Capstone Project

Toronto neighborhoods: Where to live

Table of content

1. Introduction/ Business Problem

Toronto is the most populous city in Canada. This city is an international center of business, finance, arts, and culture. Every year, there are thousands of people choose Toronto as their new destination for living and working. Before their departures, people want to get information about different neighborhoods in Toronto and then decide where they live.

The aim of this project is to describe different neighborhoods in Toronto to help people make choices of which places fit their interests. This project results can help people identify where is quiet, where is more crowded with lots of shopping centers, which places are suitable for cuisine, travelling, etc. From that, they can decide which places fits their lifestyle and interests.

Our target audience is people who is planning to move to Toronto. And people who want to travel to Toronto might care about this problem.

2. Data

2.1 Data Source

I use these **3 data sources** to create the data used for this project, as following:

1. a list of neighborhoods in Toronto is scraped from Wkipedia
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2. a link to a csv file that has the geographical coordinates of each postal code
http://cocl.us/Geospatial_data

3. a list of venue categories is requested from Foursquare API
https://developer.foursquare.com/docs/build-with-foursquare/categories/.

The description of each dataset will be given in the 2.2 part.

2.2 Data preparation and data description

To describe the different between Toronto neighborhoods, the feature of neighborhood (from Toronto neighborhood data) and feature of Venues and Main venues category (from Foursquare dataset) will be extracted.

For the Toronto neighborhood data, after scraping from Wikipedia, the list is transformed into a pandas dataframe and merged with the geographical coordinates file. This data contains 10 boroughs and 103 neighborhoods.

Using Foursquare API to explore the neighborhood and then segment them to get the data contains the name of venues, category, and corresponding latitude and longitude.

After that, these Toronto neighborhood data and Foursquare venues are merged into one as the following:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Malvern, Rouge	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
3	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Affordable Toronto Movers	43.787919	-79.162977	Moving Target
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank

I also request from Foursquare API a list of venues and main venue category and then merge with Toronto data to get the final dataset:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Main Venues Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant	Food
1	Malvern, Rouge	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop	Shop & Service
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar	Nightlife Spot
3	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Affordable Toronto Movers	43.787919	-79.162977	Moving Target	Travel & Transport
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank	Shop & Service

The feature of Neighborhood, Venue Category and Main Venues Category are extracted to describe the different between Toronto neighborhood.

3. Methodology

The methodology used in this notebook is as followed:

+ Get a list of neighborhoods in Toronto by scraping a Wikipedia page, get the geographical coordinates of each neighborhood from a given csv file.

+ Pass the list of neighborhoods with geographical coordinates through the Foursquare API to return a list of venues in the neighborhood within 500 meters.

+ Perform several analysis on dataset containing feature of neighborhood, near-by venues, and main venue categories.

+ Perform k-means clustering on the neighborhood to solve the problem in the introduction.

3.1 Exploratory data analysis

The Toronto neighborhood data is merged with the csv file of geographical coordinates. Every neighborhood is marked by a blue dot on the map of Toronto (Figure 1).

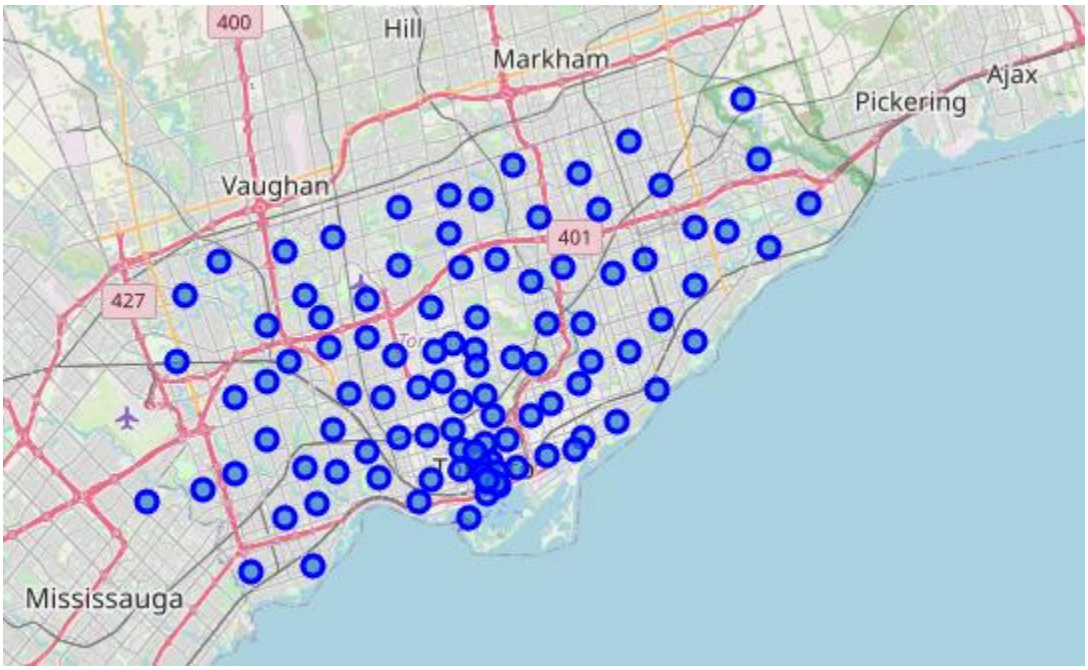

Figure 1. Toronto map

The venues data (with limit of up to 100 venues and 500 meters surrounding each neighborhood) is requested from Foursquare API. This data contains both main venue category and detailed venue category. This venues data then is merged with the Toronto neighborhood data to gain the data which contains information about main venues and detailed venues in each neighborhood. (Table 1).

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Main Venues Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant	Food
1	Malvern, Rouge	43.806686	-79.194353	Interprovincial Group	43.805630	-79.200378	Print Shop	Shop & Service
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar	Nightlife Spot
3	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Affordable Toronto Movers	43.787919	-79.162977	Moving Target	Travel & Transport
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank	Shop & Service

Table 1. The merged data of Toronto data and Venues data

Analyzing the Foursquare venue data.

A calculation the occurrences and percentage of different venue categories shows that more than half of the dataset belongs to the Food category (54.4%), followed by Shop & Service (20.6%).

Table 3 is the results calculated the 10 largest venues categories. The results show that most of the top 10 belong to Food category. Surprisingly, 'Park' appears at the rank of 4.

	Main Venues Category	Venue Category	Percent
2	Food	1148	54.4
6	Shop & Service	435	20.6
4	Outdoors & Recreation	188	8.9
3	Nightlife Spot	140	6.6
0	Arts & Entertainment	90	4.3
7	Travel & Transport	82	3.9
5	Professional & Other Places	22	1.0
1	College & University	7	0.3

Table 2. The occurrences and percentages of different venue categories.

	Venue Category	Number of venues
61	Coffee Shop	172
50	Café	99
214	Restaurant	67
196	Park	52
200	Pizza Place	48
154	Italian Restaurant	46
220	Sandwich Place	40
155	Japanese Restaurant	40
21	Bakery	39
145	Hotel	36

Table 3. The number of venues (shown for the top 10 venues)

Number of venues per neighborhood varies a lot. We observe that there are some neighborhoods that have only a few venues. (table 4)

	Neighborhood	Venue
0	Agincourt	3
1	Alderwood, Long Branch	8
2	Bathurst Manor, Wilson Heights, Downsview North	20
3	Bayview Village	4
4	Bedford Park, Lawrence Manor East	23

Table 4. The number of venues per neighborhood

3.2 Data preparation

In Toronto data, neighborhoods having 5 or less than 5 venues are removed to get reliable results.

To perform k-means clustering, the one hot encoding is performed.

3.3 Clustering

Find the optimal k

By elbow method and the silhouette score.

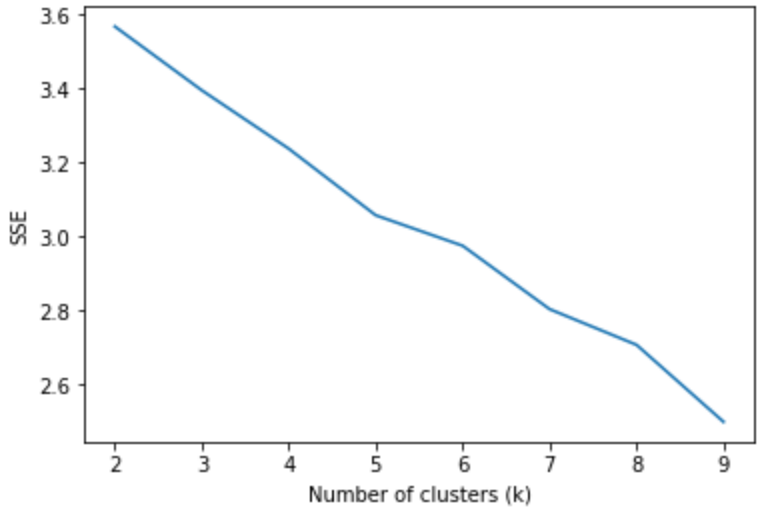

Figure 2. The elbow method

The elbow is not clear to define which k is optimal value. However, we can see that there is a slightly elbow around k = 5, 6, and 8. We need other method to find k. That is silhouette score. And here is the result of the silhouette score.

k	Silhouette score
2	0.0711
3	0.0747
4	0.1171
5	0.0701
6	0.143
7	0.1033
8	0.0602
9	0.1133

Table 5. The silhouette scores

The optimal number of clusters for the dataset is k = 6.

k-means clustering (k = 6)

k-means clustering is performed on the neighborhood venues data to cluster neighborhoods into groups that are similar by venue type. The visualization of the results is shown on the map of Toronto.

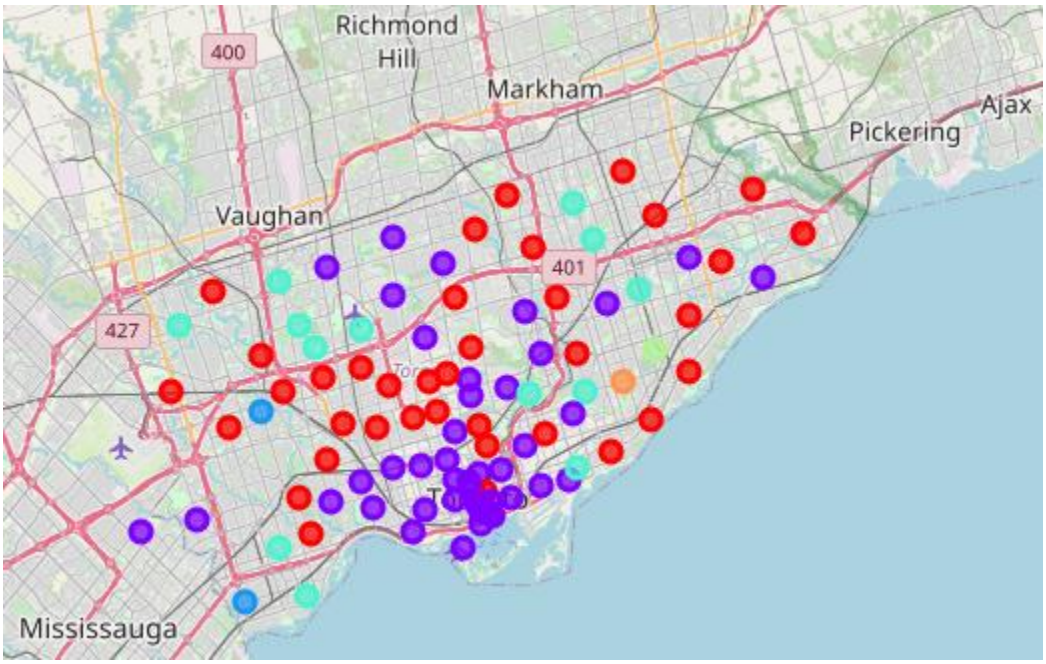

Figure 3. Toronto neighborhoods are divided into 6 clusters (colored).

Cluster	Cluster Labels	Color
1	0	Purple
2	1	Blue
3	2	Cyan
4	3	Olive
5	4	Orange
6	5	Red

Table 6. The cluster number, cluster labels and corresponding colors.

Red circles indicate the quiet neighborhoods.

We calculate the numbers and percentage of venues in each cluster.

Main venue category	Cluster Labels											
	0		1		2		3		4		5	
	No of venues	%	No of venues	%	No of venues	%	No of venues	%	No of venues	%	No of venues	%
Arts & Entertainment	7	3.5	81	4.8	0	0.0	1	0.5	0	0.0	0	0.0
College & University	1	0.5	5	0.3	0	0.0	0	0.0	0	0.0	0	0.0
Event	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Food	89	44.9	987	57.9	10	62.5	77	40.3	2	28.6	3	37.5
Nightlife Spot	6	3.0	126	7.4	1	6.2	3	1.6	0	0.0	0	
Outdoors & Recreation	10	5.1	122	7.2	3	18.8	31	16.2	0	0.0	1	12.5
Professional & Other Places	2	1.0	18	1.1	0	0.0	1	0.5	0	0.0	0	0.0
Residence	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Shop & Service	79	39.9	302	17.7	1	6.2	71	37.2	4	57.1	0	0.0
Travel & Transport	4	2.0	64	3.8	1	6.2	7	3.7	1	14.3	4	50.0

Table 7. The numbers and percentage of venues in each cluster.

We also summarize the top 5 common venues in each cluster.

0	1	2	3	4	5
Clothing Store	Coffee Shop	Pizza Place	Grocery Store	Discount Store	Bus Line
Coffee Shop	Café	Coffee Shop	Park	Hobby Shop	Bakery
Fast Food Restaurant	Restaurant	Sandwich Place	Fast Food Restaurant	Chinese Restaurant	Bus Station
Restaurant	Italian Restaurant	Intersection	Pizza Place	Coffee Shop	Ice Cream Shop
Japanese Restaurant	Japanese Restaurant	Middle Eastern Restaurant	Bank	Department Store	Intersection

Table 8. Top 5 most common venues in each cluster.

4. Results

Based on the above analysis, we can describe each cluster.

Cluster 1

This cluster is mainly located in the center of city. The most common venues categories are Food, Shop & Service, Outdoors & Recreation, and Arts & Entertainment. Within the Food category, the most common venues are Fast Food Restaurant and Japanese Restaurant.

Cluster 2

Most of restaurants in Toronto placed in these neighborhoods (987 food venues). This is a heaven for people loves food and hanging out with friends. This place is full of restaurants, shops, and universities. The most common venues categories are Food, Shop & Service, Arts & Entertainment, Nightlife Spot, and Outdoors & Recreation.

Cluster 3

The most common venues categories are Food, Outdoors & Recreation, Shop & Service, Travel & Transport, and Nightlife Spot. This cluster has an average number of Food venues. The most common venues in Food category are Pizza Place and Coffee Shops.

Cluster 4

This is the most diverse cluster. Here, we can find shops, restaurants, parks, banks, etc. This is the only cluster having 'Park' venue in the top 5 common venues. The most common venues categories are Food, Shop & Service, Outdoors & Recreation, and Travel & Transport.

Cluster 5

These neighborhoods are quite calm. There are few restaurants and coffee shops here. The most common venues categories are Shop & Service, Food, and Travel & Transport. The number of venues in this cluster is quite small.

Cluster 6

This is the quietest neighborhood. The number of venues reduce so much. There are few places for shopping or entertainments. The most common venues categories are Travel & Transport, Food, and Outdoors & Recreation.

5. Discussion

Observations and recommendation based on the results:

Base on the results, cluster 6 is highly recommended neighborhoods for people who love quiescence. Young students can rent rooms in cluster 2 for their study as well as youth life. Cluster 1, 2, 3, and 4 are great places for people who wants to know more about Toronto cuisine.

Recommendation for enhancement:

We can enhance the results by importing more data related to schools and price for renting houses. If there are these two indices in Toronto data, people can have more reliable results to make their decisions.

6. Conclusion

This notebook aims to describe different neighborhoods in Toronto, project results can help people identify where is quiet, where is more crowded with lots of shopping centers, etc. From that, they can decide which places fits their lifestyle. We can cluster Toronto neighborhoods into 6 clusters, which are differentiate by the number of venues per neighborhood and the most common venues in each cluster. Cluster 1 to 4 also show a common in high percentage of Food categories. While cluster 5 and 6 indicate

quiet neighborhoods. These results can be improved by adding more data about price of renting house and schoolings.