Toronto Neighborhood Clustering based on Citizen Comfort Level

Priambudi Pujihatma

Background

Purpose of this analysis is to cluster Toronto Area based on several factors that affect citizen comfortable living. This analysis will be useful for those who are going to relocate to Toronto and to new families who are looking for the best neighborhood to live. The comfort level is defined by the following factors:

- Apartment rental price
- Presence of supporting facilities, such as: restaurant, coffee shop, park, health facilities / gym, grocery stores, farmers market, café and bakery.

Output of this analysis is the neighborhood clusters along with plus delta aspect of each neighborhood. Thus, the potential Toronto residents can choose the best cluster based on their personal preference.

Data

This project utilizes three data sources:

- List of Toronto borough, neighborhood and borough codes as shown in https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and coordinate for each borough
- Apartment rental price. The source for apartment rental price data is taken from Kaggle through this link: https://www.kaggle.com/rajacsp/toronto-apartment-price. The data is from 2018, which consists of 1125 apartment rental price data, along with the address, number of rooms, dens, bathrooms and coordinates for each apartment.
- The source of supporting facilities is taken from Foursquare. The default maximum number of events to be retrieved from Foursquare is 100. To achieve 100 venues per borough, the search radius is increased up to 5 km.

Data Preprocessing

Using Pandas, the list of Toronto borough, neighborhood and borough codes can be extracted into neighborhood data frame. The neighborhood will be used as a key to connect the apartment rental price and facilities data. The coordinate for each neighborhood is added to the existing data frame. Google maps can be used to identify each neighborhood coordinates. The map of Toronto with each neighborhood coordinates was generated using Folium and is shown in Fig. 1

The apartment rental price data has the address and coordinates for each apartment. The address data contains information related to borough. However, the borough is not explicitly separated in a dedicated column. The typical way to write a borough code on a Canadian address is by writing the borough code after the province, as shown in Table 1, e.g.: the first data has the borough code M5V written after the province code (ON, which stands for Ontario). Thus, ON can be used as an identifier to split the address data and get the borough code value. A spreadsheet application is used to split the address using text to

column feature. The end result is an apartment data frame with dedicated column which contains the borough code for each apartment.

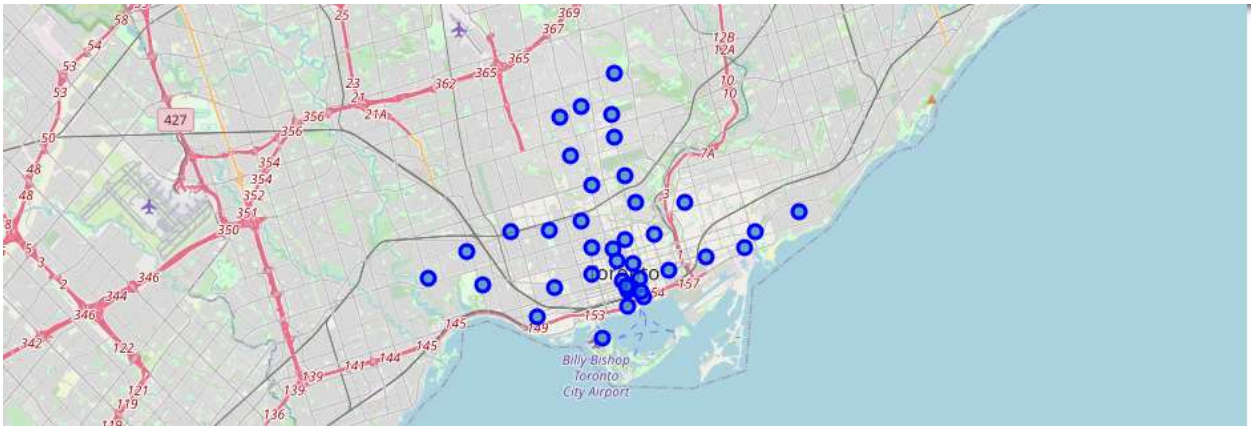


Fig 1: Map of Toronto Neighborhood

Table 1: Example of Apartment Location Data from Kaggle

Address	Lat	Long
361 Front St W, Toronto, ON M5V 3R5, Canada	43.64305	-79.3916
89 McGill Street, Toronto, ON, M5B 0B1	43.66061	-79.3786
10 York Street, Toronto, ON, M5J 0E1	43.64109	-79.3814

Once the Toronto borough code is separated, a merge function can be executed to add neighborhood and neighborhood coordinates from the neighborhood data frame to the apartment data frame. Afterwards, using group by function, the average apartment rental cost for each borough can be calculated. The head of the data frame is shown in Table 2.

Table 2: Head of Apartment Data Frame

Neighborhood	Latitude	Longitude	Price
Berczy Park	43.6447708	-79.3733064	$2,680
Brockton, Parkdale Village, Exhibition Place	43.6368472	-79.4281914	$2,057
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	43.6289467	-79.3944199	$2,474
Central Bay Street	43.6579524	-79.3873826	$1,856
Christie	43.669542	-79.4225637	$2,842

The location data is taken using Foursquare API with 5 km search radius from each neighborhood center. The API resulted in 100 venues for each neighborhood, stored in the location data frame. The search result is then categorized based on venue type. Overall there are 165 unique venue categories. The average value of each category is calculated, which represent the ratio of each venue category when compared to overall venue available in each neighborhood. Out of 165 unique categories, 39 of them are different types of Restaurants, ranging from American Restaurant to Vietnamese Restaurants. Thus, using Pandas,

the columns that contains the word “Restaurant” in the column name are added together and grouped as one category “Restaurant”.

Prior to clustering, the apartment data frame and the location data frame are merged using neighborhood as key. For better clustering accuracy, the average apartment rental price is scaled using min max scaler. The selected algorithm for clustering is k-means with 5 categories.

Result and Discussion

The clustering output was stored in a summary data frame. Summary data frame contains each neighborhood, cluster labels, the average apartment rental price (already converted back to original values) and the ratio of eight selected venues: restaurant, coffee shop, park, health facilities / gym, grocery stores, farmers market, café and bakery. The head of the summary data frame is given in Table 3.

Table 3: Head of Summary Data Frame

Cluster	**Neighborhood**	**Price**	**Restaurant**	**Coffee Shop**	**Park**	**Gym**	**Grocery Store**	**Farmers Market**	**Café**	**Bakery**
3	Berczy Park	2680	0.19	0.11	0.07	0.03	0	0.03	0.01	0.02
4	Brockton, Parkdale Village, Exhibition Place	2057	0.23	0.04	0.11	0.02	0.01	0	0.05	0.04
0	CN Tower, King and Spadina, Railway Lands	2474	0.22	0.04	0.08	0.04	0	0.02	0.04	0.04
2	Central Bay Street	1855	0.26	0.07	0.03	0.03	0	0.02	0.05	0.01
3	Christie	2841	0.22	0.06	0.07	0	0.01	0	0.07	0.04

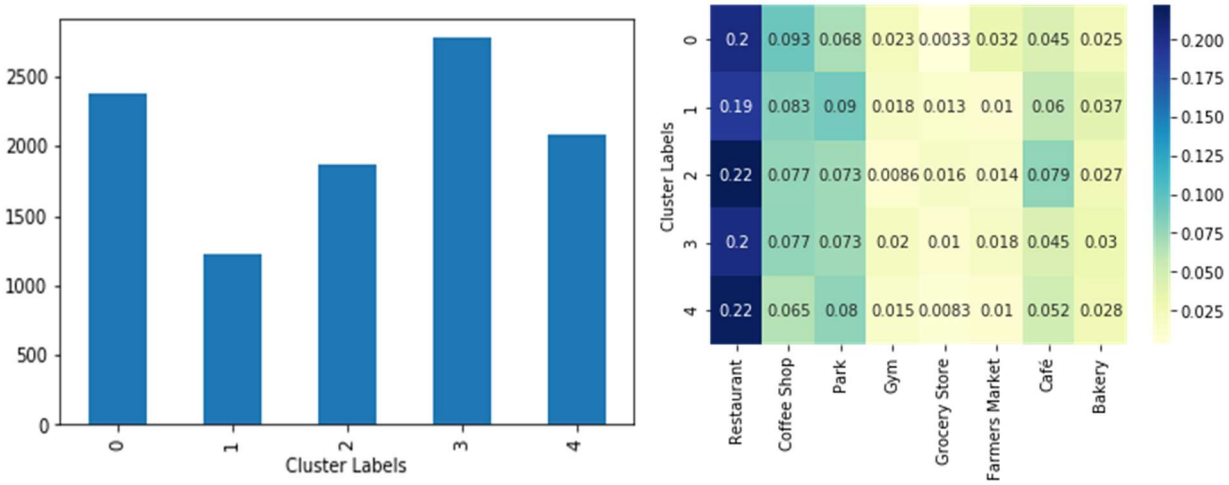
The summary data base will be used for further analysis. The clustering process resulted in 5 cluster labels, summarized in Table 4.

Cluster Labels	Neighborhood
0	CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport, Church and Wellesley, Garden District, Ryerson, Harbourfront East, Union Station, Toronto Islands, Regent Park, Harbourfront, Studio District
1	India Bazaar, The Beaches West, North Toronto West, Lawrence Park, Runnymede, Swansea, The Beaches
2	Central Bay Street, Dufferin, Dovercourt Village, Parkdale, Roncesvalles, Rosedale, St. James Town, Cabbagetown, Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park, University of Toronto, Harbord
3	Berczy Park, Christie, Davisville, Richmond, Adelaide, King
4	Brockton, Parkdale Village, Exhibition Place, Kensington Market, Chinatown, Grange Park, Little Portugal, Trinity, St. James Town, The Annex, North Midtown, Yorkville, The Danforth West, Riverdale

Using Folium, the cluster result is represented using map as shown in Fig 2. The cluster 0, 1, 2, 3, 4 are represented with red, purple, light blue, green and orange dots respectively.

Fig 2: Clustering Result

Each cluster plus and delta is analyzed using charts shown in Fig. 3(a) and Fig. 3(b) shown below.



(a) (b)

Fig 3: (a) Average Apartment Price (b) Proportion of Venue

Using the above chart, several conclusions can be drawn:

- Cluster 1 has the lowest average apartment cost. Fig. 2 shows that cluster 1 is located at the edge of downtown Toronto. In terms of venue, Cluster 1 has more Park compared to other clusters. There is not a significant difference of Restaurant between Cluster 1 and other clusters. It also has the highest proportion of bakery. However, Cluster 1 lacks of other supporting facilities, such as gym, grocery store, farmer marker and café.
- Cluster 0, 3 and 4 are “downtown clusters” that have high apartment cost. As expected, downtown area has so many supporting facilities which outweigh cluster 1.

- Cluster 2 has medium average apartment cost. The location is still on the outer ring of cluster 0, 3 and 4 however they are located closer to downtown compared to Cluster 1. Cluster 2 can also compete with Cluster 0, 3 and 4 in terms of supporting facilities, especially for restaurants, café and bakery.

Conclusion

This analysis provides Toronto area clustering based on comfort level, which is defined by apartment rental cost and the presence of supporting facilities. Each cluster has their own plus and delta which can be used as further consideration for potential Toronto residents. The results show that Cluster 1 is the area with the lowest rental cost, however it lacks the supporting facility when compared to other clusters. The best tradeoff is found in Cluster 2, with medium rental cost and supporting facility that can compete with downtown clusters.