Classifying Toronto Neighborhood

Mohd Belal

January 30, 2021

1. Introduction

1.1 Background

Buying a perfect home is a dream for every individual. We have been always trying to buy the home which would be perfect for us. For years property consultants and brokers are the people who have helped us in our endeavor

Now with the addition of new technologies stakeholders have changed their methods to improve their services and in this IT age with the help of technology and data they are trying their best to achieve better results.

I am using Foursquare API as well as scraped webpages to get the average property cost in each and every area of Toronto City. As we know cost is the major driver where a person may live.

Foursquare API would be used to get the locality of an area. The locality and venues would be helpful for individual looking to get the best place he needs. For example, A bachelor would like to live where there are nearby pubs, entertainment centers and work places. But a person having a family may want to live where there is nearby schools, shops and parks.

1.2 Problem

To get the desired location on one's desire would be our main problem. The Data of nearby venues with the addition of cost would help us to choose the best neighborhood.

1.3 Interest

All the stakeholder would be very much interested in our model which could help them to choose the best place for them.

2. Data acquisition and Cleaning

2.1 Data sources

We are using **Foursquare API**, Geocoders, and web Scraping techniques to solve our problem.

Foursquare API would be used to get nearby venues around a location. This venue data would be used to classify our neighborhood based on the locality.

Geocoders would be used to get latitude and longitude of neighborhoods. This latitude and longitude are required for maps and Foursquare API.

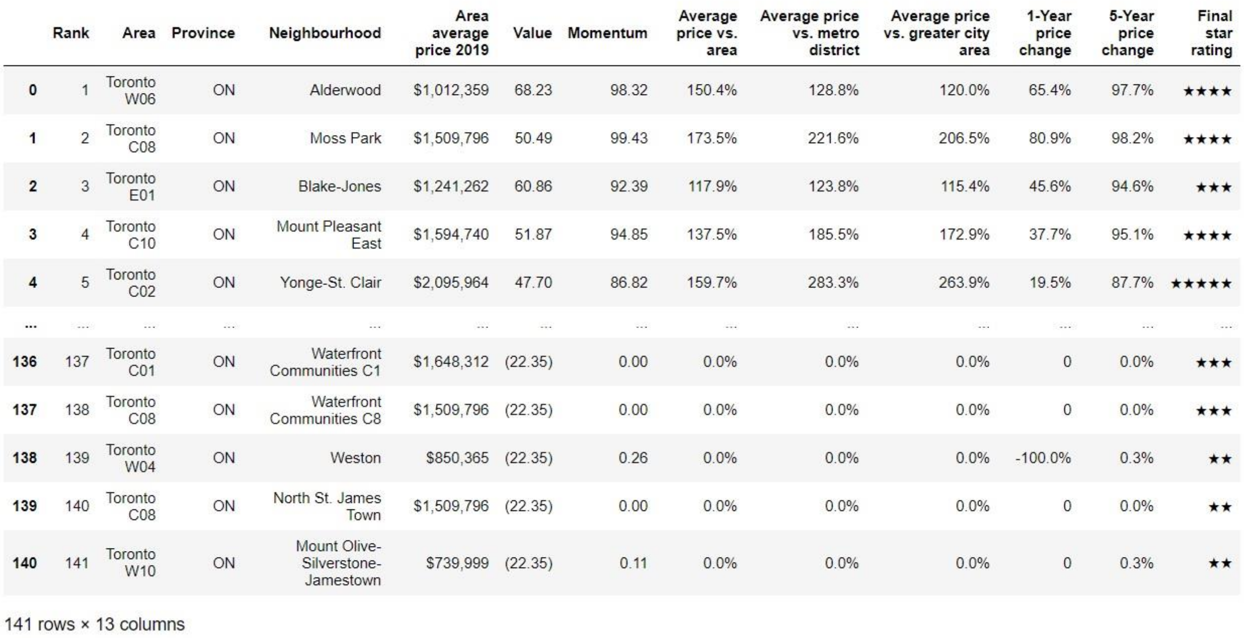
I searched but couldn't find any structured dataset to get average housing cost in a neighborhood. So, I scraped a webpage which shows the average housing cost of a neighborhood.

Click here to View webpage

2.2 Data Cleaning and Feature Selection.

Data scraped from webpage very unwanted columns which is not suitable for our problem. There are no missing rows in our data.

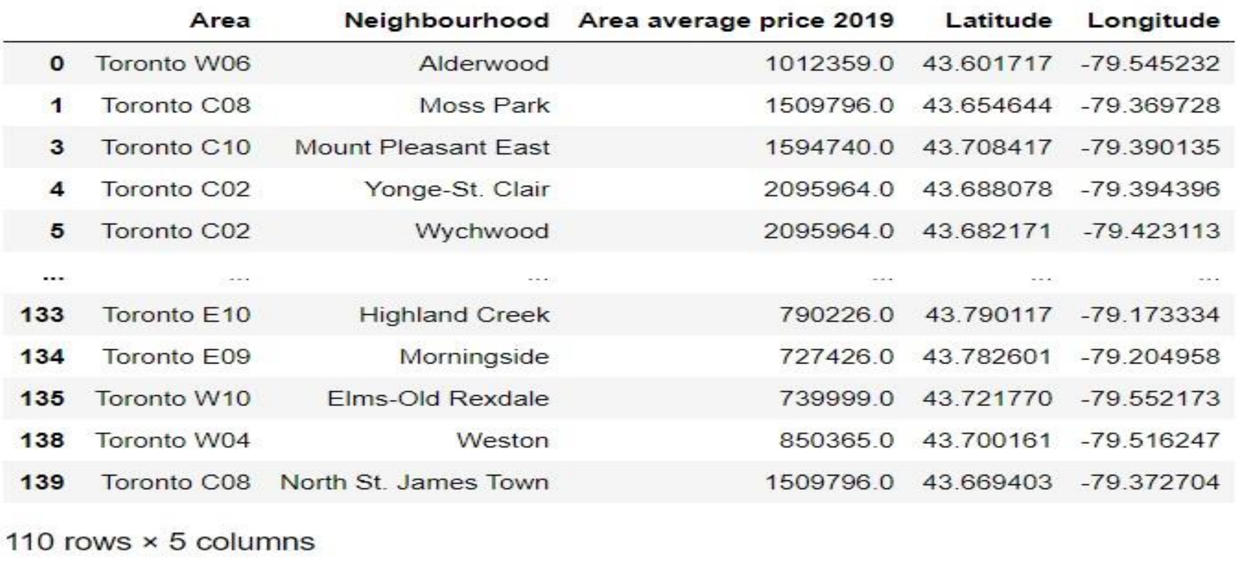
I needed the latitude and longitude of our neighborhood. So, I used geocoders to get the latitude and longitude but there was some neighborhood whose data is not available in geocoders. Hence, we deleted that rows.
The initial shape of our dataset was 141 rows and 13 columns.



	Rank	Area	Province	Neighbourhood	Area average price 2019	Value	Momentum	Average price vs. area	Average price vs. metro district	Average price vs. greater city area	1-Year price change	5-Year price change	Final star rating
0	1	Toronto W06	ON	Alderwood	$1,012,359	68.23	98.32	150.4%	128.8%	120.0%	65.4%	97.7%	★★★★
1	2	Toronto C08	ON	Moss Park	$1,509,796	50.49	99.43	173.5%	221.6%	206.5%	80.9%	98.2%	★★★★
2	3	Toronto E01	ON	Blake-Jones	$1,241,262	60.86	92.39	117.9%	123.8%	115.4%	45.6%	94.6%	★★★
3	4	Toronto C10	ON	Mount Pleasant East	$1,594,740	51.87	94.85	137.5%	185.5%	172.9%	37.7%	95.1%	★★★★
4	5	Toronto C02	ON	Yonge-St. Clair	$2,095,964	47.70	86.82	159.7%	283.3%	263.9%	19.5%	87.7%	★★★★★
...	...	...	...	...	...	...	...	...	...	...	...	...	...
136	137	Toronto C01	ON	Waterfront Communities C1	$1,648,312	(22.35)	0.00	0.0%	0.0%	0.0%	0	0.0%	★★★
137	138	Toronto C08	ON	Waterfront Communities C8	$1,509,796	(22.35)	0.00	0.0%	0.0%	0.0%	0	0.0%	★★★
138	139	Toronto W04	ON	Weston	$850,365	(22.35)	0.26	0.0%	0.0%	0.0%	-100.0%	0.3%	★★
139	140	Toronto C08	ON	North St. James Town	$1,509,796	(22.35)	0.00	0.0%	0.0%	0.0%	0	0.0%	★★
140	141	Toronto W10	ON	Mount Olive-Silverstone-Jamestown	$739,999	(22.35)	0.11	0.0%	0.0%	0.0%	0	0.3%	★★

141 rows × 13 columns

We are only interested in Area, Neighborhood and Area average price. Hence, we would drop other columns.
We will also drop the rows whose geographical data is not available. After the data looks like this.



	Area	Neighbourhood	Area average price 2019	Latitude	Longitude
0	Toronto W06	Alderwood	1012359.0	43.601717	-79.545232
1	Toronto C08	Moss Park	1509796.0	43.654644	-79.369728
3	Toronto C10	Mount Pleasant East	1594740.0	43.708417	-79.390135
4	Toronto C02	Yonge-St. Clair	2095964.0	43.688078	-79.394396
5	Toronto C02	Wychwood	2095964.0	43.682171	-79.423113
...	...	...	...	...	...
133	Toronto E10	Highland Creek	790226.0	43.790117	-79.173334
134	Toronto E09	Morningside	727426.0	43.782601	-79.204958
135	Toronto W10	Elms-Old Rexdale	739999.0	43.721770	-79.552173
138	Toronto W04	Weston	850365.0	43.700161	-79.516247
139	Toronto C08	North St. James Town	1509796.0	43.669403	-79.372704

110 rows × 5 columns

The columns that I have decided to keep are Area Neighborhood Area average price 2019.

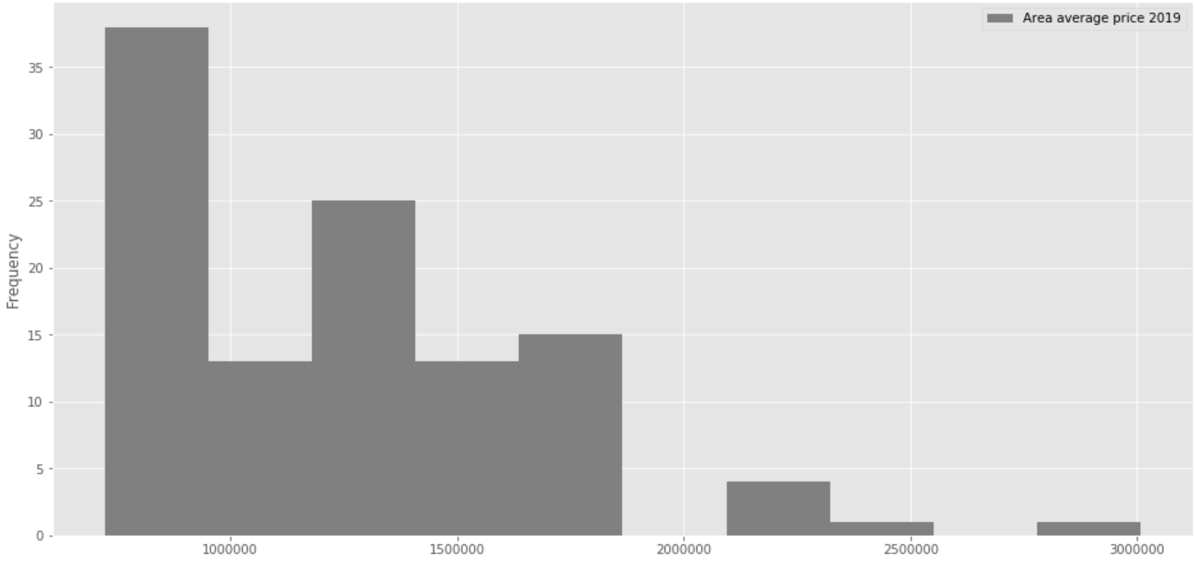
Table 1 Simple feature Selection

Kept Features	Dropped features	Reason for Keeping features	Reason for dropping features
Area, Neighborhood	Rank, Province, Value, Momentum, Average price vs. metro district, Average price vs. greater city area,	To get the latitude and longitude and to plot maps.	Not helpful in our model
Area average price 2019	1-Year price change, 5-Year price change, Final star rating	One of the main reasons for purchase	May complicate our result.

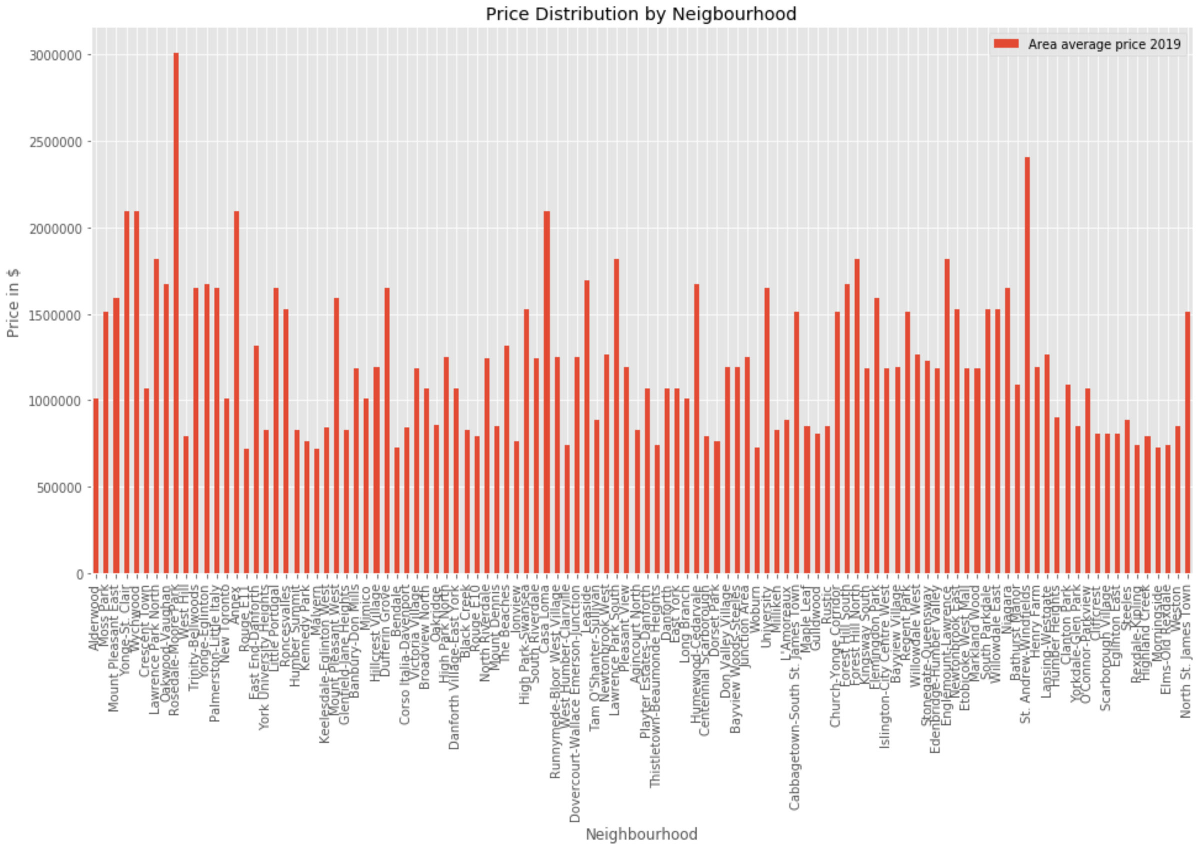
3. Exploratory Data analysis

3.1 Price Distribution

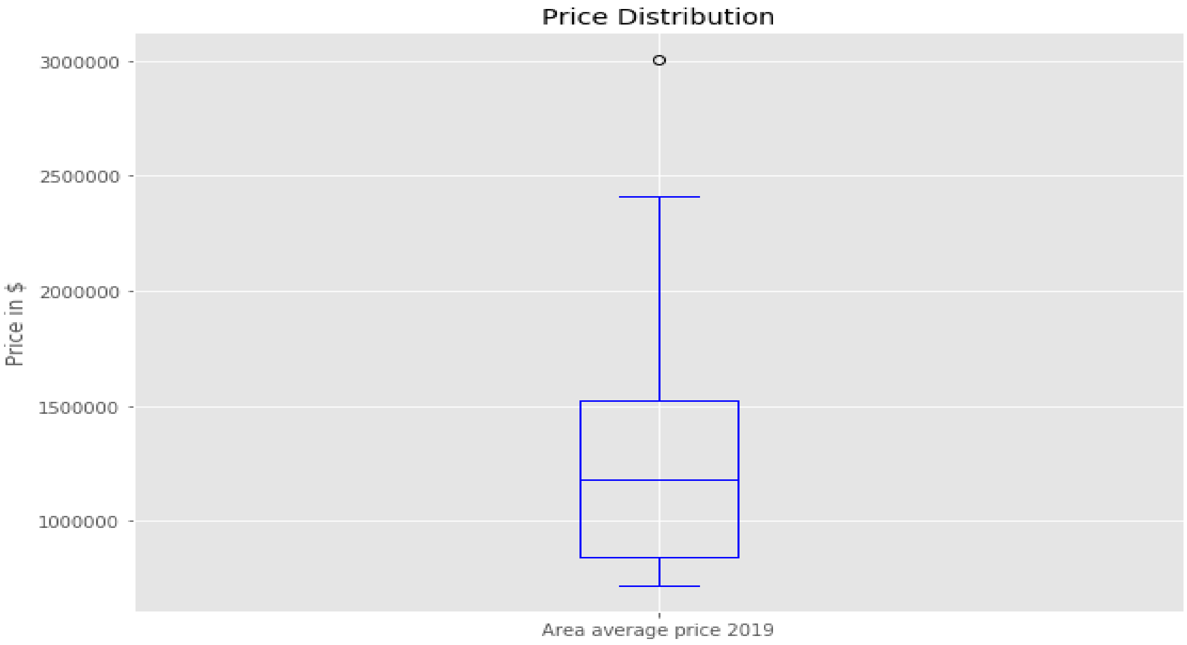
We find that most of the neighborhood is less than $100000. Between $200000 and $300000 there are only a few neighborhoods.



3.2 Price Distribution of each neighborhood



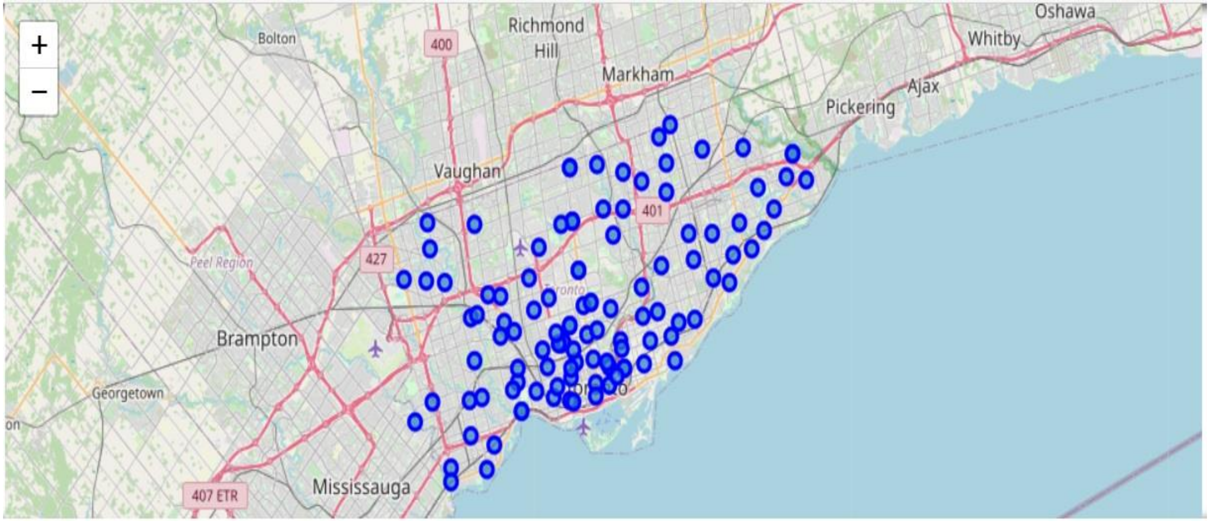
3.2 Statistical Data of Price Distribution



As we can see most common price is around $120000.

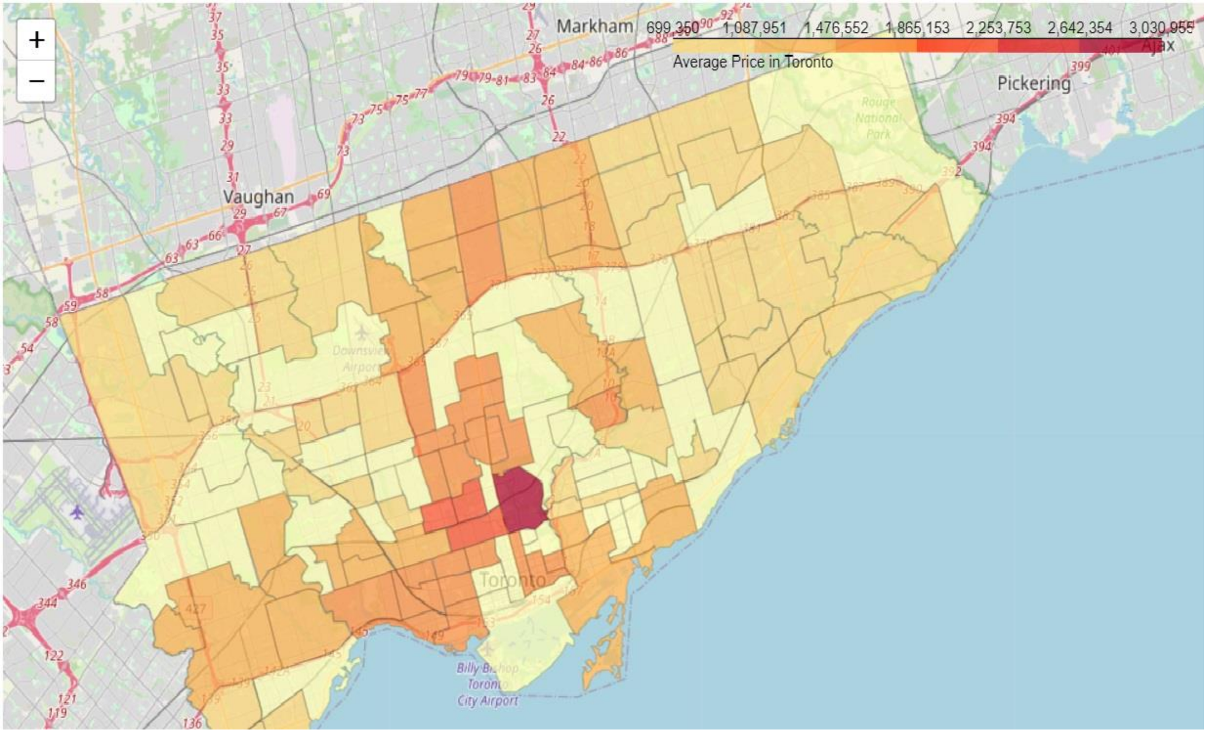
3.3 Geographical representation of Neighborhood

Now we plot the whole neighborhood in the map.



3.4 Geographical representation of neighborhood in terms of cost

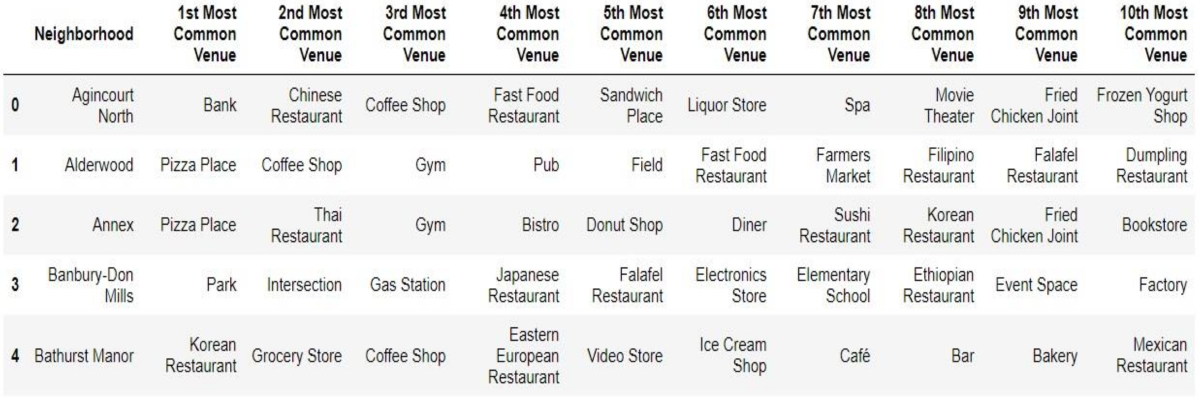
We have plotted the neighborhood in terms of cost.



As we can clearly see the epicenter of the city is costly and outskirts are comparatively cheap.

3.5 Representation of common venues

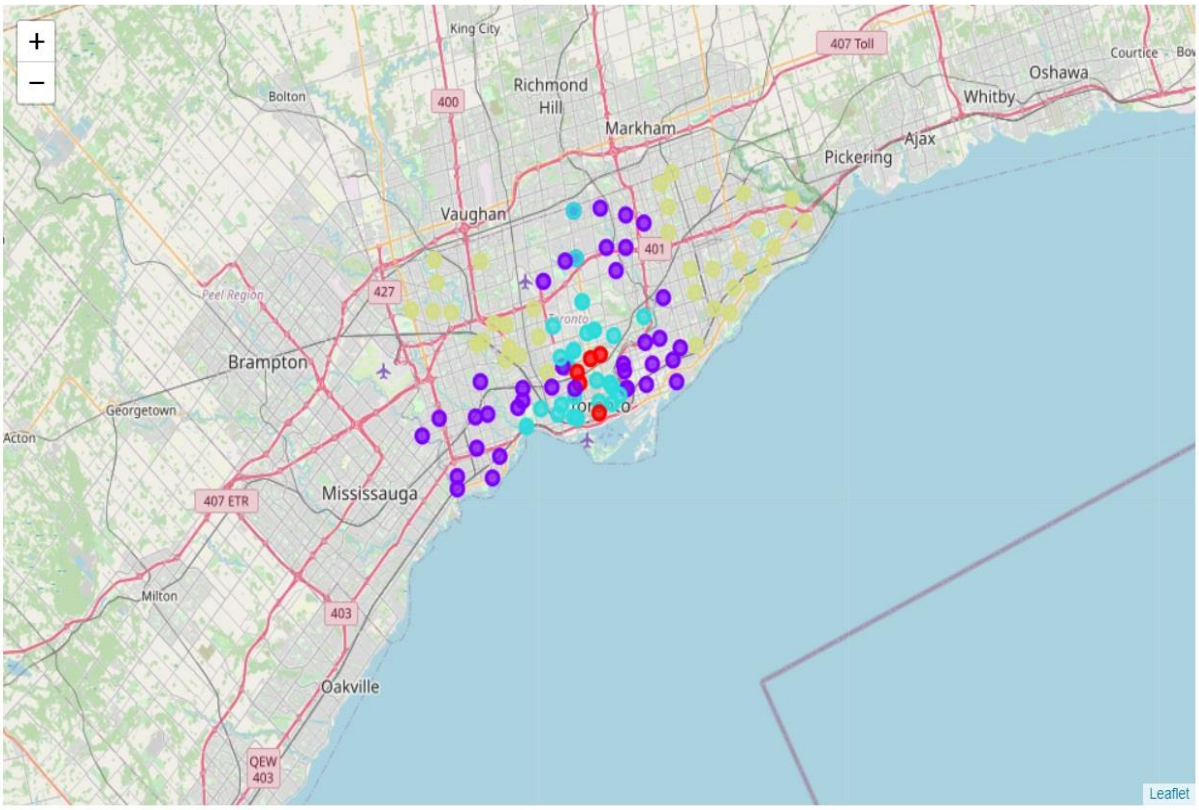
We used foursquare API to get nearby venues then grouped them together and find out the popular places near a neighborhood.



	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 North	Bank	Chinese Restaurant	Coffee Shop	Fast Food Restaurant	Sandwich Place	Liquor Store	Spa	Movie Theater	Fried Chicken Joint	Frozen Yogurt Shop
1	Alderwood	Pizza Place	Coffee Shop	Gym	Pub	Field	Fast Food Restaurant	Farmers Market	Filipino Restaurant	Falafel Restaurant	Dumpling Restaurant
2	Annex	Pizza Place	Thai Restaurant	Gym	Bistro	Donut Shop	Diner	Sushi Restaurant	Korean Restaurant	Fried Chicken Joint	Bookstore
3	Banbury-Don Mills	Park	Intersection	Gas Station	Japanese Restaurant	Falafel Restaurant	Electronics Store	Elementary School	Ethiopian Restaurant	Event Space	Factory
4	Bathurst Manor	Korean Restaurant	Grocery Store	Coffee Shop	Eastern European Restaurant	Video Store	Ice Cream Shop	Café	Bar	Bakery	Mexican Restaurant

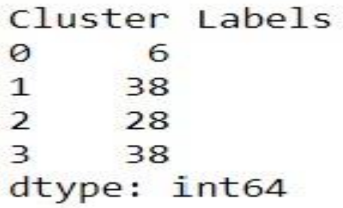
4. Modeling

We are clustering our neighborhood into various clusters to classify our neighborhood into various ways.
The model that I used is K_means clustering which converts the data into clusters.



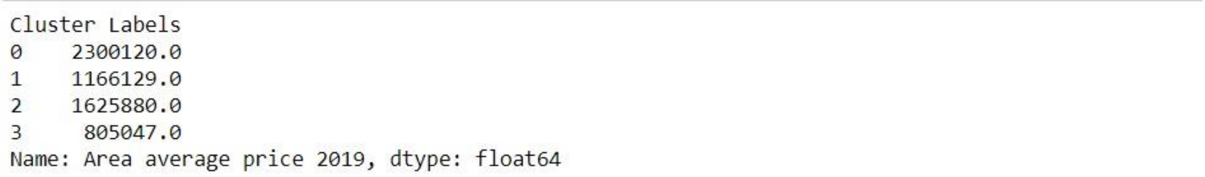
5. Result

We have clustered our neighborhood. The first cluster has 6 neighborhoods. The second has 38. Third has 28 and fourth has also 38.



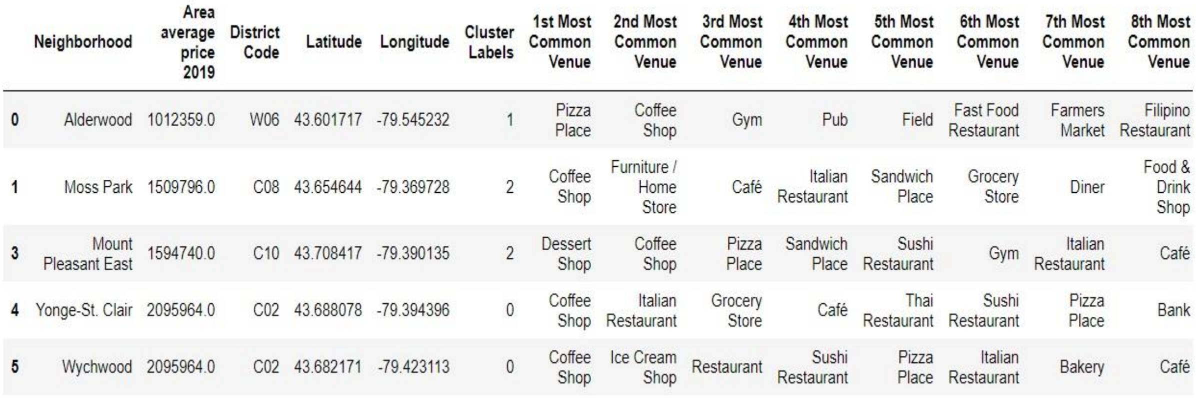
```
Cluster Labels
0      6
1     38
2     28
3     38
dtype: int64
```

The average price of first cluster is $2300120. The second cluster’s is $1166129 and third cluster is $1625880. The fourth cluster is the cheapest among them having average price $805047.



```
Cluster Labels
0    2300120.0
1    1166129.0
2    1625880.0
3     805047.0
Name: Area average price 2019, dtype: float64
```

The popular venues are restaurant and coffee shops.



	Neighborhood	Area average price 2019	District Code	Latitude	Longitude	Cluster Labels	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
0	Alderwood	1012359.0	W06	43.601717	-79.545232	1	Pizza Place	Coffee Shop	Gym	Pub	Field	Fast Food Restaurant	Farmers Market	Filipino Restaurant
1	Moss Park	1509796.0	C08	43.654644	-79.369728	2	Coffee Shop	Furniture / Home Store	Café	Italian Restaurant	Sandwich Place	Grocery Store	Diner	Food & Drink Shop
3	Mount Pleasant East	1594740.0	C10	43.708417	-79.390135	2	Dessert Shop	Coffee Shop	Pizza Place	Sandwich Place	Sushi Restaurant	Gym	Italian Restaurant	Café
4	Yonge-St. Clair	2095964.0	C02	43.688078	-79.394396	0	Coffee Shop	Italian Restaurant	Grocery Store	Café	Thai Restaurant	Sushi Restaurant	Pizza Place	Bank
5	Wychwood	2095964.0	C02	43.682171	-79.423113	0	Coffee Shop	Ice Cream Shop	Restaurant	Sushi Restaurant	Pizza Place	Italian Restaurant	Bakery	Café

6. Discussion

I have observed that places close to the main city is comparatively expensive. The area near city is moderately expensive and the neighborhood where there is nearby pubs, and entertainment center are comparatively expensive.

7. Conclusion

The average cost is clustered as:

1. The first cluster's average cost of 2300120.0
2. The second cluster's average cost of 808301.0
3. The third cluster's average cost of 1631649.0
4. The fourth cluster's average cost of 1171028.0

The neighborhood is clustered as:

1. First cluster's nearby venues consist of restaurant and grocery shops

2. Second cluster's nearby venues consist of all types of restaurant and basic amenities. It looks a great neighborhood to live as the cost is also low.
3. Third Cluster's nearby venues consist of pubs, restaurant, parks, entertainment and leisure places.
4. Fourth Cluster's nearby venues consist of Parks, restaurant, Coffee shops, Banks.

The **conclusion** I got from the above result is the first cluster is the most expensive one. It consists of restaurant and different types of shops and museum. But this cluster only has six neighborhoods and by analyzing the map we find that it is in the epicenter of the city.

The **second** cluster is the least expensive one and it has restaurant and basic amenities around which could be a great place for people having families by analyzing the map we could know that these neighborhoods surround the epicenter of the city.

The **third** cluster is all around the city and it is a second most expensive neighborhood. It consists of all the leisure center and entertainment places hence would be suitable for bachelors. It may also be suitable for opening new stores.

The **fourth** cluster is away from the epicenter of the city and it also a second least expensive place. It has Parks, restaurant and coffee shops all around it.