IBM Applied Data Science Capstone Project

Features of Toronto Neighbourhoods for Prospective Home Owners

By: Jerry Tan

Submitted: 20/03/2019

1. Introduction

For many people, choosing the right area to live in can be a difficult decision. While buying a house or signing a rental lease are likely to be some of the biggest decisions in an individual's life, such decisions are often made without enough knowledge of differences between the different housing areas in terms of restaurants, sports facilities available and so on.

This project will as such focus on identifying and comparing the features of different neighbourhoods in Toronto, in terms of the types facilities that are within close proximity of each neighbourhood. The target audience for this project are to-be home owners and renters.

For example, a person who is into food and dining might want to live in an area where there are more nice restaurants. In the same vein, a sports enthusiast may prefer to live in an area where there are many different types of sporting facilities.

Such comparisons are hard to conduct due to the impracticality of it – a person who wanted to physically scout every single housing area comparing the restaurants and facilities available would not quite have the time or energy required to do so.

Foursquare's location data provides a good solution to this problem – by aggregating the different facilities in each housing area and comparing them across the board, a person deciding where to stay can evaluate which areas may be more suitable for him or her, and hence make better housing purchases or rental decisions.

2. Data

The main data required for this analysis is data pertaining to neighbourhoods in Toronto. We first scraped the list of neighbourhoods in Toronto from Wikipedia at this page:

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

Using BeautifulSoup, we pulled data that looked like this, with the shape of the dataframe being (289, 3):

	Postcode	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
5	M5A	Downtown Toronto	Regent Park

This dataframe was cut to (211, 3) after removing rows whereby the 'Neighbourhood' column was not assigned.

We then created a function that used the Geopy module to pull the latitudes and longitudes of each neighbourhood, achieving a dataframe that looked like the following:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge	43.804930	-79.165837
1	M1B	Scarborough	Malvern	43.809196	-79.221701
2	M1C	Scarborough	Highland Creek	43.790117	-79.173334
3	M1C	Scarborough	Port Union	43.775504	-79.134976
4	M1C	Scarborough	Rouge Hill	43.780271	-79.130499
5	M1E	Scarborough	Guildwood	43.754899	-79.197776
6	M1E	Scarborough	Morningside	43.782601	-79.204958

Locations that were unable to be geocoded were assigned 0 for both latitude and longitude. After removing such locations, the dataframe was reduced to 198 rows (from 211).

The Foursquare API was then used to call for venues that are within the vicinity of each neighbourhood (within 500m radius), with a limit of 100 venues per neighbourhood. The eventual dataframe looked like this:

	Neighbourhood	Latitude	Longitude	Venue_Name	Venue_Latitude	Venue_Longitude	Venue_Category_1
0	Rouge	43.804930	-79.165837	Dean Park	43.804364	-79.169159	Park
1	Rouge	43.804930	-79.165837	Paul's Breakfast & Burgers	43.803835	-79.169825	Fast Food Restaurant
2	Malvern	43.809196	-79.221701	Shoppers Drug Mart	43.809202	-79.223320	Pharmacy
3	Malvern	43.809196	-79.221701	Subway	43.806805	-79.222515	Sandwich Place
4	Malvern	43.809196	-79.221701	Pizza Hut	43.808326	-79.220616	Pizza Place

This will give us the names and category of facilities within the surrounding of each neighbourhood. These venues were then manually coded into higher level venue categories, such as dining venues, sports facilities, entertainment venues and so on – which will be discussed in the following section.

3. Methodology

In order to give prospective home owners better information regarding potential neighbourhoods they can purchase their new homes in, the following methodology was adopted:

1. Categorize each venue into higher-level categories i.e. dining venues, nightlife venues, education venues etc. This was done manually, and ultimately resulted in the following higher-level categories:

```
['Sports' 'Dining' 'Healthcare' 'Shopping' 'Entertainment' 'Transport'
 'Business' 'Nightlife' 'Household' 'Education']
```

2. Summarize and find the neighbourhoods with the highest counts for each higher-level category
3. Cluster the neighbourhoods into unique clusters using k-means clustering for ease of comparison across similar neighbourhoods

4. Results

4.1 Summary of Higher-Level Categories

Overall, the count for each higher-level category was as follows:

```
Venue_Category_2
Business          39
Dining          3132
Education         10
Entertainment    432
Healthcare        74
Household         76
Nightlife        466
Shopping        1027
Sports           355
Transport        103
dtype: int64
```

The dataframe was grouped by neighbourhood into a new dataframe that looked like the following:

Venue_Category_2 Neighbourhood	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Adelaide	1	54	0	11	0	1	8	20	4	1
Agincourt	0	9	0	0	0	0	0	1	0	2
Agincourt North	0	18	0	2	1	1	2	7	0	0
Albion Gardens	0	1	0	3	0	0	0	1	1	0
Alderwood	0	4	0	0	1	0	1	0	3	0

The top neighbourhoods for each higher-level category was then found to be as follows:

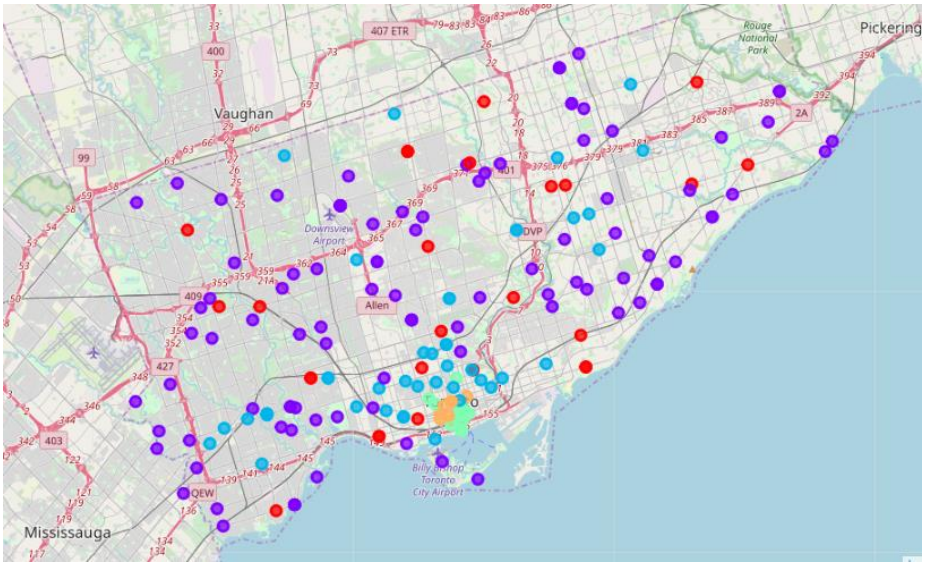
Venue_Category_2 Rank	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
1	Maple Leaf Park	Kensington Market	Jamestown	CN Tower	Deer Park	The Beaches West	Toronto Dominion Centre	Lawrence Heights	CN Tower	Runnymede
2	Adelaide	First Canadian Place	King and Spadina	Studio District	New Toronto	The Beaches	King	Yorkville	Harbourfront West	The Danforth West
3	First Canadian Place	Chinatown	Fairview	Princess Gardens	Malvern	Parkwoods	Commerce Court	Golden Mile	Harbourfront	Bathurst Quay
4	Flemingdon Park	Design Exchange	Mount Olive	Harbourfront West	Runnymede	The Junction North	First Canadian Place	Garden District	Harbourfront East	Wexford
5	Golden Mile	Toronto Dominion Centre	Princess Gardens	Harbourfront East	Lawrence Park	Runnymede	Little Portugal	Trinity	King and Spadina	Guildwood
6	Grange Park	Commerce Court	Martin Grove	Harbourfront	St. James Town	Lawrence Park	Design Exchange	Grange Park	Richmond	Riverdale
7	Guildwood	Willowdale	Richmond	Commerce Court	Parkdale Village	St. James Town	Church and Wellesley	Scarborough Town Centre	Studio District	Maryvale
8	Harbourfront	Willowdale South	Garden District	King	Parkdale	The Junction South	North Toronto West	Adelaide	The Beaches	Parkdale
9	Harbourfront East	Willowdale West	Grange Park	Trinity	Maryvale	First Canadian Place	Northwest	South Niagara	The Beaches West	Harbourfront West
10	Harbourfront West	King	Studio District	Toronto Dominion Centre	West Hill	Martin Grove	Studio District	Richmond	Design Exchange	Thorncliffe Park

4.2 Clustering Analysis

The data was then normalized, and k-means clustering was performed for k = 5:

Venue_Category_2 Neighbourhood	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
Adelaide	0.4	0.541978	-0.051282	0.366026	-0.189744	0.203419	0.400733	0.288889	0.217949	0.117949
Agincourt	-0.1	-0.100879	-0.051282	-0.092308	-0.189744	-0.129915	-0.170696	-0.083660	-0.182051	0.367949
Agincourt North	-0.1	0.027692	-0.051282	-0.008974	0.310256	0.203419	-0.027839	0.033987	-0.182051	-0.132051
Albion Gardens	-0.1	-0.215165	-0.051282	0.032692	-0.189744	-0.129915	-0.170696	-0.083660	-0.082051	-0.132051

The neighbourhoods were then mapped out by clusters, creating a map of Toronto that looked as such:



Legend {Cluster: Color} = {0: Red, 1: Purple, 2: Blue, 3: Green, 4: Orange}

The means of count for each category by cluster are summarized in the table below:

Venue_Category_2 Cluster	Business	Dining	Education	Entertainment	Healthcare	Household	Nightlife	Shopping	Sports	Transport
0	0.038462	21.807692	0.0	0.769231	1.461538	1.115385	1.730769	5.653846	2.038462	1.000000
1	0.150000	2.560000	0.0	0.520000	0.120000	0.070000	0.310000	1.190000	1.170000	0.410000
2	0.136364	23.068182	0.0	1.931818	0.500000	0.568182	3.545455	9.954545	1.409091	0.454545
3	0.800000	56.066667	0.0	10.666667	0.066667	0.533333	11.000000	11.266667	5.533333	1.000000
4	0.500000	45.300000	1.0	11.500000	0.100000	0.700000	6.900000	15.400000	4.000000	0.100000

5. Discussion

This research provides some key insights for prospect home owners and renters. For example, people who have kids may want to find housing in cluster 4 neighbourhoods, where there are educational facilities nearby.

Cluster 3 and 4 neighbourhoods have the highest concentration of dining, shopping and entertainment venues, suggesting them to be around the city centre – where property and rental prices may also be more expensive.

In general, home owners and renters may want to avoid cluster 1 neighbourhoods which are seemingly lacking in all sorts of facilities.

Cluster 0 and cluster 2 are likely to be the cheaper alternatives for prospect home owners, in which case families may wish to live in cluster 0 neighbourhoods given the higher concentration of healthcare, household and sports facilities. On the same note, younger people may want to consider cluster 2 neighbourhoods over cluster 0 neighbourhoods, given the higher availability of entertainment, nightlife and shopping venues.

6. Conclusion

There are ways to improve on this work, some of which are listed below:

1. Increase the radius and limit when pulling the Foursquare venue data – for deeper analysis, we could pull locations that are within 500m, 1000m, 2000m (and so on), so that the audience can further consider whether venues within the range of radiuses can be satisfactory enough for their living.
2. Augment the research with property and rental prices – while information regarding nearby facilities are important, price plays a huge role in the audiences' decisions.

Overall, this report gives a simple but good insight into what can be achieved with simple use of python modules and APIs such as the Foursquare API.