Exploring Segmentation of Toronto Neighborhoods for Prospective Tenants

Coursera Capstone – Aug 2020

Ameya Sohoni *ameyasohoni@outlook.com*

1 Problem Introduction

The city of Toronto is the most populous city in Canada with a recorded population of 2,731,571 in 2016. In addition to being the most populous city and the economic capital of Canada, Toronto is also home to the largest migrant population of any Canadian city. In 2019, Toronto saw an influx of 313,580 immigrants from various corners of the world. For newcomers to a metropolitan area, an essential decision is finding the right place to stay. This decision can be influenced by a plethora of factors including rent rates, safety, convenience, transit to name a few. While individual preferences vary on what weight to place on which parameter, it would be greatly useful to have a scientific segmentation of neighborhoods based on key deciding features, to facilitate the decision.

In this report, I have attempted to segment Toronto neighborhoods based on a few key features which would allow prospective tenants to make an informed decision on which location to select. While this analysis is geared to suit a decision on place of stay, the findings are likely suitable for other scenarios such as places of business and erection of public utilities as well.

2 Dataset and Feature Selection

The data required for this analysis was extracted from multiple sources based on the required feature-set. To make a decision on place of stay the key independent parameters considered are:

1. Average rent (for 2-bedroom apartment) - CA$/month
2. Crime rates – Major Crimes/100k population
3. Convenience – Availability of shops, offices, stores
4. Cultural/ Community – Proximity to movie halls, parks, outdoor recreation venues
5. Social Life – Proximity to dining and nightlife venues
6. Education – Availability of schools, college/ university
7. Transit – Proximity to travel options

The primary list of neighborhoods was extracted from the Toronto police website [1] which segregates the city into 140 neighborhoods. While there are other ways to divide the city into neighborhoods, this particular division is a standard method used by the city of Toronto which provided a very convenient and unbiased dataset. The Toronto police website also provided data for crime rates (see pt. 2 above).

Data for average rent by neighborhood was taken from zumper.com [2] which is a popular portal for rental seekers. However, the neighborhoods used by zumper (let's call them rent neighborhoods) do not match with the standard neighborhood divisions (let's call them base neighborhoods) considered in my primary dataset. As a result, I had to make an approximation using lat/long coordinates. I did this by:

1. Checking which rent neighborhoods fall within the base neighborhood and using the corresponding rent or mean rent (if more than one fall within)
2. For the base neighborhoods within which NO rent neighborhood falls, the nearest rent neighborhood to the base neighborhood center is considered

While this approach is not perfect, in relative terms it provides a good approximation of the average rent for each neighborhood, considering the lack of freely available data.

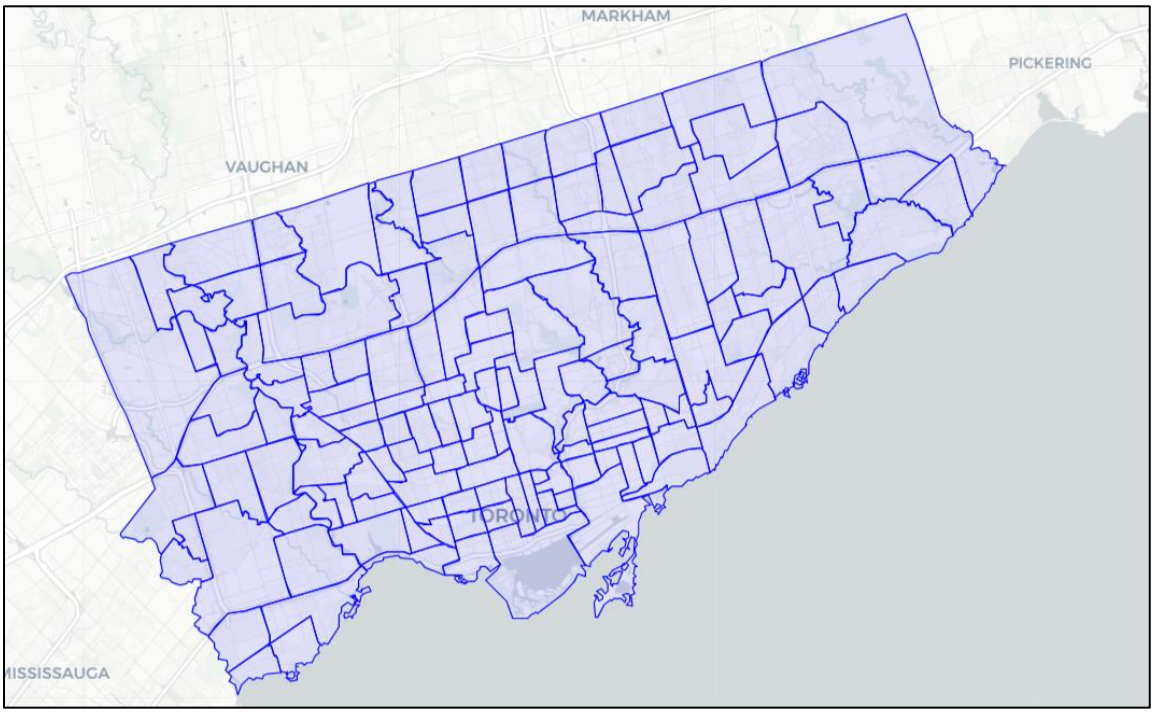

Fig 1: Base Neighborhoods in Toronto

For the remaining parameters *viz. Convenience, Cultural/ Community, Social Life, Education & Transit*, I used the Foursquare.com Places API to extract (upto 100) top recommended venues around the neighborhood center. This provided a representative sample of the neighborhood's accessibility to similar venues.

The original results of the API GET query to Foursquare however, grouped venues in too many disparate categories. For example, a venue such as a restaurant can be categorized as Afghan Restaurant or Italian Restaurant or Pizza Place. While this is useful in knowing granular details about the venue; it is counter-productive if you are trying to segment the data into decision support parameters such as in our case.

To overcome this problem, I have used Foursquare's own categorization of its venues. This grouping can be queried with a GET request to the API [3]. This grouping reduces the number of venue categories from hundreds to a more manageable ten.

Feature Category	
Arts & Entertainment	66
College & University	38
Event	12
Food	357
Nightlife Spot	26
Outdoors & Recreation	107
Professional & Other Places	110
Residence	5
Shop & Service	178
Travel & Transport	56

Table 1: Truncated list of categories from Foursquare's API call

With this reduced list we can clearly see which categories fit within our pre-defined parameters.

1. **Culture and Community** - Arts & Entertainment, Event, Outdoors & Recreation
2. **Education** - College & University
3. **Socializing** - Food, Nightlife Spot
4. **Convenience** - Professional & Other Places, Residence, Shop & Service
5. **Transit** - Travel & Transport

This gives us our final set of features and the associated venues.

Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category
Yonge-St.Clair	The Bagel House	43.687374	-79.393696	Bagel Shop
Yonge-St.Clair	Capocaccia Café	43.685915	-79.393305	Italian Restaurant
Yonge-St.Clair	The Market By Longo's	43.686711	-79.399536	Supermarket
Yonge-St.Clair	Union Social Eatery	43.687895	-79.394916	American Restaurant
Yonge-St.Clair	LCBO	43.686991	-79.399238	Liquor Store
...	...	...	...	...
Briar Hill-Belgravia	Adrenalin Fitness	43.705790	-79.455548	Gym / Fitness Center
Briar Hill-Belgravia	Fairbank Memorial Park	43.692028	-79.448924	Park
Briar Hill-Belgravia	Andrew's Formals	43.697521	-79.442088	Men's Store
Briar Hill-Belgravia	Nairn Park	43.690654	-79.456300	Park
Briar Hill-Belgravia	Maximum Woman	43.690651	-79.456333	Women's Store

Table 2: Sample table of venues from GET request to Foursquare API call

3 Methodology

The preliminary dataset consists of data on Rent (Average Rent), Crime (Crime Rate) and Category (*Convenience, Cultural/ Community, Social Life, Education & Transit*).

Neighborhood	Venue	Total Crime	Rent	Category
Yonge-St.Clair	The Bagel House	646.5	3300.0	Socializing
Yonge-St.Clair	Capocaccia Café	646.5	3300.0	Socializing
Yonge-St.Clair	The Market By Longo's	646.5	3300.0	Convenience
Yonge-St.Clair	Union Social Eatery	646.5	3300.0	Socializing
Yonge-St.Clair	LCBO	646.5	3300.0	Convenience
...	...	...	...	...
Briar Hill-Belgravia	Adrenalin Fitness	1276.6	2250.0	Culture and Community
Briar Hill-Belgravia	Fairbank Memorial Park	1276.6	2250.0	Culture and Community
Briar Hill-Belgravia	Andrew's Formals	1276.6	2250.0	Convenience
Briar Hill-Belgravia	Nairn Park	1276.6	2250.0	Culture and Community
Briar Hill-Belgravia	Maximum Woman	1276.6	2250.0	Convenience

Table 3: Preliminary dataset for clustering

Taking a look at the spread of crime and rent data over Toronto city.

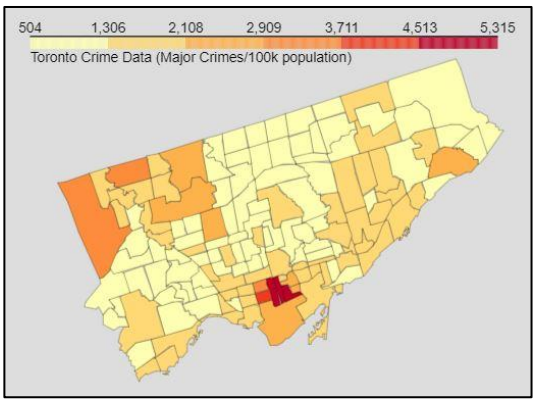

Fig 2: Crime rates (Major Crimes/ 100k)

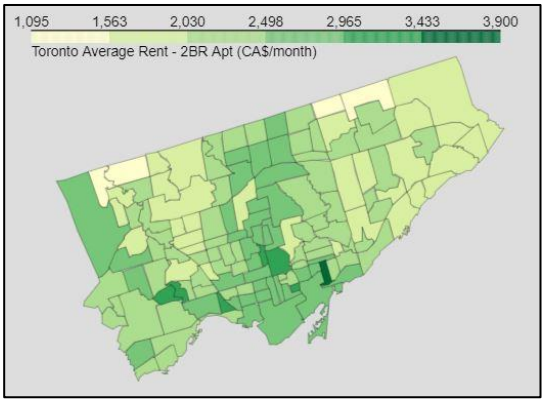

Fig 3: Average rent – 2BR Apt (CA$/month)

To segment these neighborhoods, I have used a clustering algorithm to group together similar neighborhoods. Using an unsupervised clustering algorithm (k-Means Clustering), neighborhoods with similar intra-group traits and dissimilar inter-group traits are identified.

To use k-means clustering the following pre-requisites must be satisfied:

1. Data available with *f* independent variables (features) and *n* observations available in *(f,n)* array

2. Normalized data (std = 0) (optional but preferable)
3. Number of clusters to group data into *(k)*

Using pandas to re-arrange data: The preliminary dataset (Fig. 2) was converted into a 7 feature (column) dataframe using a method called onehot encoding. This can be achieved in pandas with the pd.dummies() function. This in combination with grouping the dataframe by 'Neighborhood name' gives us a dataframe that satisfies pt 1 above.

Neighborhood	Total Crime	Rent	Convenience	Culture and Community	Education	Socializing	Transit
Agincourt North	735.2	2100.0	0.365854	0.024390	0.0	0.609756	0.000000
Agincourt South-Malvern West	1384.8	2100.0	0.184211	0.105263	0.0	0.710526	0.000000
Alderwood	730.1	2540.0	0.473684	0.157895	0.0	0.315789	0.052632
Annex	1978.8	2744.0	0.180000	0.110000	0.0	0.690000	0.020000
Banbury-Don Mills	798.1	2250.0	0.304348	0.108696	0.0	0.586957	0.000000

Table 4: Using pandas to create (140,7) dataset

As can be seen the data for Crime and Rent is on a very different scale from the data for venue Categories. The Categories data stipulates the probability of finding a category of venue e.g. Education in the neighborhood, whereas, the crime and rent data are absolute values for Major Crimes/ 100k people and Average CA$/month rent respectively.

This is not a hindrance in clustering analysis since the cluster centroids are formed using relative data for each feature. However, it does help to normalize the data (std = 0).

```
from sklearn.preprocessing import StandardScaler

to_clustering = StandardScaler().fit_transform(to_grouped.drop(columns='Neighborhood'))
```

Fig 3: Normalize the data using StandardScaler()

To determine how many clusters the data should be grouped into, I have used the "elbow method".

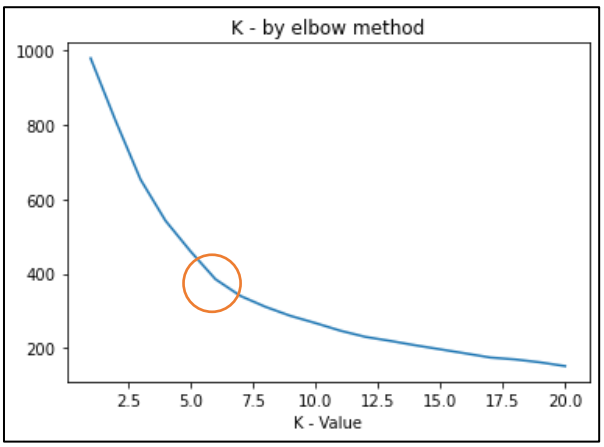

Fig 4: Finding no. of clusters using the elbow method

*Here we can clearly see an inflexion point at **k = 6**, which is selected as the optimal cluster number.*

4 Results

The k-means algorithm results are summarized below:

Cluster Labels	Total Crime	Rent	Convenience	Culture and Community	Education	Socializing	Transit
0	1284.763265	2079.418367	0.402792	0.122121	0.000000	0.458487	0.016601
1	1085.800000	2417.888889	0.167948	0.692829	0.013177	0.126047	0.000000
2	1304.528571	2612.275510	0.181470	0.151817	0.000305	0.650608	0.014185
3	1316.700000	2750.000000	0.000000	0.222222	0.222222	0.444444	0.111111
4	4527.020000	2729.333333	0.212000	0.138000	0.010000	0.626000	0.014000
5	1199.111111	2104.888889	0.258955	0.168825	0.000000	0.461577	0.110082

No. of clusters [k] = 6 | Inertia [$\sum (\bar{x} - xi)^2$] = 385.3 | No. of iterations [n] = 8

Table 5: k-means clustering results

The algorithm grouped the data into 6 distinct clusters. Taking a look at the size of each cluster, we see that Clusters 0 and 2 are the most popular with the highest count of neighborhoods:

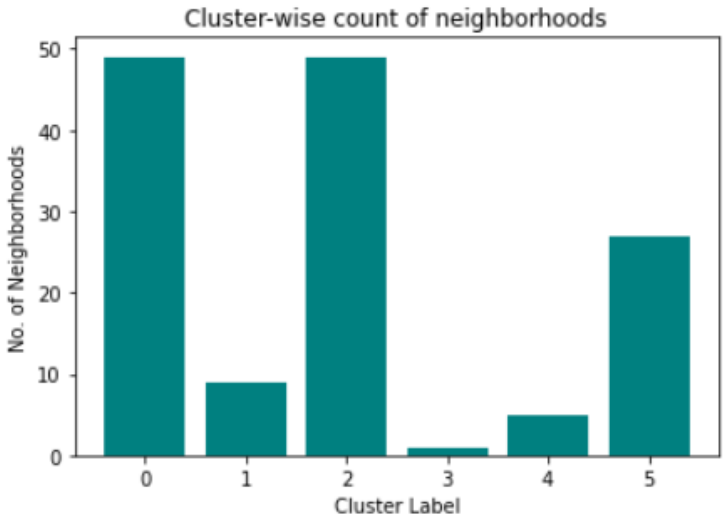

Fig 5: Cluster-wise count of neighborhoods

We can get a better idea of the salient characteristics of each cluster if the results are converted to rankings instead of absolute values.

Cluster Labels	Total Crime	Rent	Convenience	Culture and Community	Education	Socializing	Transit
0	4.0	1.0	1.0	3.0	3.0	6.0	6.0
1	1.0	4.0	4.0	1.0	3.0	5.0	5.0
2	5.0	5.0	5.0	2.0	1.0	3.0	2.0
3	6.0	6.0	3.0	6.0	2.0	1.0	4.0
4	3.0	3.0	6.0	4.0	3.0	4.0	1.0
5	2.0	2.0	2.0	5.0	3.0	2.0	3.0

Table 6: k-means results viewed as ranks instead of absolute values

5 Discussion

From a cursory look at the results the clustering looks to have been effective. While machine learning algorithms may not be essential for a sample size of 140 neighborhoods, the methodology followed can easily be extended to an entire province or even larger area.

The primary introduction of error can be the somewhat low-quality data on average rent. This can be circumvented by scraping data from real estate portals using python libraries such as Scrapy or Selenium. However, this was considerably beyond the scope of this report.

The clusters formed do provide a good overall idea of the type of neighborhood, which was the key problem statement. Preferences vary from individual to individual - a young unmarried person may want to live in lively location with a lot of eat-out options, while someone with a large family might prefer a location with more community activities such as parks & community centers. The segmentation done here can at least provide a newcomer with a partial understanding of what to expect in a particular neighborhood.

6 Conclusion

Comparing the neighborhood clusters from Table 6, we can conclude:

1. Cluster 0: Low Rent, High Convenience, Poor Transit and Dining options
2. Cluster 1: Best Cultural activities (Community Centers, Parks, Events) and most safe
3. Cluster 2: Best for Dining and Nightlife
4. Cluster 3: Best for Educational institutes and transit/ stay options
5. Cluster 4: Good for Dining/ Nightlife but mediocre/poor for everything else
6. Cluster 5: Good for most parameters, well balanced

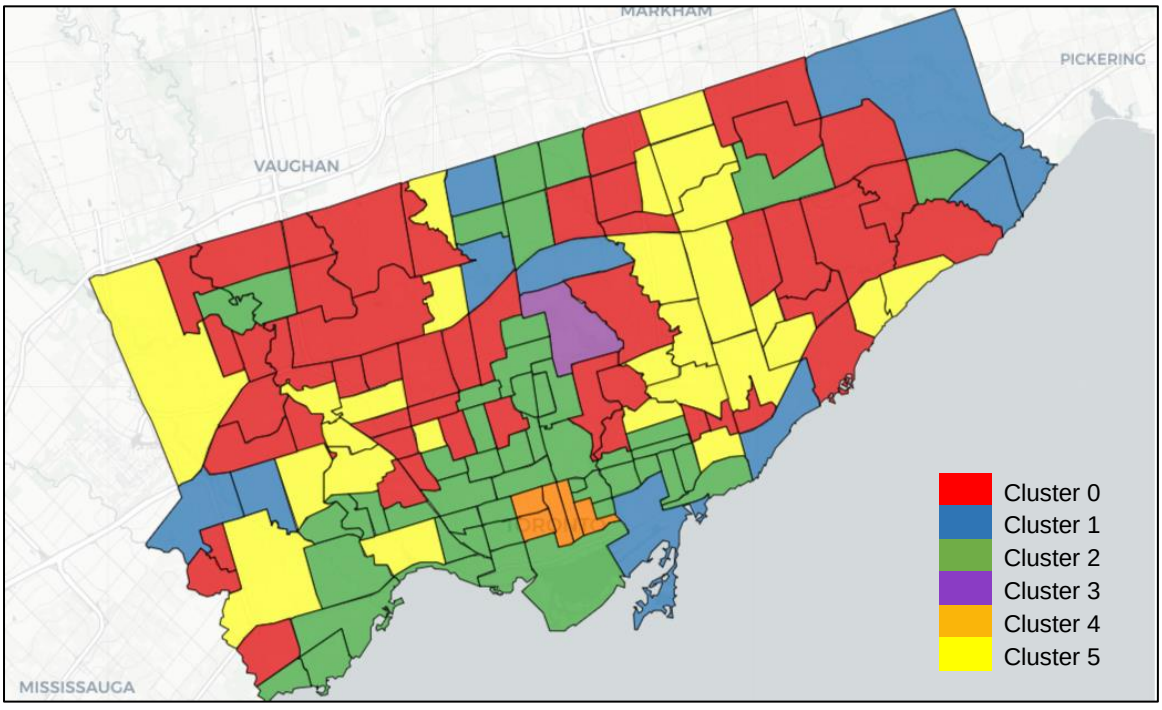

Fig 6: Segmentation of Toronto Neighborhoods

Key insights: With this method I was able to segment the city neighborhoods into 6 distinct clusters with significant representative characteristics. On closer inspection of the cluster map (Fig 6) we can see that Cluster 5 which is the most well-balanced cluster as per our parameters is quite even distributed across the city. This spells good news for prospective tenants as this would indicate there is a sensible location option in most parts of the city.

Another interesting insight to look into is that Cluster 3 which is home to the most educational institutions is also surrounded by areas in Cluster 0 and Cluster 1. This means what students who are willing to undertake a short transit have low rent and low crime options available.

Future: While this analysis does provide some worthwhile insights, the possibilities with machine learning clustering algorithms are endless. With a more granular division of the city and better data on rent we can achieve an even better understanding of the defining characteristics of neighborhoods and make more informed decisions.

7 References

1. Neighbourhood Crime Rates (Boundary File): https://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file-/data
2. Average Rent in Toronto, ON and Cost Information: https://www.zumper.com/rent-research/toronto-on
3. Venue Categories | Build with Foursquare: https://developer.foursquare.com/docs/build-with-foursquare/categories/
4. Toronto City GeoJson file (Courtesy github user jasonicarter): https://raw.githubusercontent.com/jasonicarter/toronto-geojson/master/toronto_crs84.geojson