Cousrera Capstone Project: Final Report

Finding better residential places in Scarborough, Toronto

1. Introduction:

Stable, affordable housing is a key determinant of health. The house and neighborhood where one grows up impacts the health and longevity of one's life. Having a safe, decent, stable, affordable home also impacts a child's success in school. Therefore people good residential places with good neighborhoods to live a happy life.

From schools, restaurants, sports activities to social gatherings all comes under a healthy neighborhood and for which any person looks into while finding a home for himself.

The purpose of this Project is to help people in exploring better facilities around their neighborhood. It will help people making smart and efficient decision on selecting great neighbourhood out of numbers of other neighborhoods in Scarborough, Toronto.

Lots of people are migrating to various states of Canada and needed lots of research for good housing prices and reputed schools for their children. This project is for those people who are looking for better neighborhoods. For ease of accessing to Cafe, School, Super market, medical shops, grocery shops, mall, theatre, hospital, likeminded people, etc.

This Project aim to create an analysis of features for a people migrating to Scarborough to search a best neighborhood as a comparative analysis between neighborhoods. The features include median housing price and better school according to ratings, crime rates of that particular area, road connectivity, weather conditions, good management for emergency, water resources both fresh and waste water and excrement conveyed in sewers and recreational facilities.

It will help people to get awareness of the area and neighborhood before moving to a new city, state, country or place for their work or to start a new fresh life.

2. Data Section

Data Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Will use Scarborough dataset which we scrapped from Wikipedia on Week 3. Dataset consisting of latitude and longitude, zip codes.

Foursquare API Data:

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, thefoursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meters.

The data retrieved from foursquare contained information of venues within a specified distance ofthe longitude and latitude of the postcodes. The information obtained per venue as follows:

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

Data Collection/Gathering:

The data used was the location of schools that was acquired using the foursquare website. To gather the data, foursquare API was used along with the foursquare credentials Client ID and Client Secret.

Data Cleaning:

To fit the model, one needs toget rid of the null values. Hence, the firstly, the columns with null, none or NaN values, were identified. The cleaning of data was done by removing the columns with NaN or null values.

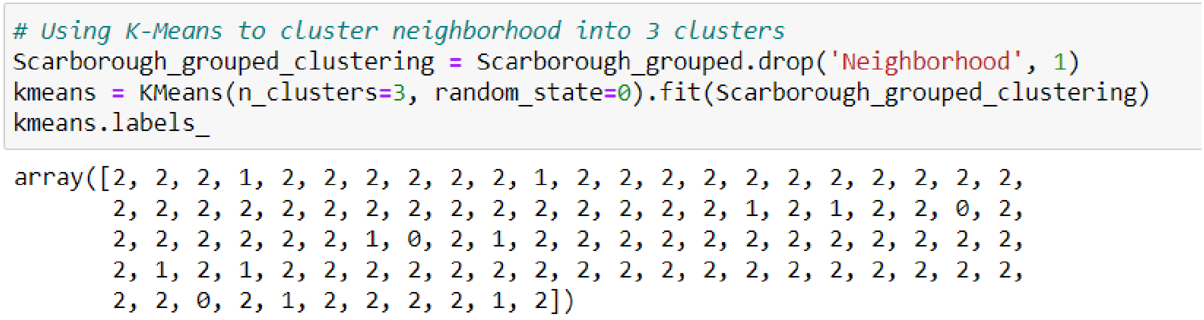
The columns were dropped keeping in mind whether they were really useful for analysis purpose or not. Heat map is generated for visualizing all the null values of the columns.

3. Methodology

k-means Clustering

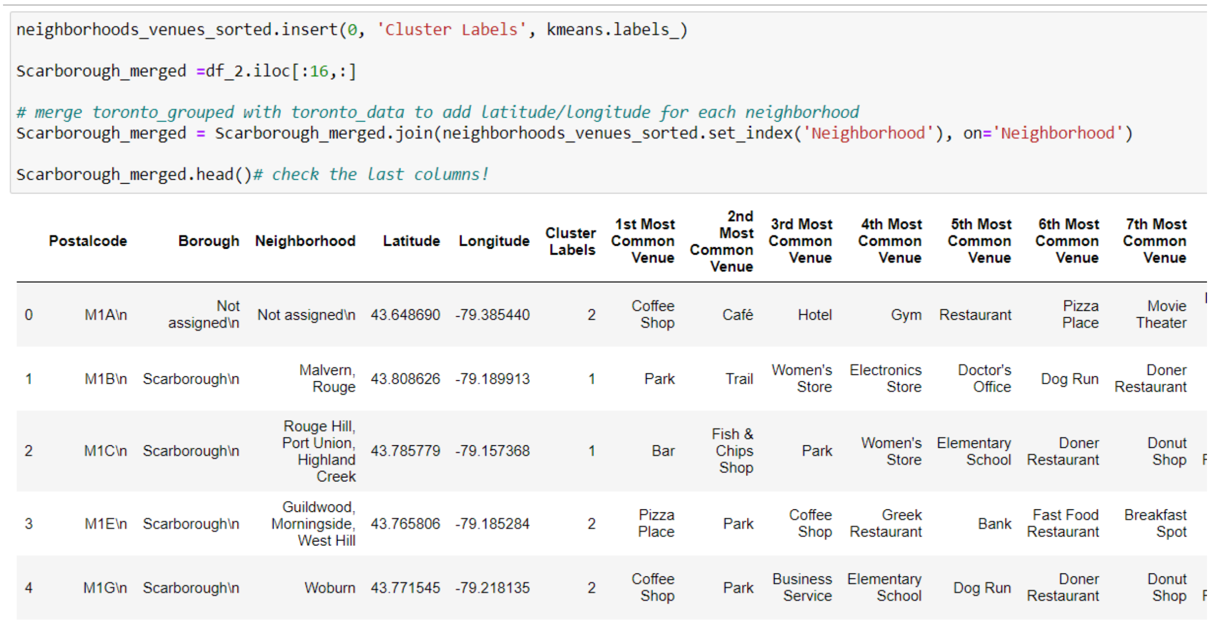
For this project we will be using the k-means clustering approach. K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

The aim of this approach is to compare the similarities of two cities and to decide to explore neighbourhoods, segment them, and group them into clusters to find similar neighbourhoods in a big city like New York and Toronto. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.



```
# Using K-Means to cluster neighborhood into 3 clusters
Scarborough_grouped_clustering = Scarborough_grouped.drop('Neighborhood', 1)
kmeans = KMeans(n_clusters=3, random_state=0).fit(Scarborough_grouped_clustering)
kmeans.labels_
```

```
array([2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 0, 2,
       2, 2, 2, 2, 2, 2, 1, 0, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 0, 2, 1, 2, 2, 2, 2, 1, 2])
```

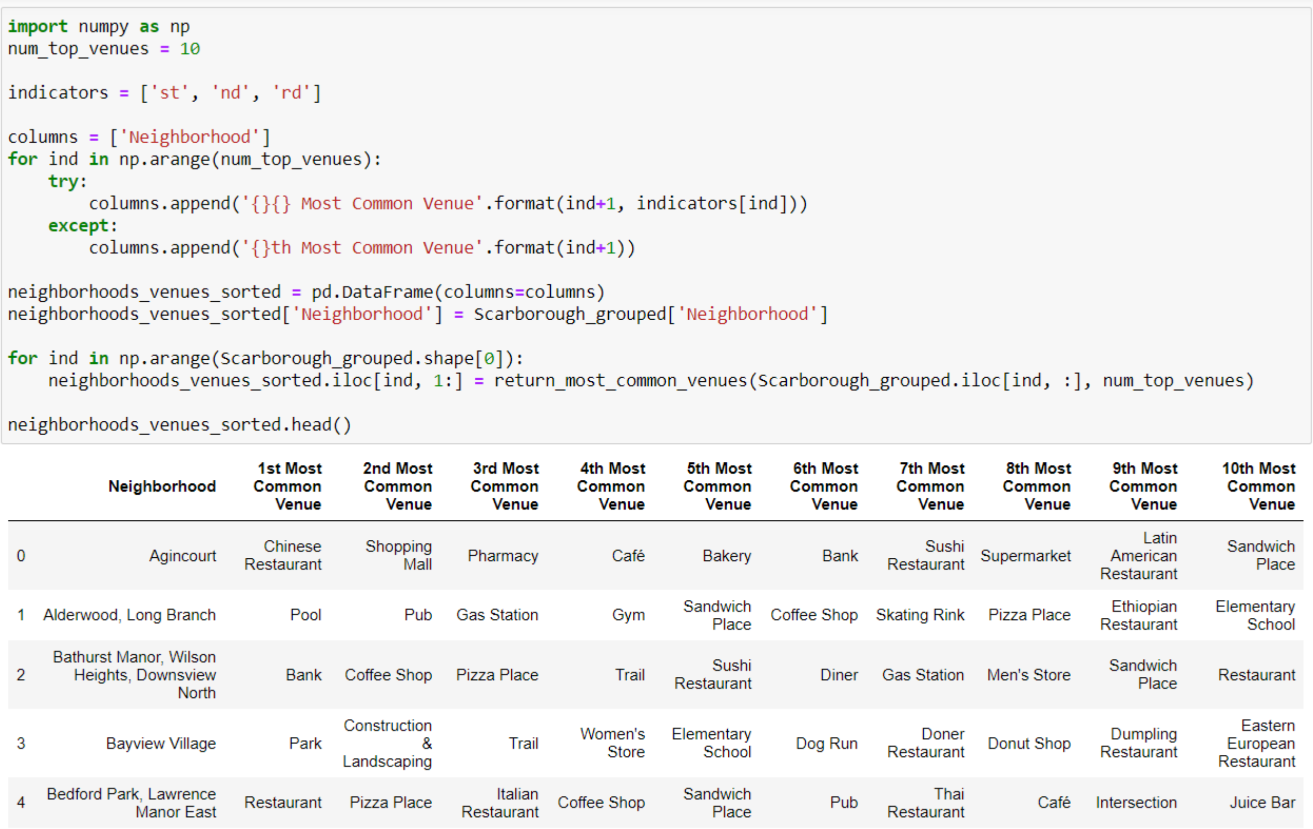


```
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

Scarborough_merged =df_2.iloc[:16,:]

# 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!
```

	Postalcode	Borough	Neighborhood	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
0	M1A\n	Not assigned\n	Not assigned\n	43.648690	-79.385440	2	Coffee Shop	Café	Hotel	Gym	Restaurant	Pizza Place	Movie Theater
1	M1B\n	Scarborough\n	Malvern, Rouge	43.808626	-79.189913	1	Park	Trail	Women's Store	Electronics Store	Doctor's Office	Dog Run	Doner Restaurant
2	M1C\n	Scarborough\n	Rouge Hill, Port Union, Highland Creek	43.785779	-79.157368	1	Bar	Fish & Chips Shop	Park	Women's Store	Elementary School	Doner Restaurant	Donut Shop
3	M1E\n	Scarborough\n	Guildwood, Morningside, West Hill	43.765806	-79.185284	2	Pizza Place	Park	Coffee Shop	Greek Restaurant	Bank	Fast Food Restaurant	Breakfast Spot
4	M1G\n	Scarborough\n	Woburn	43.771545	-79.218135	2	Coffee Shop	Park	Business Service	Elementary School	Dog Run	Doner Restaurant	Donut Shop

The above table shows the clusters in which the neighbourhoods are distributed.

After finding the clusters we found the most common venues near those neighbourhoods. This will help the people to analyse the neighbourhood based on the venues near their place.



```
import numpy as np
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = Scarborough_grouped['Neighborhood']

for ind in np.arange(Scarborough_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(Scarborough_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

	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	Chinese Restaurant	Shopping Mall	Pharmacy	Café	Bakery	Bank	Sushi Restaurant	Supermarket	Latin American Restaurant	Sandwich Place
1	Alderwood, Long Branch	Pool	Pub	Gas Station	Gym	Sandwich Place	Coffee Shop	Skating Rink	Pizza Place	Ethiopian Restaurant	Elementary School
2	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Pizza Place	Trail	Sushi Restaurant	Diner	Gas Station	Men's Store	Sandwich Place	Restaurant
3	Bayview Village	Park	Construction & Landscaping	Trail	Women's Store	Elementary School	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant
4	Bedford Park, Lawrence Manor East	Restaurant	Pizza Place	Italian Restaurant	Coffee Shop	Sandwich Place	Pub	Thai Restaurant	Café	Intersection	Juice Bar

Using credentials of Foursquare API features of near-by places of the neighbourhoods would be mined. Due to http request limitations the number of places per neighbourhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

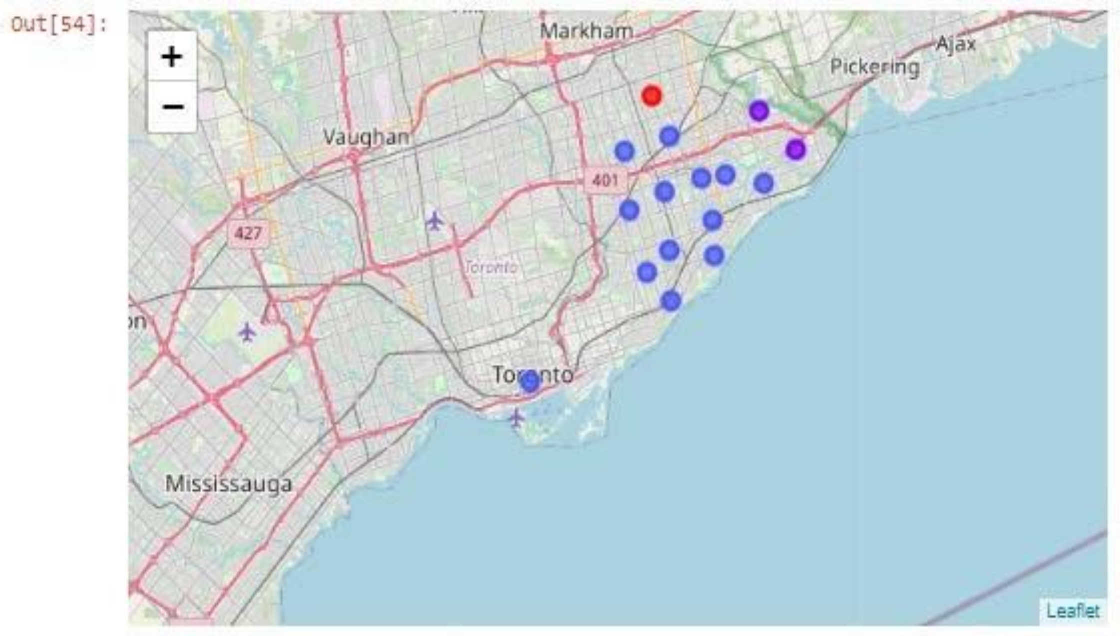
4. Result Analysis

Scarborough is a popular destination for new immigrants in Canada to reside. As a result, it is oneof the most diverse and multicultural areas in the Greater Toronto Area, being home to various religious groups and places of worship. Although immigration has become a hot topic over the pastfew years with more governments seeking more restrictions on immigrants and refugees, the general trend of immigration into Canada has been one of on the rise.

This project has used Four-square API as its prime data gathering source as it has a database ofmillions of places, especially their places API which provides the ability to perform location search,location sharing and details about a business.

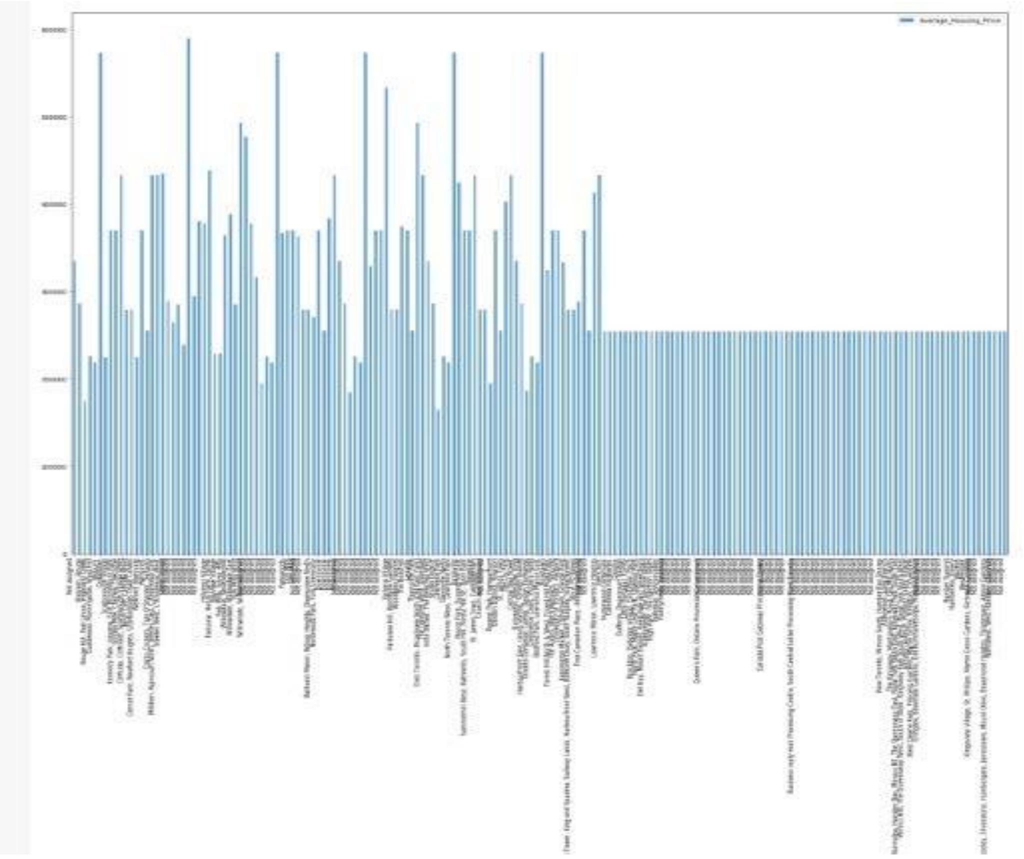
In this section we will be displaying the results based on which a suitable place in Scarborough can be found.

Mapping the clusters of neighborhood in Scarborough

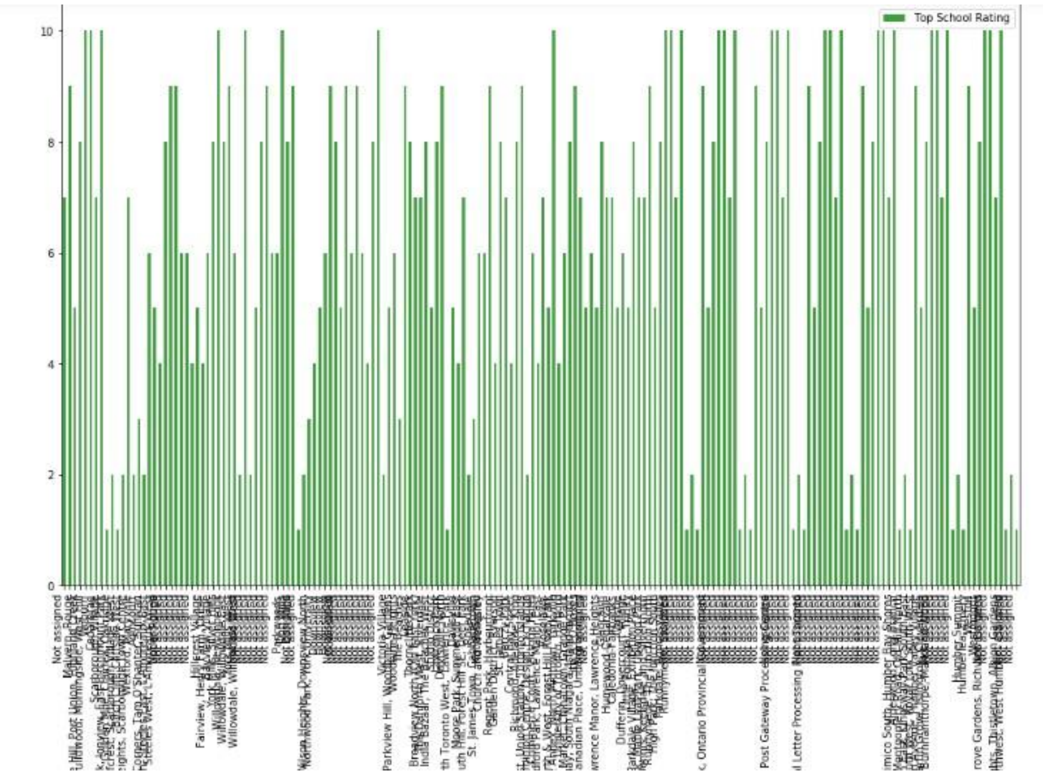


Out[54]:

Plot of Average Housing price in Scarborough



Plot of School ratings in Scarborough



5. Discussion

The major purpose of this project, is to suggest a better neighborhood in a new city for the personwho are shifting there. Social presence in society in terms of likeminded people. Connectivity to the airport, bus stand, city center, markets and other daily needs things nearby.

1. Sorted list of houses in terms of housing prices in an ascending or descending order
2. Sorted list of schools in terms of location, fees, rating and reviews

6. Conclusion and Future Work

In this project, using k-means cluster algorithm I separated the neighborhood into 10(Ten) different clusters and for 103 different latitude and longitude from dataset, which have very-similar neighborhoods around them. Using the charts above results presented to a particular neighborhood based on average house prices and school rating have been made.

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy ofappreciation. This project has shown me a practical application to resolve a real situation that hasimpacting personal and financial impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.
This project can be continued for making it more precise in terms to find best house in Scarborough. Best means on the basis of all required things (daily needs or things we need to livea better life) around and also in terms of cost effective.