Exploring and Cluster Toronto's borough for Homebuyers

Phuong Linh Nguyen

Feb 22 ,2020

1.Introduction:

1.1 Background: With a population of 6,139,000, Toronto is the most populous city in Canada. It is Canada's commercial and financial centre. It is also a very diverse city. 51% of its residents were born outside of Canada. It is home to 230 different nationalities. Due to its characteristics, 17% of Canada immigrants choose Toronto as their final destination.

1.2 Problem : With large amounts of immigration per year, housing is one of the biggest concerns for newcomers . Different households will have different budgets and needs. There are several factors that home buyers can look into when selecting a neighborhood for their home like: common venues in the areas and house price.

1.2 Interested Audience: This project will help home-owners and real estate agents gain insight about different neighborhoods in Toronto considering these following factors: common venues and house price.

2. Data Acquisition and Cleaning

2.1. Data Source Description:

The following data will be used to support this report:

- List of boroughs in Toronto with its latitude and longitude. The original data got 3 columns (namely postal code, borough and neighborhood) and 288 rows. The latitude and longitude for each group of neighborhood is retrieved through Foursquare.

- Average house price across areas in Toronto:There are 35 districts with house price
- List of districts and neighborhoods: There are 33 districts with its neighborhoods.

2.2 How the data will solve the problem:

Use Foursquare and geopy to find top 10 venues for each neighborhood: People who love eating and dining out might be interested in neighborhoods with lots of restaurants and coffee shops. Families who have kids usually prefer a neighborhood with lots of parks. Individuals who pursue an active lifestyle might be interested in looking into areas with gym and trails.

Average house prices across different regions will help home-buyers set their budget and financial plans to meet their needs.

Ethnicities for residents in different neighborhoods of Toronto will help home buyers select a suitable property. Living in an area with people who have similar backgrounds can create a sense of belonging.

2.3 Data cleaning:

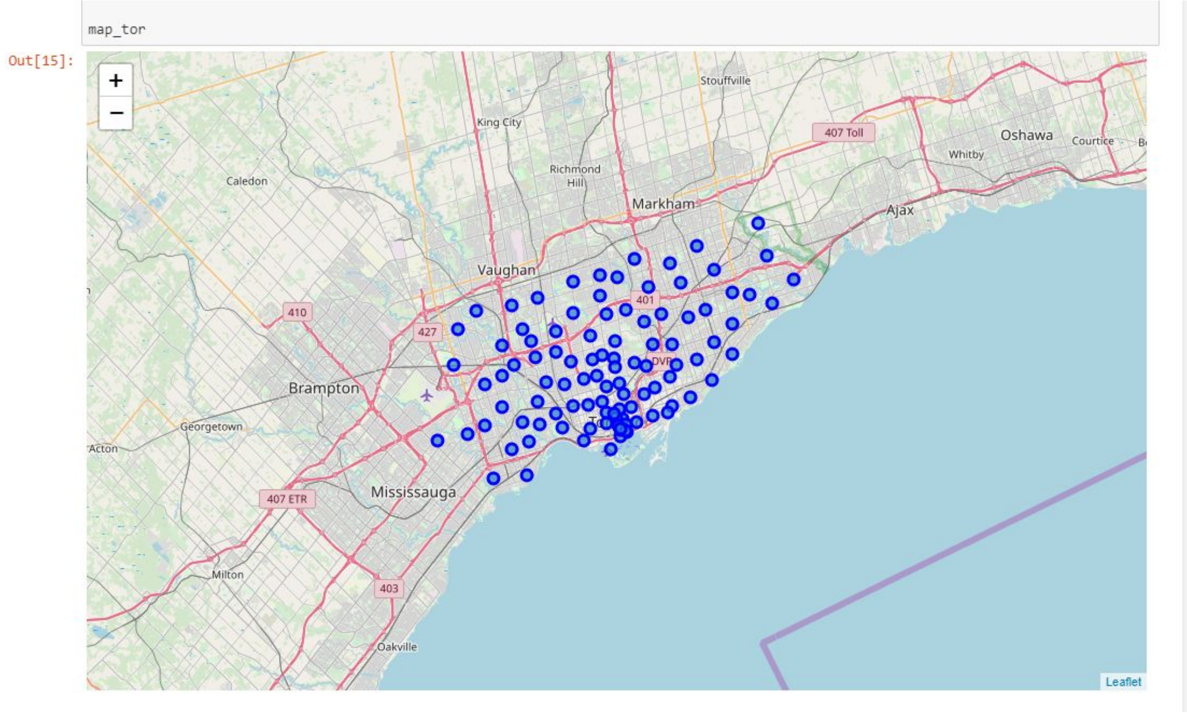
List of boroughs in Toronto is retrieved here. Originally, the data frame has 3 columns (Neighborhood, Borough and postal code) with 288 rows. There are 78 ‘Not assigned’ values for neighborhoods and 77 “Not assigned’ values for boroughs. I dropped all the postal codes that have missing Borough. If a postal code has borough value but does not have a value for neighbor, I assign the neighborhood the same as its borough. Neighborhood is grouped if they share the same Boroughs. I ended up with a data frame with the original 3 columns and 103 rows.

The list of districts and neighborhoods in Toronto is retrieved from here.
Regarding house price data,I created a data frame for here. Since this data is under PDF format, I do data entry for each district of Toronto. Since the data frame on wikipedia has 33 districts while the data frame on house price has 35 districts. I dropped those 2 districts that are not on wikipedia data frame. I join these two tables together and end up with a data frame on house prices for 33 districts with a group of neighborhoods in each district.

3 Methodology

3.1 Explore and cluster the neighborhood of Toronto:

I visualized all the borough superimposed through folium. I created the map by using boroughs latitude and longitude.



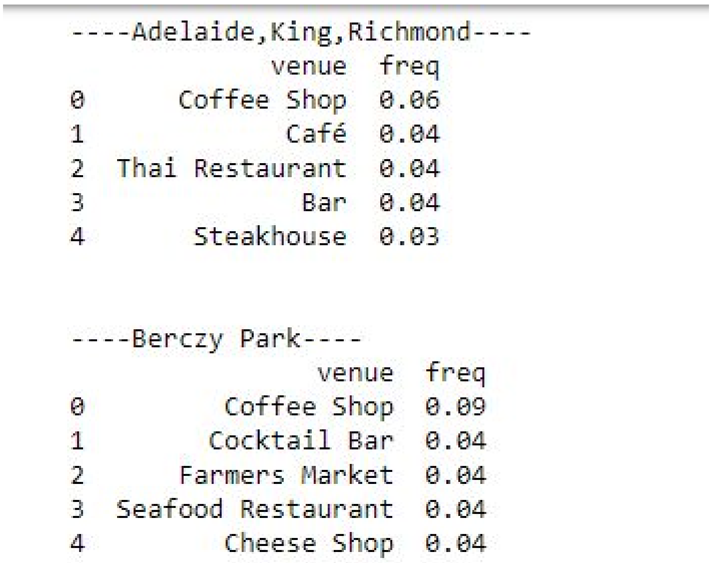
```
map_tor
```

Out[15]:

I decided to focus on the 4 main boroughs of Toronto for this part: Downtown Toronto, East Toronto, Central Toronto and West Toronto. I dropped all other boroughs on the data frame.Then I retrieve all the venues (including venues name, its latitude & longitude and its category) that are located in each cluster of neighborhoods belonging to those 4 boroughs. There are 1717 venues that are retrieved.

I count the number of venues for each neighborhood. The amount of venues retrieved for each group of neighborhood is ranging on a wide spectrum. Some neighborhoods have a high number of venues located (100). Some neighborhoods have a very low number of venues (4) . These numbers of venues might indicate the level of development for each group of neighborhoods.

I created the top 5 common venue categories for each neighborhood. Below is the example of 2 neighborhoods with top 5 venue categories:

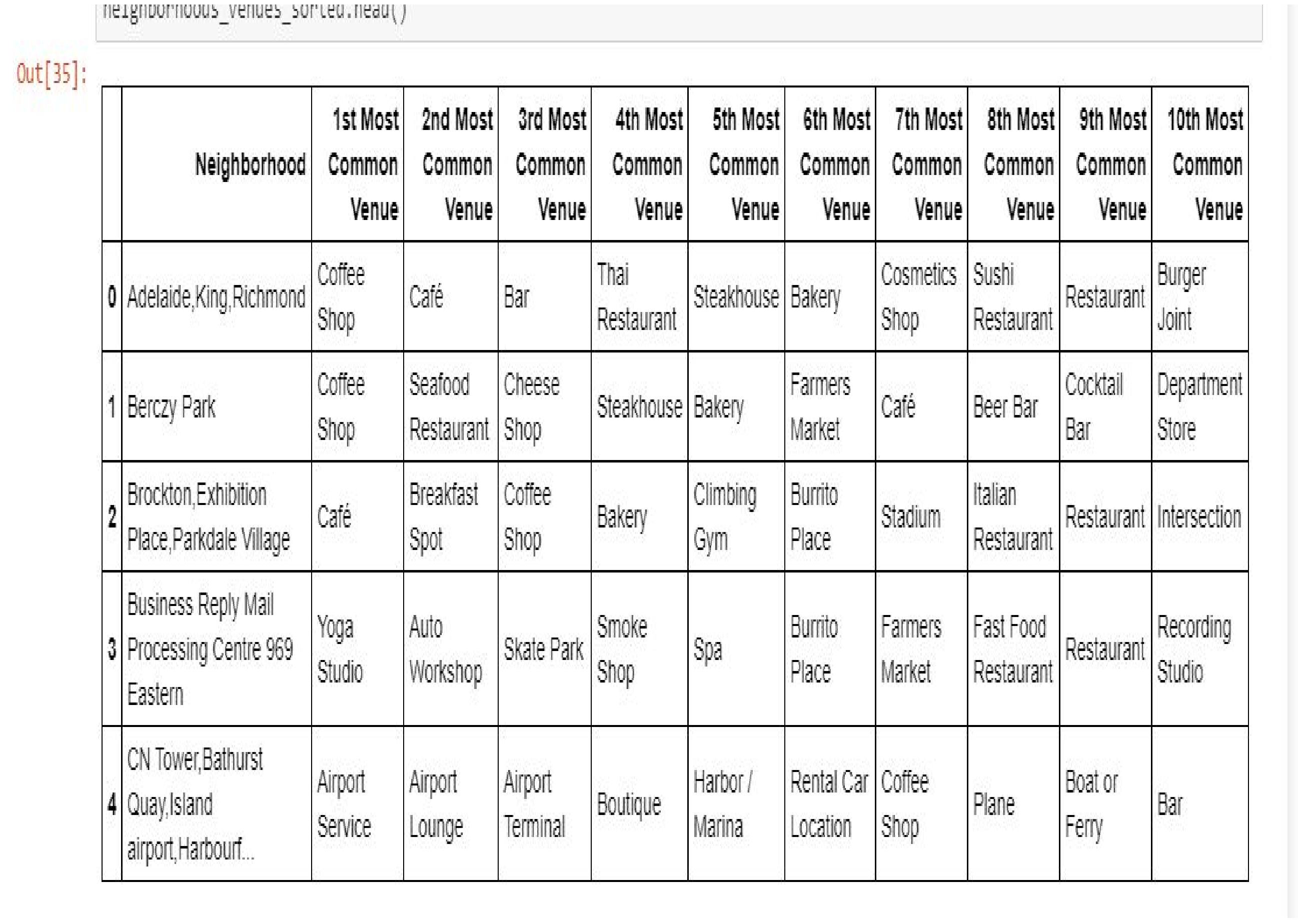


```
----Adelaide,King,Richmond----
             venue  freq
0      Coffee Shop  0.06
1             Café  0.04
2  Thai Restaurant  0.04
3              Bar  0.04
4       Steakhouse  0.03


----Berczy Park----
                venue  freq
0         Coffee Shop  0.09
1        Cocktail Bar  0.04
2      Farmers Market  0.04
3  Seafood Restaurant  0.04
4         Cheese Shop  0.04
```

It also showed the difference in frequency between 5 most common venues in each neighborhood. For example, in Berczy Park, the frequency for Coffee Shop is the highest, more than two times higher than the second most common venu (Cocktail Bar)

Then I created the top 10 common venue categories for each neighborhood. The common venue is ranked based on the frequency of venue on venue categories. Below is the example of 4 neighborhoods with its most common venues categories

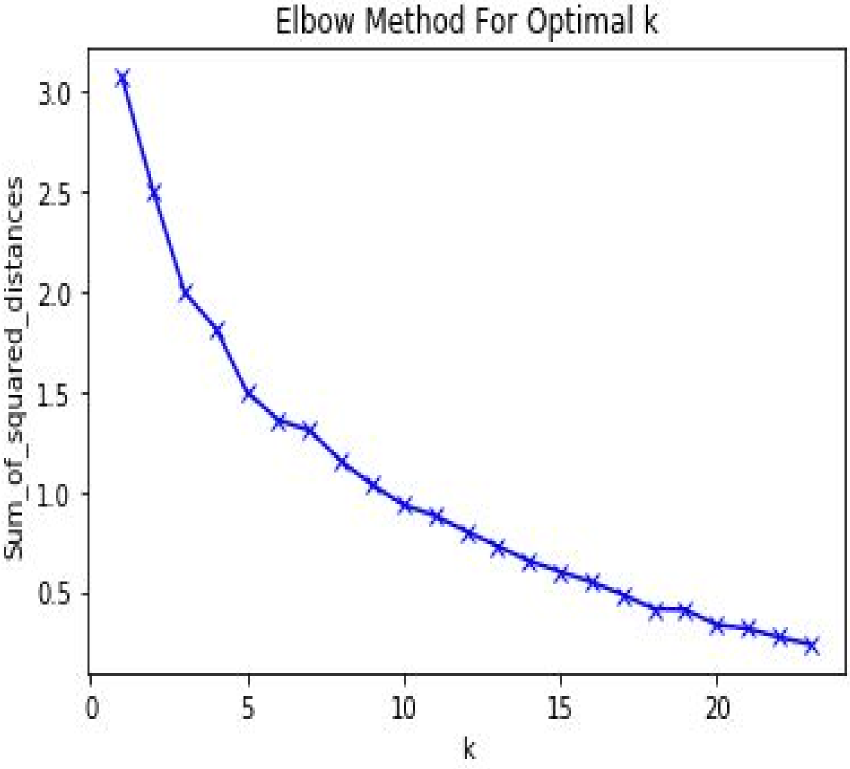


```
neighborhoods_venues_sorted.head()
```

Out[35]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adelaide,King,Richmond	Coffee Shop	Café	Bar	Thai Restaurant	Steakhouse	Bakery	Cosmetics Shop	Sushi Restaurant	Restaurant	Burger Joint
1	Berczy Park	Coffee Shop	Seafood Restaurant	Cheese Shop	Steakhouse	Bakery	Farmers Market	Café	Beer Bar	Cocktail Bar	Department Store
2	Brockton,Exhibition Place,Parkdale Village	Café	Breakfast Spot	Coffee Shop	Bakery	Climbing Gym	Burrito Place	Stadium	Italian Restaurant	Restaurant	Intersection
3	Business Reply Mail Processing Centre 969 Eastern	Yoga Studio	Auto Workshop	Skate Park	Smoke Shop	Spa	Burrito Place	Farmers Market	Fast Food Restaurant	Restaurant	Recording Studio
4	CN Tower,Bathurst Quay,Island airport,Harbourf...	Airport Service	Airport Lounge	Airport Terminal	Boutique	Harbor / Marina	Rental Car Location	Coffee Shop	Plane	Boat or Ferry	Bar

As I have common venues in each neighborhood, I use the K-Means algorithm to cluster all the groups of neighborhoods. I find optimal k by Elbow Method. The result is below:

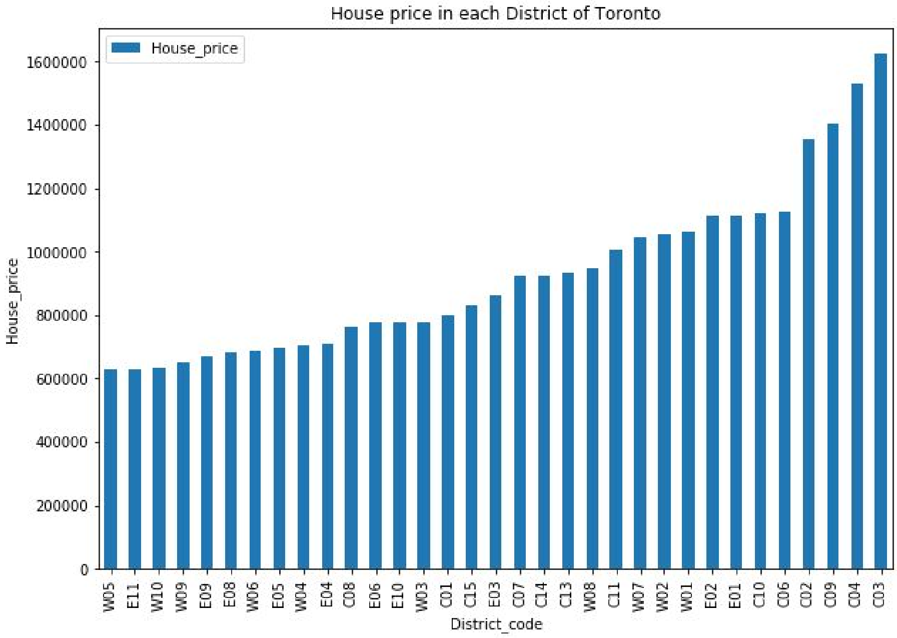


```
Optimal k  is 5
```

I run the model with K-means = 5. I got 5 clusters of neighborhoods. This would help interested audiences to compare the similarity and dissimilarity of each group of neighborhoods in Toronto.

3.2 House price for districts in Toronto:

There was no available data for house prices in each district in Toronto. So I collect the data and create the data frame from scratch. I sorted the data frame based on house price and got the bar chart below



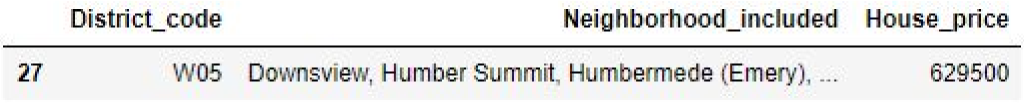
4. Result:

4.1 Examine each cluster in Toronto:

There are 5 clusters of neighborhoods in Toronto with k = 5 through k means model. Each cluster of neighborhoods are shown below:

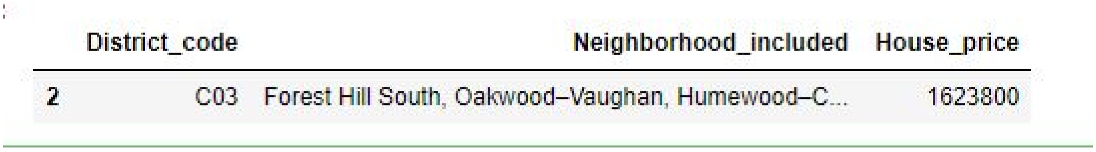
4.2. *House price for each districts in Toronto:*

The district that has the lowest house price is W05. Its average house price is $629,500. Neighborhoods included in this area are Downview, Humber Summit, Humbermede (Emery), Jane and Finch (Black Creek or Glenfield - Jane Heights), York University Heights.



	District_code	Neighborhood_included	House_price
27	W05	Downsview, Humber Summit, Humbermede (Emery), ...	629500

The district that has the highest house price is C03. Its average house price is $1,623,800. Neighborhoods included in this area are Forest Hill South, Oakwood-Vaughan,...



	District_code	Neighborhood_included	House_price
2	C03	Forest Hill South, Oakwood–Vaughan, Humewood–C...	1623800

The average house price in Toronto across Districts is $926,360

5. Discussion:

Regarding k-mean model, 5 different clusters have different characteristics. Real estate agents can segment their customers based on their needs to recommend appropriate clusters of neighborhoods. For example, families will prefere clusters of neighborhoods that have lots of parks for their children. Clusters of neighborhoods like Rosedale, Forrest Hill North , Forrest Hill North.

Regarding house prices in Toronto, real estate agents and home-buyers can find appropriate neighbors with a suitable price range. For example, home -buyers with a high budget can look into clusters of neighborhoods like Forrest Hill South, Lawrence Manor, Bedford Park. Home buyers on a low budget can look into more affordable areas like Downsview , Humber Summit, Malvern….

These two results can be further analyzed and applied if real estate companies have a range of customer profiles with specific budgets and preference in the neighborhood. The report will be the first foundation for home-owners looking to buy a suitable property in Toronto.

6. Conclusion:

I have done exploratory analysis on Toronto neighborhoods and house prices in each neighborhood. I use the k-means model (unsupervised classification model) to cluster neighborhoods in Toronto. Using elbow method, I have found out that the optimal k-mean for this model is 5. Neighborhoods are clustered based on the the frequency of venues categories (data retrieved by Foursquare). This will help home-buyers to find similarity and dissimilarity between group of neighborhoods.

I have created a dataframe for average house price across Toronto districts. Basic analysis for house price will help the customers find appropriate neighborhoods that fit their budget.