwest	43.706748	-79.594054

103 rows × 5 columns

This City of Toronto map shows the locations of each postcode.



3.2. Load Venues in Each City of Toronto's Postcode

I fetched the top venues in each City of Toronto's postcode from Foursquare. The following information of each returned venue were parsed (from Foursquare's json response):

- Venue name
- Venue latitude
- Venue longitude
- Venue category

	Postcode	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M1B	43.806686	-79.194353	Images Salon & Spa	43.802283	-79.198565	Spa
1	M1B	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
2	M1B	43.806686	-79.194353	Wendy's	43.802008	-79.198080	Fast Food Restaurant
3	M1B	43.806686	-79.194353	Staples Morningside	43.800285	-79.196607	Paper / Office Supplies Store
4	M1B	43.806686	-79.194353	Harvey's	43.800106	-79.198258	Fast Food Restaurant
...	...	...	...	...	...	...	...
3940	M9V	43.739416	-79.588437	Canadian Tire	43.741844	-79.582294	Hardware Store
3941	M9V	43.739416	-79.588437	Pizza Pizza	43.733202	-79.592538	Pizza Place
3942	M9V	43.739416	-79.588437	Tim Hortons	43.742015	-79.589690	Coffee Shop
3943	M9W	43.706748	-79.594054	Economy Rent A Car	43.708471	-79.589943	Rental Car Location
3944	M9W	43.706748	-79.594054	227 Lounge	43.703184	-79.588193	Lounge

3945 rows × 7 columns

In summary, we had the following returned venues per postcode:

```
Postcode
M1B     12
M1C      5
M1E     15
M1G      6
M1H     22
        ..
M9N      6
M9P     10
M9R     14
M9V     13
M9W      2
```

The top 10 venue categories were:

```
Coffee Shop           303
Café                  169
Park                  117
Pizza Place           113
Restaurant            102
Italian Restaurant     97
Bakery                 94
Sandwich Place         78
Hotel                  68
Bar                    68
```

3.3. School Ratings in Each City of Toronto's Postcode

First, I used Foursquare to find out elementary and secondary schools in each postcode area. Then each school from Foursquare response was matched to the school ratings posed on Fraser Institute web site as listed on section 2.3 above.

Note that I did not take into account the school boundaries which would not be easy to determinate. However, I was interested the over all school rating trends in each postcode; not the exact values.

I tried to match school name removing common endings (such as "Junior High", "JHS", "Jr Public School", etc.). If there was no match, I tried to match by postal code if Foursquare response had this school information.

Schools would be ignored if they could not be found in Fraser Institute's School Ratings list.

School ratings average (total ratings / # of school found) for elementary and secondary schools in each postcode were calculated:

	Postcode	Latitude	Longitude	Avg Rating Elementary Schools	Avg Rating High Schools
0	M1B	43.806686	-79.194353	6.240000	5.933333
1	M1C	43.784535	-79.160497	6.950000	4.125000
2	M1E	43.763573	-79.188711	5.033333	5.033333
3	M1G	43.770992	-79.216917	4.075000	5.750000
4	M1H	43.773136	-79.239476	5.200000	3.800000
...	...	...	...	...	...
98	M9N	43.706876	-79.518188	6.500000	0.000000
99	M9P	43.696319	-79.532242	5.900000	7.800000
100	M9R	43.688905	-79.554724	5.900000	7.266667
101	M9V	43.739416	-79.588437	5.200000	7.966667
102	M9W	43.706748	-79.594054	3.850000	7.450000

103 rows × 5 columns

3.4. Load Transportation in Each City of Toronto's Postcode

I looked up of transportation services (like bus and train stops) in each postcode within walking distance (800m radius); expecting high counts signified easy transportation accesses.

	Postcode	Latitude	Longitude	Transportation Count
0	M1B	43.806686	-79.194353	3
1	M1C	43.784535	-79.160497	1
2	M1E	43.763573	-79.188711	5
3	M1G	43.770992	-79.216917	0
4	M1H	43.773136	-79.239476	1
...	...	...	...	...
98	M9N	43.706876	-79.518188	0
99	M9P	43.696319	-79.532242	1
100	M9R	43.688905	-79.554724	7
101	M9V	43.739416	-79.588437	3
102	M9W	43.706748	-79.594054	0

103 rows × 4 columns

3.4. Statistics Canada - 2016 Census

I scrapped 2016 Census HTML pages from Statistic Canada for the following data. Each of them was identified by the HTML "headers" attribute value in the web page and they were all numeric values.

Data	HTML headers attribute value
Population size	'L1000 geo1 total1'
Median age	'L2033 geo1 total1'
Average household size	'L3017 geo1 total1'
Without children in a census family	'L6003 geo1 total1'
With children in a census family	'L6004 geo1 total1'
Number of households	'L6001 geo1 total1'
Median total income	'L13001 geo1 total1'
Non-immigrants	'L18001 geo1 total1'
ImmigrantsCensus	'L18002 geo1 total1'
Non-permanent residents	'L18010 geo1 total1'

There were many more data/features per postcode from Statistic Canada; however, the above features were the most asked when someone was looking for housing.

Excerpt of scrapped results:

	Postcode	Census Population Size	Census Median Age	Census Average Household Size	Census Households With Children (%)	Census Number of Households	Census Median Family Income	Census Non Immigrants (%)	Census Immigrants (%)	Census Non Permanent Residents (%)
0	M1B	66108	38.2	3.3	80.10	20230	69126.0	38.21	59.92	1.87
1	M1C	35626	44.0	3.1	70.14	11275	109785.0	54.12	44.89	0.99
2	M1E	46943	42.2	2.7	72.57	17160	62047.0	50.74	47.79	1.47
3	M1G	29690	37.2	3.0	76.57	9765	54450.0	40.54	56.03	3.43
4	M1H	24383	38.1	2.7	70.77	8985	58492.0	37.27	57.95	4.77
...	...	...	...	...	...	...	...	...	...	...
98	M9N	25074	39.0	2.4	73.68	10170	50545.0	49.56	47.74	2.71
99	M9P	20874	45.9	2.6	67.12	7905	73425.0	55.01	43.82	1.17
100	M9R	33743	40.6	2.7	70.89	12335	63032.0	47.34	50.69	1.97
101	M9V	55959	35.9	3.3	80.15	16805	59760.0	34.74	62.55	2.71
102	M9W	40684	38.3	2.9	74.47	13705	65826.0	44.18	52.79	3.02

103 rows × 10 columns

Note that Statistic Canada's 2016 Census did not have information for the following postcodes, probably due to the fact of very small population size in these areas:

	Postcode	Census Population Size	Census Median Age	Census Average Household Size	Census Households With Children (%)	Census Number of Households	Census Median Family Income	Census Non Immigrants (%)	Census Immigrants (%)	Census Non Permanent Residents (%)
60	M5K	0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0
61	M5L	0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0
69	M5W	15	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0
70	M5X	10	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0
85	M7A	10	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0
86	M7R	0	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0
87	M7Y	10	0.0	0.0	0.0	0	0.0	0.0	0.0	0.0

4. Methodology and Results

4.1. Data Preparations

- Took only top 10 venue categories (see section 3.2 above) for each postcode to reduce bias on clustering training toward venues. Normalized venue as average by postcode.
- Normalized elementary and secondary school ratings (divided by 10)
- Normalized transportation counts

- Normalized census data. Data in percentage (like Census Household with Children) were divided by 100.

	Postcode	Coffee Shop	Café	Park	Pizza Place	Restaurant	Italian Restaurant	Bakery	Sandwich Place	Bar	Hotel	Avg Rating Elementary Schools	Avg Rating High Schools	Transportation Count	Census Population Size
0	M1B	0.083333	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.624000	0.593333	0.061224	0.871023
1	M1C	0.000000	0.0	0.000000	0.000000	0.0	0.2	0.000000	0.000000	0.2	0.0	0.695000	0.412500	0.020408	0.469399
2	M1E	0.133333	0.0	0.000000	0.200000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.503333	0.503333	0.102041	0.618509
3	M1G	0.333333	0.0	0.333333	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.407500	0.575000	0.000000	0.391188
4	M1H	0.090909	0.0	0.000000	0.000000	0.0	0.0	0.090909	0.000000	0.0	0.0	0.520000	0.380000	0.020408	0.321264
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98	M9N	0.000000	0.0	0.000000	0.166667	0.0	0.0	0.000000	0.000000	0.0	0.0	0.650000	0.000000	0.000000	0.330369
99	M9P	0.100000	0.0	0.000000	0.100000	0.0	0.0	0.000000	0.100000	0.0	0.0	0.590000	0.780000	0.020408	0.275031
100	M9R	0.071429	0.0	0.000000	0.071429	0.0	0.0	0.000000	0.071429	0.0	0.0	0.590000	0.726667	0.142857	0.444589
101	M9V	0.076923	0.0	0.000000	0.230769	0.0	0.0	0.000000	0.076923	0.0	0.0	0.520000	0.796667	0.061224	0.737302
102	M9W	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.385000	0.745000	0.000000	0.536042

103 rows × 23 columns

4.2. K-Means Clustering

I applied K-Means clustering with K=7.

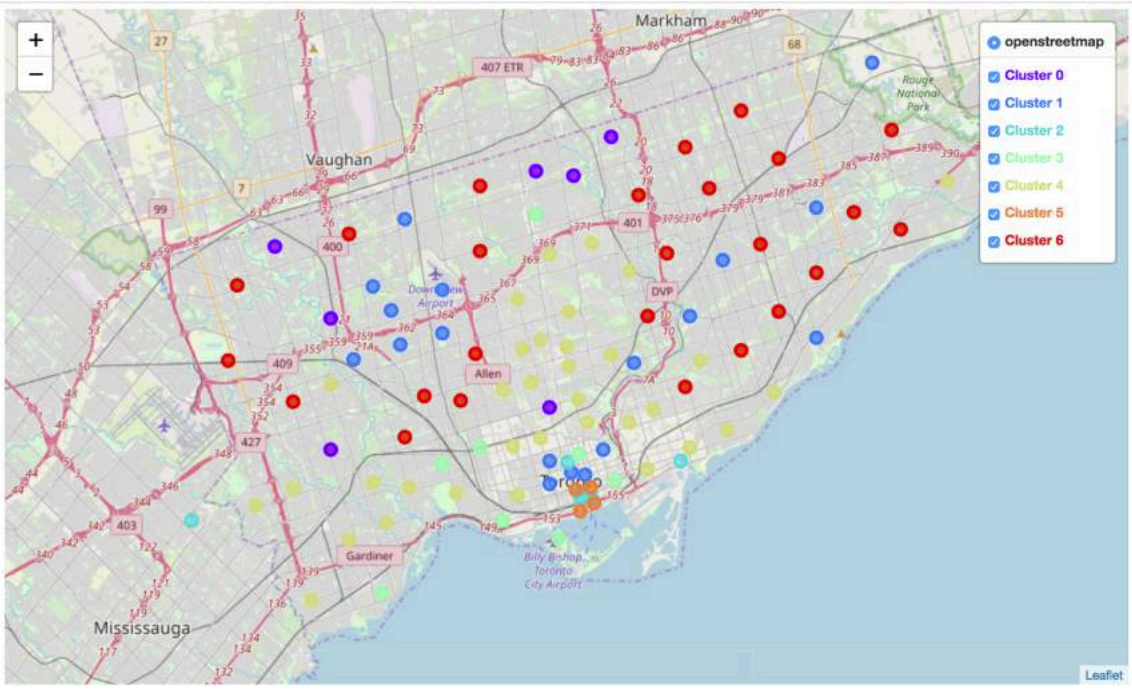
Here are number of postcodes per cluster:

```
Cluster label 0: 7
Cluster label 1: 18
Cluster label 2: 7
Cluster label 3: 8
Cluster label 4: 34
Cluster label 5: 4
Cluster label 6: 25
```

Averages of each cluster:

Cluster Labels	Avg Rating Elementary Schools	Avg Rating High Schools	Transportation Count	Census Population Size	Census Median Age	Census Average Household Size	Census Households With Children (%)	Census Number of Households	Census Median Family Income	Census Non Immigrants (%)	Census Immigrants (%)	Census Non Permanent Residents (%)
0	0.000000	6.909841	2.428571	24102.428571	42.142857	2.571429	63.782857	9640.000000	69550.285714	42.551429	53.604286	3.844286
1	6.426481	4.830992	3.888889	18951.388889	37.816667	2.511111	67.159444	7534.722222	56140.000000	43.737778	51.050000	5.212222
2	5.076531	5.841111	30.428571	6.428571	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	7.224896	6.532054	3.875000	45069.875000	36.225000	1.887500	47.798750	23288.125000	61639.250000	54.820000	39.465000	5.716250
4	7.443186	6.889511	2.500000	22639.000000	41.911765	2.338235	61.495588	9636.323529	88599.441176	65.060294	32.721765	2.217941
5	5.737500	5.771250	38.500000	7154.750000	33.975000	1.625000	31.440000	4322.500000	86254.750000	55.617500	38.465000	5.920000
6	6.014133	6.015467	3.040000	42533.920000	40.116000	2.776000	72.959600	15151.600000	60056.360000	40.668800	56.274000	3.056400

Map of clusters over City of Toronto:



5. Data Analysis

Cluster 0: Good Schools and Families with Children

- Top venues: Eateries and parks
- Good secondary schools but no elementary school rating found
- Many families with children (63.7%)
- Medium population density

Cluster 1: Good Elementary Schools and Young Families with Children

- Top venues: Eateries and parks
- Good elementary schools but poor secondary school ratings
- Young families in the neighbourhoods (Census Median Age < 40)
- Many families with children (67%)
- Low population density
- Look like these were new neighbourhoods in development

Cluster 2: Postcodes with insufficient census data

- There were no census data (zero most of the cases) in these postcode areas because the census population size was very small (less than 15)
- No characteristics could be determined

Cluster 3: Downtown Toronto, Average Schools, and High Population Density

- Top venues: Eateries (coffee, pizza, and restaurants) and hotels
- Household Size was small, at 1.89 persons per household
- Both Elementary and Secondary School Ratings were average to good
- High Population Density due to the highest Census Population Size with small Household Size, most likely high rise apartments

Cluster 4: Good Schools, High Family Income, and Non-Immigrants

- Top venues: Eateries and parks
- Good elementary and secondary school ratings
- High family income
- Predominantly non-immigrants (Census non-Immigrants = 65%)
- Older neighbourhoods located South of Highway 401

Cluster 5: Downtown Toronto, High Family Income, and Young Families

- Top venues: Eateries and hotels
- Schools were poor with both ratings below 6
- Household Size was small, at 1.62 persons per houshold, which was consistent with the low Family With Children (31%)
- High Census Median Family Income (over $82k)
- Look like mostly young (Census Median Age = 34) professional couples without children lived in these neighbourhood

Cluster 6: Newer Neighbourhoods with High Population Density and Families with Children

- To venues: Eateries and parks
- Schools were average
- High population density with large Census Population Size (over 42k persons) and the highest Census Average Household Size (2.78 persons per household)
- 73% families with children, significantly higher than other clusters

Additional Observations in General about Toronto

- It seems elementary and secondary school ratings were poor in general, mostly 6 or below
- Non-immigrants and immigrants ratios was close to 50:50 in average
- Household size was small with average below 3

6. Discussion

There were few issues and difficulties on data quality and data acquisition that could impact our data analysis and possibly our results.

6.1. Issues On Data Quality

We can observe that the Avg Rating Elementary Schools for Cluster 0 is zero. This could be:

1. There were no elementary schools in postcode areas of this cluster; and/or
2. Foursquare did not have complete school information; and/or
3. The elementary schools in the area were not listed in the data source Fraser Institute

Taking the postcode M2K in Cluster 0, there was an elementary school "Pineway Public School" from Foursquare; however, Fraser did not have school rating for this school.

On alternative could be assigned the same value from Avg Rating Secondary Schools as approximation for the missing Avg Rating Elementary Schools.

Also, the transportation information from Foursquare was also incomplete in many areas. Using the same postcode M2K of Bayview Village in North York as an example, there was only one bus stop in 10 minute walking distance from Foursquare. However, looking from Toronto Transit Commission (TTC) web site, there were many TTC bus stops in the area.

Foursquare venues data were leaning toward eateries and places where people would spend their leisure times. It provided poor information on transportations for example.

In addition, although 2016 Census data from Statistic Canada have a lot of details but these data are already at least 3 years old. Demographic information in neighbourhoods with high immigrants and non-permanent residents and with busy recent real estate development would make the census data inaccurate until next Census in 2021.

6.2. Difficulties On Data Acquisition

Originally, I was looking to scrap HTML from HouseSigma for median pricing for residential properties; however this could not be done because the listed communities could not be matched to the neighbourhoods from Wikipedia.

I could not find another data source for Toronto residential pricing; therefore, one of the most important features, house pricing, could not be included in the analysis.

7. Conclusion

Answering the question stated of this study: Which neighbourhoods have the similar features from where a family with young children can purchase a detached residential property?

We grouped the neighbourhoods into 7 clusters using K-Means clustering algorithm; having one of them (with 7 postcode areas) without enough census and school data. A family can determine which features are the most important and then search the neighbourhoods in the cluster, which has the features the family is looking for.

Obviously, there are many improvements can be made to improve the data quality to get better clustering. As a matter of fact, it is expected that using local transit agencies data can provide better data than Foursquare. For this study, using data from TTC (Toronto Transit Commission) and Go Transit can get much accurate information on bus and train stops at any location.
However, there are "human factors" that cannot include in "mathematical analysis" which sway the house purchasing decision. Examples are: nearness to other family members or friends, closeness to employment sites, ambient of the neighbourhood (such as country style) , ravines or old large trees, etc.