Exploring Toronto, Finding the Best Neighborhoods to Live In as well as to Open an Asian Restaurant

Capstone Project

Muhammad Umer Abbasi

In Partial Fulfillment
of the Requirements for the Specialization
IBM Data Science Professional Certificate

COURSERA

INTRODUCTION

Moving to a new city is often a big challenge because of many reasons. One such reason is that you do not know much about the city, you do not know about its neighborhoods, you do not know about its demographics. Because of this reason you need someone to help you explore the city, someone to find you the most suitable neighborhood to live in, someone to suggest you the optimal location for starting a new business.

In this project, I replaced that someone by a Data Science based algorithm which uses your present neighborhood data and your target city's data to help you find the best neighborhood to live in this new city as well as to help you explore this city. In addition to this, if you plan to open a specific type of business in this new city but do not know where to open, this algorithm will help you in this too.

Keeping in mind the aforementioned problems, I created a hypothetical scenario and used my algorithm on it. In this scenario, a client wants to move from 'Clifton, Karachi' to 'Toronto, Canada' and he wants me to help him with following problems:

1. Help him explore the city.
2. Find him the most suitable neighborhood to live in. He wants to live in a neighborhood as similar as possible to his old neighborhood in Karachi, Pakistan.
3. Find him the optimal neighborhood to open his high-end Asian restaurant.

Importance

This project is helpful for several people, for instance:

1. Anyone who is planning to move to Toronto. This is an example for people moving from Karachi to Toronto, but the similar analysis can be done for any other pair of cities.

2. Anyone who wants to explore any city.

3. Anyone who wants to open a business in a city. We did analysis for a high-end Asian restaurant in Toronto, but similar analysis can be done for any other kind of restaurant as well as for any other kind of business in any city.

Project Goal

As our problem has three parts, so, let us tackle each part separately. I proposed following solutions, respectively, for each part:

1. Build a data set comprising of venues located in each neighborhood of 'Toronto, Canada' and then cluster neighborhoods based on common venue types present there. This would give client quite good understanding about distribution of venues throughout the Toronto and would help him explore the city.

2. Include client's present neighborhood (along with venues located there) in the aforementioned data set and cluster this data set on basis of common venue categories (like we did for problem 1). Now filter out the cluster which includes client's present neighborhood (let's call it X) from this data set and cluster it (X). Repeat the same steps again until the number of neighborhoods in X fall below a threshold value (i.e. 15 or 20 etc.). This would give client a good set of suitable neighborhoods to live in.

3. Best neighborhood for a restaurant depends on multiple factors, i.e. demographic overview of the location i.e. average age, average income, population, distribution of ethnic groups, plans for neighborhood's future development, and competition.

As we are opening a high-end Asian restaurant, so an ideal neighborhood would be the one with:

* high average household income (indicates high buying power),

* high percentage of Asian population (assuming that Asian population is more inclined towards Asian cuisine),

* moderate number of competitors because if number of competitors are low in your restaurant's neighborhood, it's very unlikely that people would come there to try your cuisine especially when the restaurant would be new and if the number of competitors are very high, it would be relatively difficult to compete with them i.e. you would have to come up with a very strong USP (unique selling point). So, moderate number of competitors mean that people come to this neighborhood for Asian cuisine and it is also not very difficult to deal with the completion here.

Data Sources

1. Client provided his present address.
2. Neighborhoods of Toronto, along with their postal codes would be scraped from Wikipedia.
3. A geocoding library Geopandas will be used to get geocoordinates of each neighborhood.
4. These geocoordinates would then be used to extract top 100 venues in each neighborhood within 500m radius (as 500m is a reasonable walking distance) using Foursquare API.
5. Toronto's demographic data is obtained from Toronto's Open Data Portal.

How We Will Use This Data?

For problem 1:

1. Using postal codes (scraped from Wikipedia) and Geocoder library we will get geocoordinates of all the neighborhoods.

2. Then providing Foursquare API with these geocoordinates, we will get top 100 venues in each neighborhood within 500m radius.

3. We will then group this data by 'venue category' to get an estimate of frequency of each venue category in each neighborhood.

4. Then, we will cluster neighborhoods based on this frequency of venue categories to group similar neighborhoods together.

For problem 2:

1. Add client's present neighborhood data into the data set built for problem 1 and cluster it.

2. Filter out the cluster from this data set which includes client's present neighborhood (let us call it X).

3. Now, cluster X and go to the step 2 again.

4. Repeat this process until number of neighborhoods in X fall below a threshold value (i.e. 15 or 20)

For problem 3:

1. From the data collected for problem 1, we will build a new data set comprising of total number of Asian restaurants in each neighborhood along with neighborhood names and postal codes.

2. From Toronto's demographics data, we will get Asian population, total population, number of high earning households, and average household income of each neighborhood.

3. We will then cluster neighborhoods on basis of these parameters and will choose a cluster of neighborhoods closest to the ideal neighborhood (mentioned earlier).

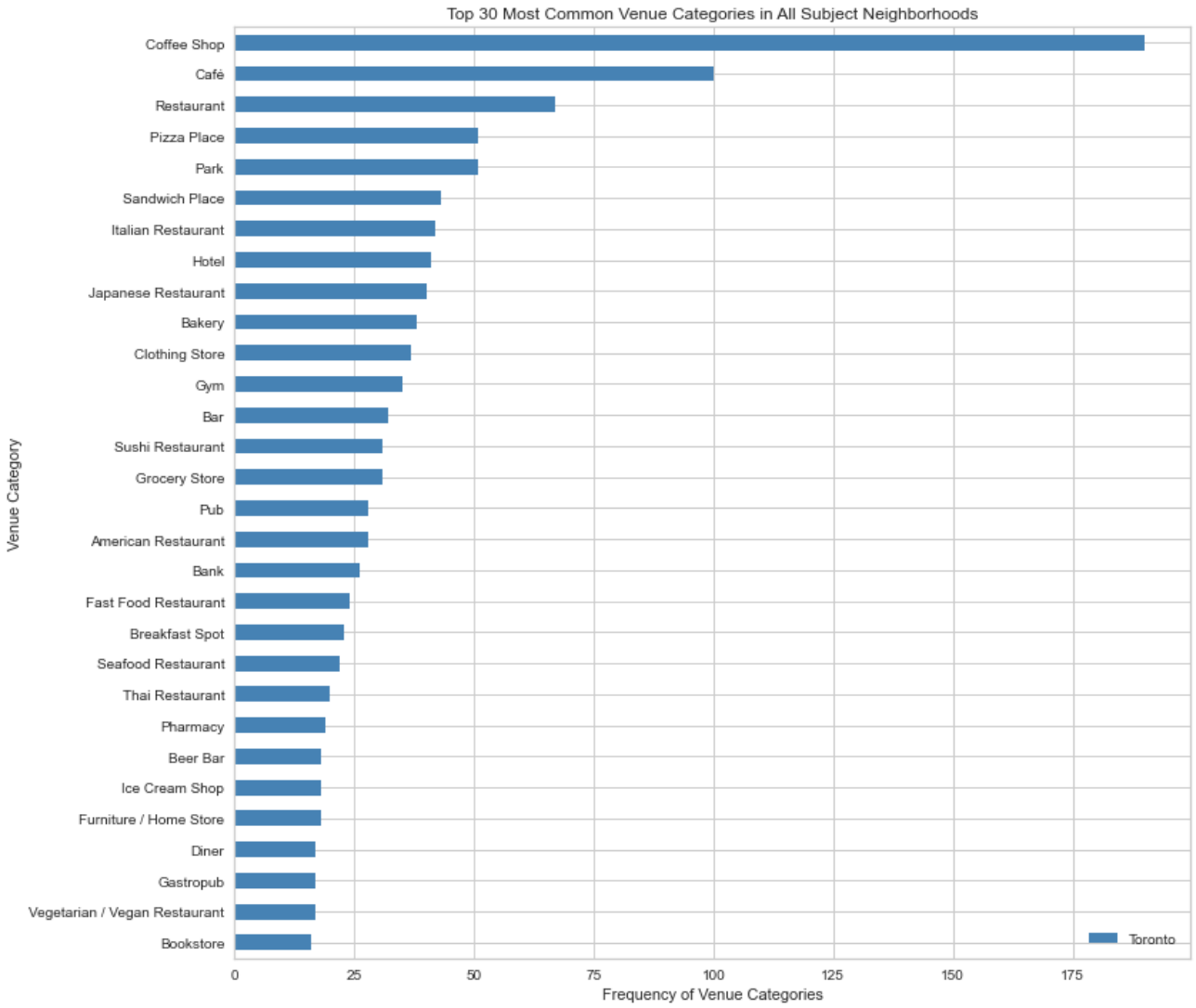
Chapter 1

This chapter tackles problem 1 and help client explore the city of Toronto.

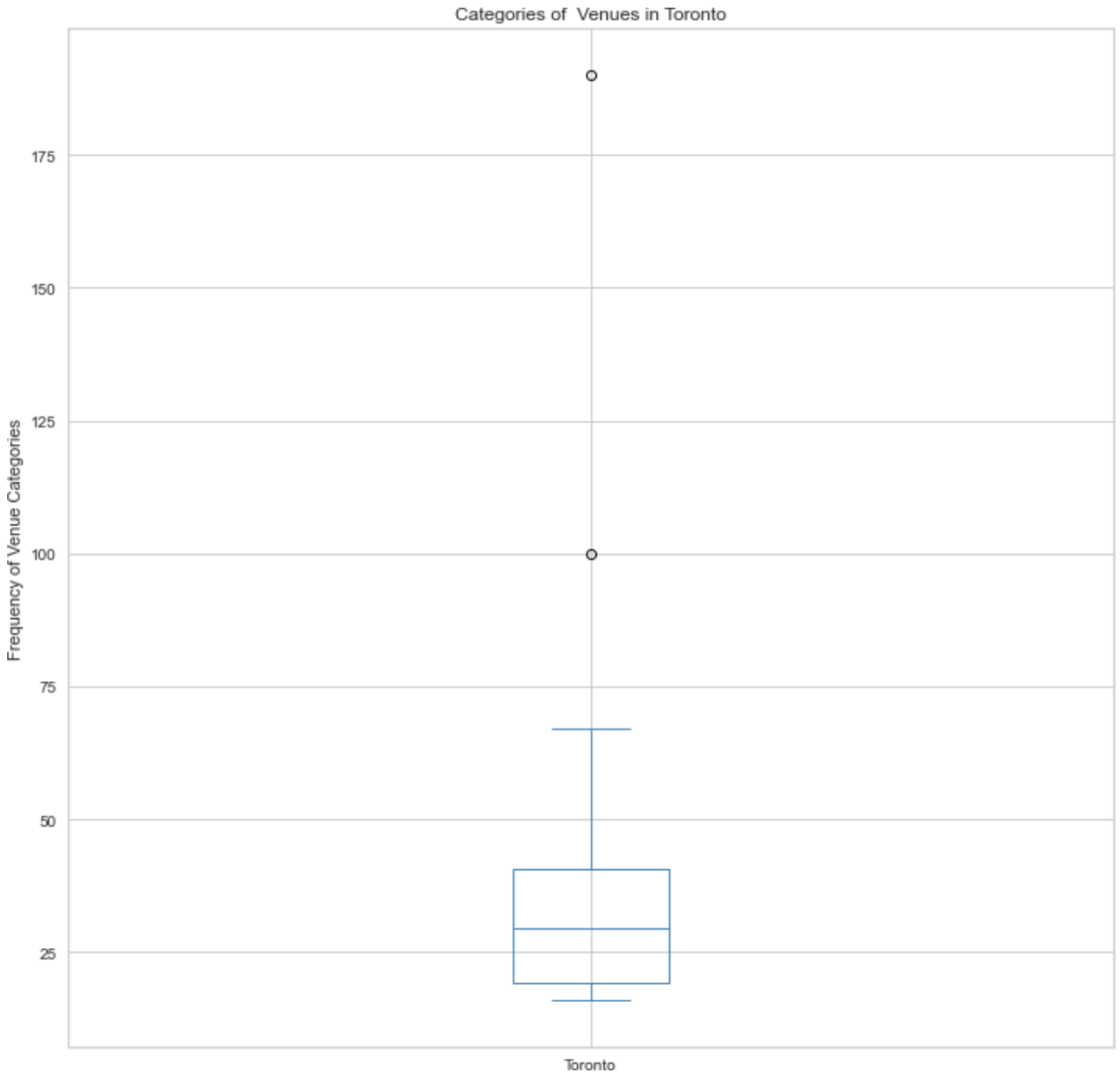
Data Collection and Exploratory Analysis

I gathered Toronto's neighborhoods' data using Wikipedia, Geopandas, and Foursquare to build a data set that contains top 100 venues located in each neighborhood within 500m radius.

You can see here the top 30 most common venues in Toronto:



Also, you can see that frequency of occurrence of the most venues are around 35-40 but there are some exceptional cases too:



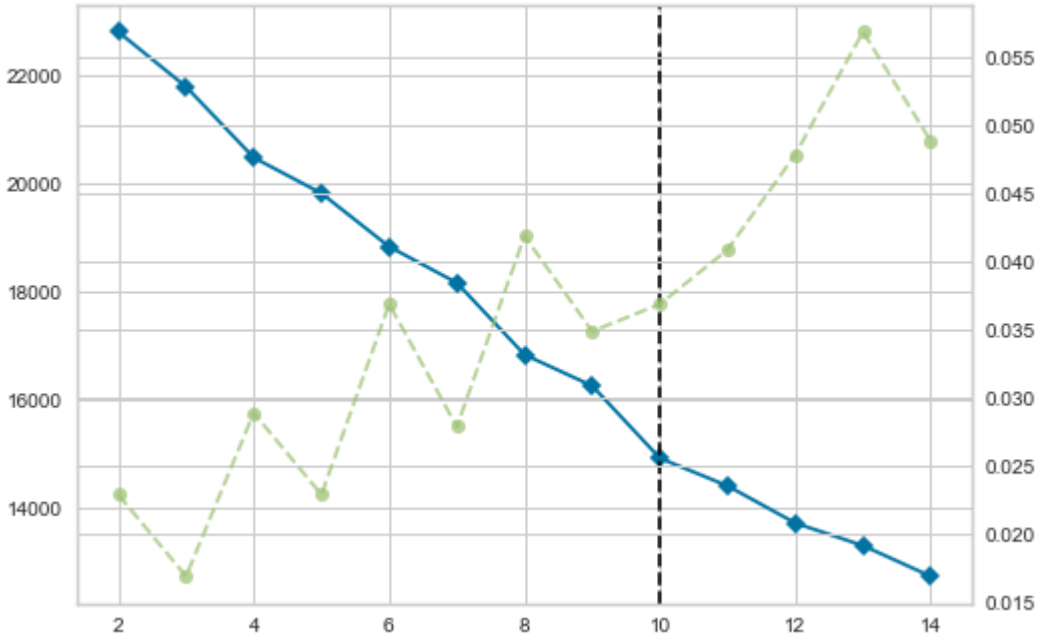
I extracted venue category of each venue and then extracted the top 10 most common venue categories in each neighborhood as shown below:

	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	Agincourt	Breakfast Spot	Latin American Restaurant	Lounge	Skating Rink	Doner Restaurant	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run
1	Alderwood, Long Branch	Pizza Place	Gym	Coffee Shop	Pub	Sandwich Place	Discount Store	Department Store	Dessert Shop	Dim Sum Restaurant	Diner
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Gift Shop	Fried Chicken Joint	Sandwich Place	Bridal Shop	Diner	Restaurant	Deli / Bodega	Supermarket
3	Bayview Village	Japanese Restaurant	Café	Bank	Chinese Restaurant	Dessert Shop	Diner	Discount Store	Distribution Center	Dog Run	Women's Store
4	Bedford Park, Lawrence Manor East	Restaurant	Sandwich Place	Coffee Shop	Italian Restaurant	Breakfast Spot	Indian Restaurant	Butcher	Café	Sushi Restaurant	Cupcake Shop

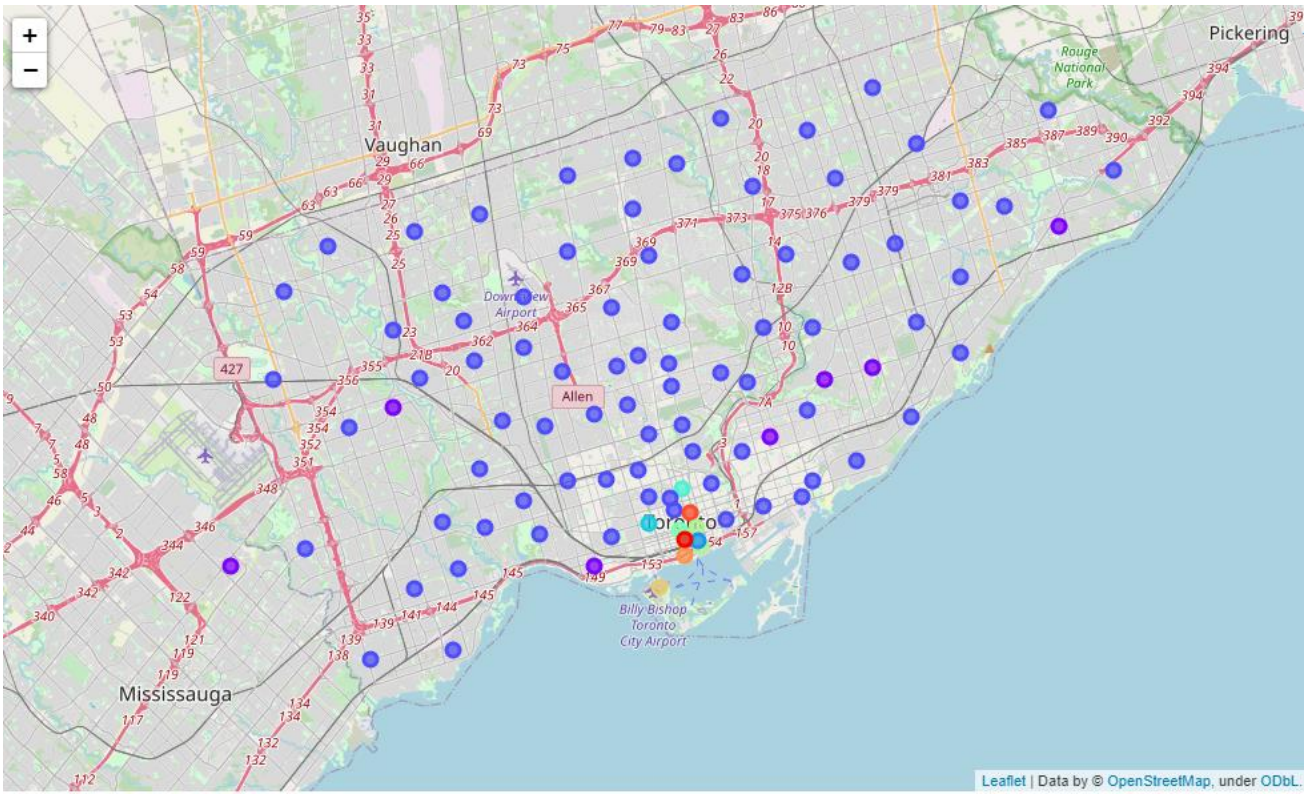
Clustering

I clustered neighborhoods based on common type of venues located in each neighborhood using an unsupervised machine learning algorithm, K-Means (it is basically a clustering algorithm).

In KMeans, we must choose k i.e. the number of clusters we want our dataset to be divide into. I chose elbow method to find the best k as you can see below:



We can see here that the best k is 10, i.e. we will find the best result if we divide our dataset into 10 clusters. So, I did the same and did show the clustered neighborhoods on map using color coding, as you can see below:



Conclusion

We can see that a big majority of neighborhoods fall under one cluster (cluster 1). This cluster's most common venue categories are the ones fulfilling basic human needs. So, you can find almost anything, you need on daily basis, in these neighborhoods. Other neighborhoods include, along with basic needs providing venues, some sort of special venues like bars or hotels etc. So, client can always refer to these clusters whenever he wants to visit a specific type of venue or wants to know the category of his target neighborhood.

Chapter 2

This chapter tackles problem 2 i.e. finding client the best neighborhood to live in.

Data Collection and Exploratory Analysis:

We used the dataset built for problem 1 along with client's present address to tackle this problem.

Clustering and Filtering:

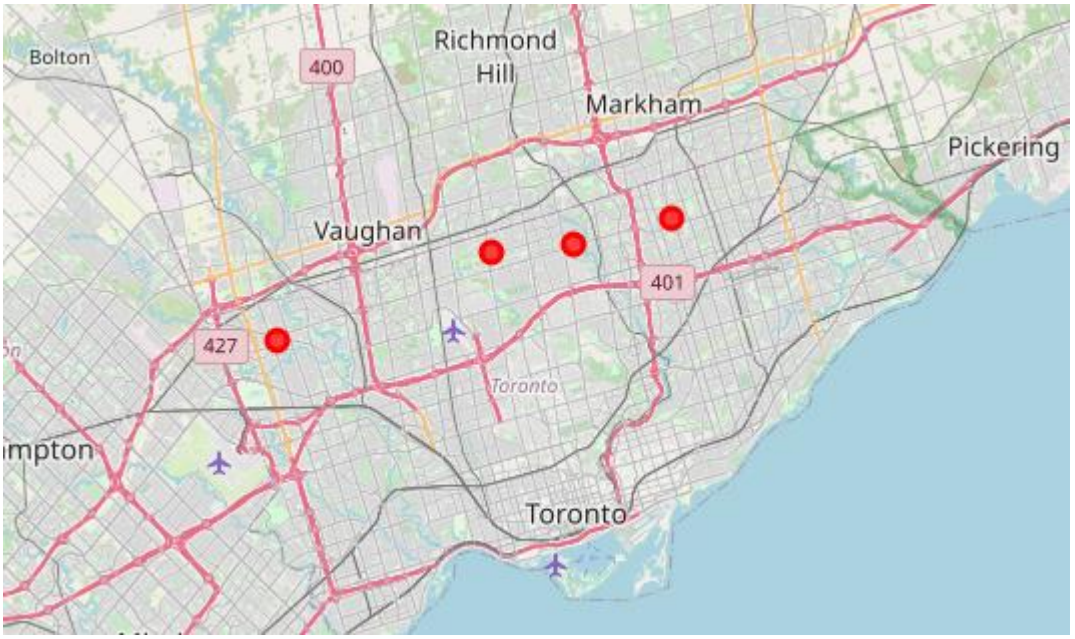
As discussed in the section 'How we will use this data?' I performed repeated clustering and filtering on this data set until number of neighborhoods in the filtered data set fell below 20.

Conclusion:

We found client the optimal neighborhoods to live in, as shown below:

Neighborhood	Latitude	Longitude	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
Bayview Village	43.786947	-79.385975	Japanese Restaurant	Café	Bank	Chinese Restaurant	Dessert Shop	Diner	Discount Store	Distribution Center	Dog Run	Women's Store
South Steeles, Silverstone, Humbergate, Jamestown, Mount Olive, Beaumond Heights, Thistletown, Albion Gardens	43.739416	-79.588437	Grocery Store	Fried Chicken Joint	Pizza Place	Sandwich Place	Beer Store	Fast Food Restaurant	Pharmacy	Donut Shop	Doner Restaurant	Dance Studio
Steeles West, L'Amoreaux West	43.799525	-79.318389	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Pizza Place	Breakfast Spot	Sandwich Place	Bank	Electronics Store	Pharmacy	Grocery Store
Willowdale, Willowdale West	43.782736	-79.442259	Grocery Store	Coffee Shop	Pharmacy	Pizza Place	Bank	Dog Run	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store

Same neighborhoods on Toronto's map are also shown below:



Chapter 3

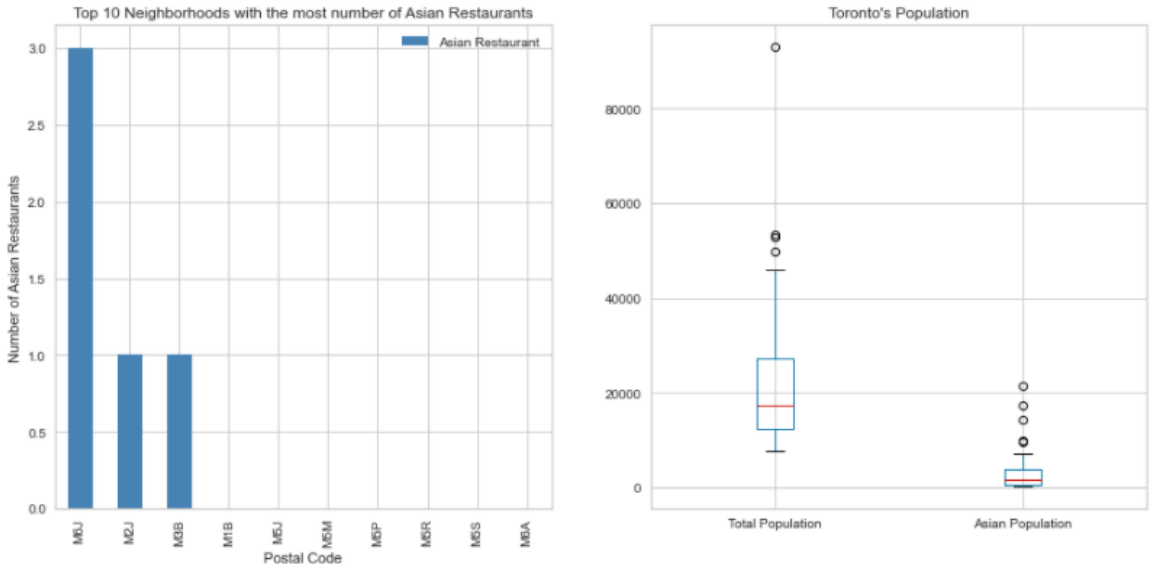
This chapter tackles problem 3 i.e. finding client the optimal location to open his high-end Asian restaurant.

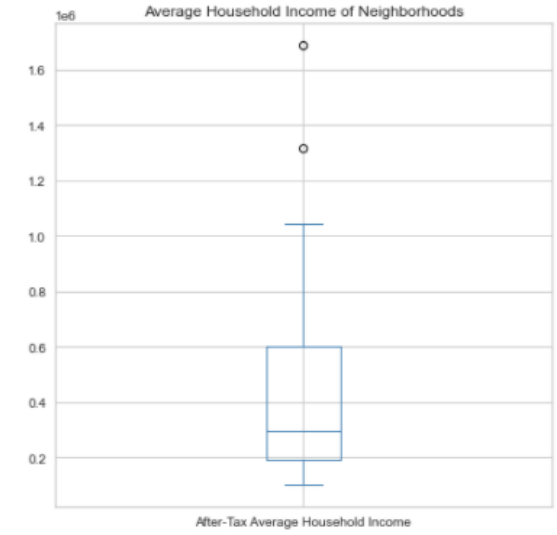
Data Collection and Exploratory Analysis:

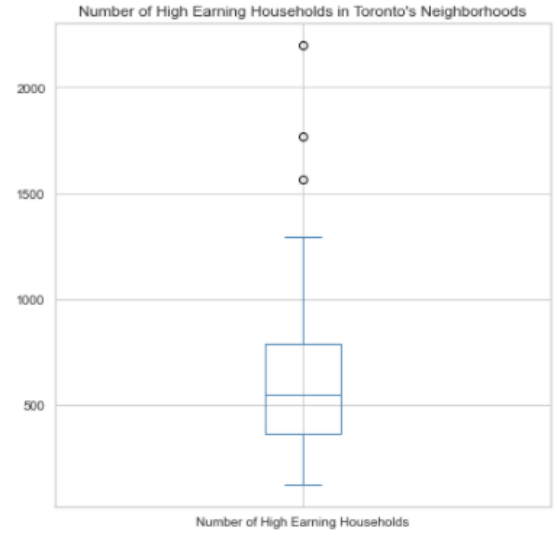
I used Toronto's Neighborhood profile from Toronto's Open Data Portal and extracted total population, Asian population, number of high earning household, and average household income of each neighborhood. From the venues data, obtained using Foursquare in chapter 1, I extracted number of Asian restaurants in each neighborhood. Then I combined these two datasets as shown below:

	Postal Code	Neighborhood	Latitude	Longitude	Total Population	Asian Population	Number of High Earning Households	After-Tax Average Household Income	Asian Restaurant
0	M1B	Malvern, Rouge	43.806686	-79.194353	43794	17465	765	533202	0
1	M1C	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	46087	1750	1770	980578	0
2	M1E	Guildwood, Morningside, West Hill	43.763573	-79.188711	9917	815	440	177062	0
3	M1G	Woburn	43.770992	-79.216917	53485	21545	855	629030	0
4	M1J	Scarborough Village	43.744734	-79.239476	16724	5520	230	185967	0

Then I performed some preliminary analysis on this data set as shown below:



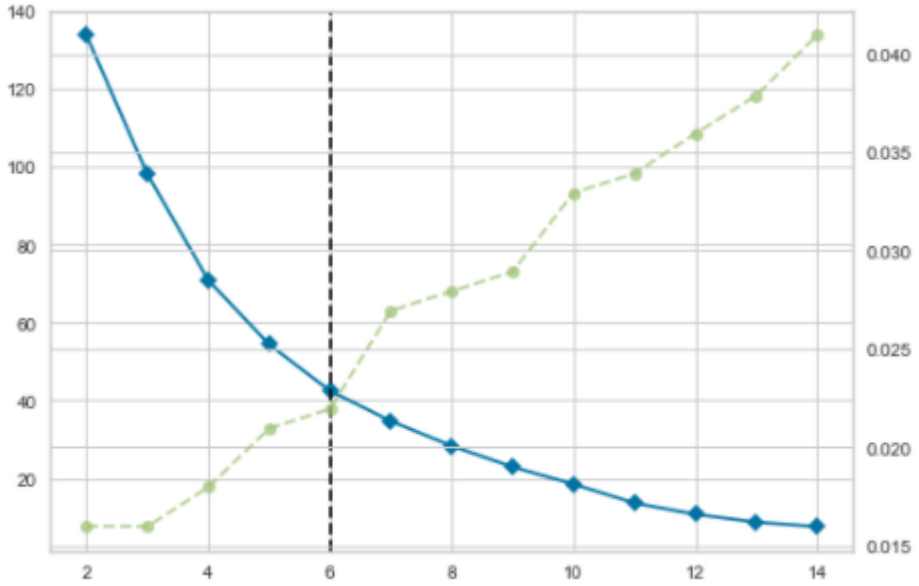




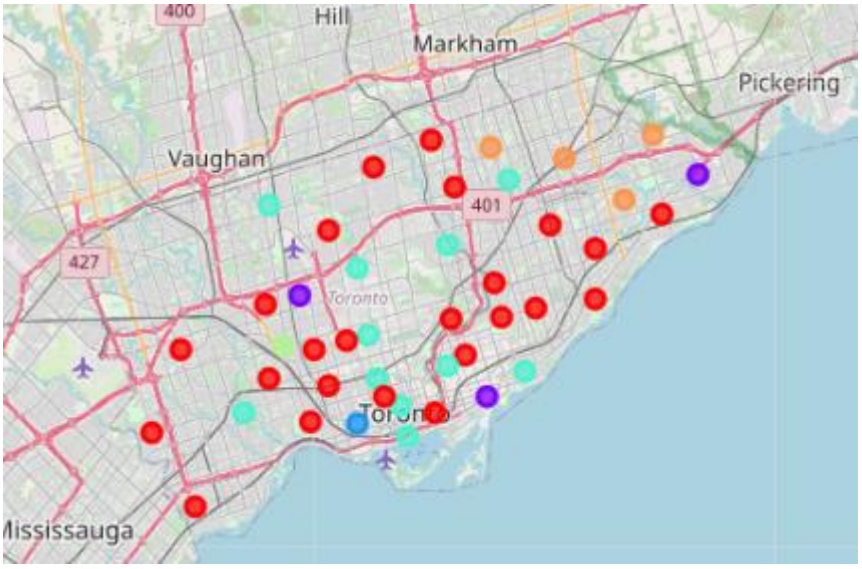
Here we can see that we have only 3 neighborhoods which have Asian restaurants. Also, the number of Asian restaurants is quite low i.e. 5 in the whole Toronto. We can also see the distributions of population, Asian population, average household income and number of high earning household in the whole city,

Clustering and Filtering:

This time I clustered neighborhoods on basis of the aforementioned features using K-Means clustering. For choosing the optimal k, I once again relied on elbow method (as it is quite reliable). You can see its graph below:



Here we can see that the best k is 6. So, I clustered neighborhoods into 6 groups and did show them on map as shown below:



Conclusion:

From all the clusters, cluster 3 was the closest to our ideal neighborhoods because it has following features:

1. Medium number of high earning households.
2. Medium average household income.
3. High percentage of Asian population.

4. Low to medium competition (compared to all other neighborhoods)

Here is the snapshot of cluster 3:

	Postal Code	Neighborhood	Latitude	Longitude	Total Population	Asian Population	Number of High Earning Households	After-Tax Average Household Income	Asian Restaurant	Label
0	M3B	Don Mills	43.745906	-79.352188	27051	2890	645	414411	1	3
1	M1T	Clarks Corners, Tam O'Shanter, Sullivan	43.781638	-79.304302	27446	4690	585	475647	0	3
2	M3J	Northwood Park, York University	43.767980	-79.487262	30531	4175	730	462928	0	3
3	M4E	The Beaches	43.676357	-79.293031	21567	645	955	659192	0	3
4	M4K	The Danforth West, Riverdale	43.679557	-79.352188	26846	1735	1295	594312	0	3

After analysis of this cluster, I reached conclusion that the client can consider any of top 3 neighborhoods for opening his restaurant.

Future Directions

This project was a specific example, but this approach can be used for any other similar problem too. For instance, if you want to explore some other city, if you want to find the most suitable neighborhood for you to live in, in some other city or, if you want to find the best location to open a specific business, you can use this approach. For this specific example results can be improved if we have more insight about the Asian population in each neighborhood, for instance, the number of high earning Asian households and the average income of Asian households can improve results significantly.