Final Report

on

Data Science Capstone Project : 5

Topic : The Battle of Neighborhood – Finding a better place in Scarborough , Toronto .

Instructor : Sir Alex Aklson , Senior Data Scientist at IBM .

Made by : Rishi Kumar

Table of Content :

- Introduction
- Data Section
- Methodology Section
- Result Section
- Discussion Section
- Conclusion Section

Introduction

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 Neighborhood out of numbers of other Neighborhood in Scarborough, Toranto.

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 Neighborhood . For ease of accessing to Cafe, School, Super market, medical shops, Grocery shops, mall, theatre, hospital, like Minded 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 Neighborhood . 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.

Figure : Scarborough , Toronto

Data Section

Source of data

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

We 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 Neighborhood of that specific borough. In order to gain that information we will use "Foursquare" location 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, the foursquare 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 Neighborhood and 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 meter.

The data retrieved from Foursquare contained information of venues within a specified distance of the 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

Map of Scarborough

We will create the Map of Scarborough through the use of folium library .

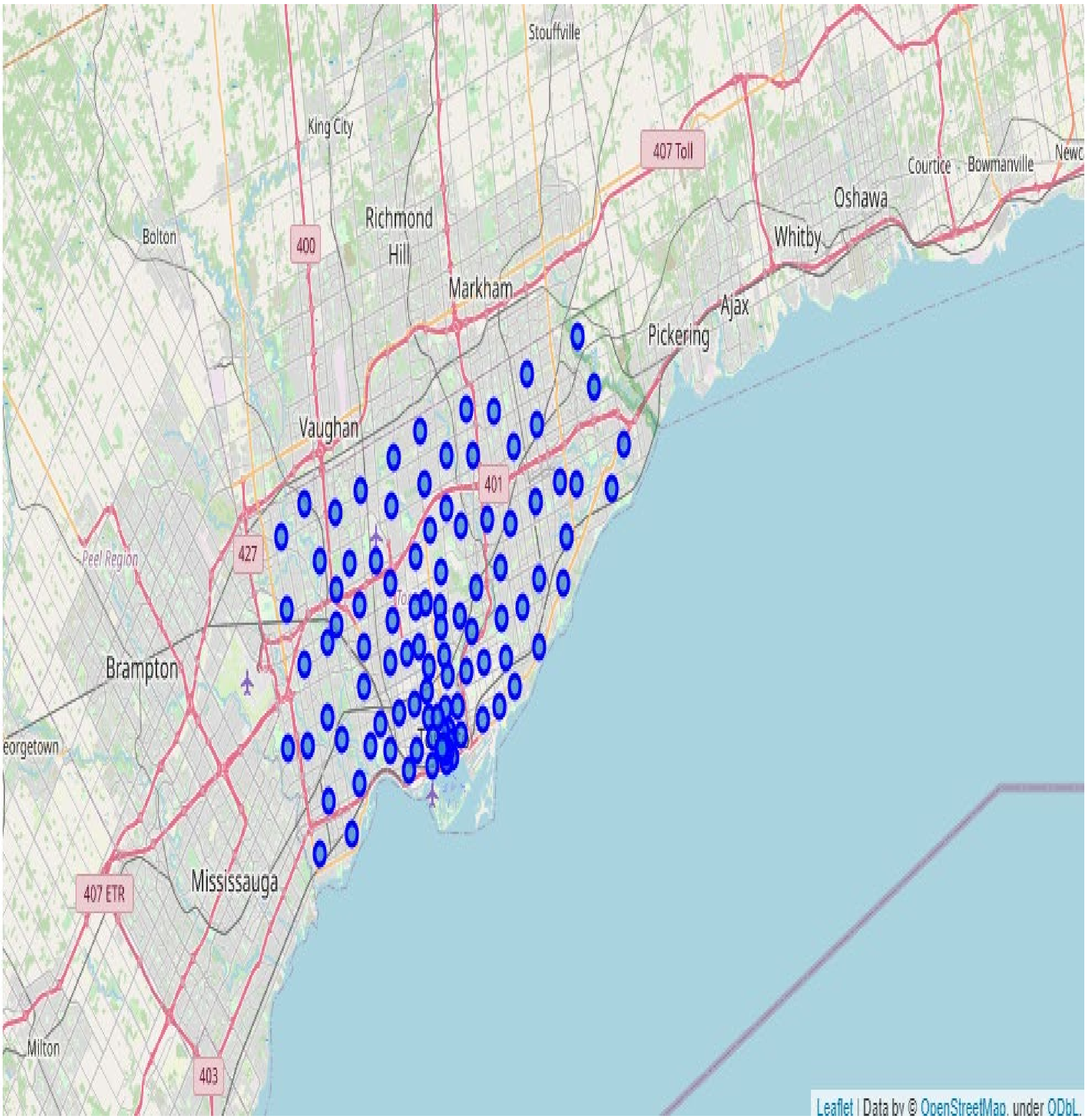


Figure 2 : Map of Scarborough

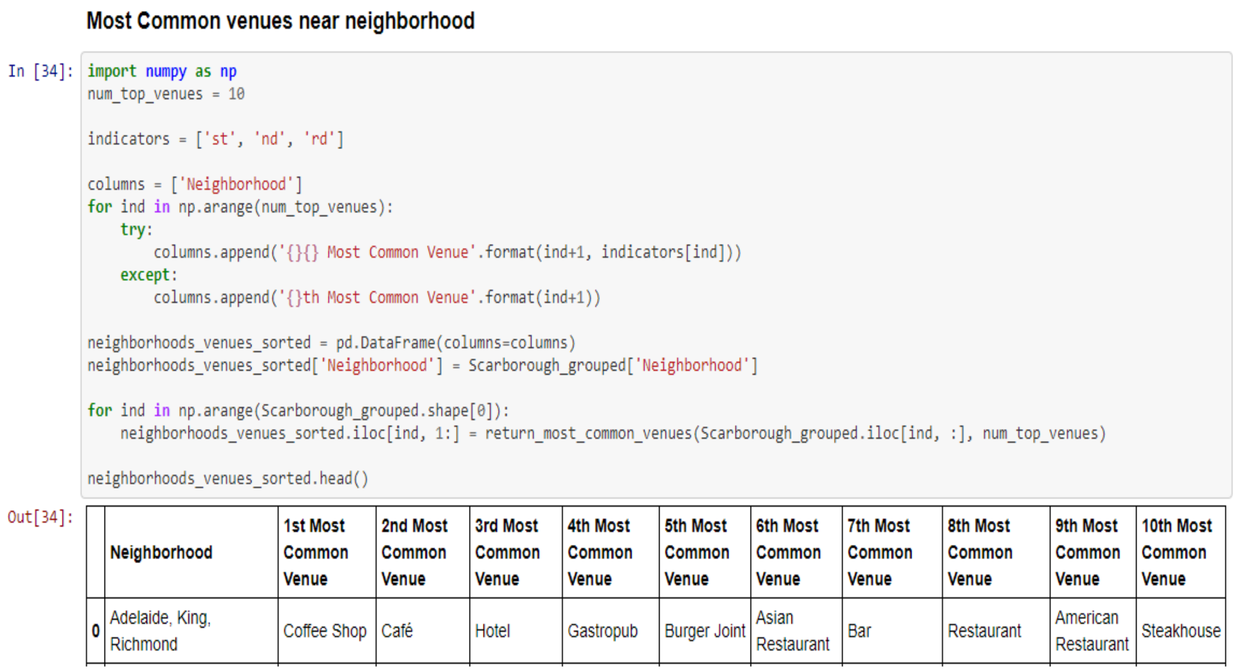
Methodology Section

Clustering Approach

To compare the similarities of two cities, we decided to explore Neighborhood , segment them, and group them into clusters to find similar Neighborhood 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.

Most Common Venues near by Neighborhood



Most Common venues near neighborhood

In [34]:

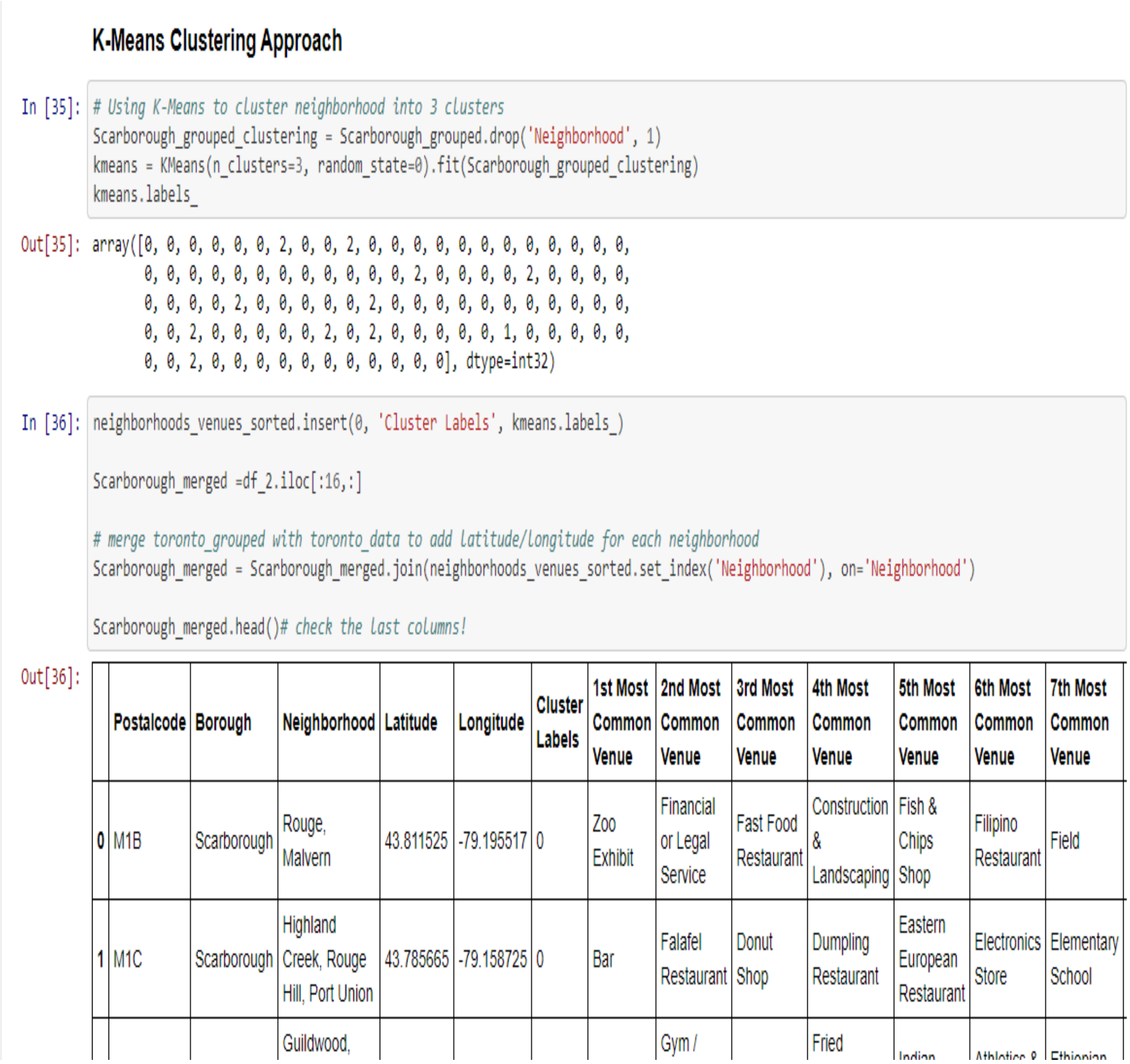
```
import numpy as np
num_top_venues = 10

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()
```

Out[34]:

	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	Adelaide, King, Richmond	Coffee Shop	Café	Hotel	Gastropub	Burger Joint	Asian Restaurant	Bar	Restaurant	American Restaurant	Steakhouse

Using K – Means Clustering approach we will create the map of clusters .

Using K – Means Clustering Approach



K-Means Clustering Approach

In [35]:
```
# 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_
```

Out[35]:
```
array([0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 2, 0, 0, 0, 0,
       0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 2, 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int32)
```

In [36]:
```
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!
```

Out[36]:

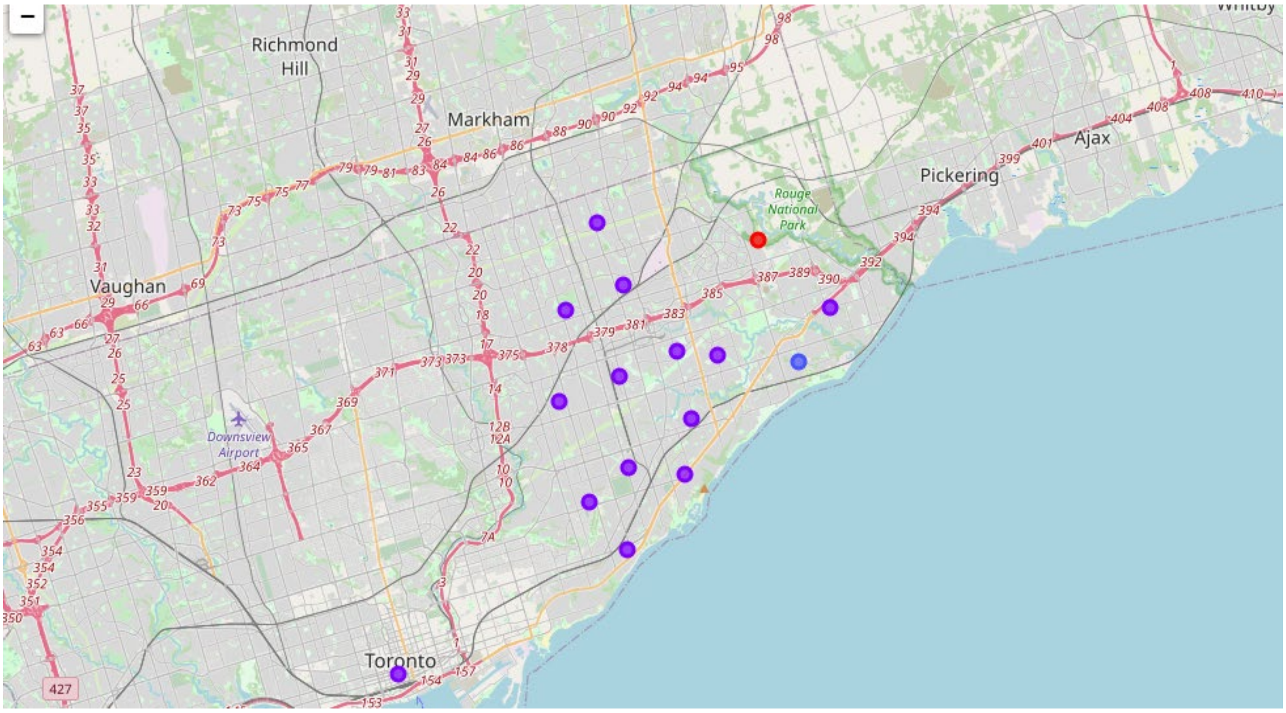
	Postalcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	M1B	Scarborough	Rouge, Malvern	43.811525	-79.195517	0	Zoo Exhibit	Financial or Legal Service	Fast Food Restaurant	Construction & Landscaping	Fish & Chips Shop	Filipino Restaurant	Field
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.785665	-79.158725	0	Bar	Falafel Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Elementary School
			Guildwood,					Gym /		Fried	Indian	Athletics &	Ethiopian

Work Flow

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

Result Section

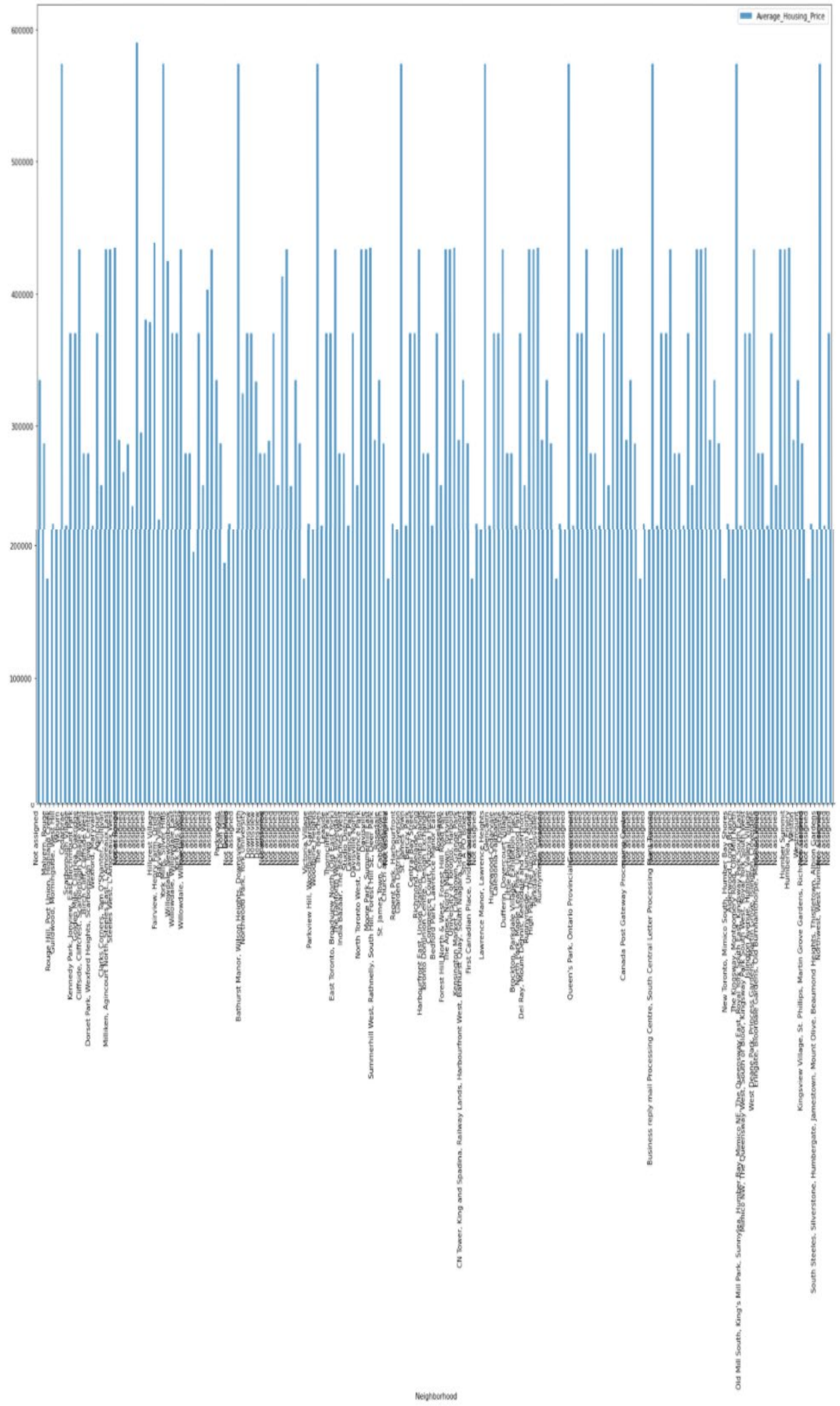
Map of Clusters in Scarborough



Average Housing Price by Clusters in Scarborough

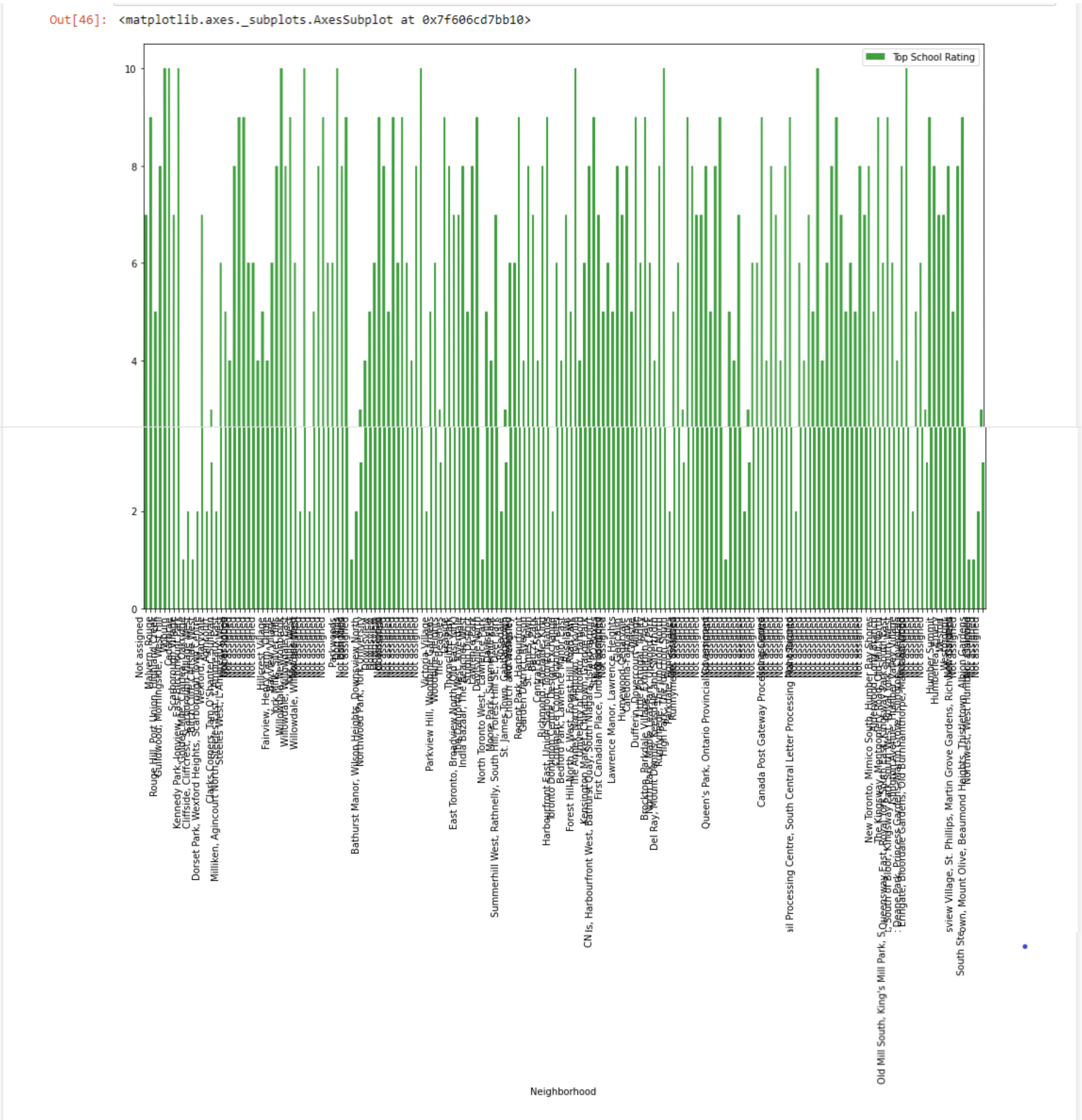
- We calculate the average housing price by using clusters in the Scarborough
- In this case we have calculated and see the average house price of 306 cities .
- In order to create the clusters in the graph using K – Means Clustering algorithm .
- In the below figure we can see the clusters of the city locations Scarborough .

```
Out[42]: <matplotlib.axes._subplots.AxesSubplot at 0x7f606e0e0d10>
```



Average_Housing_Price
600000
500000
400000
300000
200000
100000
0
Not assigned
Malvern, Rouge
Rouge Hill, Port Union, Highland Creek
Guildwood, Morningside, West Hill
Kennedy Park, Ionview, East Birchmount Park
Cliffside, Cliffcrest, Scarborough Village West
Dorset Park, Wexford Heights, Scarborough Town Centre
Clarks Corners, Tam O'Shanter, Sullivan
Milliken, Agincourt North, Steeles East, L'Amoreaux East
Hillcrest Village
Fairview, Henry Farm, Oriole
Willowdale, Willowdale East
Bathurst Manor, Wilson Heights, Downsview North
Parkview Hill, Woodbine Gardens
East Toronto, Broadview North (Old East York)
North Toronto West, Lawrence Park
Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park
St. James Town, Cabbagetown
Regent Park, Harbourfront
Harbourfront East, Union Station, Toronto Islands
Forest Hill North & West, Forest Hill Road Park
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport
First Canadian Place, Underground city
Lawrence Manor, Lawrence Heights
Brockton, Parkdale Village, Exhibition Place
Del Ray, Mount Dennis, Keelsdale and Silverthorn
Queen's Park, Ontario Provincial Government
Canada Post Gateway Processing Centre
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto
New Toronto, Mimico South, Humber Bay Shores
Old Mill South, King's Mill Park, Sunnylea, Humber Bay, Mimico NE, The Queensway East, Royal York South East, Kingsway Park South East
Kingsview Village, St. Phillips, Martin Grove Gardens, Richview Gardens
South Steeles, Silverstone, Humbergate, Jamestown, Mount Olive, Beaumond Heights, Thistletown, Albion Gardens
Neighborhood

School Ratings by Clusters in Scarborough



```
Out[46]: <matplotlib.axes._subplots.AxesSubplot at 0x7f606cd7bb10>
```

The Location : Scarborough is a popular destination for new immigrants in Canada to reside. As a result, it is one of 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 past few 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.

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

Discussion Section

Problem Which Tried to Solve

The main purpose of this project is to solve the following purpose -

- Suggest a better Neighborhood in a new city for the person who are shifting there.
- Social presence in society in terms of like minded person .
- Connectivity to the airport, bus stand, city centre, markets and other daily needs things nearby.

And

- Sorted list of house in terms of housing prices in a ascending or descending order
- Sorted list of schools in terms of location, fees, rating and reviews

Conclusion Section

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 Neighborhood around them. Using the charts above results presented to a particular Neighborhood based on average house prices and school ratings have been made.

I feel rewarded with the efforts and believe this course with all the topics covered is well worth of appreciation. This project has shown me a practical application to resolve a real situation that has impacting 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.

Future Works

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 live a better life) around and also in terms of cost effective.

Libraries which are used to develop the Project

- **Pandas :** For creating and manipulating data frames .
- **Folium :** Python visualization library would be used to visualize the Neighborhood cluster distribution of using interactive leaflet map .
- **Scikit Learn** : For importing k-means clustering .
- **JSON** : Library to handle JSON files .
- **XML :** To separate data from presentation and XML stores data in plain text format.
- **Geocoder :** To retrieve Location Data .
- **Beautiful Soup and Request :** To scrap and library to handle http requests .
- **Matplotlib :** Python Plotting Module.