Analysis of the correlation between Housing Sales Prices and Venues in the city of Toronto using k - means

By Nathan Monteiro

1. Introduction

A Real Estate agency based in the city of Toronto currently acts as a broker between the seller of a house and a buyer. Their business is doing well and they would like to expand their business into the construction of new houses as well. While they know the cost per square foot of a house in each area, they would also like to know whether different venues surrounding the house affect the sale price so that they can effectively price their new developments.

The data for sales prices of houses in various locations around Ontario is available on Kaggle and the agency plans to use Data Science to solve this problem.

2. Data

2.1 Importing the data

To carry out the analysis a dataset on House Sales needs to be imported which is available here:
https://www.kaggle.com/mnabaee/ontarioproperties/downloads/ontarioproperties.zip/1
This dataset includes the listing prices for the sale of properties in Ontario which we can use to filter out only the entries for Toronto.

Calls to the Foursquare API will also need to be made to get access to the venues in a particular area which can be accessed by following the steps here:
https://foursquare.com/

2.2 Cleaning the data

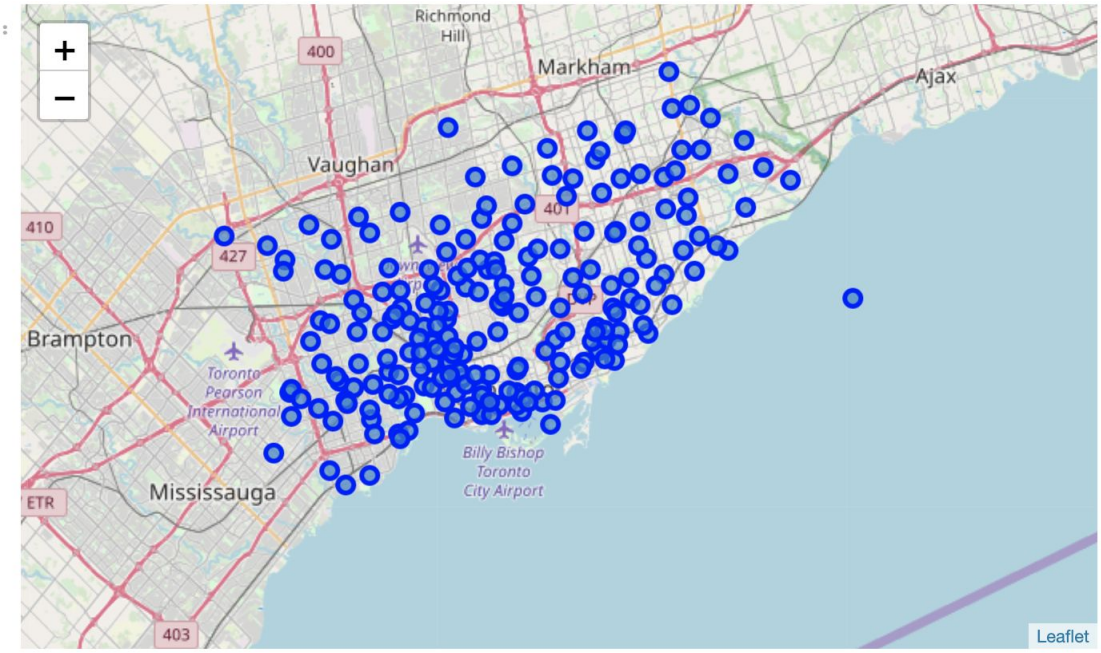
Once the data has been imported it is time to clean the data. The House Sales dataframe is filtered out to only contain entries which are located in Toronto.

	Address	Area	Price	Latitude	Longitude
0	86 Waterford Dr Toronto, ON	Richview	999888.0	43.679882	-79.544266
1	#1409 - 230 King St Toronto, ON	Downtown	362000.0	43.651478	-79.368118
2	254A Monarch Park Ave Toronto, ON	Old East York	1488000.0	43.686375	-79.328918
3	532 Caledonia Rd Toronto, ON	Fairbank	25.0	43.691193	-79.461662
4	47 Armstrong Ave Toronto, ON	Wallace Emerson	113.0	43.664101	-79.439751

A new dataset is created which contains the mean price of houses in each area by grouping the previous dataframe by unique areas and getting the mean.

	Area	Price	Latitude	Longitude
0	Agincourt	4.290126e+05	43.788408	-79.278037
1	Agincourt North	2.200000e+06	43.803215	-79.242554
2	Alderwood	9.739375e+05	43.603299	-79.545057
3	Amesbury	7.945000e+04	43.704548	-79.482700
4	Armdale	5.913167e+04	43.828528	-79.251296

By visualizing the data we notice some incorrect data and remove it from our dataframe.



A call is made to the Foursquare API to find the venues in the areas mentioned in the new dataset that we created. This information is used to create a new dataframe with all the surrounding venues in a particular area. This data is used to perform the analysis.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agincourt	43.788408	-79.278037	One2 Snacks	43.787048	-79.276658	Asian Restaurant
1	Agincourt	43.788408	-79.278037	In Cheon House Korean & Japanese Restaurant 인천관	43.786468	-79.275693	Korean Restaurant
2	Agincourt	43.788408	-79.278037	Tim Hortons	43.785637	-79.279215	Coffee Shop
3	Agincourt	43.788408	-79.278037	Maple Yip Seafood 陸羽海鮮酒家	43.784752	-79.277787	Chinese Restaurant
4	Agincourt	43.788408	-79.278037	Beef Noodle Restaurant 老李牛肉麵	43.785937	-79.276031	Chinese Restaurant

3. Methodology

3.1 Encoding Data

In order to be able to fit the data using the k-means algorithm, it needs to encoded which is done using a one-hot encoder.

	Neighborhood	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Aquarium	A
0	Agincourt	0	0	0	0	0	0	0	0	
1	Agincourt	0	0	0	0	0	0	0	0	
2	Agincourt	0	0	0	0	0	0	0	0	
3	Agincourt	0	0	0	0	0	0	0	0	
4	Agincourt	0	0	0	0	0	0	0	0	

And the rows are grouped by neighborhood and by taking the mean of the frequency of occurrence of each category.

	Neighborhood	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Aquarium
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	Agincourt North	0.0	0.0	0.0	0.0	0.0	0.333333	0.0	0.0
2	Alderwood	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	Amesbury	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Armdale	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

3.2 Adding the price variable

In order to be able to create clusters which take house sales price into consideration as well, the Price column needs to be added to the dataframe. However, it needs to be normalized first so that it does not overpower the other variables.

	Neighborhood	Price
0	Agincourt	0.030023
1	Agincourt North	0.153957
2	Alderwood	0.068157
3	Amesbury	0.005560
4	Armdale	0.004138

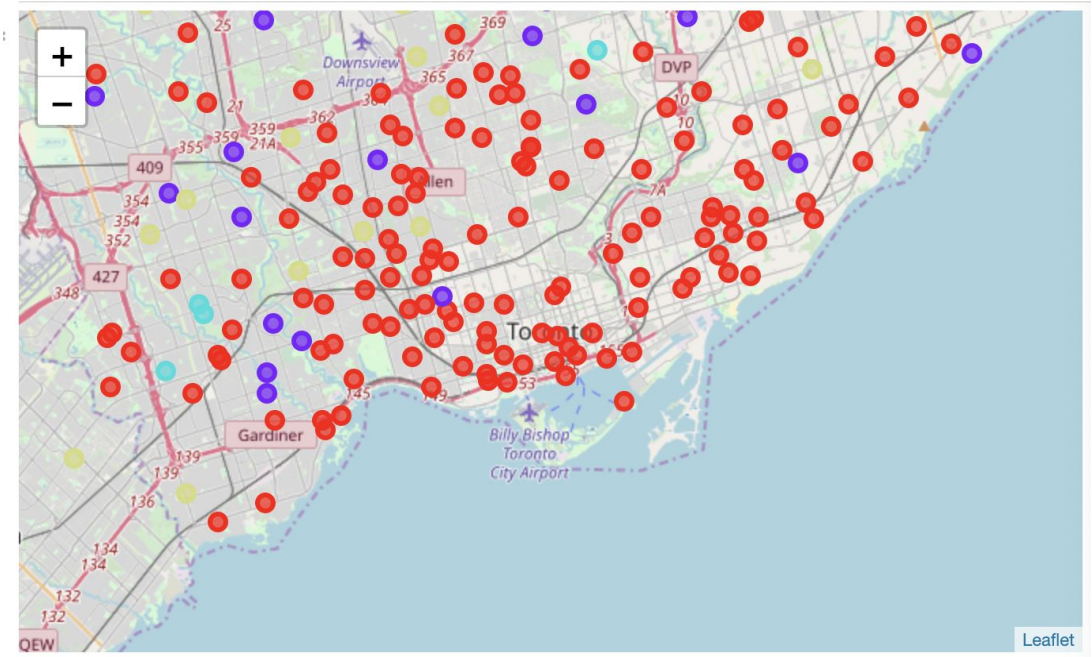
Now the price can be added to the previous dataframe.

3.3 Create clusters using k-means

Drop any column with categorical variables and then the data is ready to be fit by k-means.

	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Aquarium	Arcade	Arge Rest
0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.333333	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	

The data is fitted to create four clusters.



4. Results

4.1 Cluster 1

The first cluster has an average house sales price of approximately $1,202,752. The two most common venues are both Parks indicating that Parks are of significance to this cluster.

```
1st Most Common Venue(s) 0    Park
dtype: object
2nd Most Common Venue(s) 0    Park
dtype: object
Average Price 1202752.597849496
The number of rows are 22
```

4.2 Cluster 2

The first cluster has an average house sales price of approximately $2,570,911. The two most common venues are Tennis Courts and Yoga Studios which are expected in an upscale neighborhood.

```
1st Most Common Venue(s) 0    Tennis Court
dtype: object
2nd Most Common Venue(s) 0    Yoga Studio
dtype: object
Average Price 2570911.9404761903
The number of rows are 4
```

4.3 Cluster 3

The first cluster has an average house sales price of approximately $679,061. The two most common venues are both Pizza Places indicating that Pizza Places are of significance to this cluster. However, we also see that Wings joints are also very common.

```
1st Most Common Venue(s) 0    Pizza Place
dtype: object
2nd Most Common Venue(s) 0    Pizza Place
1    Wings Joint
dtype: object
Average Price 679061.2439325323
The number of rows are 15
```

4.4 Cluster 4

The first cluster has an average house sales price of approximately $876,866. The two most common venues are both Coffee Shops indicating that Parks are of significance to this cluster.

```
1st Most Common Venue(s) 0    Coffee Shop
dtype: object
2nd Most Common Venue(s) 0    Coffee Shop
dtype: object
Average Price 876866.1095713675
The number of rows are 173
```

5. Discussion

If the clusters are sorted based on the average house sales price, Cluster 3 has the lowest average price. It is found that houses on the lower end of the price scale are situated near food joints, particularly Pizza Places and Wings Joints. The next cluster is Cluster 4 with a slightly higher average price. Most houses are situated in this cluster. Houses in this cluster are situated near Coffee Shops. A little higher on the price scale is Cluster 1 which is situated near Parks. And at the highest end of the price scale is Cluster 2 which is situated near Tennis Courts and Yoga Studios.

6.Conclusion

It may be concluded that in the city of Toronto there may be a correlation between higher-priced houses and Tennis Courts, Yoga Studios and Parks in contrast with the relatively lower-priced houses which may be correlated with Pizza Places, Wings Joints, and Coffee Shops.