Peer-graded Assignment:

Capstone Project - The Battle of Neighbourhoods (W2)

Final Report

Fabio Melle

Background

Scarborough (Ontario), is a municipality in Canada and a borough of Toronto. Located above the Scarborough Bluffs, occupying the eastern part of the city. Scarborough is bordered by Victoria Park Avenue to the west, Steeles Avenue to the north, Rouge River and Pickering City to the east and Lake Ontario to the south. The city is named after Scarborough (United Kingdom).

The city is one of the most multicultural of the Greater Toronto Area, hosting various religions and cultures. The city has also been declared the greenest of the Greater Toronto Area [1].

Description

Some Italian people have realized that new job opportunities have opened up in Scarborough, having relatives living there. So they decided to move to this city. Before leaving, they would like to know more about this city and its neighborhoods so that they can evaluate where to live. Before leaving, they would like to have more information about this borough and its neighborhoods in order to evaluate where to go to live.

Goal

The aim of this work is to give these families the opportunity to settle in a good neighbourhood by taking into account some evaluation parameters such as the services that are inside, the quality of education guaranteed by schools and the average cost of housing.

Summary Libraries:

Pandas: For creating and manipulating dataframes.
Numpy: is a math library to work with N-dimensional arrays.
Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.
Scikit Learn: is a collection of numerical algorithms and domain specific toolboxes in this case for importing k-means clustering.
JSON: Library to handle JSON files.
Geopy: To retrieve Location Data
Requests: Library to handle http requests
Matplotlib: Very popular plotting package that provides 2D plotting, as well as 3D plotting module

Methodology

I used GitHub repository as a database. My master data which has the main components *Borough, Average House Price, School Ratings and Latitude* and *Longitude* informations of the city.

I used python **folium** library to visualize geographic details of Scarborough and its neighborhood and I created a map of Scarborough with boroughs superimposed on top. I used latitude and longitude values to get the visual as below:

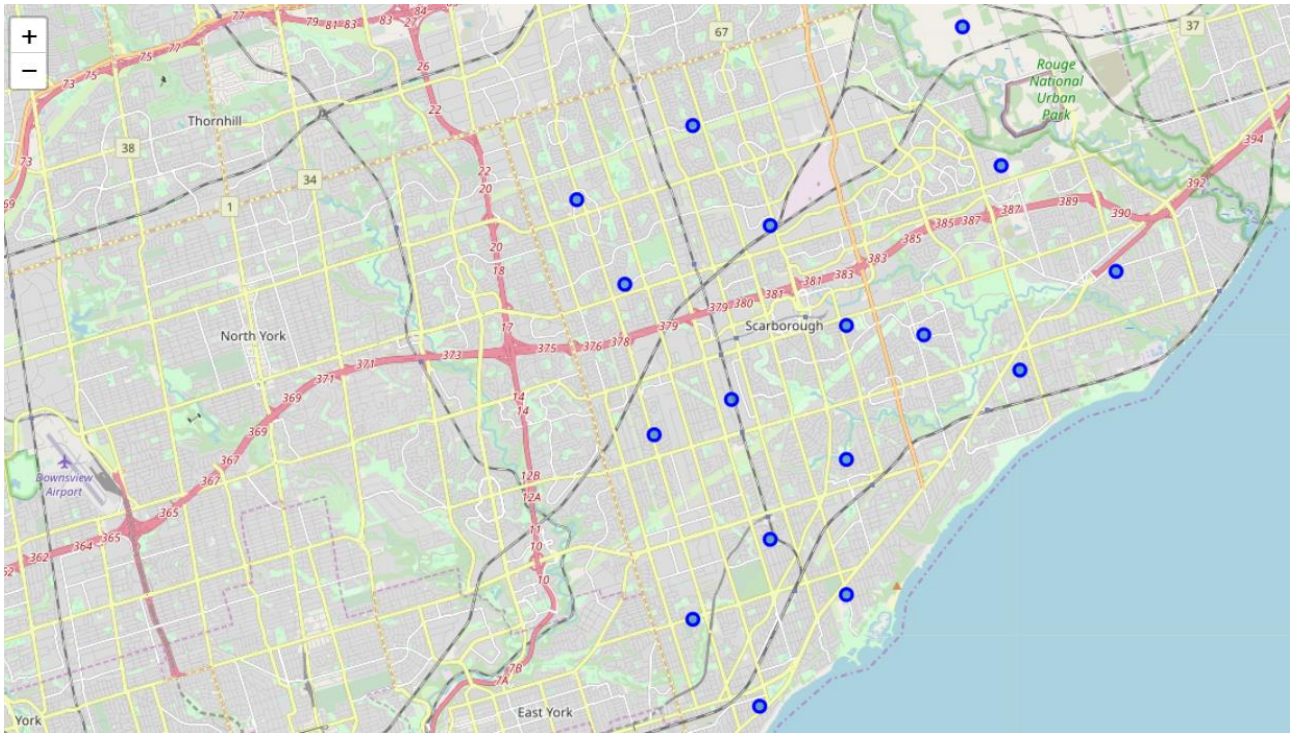


Figura 1 - Scarborough and its Neighborhoods

Geo-location data:

This project is focused on the city of Scarborough, a district of Toronto. In order to implement this work, the geo-location of this district and its neighborhoods will have to be obtained. It's "Scarborough" in Toronto. This project will also require more information about the different districts of Scarborough, average house prices and school evaluations. Below are the required data for each neighborhood:

1. Location of the neighborhood in terms of latitude and longitude

2. Average prices of the apartments

3. School Ratings

The data set containing position data and postcodes is present in the previous project (Segmenting and Clustering Neighborhoods2.ipynb). The location of Scarborough and its neighborhoods will be obtained by filtering the available information:

https://github.com/FMelle-DataScientist/Applied-Data-Science-Capstone/blob/master/Segmenting%20and%20Clustering%20Neighborhoods2.ipynb

	Postal Code	Latitude	Longitude	Borough	Neighborhood
0	M1B	43.806686	-79.194353	Scarborough	Rouge, Malvern
1	M1C	43.784535	-79.160497	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	43.763573	-79.188711	Scarborough	Guildwood, Morningside, West Hill
3	M1G	43.770992	-79.216917	Scarborough	Woburn
4	M1H	43.773136	-79.239476	Scarborough	Cedarbrae
5	M1J	43.744734	-79.239476	Scarborough	Scarborough Village
6	M1K	43.727929	-79.262029	Scarborough	East Birchmount Park, Ionview, Kennedy Park
7	M1L	43.711112	-79.284577	Scarborough	Clairlea, Golden Mile, Oakridge
8	M1M	43.716316	-79.239476	Scarborough	Cliffcrest, Cliffside, Scarborough Village West
9	M1N	43.692657	-79.264848	Scarborough	Birch Cliff, Cliffside West
10	M1P	43.757410	-79.273304	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...
11	M1R	43.750071	-79.295849	Scarborough	Maryvale, Wexford
12	M1S	43.794200	-79.262029	Scarborough	Agincourt
13	M1T	43.781638	-79.304302	Scarborough	Clarks Corners, Sullivan, Tam O'Shanter
14	M1V	43.815252	-79.284577	Scarborough	Agincourt North, L'Amoreaux East, Milliken, St...
15	M1W	43.799525	-79.318389	Scarborough	L'Amoreaux West
16	M1X	43.836125	-79.205636	Scarborough	Upper Rouge

Figura 2 - Info geo-location of neighborhoods of Scarborough

Foursquare API:

As the main source of data collection I used the Foursquare API as it has a database with a capacity of millions of places, especially the place API which offers the ability to perform location searches, location sharing and detailed information about a particular company.

By using the Foursquare Developer API [2] with my account I will be able to get places close to the districts in the district. Due to the limitations related to my account, the number of venues per neighborhood parameter should be set to 100 and the radius parameter to 500.

The data retrieved from Foursquare contained venues information within a certain distance from the longitude and latitude of the postcodes. The information obtained for each venue is as follows:

1. Neighborhood
2. Neighborhood Latitude

3. Neighborhood Longitude
4. Venue
5. Name of the venue e.g. the name of restaurant
6. Venue Category
7. Venue Latitude
8. Venue Longitude

The Knowledge Discovery in Database (KDD) process:

KDD is an automatic iterative process of data exploration and analysis that aims to extract from data new knowledge useful to support decision-making processes.

As an input to the process you have the data coming from the sources mentioned here, as an output from the process you have a new acquired knowledge, useful for Italian families who are about to move to Scarborough. Steps taken were:

1. Selection: The data is extracted according to certain criteria that depend on our goal. In this case the first goal is to obtain postcodes for the Toronto neighbourhoods and venues within these neighborhoods.

2. Pre-processing: During this process the data is cleaned up by information considered unnecessary (e.g. NaN). At this stage, data may be transformed to avoid inconsistencies due to similar data from different sources and managed with slightly different metadata.

3. Transformation: Data can be transformed in such a way as to add to them further fields, thus making them usable and navigable for the specification application.

4. Data mining: At this stage the models are extracted from the data. I will use rules and association discovery algorithms to generate models. Once generated, the models will be interpreted. In this study I will refer to a clustering algorithm (k-means).

5. Interpretation and evaluation: in this phase the models identified by the system are used and interpreted with the aim of using the acquired knowledge. This knowledge will be used to support decisions, making forecasts and describing the phenomena observed.

Data Mining process

In order to make the evaluations, we will explore the neighbourhoods, segmenting them and grouping them into clusters to assess which are the most common local ones in the different neighbourhoods and then take decisions also taking into account these evaluation parameters. We will then proceed to cluster data thanks to an unsupervised machine learning partition algorithm: k-means clustering algorithm.

Below is my merged table with cluster labels for each neighbourhood.

K-means Clustering

```
# set number of clusters
# Using k-means to cluster the neighborhood into 3 clusters.
kclusters = 3

Scarborough_grouped_clustering = Scarborough_grouped.drop('Neighborhood', 1)
#Scarborough_grouped_clustering = Scarborough_grouped

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(Scarborough_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2], dtype=int32)
```

```
Scarborough_merged =Scarborough_data.iloc[:16,:]

# add clustering labels
Scarborough_merged['Cluster Labels'] = kmeans.labels_

# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
Scarborough_merged = Scarborough_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

Scarborough_merged.head()# check the last columns!
```

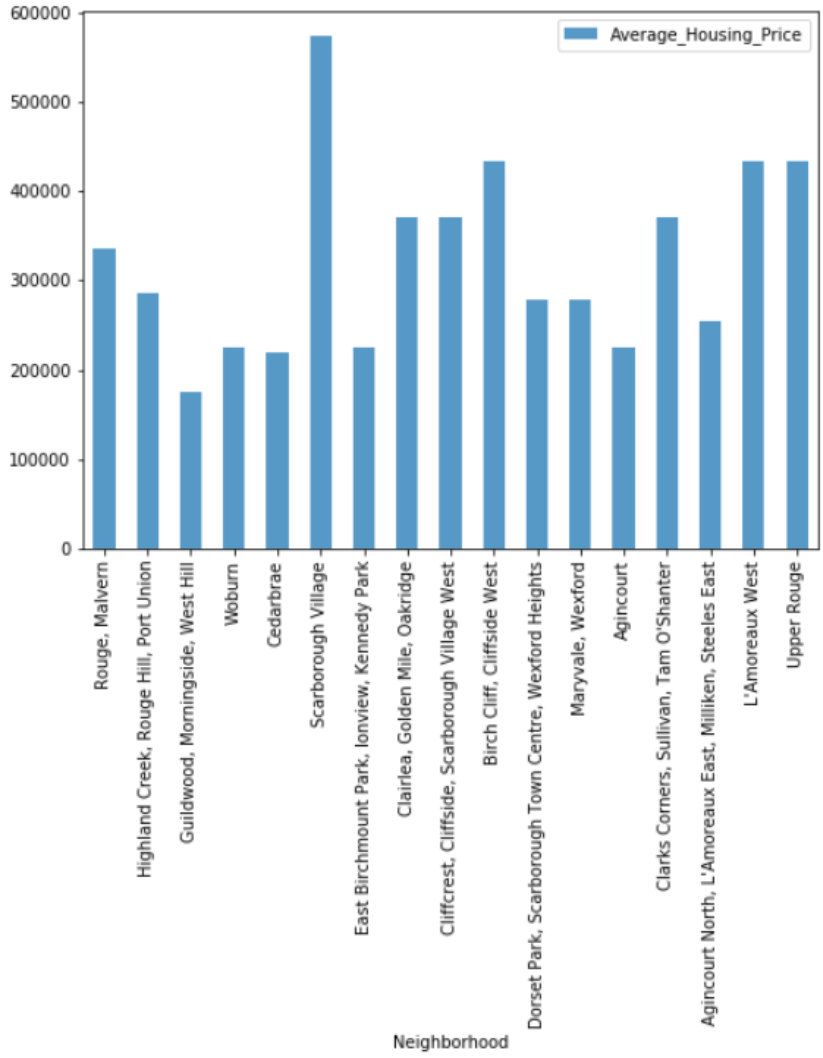
	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353	0	Fast Food Restaurant	Hobby Shop	Spa	Coffee Shop
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	0	Breakfast Spot	Bar	Burger Joint	Fried Chicken Joint
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	0	Fast Food Restaurant	Park	Laundromat	Pizza Place
3	M1G	Scarborough	Woburn	43.770992	-79.216917	0	Coffee Shop	Park	Business Service	Fish Market
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	0	Indian Restaurant	Bakery	Coffee Shop	Asian Restaurant

Figura 3 - Part 1

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
Hobby Shop	Spa	Coffee Shop	Grocery Store	Gym / Fitness Center	College Stadium	Convenience Store	Department Store	Diner
Bar	Burger Joint	Fried Chicken Joint	Department Store	Diner	Discount Store	Electronics Store	Fast Food Restaurant	Fish Market
Park	Laundromat	Pizza Place	Intersection	Rental Car Location	Restaurant	Medical Center	Breakfast Spot	Electronics Store
Park	Business Service	Fish Market	College Stadium	Convenience Store	Department Store	Diner	Discount Store	Electronics Store
Bakery	Coffee Shop	Asian Restaurant	Board Shop	Gas Station	Fried Chicken Joint	Flower Shop	Lounge	Hakka Restaurant

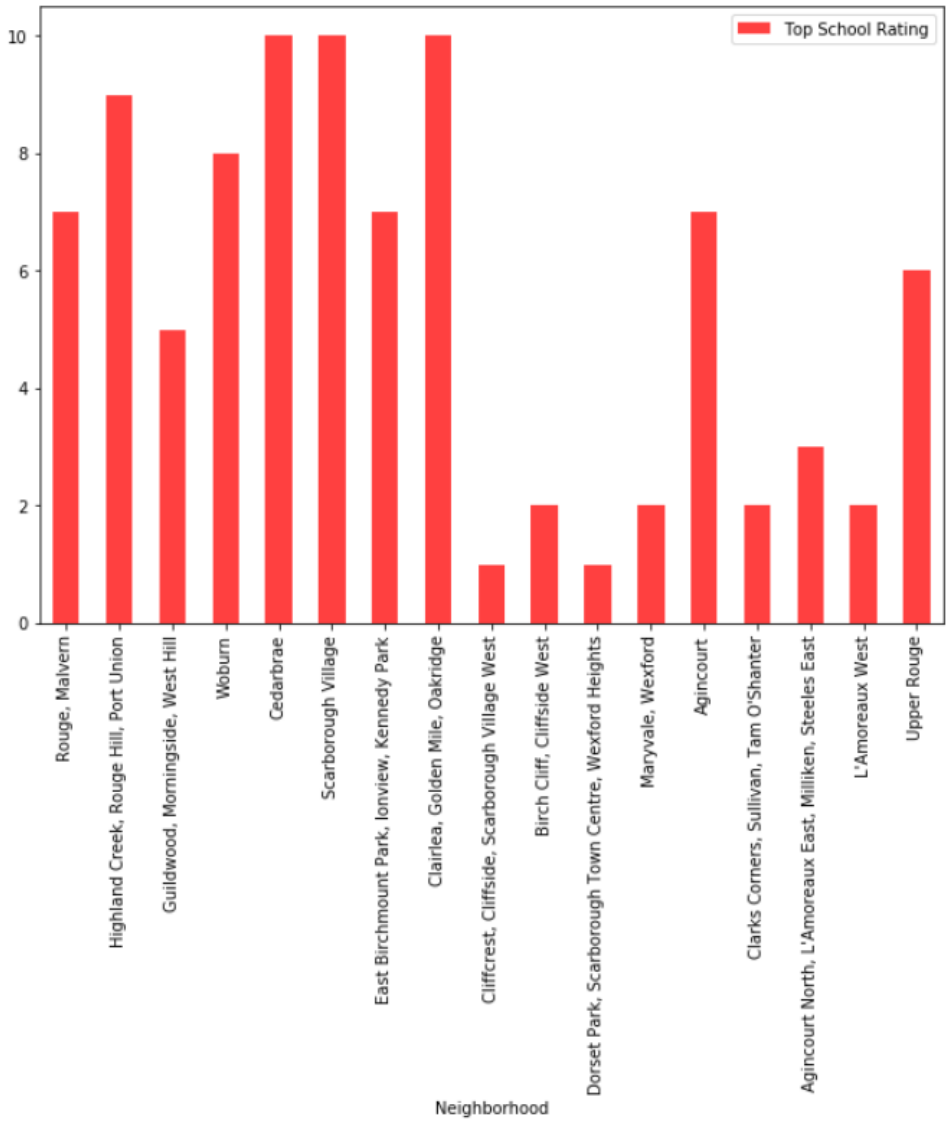
Figura 4 - Part 2

Neighborhood Average Sales Housing Prices



As you can see in the histogram above, I can define the different price ranges according to the Neighborhood of Scarborough. It is therefore clear that depending on the economic availability of a family, I will be able to choose the neighborhood to move to. It is clear that Scarborough Village is the most expensive, while the districts of Guildwood, Morningside and West Hill are the cheapest. You will also be able to compare the average prices of the accommodations with the frequency of services and venues that that neighborhood has, as shown in the list above, to see whether to move to the cheapest apartment taking into account the services offered in that neighborhood.

School rating by clusters

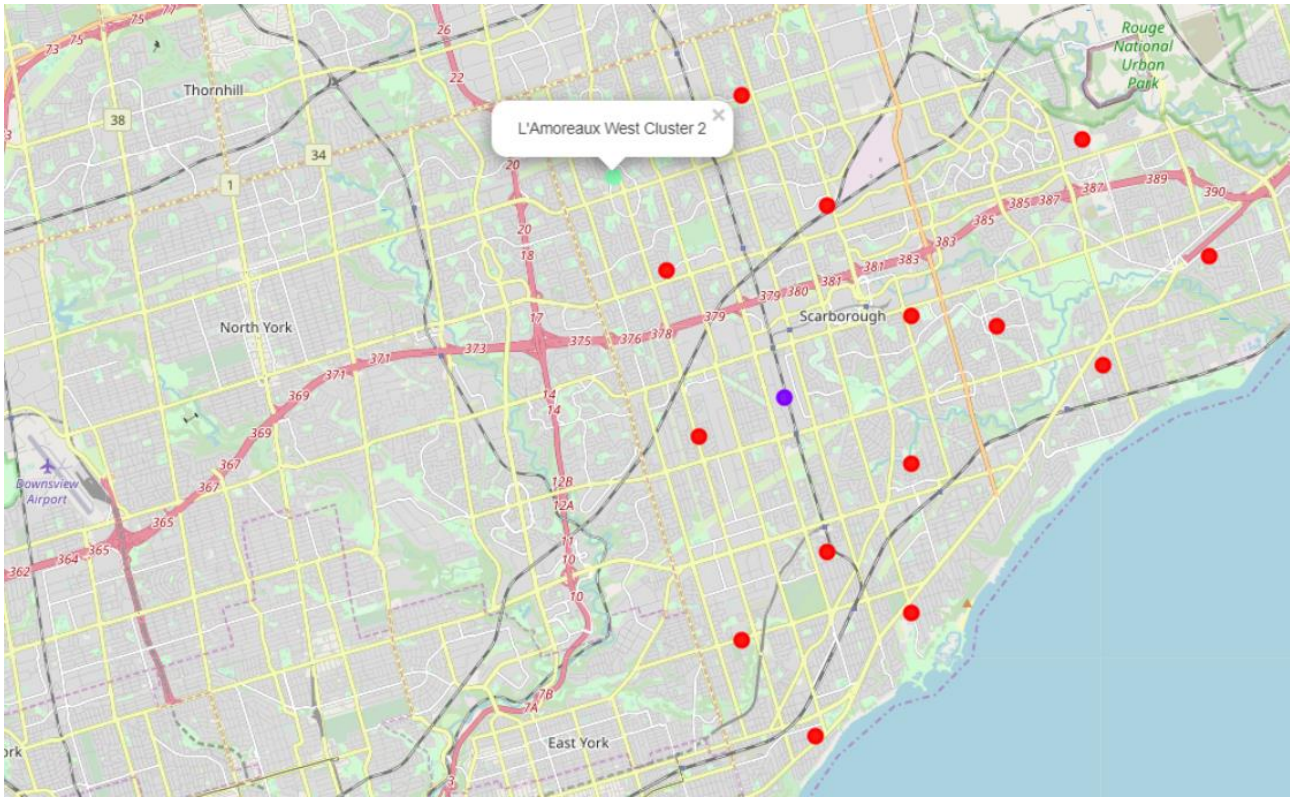


As you can see in the histogram above, I can define the different school rating per neighborhood. In this case you can see that unlike housing prices, school rating is distributed differently, e.g. the neighborhoods of Guildwood, Morningside and West Hill have the cheapest average housing prices but their school rating is not the lowest ever. So this could be a good compromise to buy a house in one of these neighborhoods.

In final section, I created choropleth map which also has the below informations for each neighborhood:

Neighborhood name,

Cluster label,



Discussion:

As I said before, Scarborough is a big city that has been attracting a lot of people in recent years as a new residential destination, with its 17 districts offering different services and venues. Because there is such complexity, you can try very different approaches in clustering studies.

I used the k-means algorithm as part of this clustering study. I set the k value to 3 and then I divide the borough into 03 clusters, which have similar neighbourhoods around them.

I also performed the data analysis through this information by adding the district coordinates and average house prices and school ratings as static data on GitHub.

I concluded the study by displaying the data and clustering information on the map of Scarborough. In the future, web or telephone applications can be developed for direct investors

Conclusion:

So, a lot of people are heading to big cities to change their lives, find a job or start a business. For this reason, people can get better results through access to platforms where this kind of information is provided.

Not only investors, but also mayors of cities can manage the city more effectively and efficiently by exploiting this kind of analysis, visualization and knowledge extraction.

Sitography:

[1] Scarborough - Wikipadia: https://en.wikipedia.org/wiki/Scarborough,_Toronto

[2] Foursquare: https://developer.foursquare.com/