CAPSTONE PROJECT

Battle of Neighbourhoods

Finding a better place in Scarborough, Toronto

Muskan

11 June 2020

1. INTODUCTION

1.1 Background

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

This Capstone Project aim to create an analysis of features for a people migrating to Scarborough to search a best neighbourhood as a comparative analysis between neighbourhoods. 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

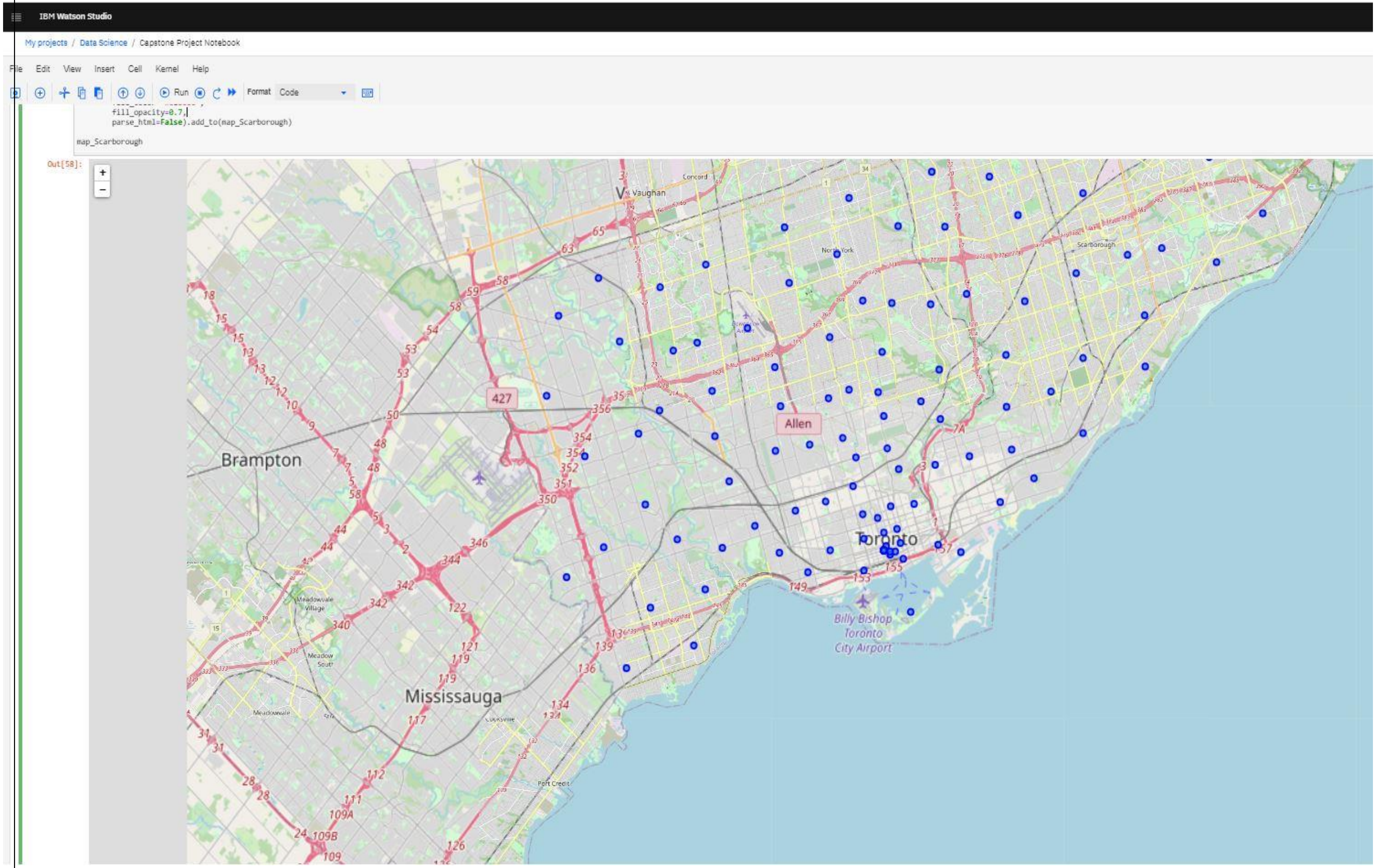
1.2 Problem

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 neighbourhoods. For ease of accessing to Cafe, School, Super market, medical shops, grocery shops, mall, theatre, hospital, like-minded people, etc.

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

1.3 Interests:

New people to the city would get awareness of the area and neighbourhood before moving to a new city, state, country or place for their work or to start a new fresh life.



```
        fill_opacity=0.7,
        parse_html=False).add_to(map_Scarborough)

map_Scarborough
```

2. Data acquisition and cleaning

2.1 Data sources

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

Used Scarborough dataset which earlier scrapped from Wikipedia. Dataset consisting of latitude and longitude, zip codes.

Foursquare API Data:

We will need data about different venues in different neighbourhoods 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, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

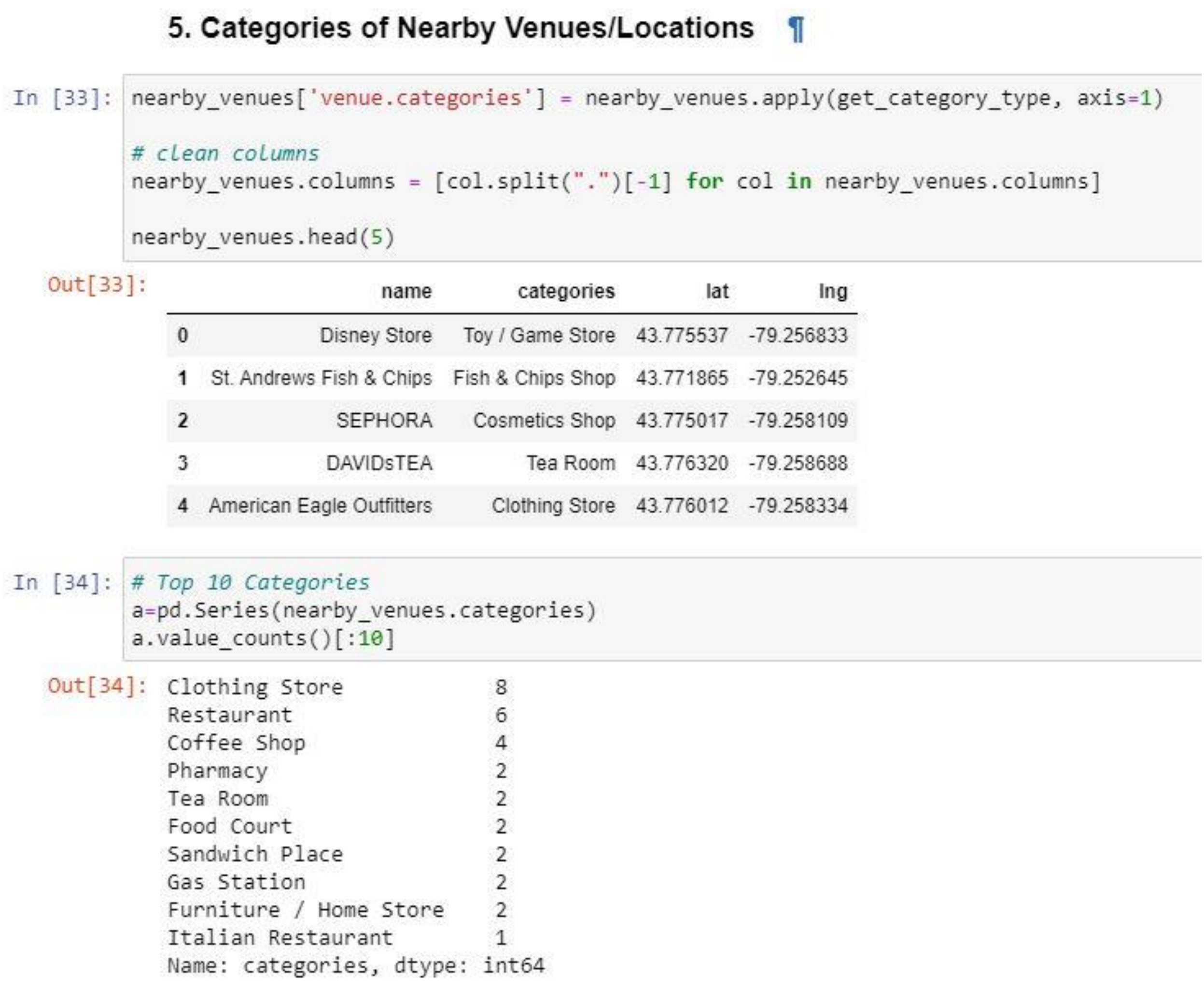
After finding the list of neighbourhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighbourhood. For each neighbourhood, we have chosen the radius to be 100 meter.

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. Neighbourhood
2. Neighbourhood Latitude
3. Neighbourhood 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

2.2 Data cleaning

Data downloaded or scraped from sources was combined into one table. There were a lot of missing values from earlier seasons, because of lack of record keeping. The cleaned tables are as follows:



5. Categories of Nearby Venues/Locations ¶

```
In [33]: nearby_venues['venue.categories'] = nearby_venues.apply(get_category_type, axis=1)

         # clean columns
         nearby_venues.columns = [col.split(".")[-1] for col in nearby_venues.columns]

         nearby_venues.head(5)
```

Out[33]:

	name	categories	lat	lng
0	Disney Store	Toy / Game Store	43.775537	-79.256833
1	St. Andrews Fish & Chips	Fish & Chips Shop	43.771865	-79.252645
2	SEPHORA	Cosmetics Shop	43.775017	-79.258109
3	DAVIDsTEA	Tea Room	43.776320	-79.258688
4	American Eagle Outfitters	Clothing Store	43.776012	-79.258334

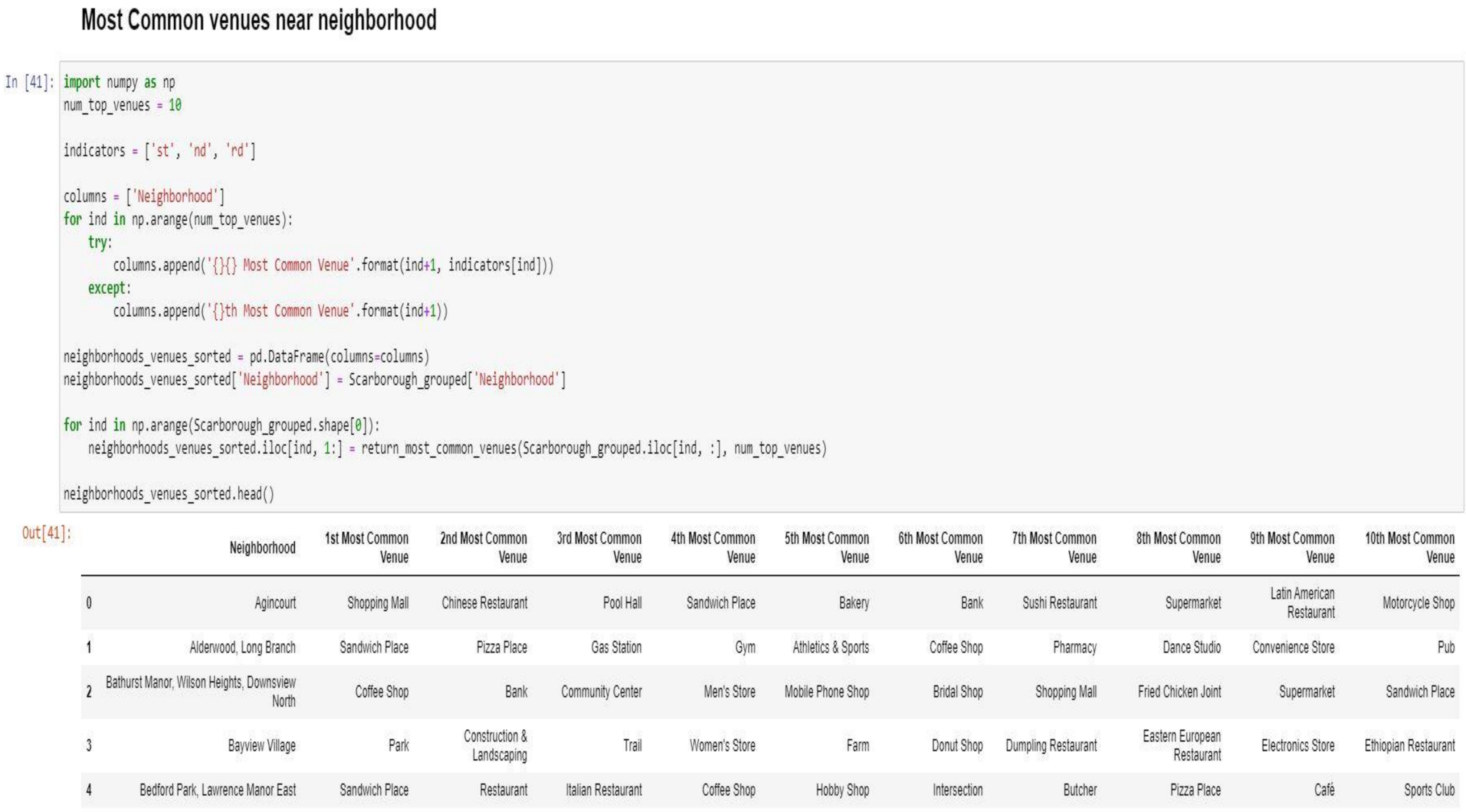
```
In [34]: # Top 10 Categories
         a=pd.Series(nearby_venues.categories)
         a.value_counts()[:10]
```

```
Out[34]: Clothing Store            8
         Restaurant                6
         Coffee Shop               4
         Pharmacy                  2
         Tea Room                  2
         Food Court                2
         Sandwich Place            2
         Gas Station               2
         Furniture / Home Store    2
         Italian Restaurant        1
         Name: categories, dtype: int64
```

Table 1 [33] shows latitude and longitudinal values of all nearby venues with its category

Table 2 [34] shows top 10 most visited venues and its frequency alongside.

Table 3 [41] shows top most visited venues of each neighbourhoods. (Next page)



Most Common venues near neighborhood

In [41]:

```
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[41]:

	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	Shopping Mall	Chinese Restaurant	Pool Hall	Sandwich Place	Bakery	Bank	Sushi Restaurant	Supermarket	Latin American Restaurant	Motorcycle Shop
1	Alderwood, Long Branch	Sandwich Place	Pizza Place	Gas Station	Gym	Athletics & Sports	Coffee Shop	Pharmacy	Dance Studio	Convenience Store	Pub
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Bank	Community Center	Men's Store	Mobile Phone Shop	Bridal Shop	Shopping Mall	Fried Chicken Joint	Supermarket	Sandwich Place
3	Bayview Village	Park	Construction & Landscaping	Trail	Women's Store	Farm	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Ethiopian Restaurant
4	Bedford Park, Lawrence Manor East	Sandwich Place	Restaurant	Italian Restaurant	Coffee Shop	Hobby Shop	Intersection	Butcher	Pizza Place	Café	Sports Club

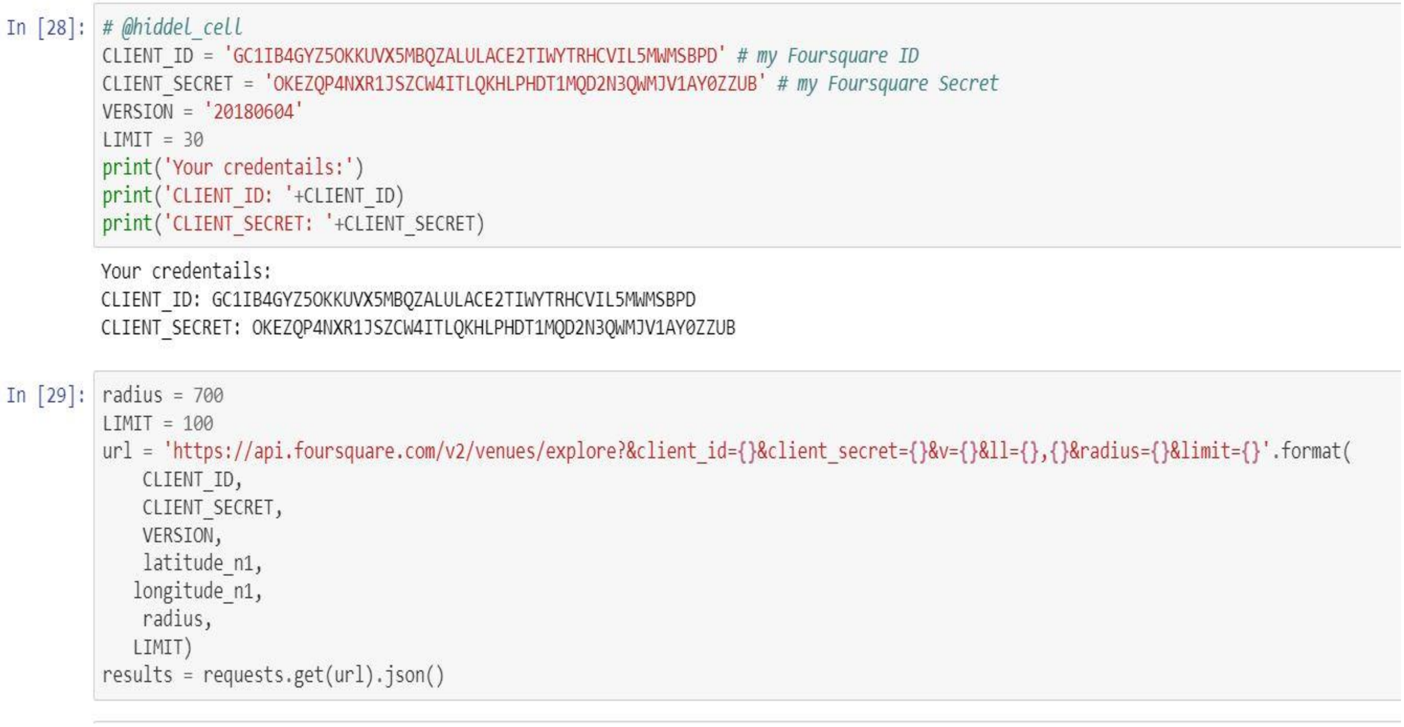
3. METHODOLOGY SECTION

3.1 Clustering Approach:

To compare the similarities of two cities, we decided 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.

3.2 Using K-means Clustering Approach

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.

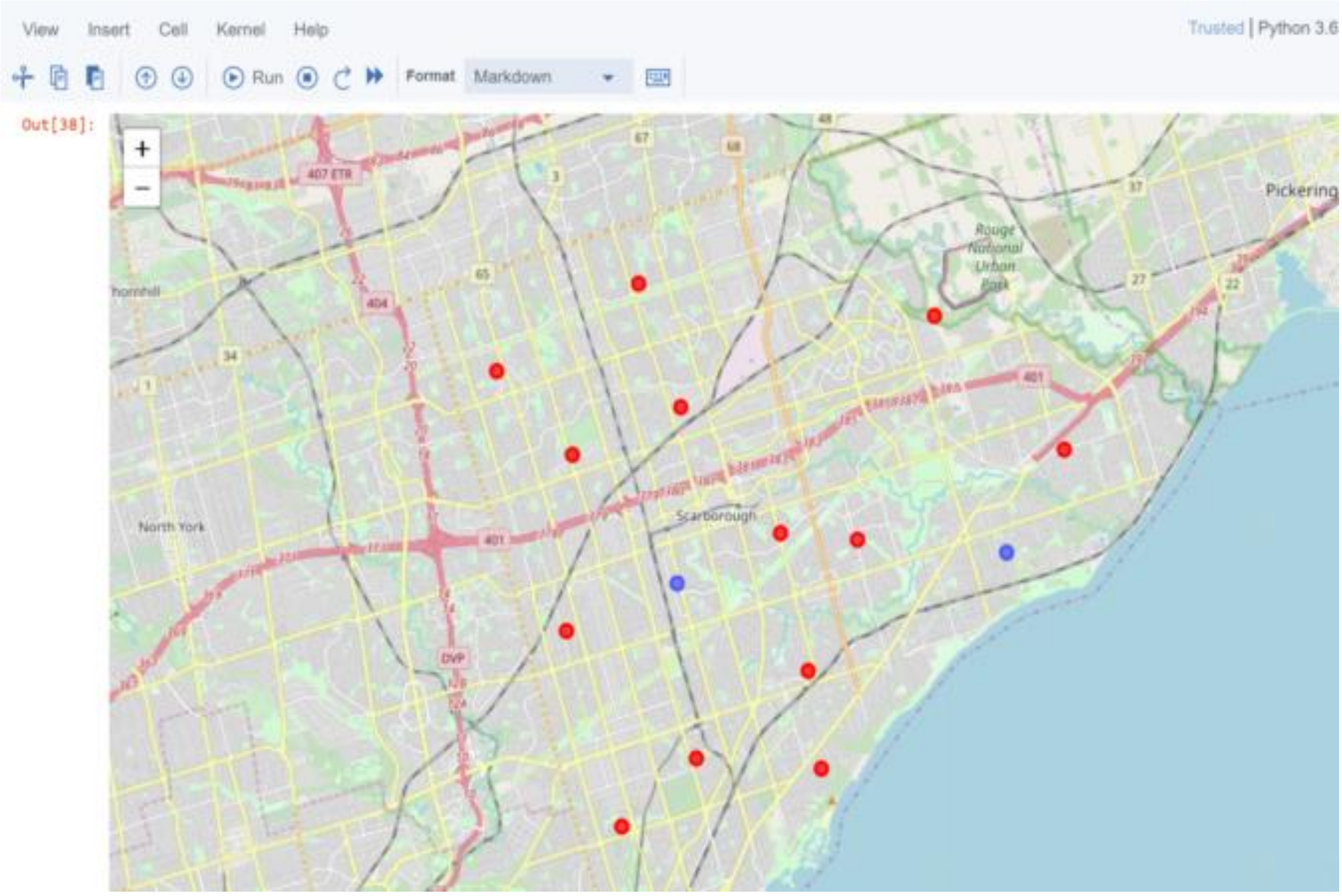


```
In [28]: # @hiddel_cell
         CLIENT_ID = 'GC1IB4GYZ5OKKUVX5MBQZALULACE2TIWYTRHCVIL5MWMSBPD' # my Foursquare ID
         CLIENT_SECRET = 'OKEZQP4NXR1JSZCW4ITLQKHLPHDT1MQD2N3QWMJV1AY0ZZUB' # my Foursquare Secret
         VERSION = '20180604'
         LIMIT = 30
         print('Your credentails:')
         print('CLIENT_ID: '+CLIENT_ID)
         print('CLIENT_SECRET: '+CLIENT_SECRET)

         Your credentails:
         CLIENT_ID: GC1IB4GYZ5OKKUVX5MBQZALULACE2TIWYTRHCVIL5MWMSBPD
         CLIENT_SECRET: OKEZQP4NXR1JSZCW4ITLQKHLPHDT1MQD2N3QWMJV1AY0ZZUB

In [29]: radius = 700
         LIMIT = 100
         url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
             CLIENT_ID,
             CLIENT_SECRET,
             VERSION,
             latitude_n1,
            longitude_n1,
             radius,
            LIMIT)
         results = requests.get(url).json()
```

Mapping Clusters:



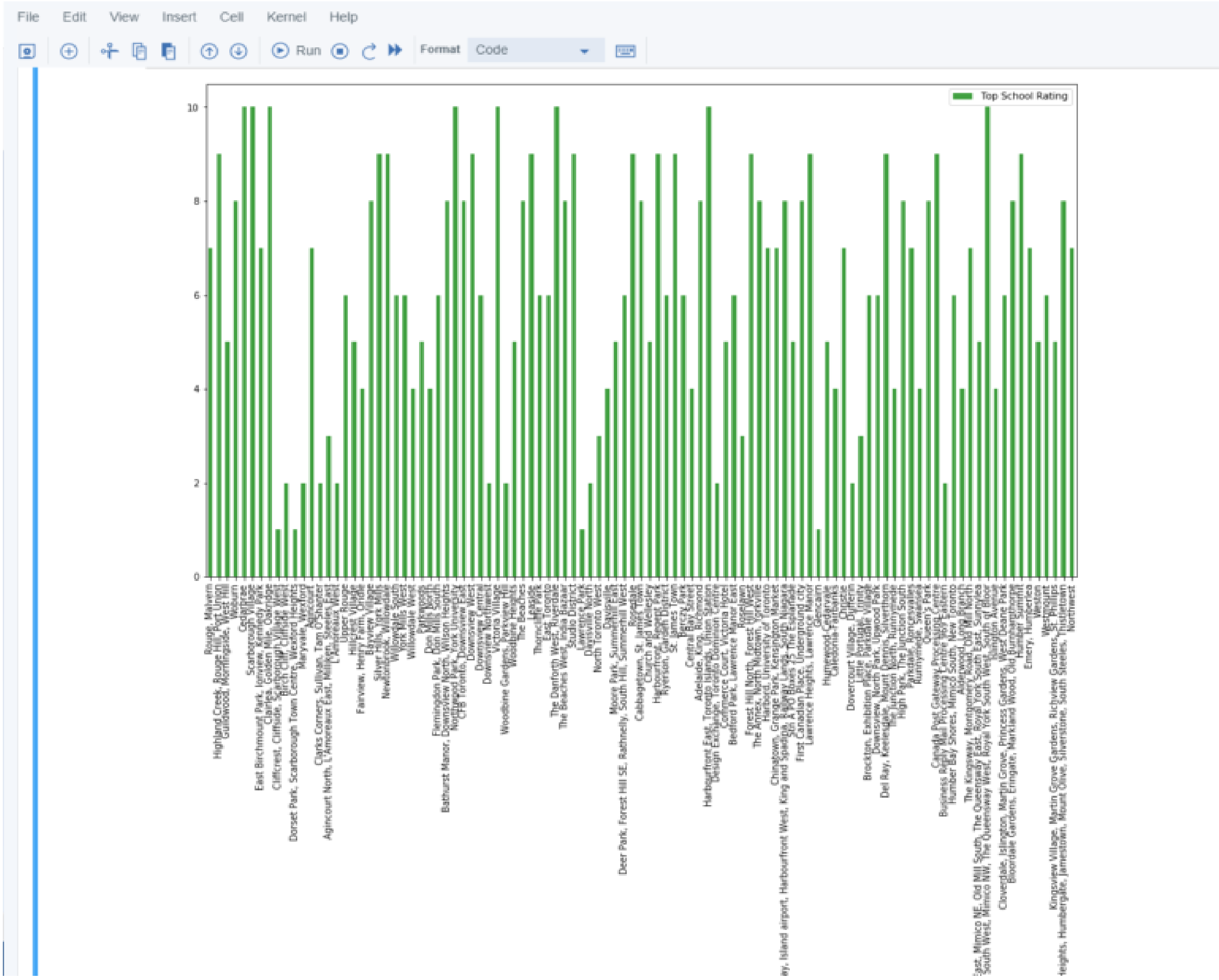
4. RESULTS SECTION

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 Capstone 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.

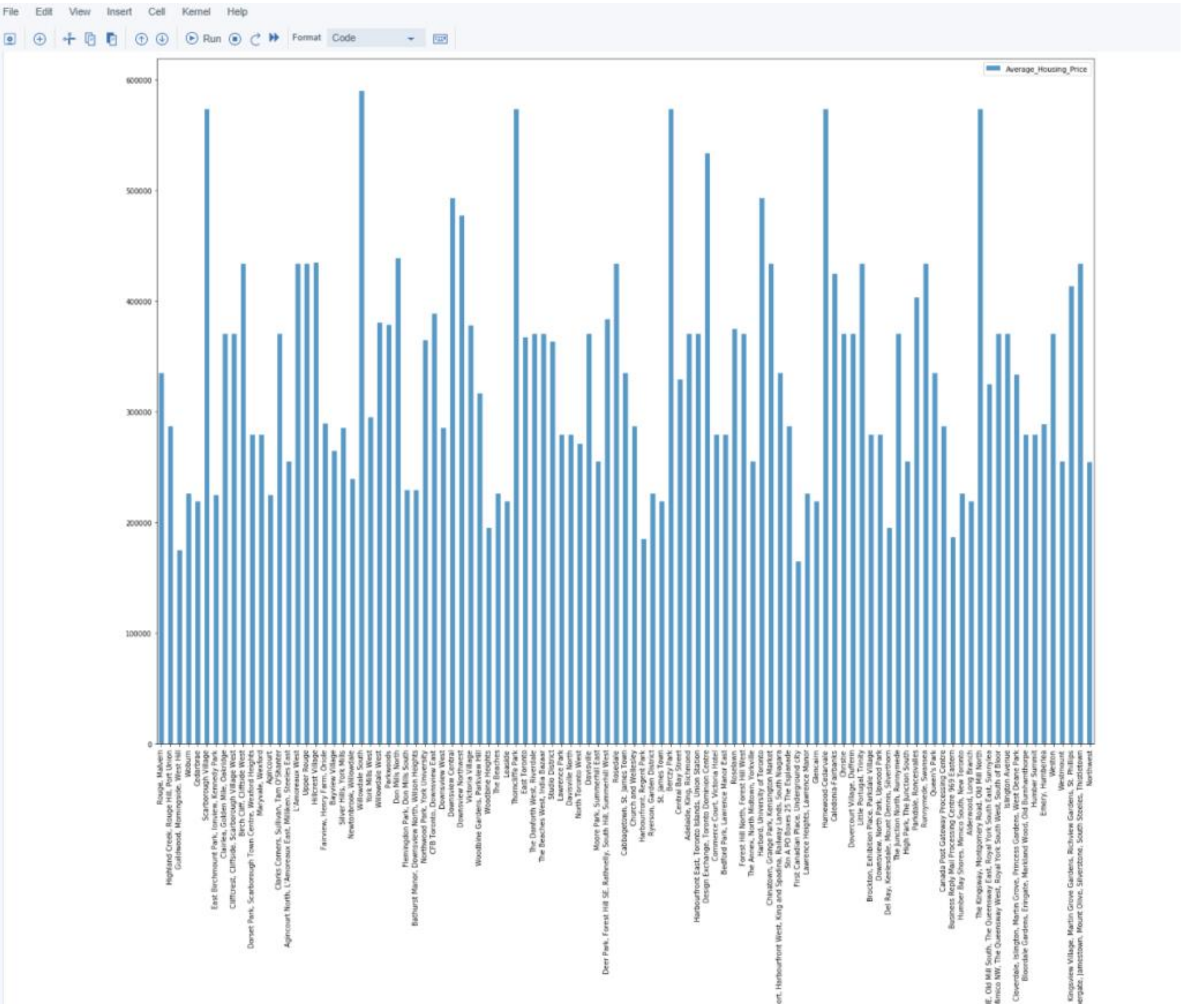
Following is the ‘Top School rating Area-wise’ bar graph,

New people who are parents, would find it helpful in analysing it their comfortabilities:


File
Edit
View
Insert
Cell
Kernel
Help
Run
Format
Code
Top School Rating
0
2
4
6
8
10
Rouge, Malvern
Highland Creek, Rouge Hill, Port Union
Guildwood, Morningside, West Hill
Woburn
Cedarbrae
Scarborough Village
East Birchmount Park, Ionview, Kennedy Park
Clairlea, Golden Mile, Oakridge
Cliffcrest, Cliffside, Scarborough Village West
Birch Cliff, Cliffside West
Dorset Park, Scarborough Town Centre, Wexford Heights
Maryvale, Wexford
Agincourt
Clarks Corners, Sullivan, Tam O'Shanter
Agincourt North, L'Amoreaux East, Milliken, Steeles East
L'Amoreaux West
Upper Rouge
Hillcrest Village
Fairview, Henry Farm, Oriole
Bayview Village
Silver Hills, York Mills
Newtonbrook, Willowdale
Willowdale South
York Mills West
Willowdale West
Parkwoods
Don Mills North
Flemingdon Park, Don Mills South
Bathurst Manor, Downsview North, Wilson Heights
Northwood Park, York University
CFB Toronto, Downsview East
Downsview West
Downsview Central
Downsview Northwest
Victoria Village
Woodbine Gardens, Parkview Hill
Woodbine Heights
The Beaches
Leaside
Thorncliffe Park
East Toronto
The Danforth West, Riverdale
The Beaches West, India Bazaar
Studio District
Lawrence Park
Davisville North
North Toronto West
Davisville
Moore Park, Summerhill East
Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West
Rosedale
Cabbagetown, St. James Town
Church and Wellesley
Harbourfront, Regent Park
Ryerson, Garden District
St. James Town
Berczy Park
Central Bay Street
Adelaide, King, Richmond
Harbourfront East, Toronto Islands, Union Station
Design Exchange, Toronto Dominion Centre
Commerce Court, Victoria Hotel
Bedford Park, Lawrence Manor East
Roselawn
Forest Hill North, Forest Hill West
The Annex, North Midtown, Yorkville
Harbord, University of Toronto
Chinatown, Grange Park, Kensington Market
ay, Island airport, Harbourfront West, King and Spadina, Railway Lands, South Niagara
Stn A PO Boxes 25 The Esplanade
First Canadian Place, Underground city
Lawrence Heights, Lawrence Manor
Glencairn
Humewood-Cedarvale
Caledonia-Fairbanks
Christie
Dovercourt Village, Dufferin
Little Portugal, Trinity
Brockton, Exhibition Place, Parkdale Village
Downsview, North Park, Upwood Park
Del Ray, Keelesdale, Mount Dennis, Silverthorn
The Junction North, Runnymede
High Park, The Junction South
Parkdale, Roncesvalles
Runnymede, Swansea
Queen's Park
Canada Post Gateway Processing Centre
Business Reply Mail Processing Centre 969 Eastern
Humber Bay Shores, Mimico South, New Toronto
Alderwood, Long Branch
The Kingsway, Montgomery Road, Old Mill North
ast, Mimico NE, Old Mill South, The Queensway East, Royal York South East, Sunnylea
South West, Mimico NW, The Queensway West, Royal York South West, South of Bloor
Islington Avenue
Cloverdale, Islington, Martin Grove, Princess Gardens, West Deane Park
Bloordale Gardens, Eringate, Markland Wood, Old Burnhamthorpe
Humber Summit
Emery, Humberlea
Weston
Westmount
Kingsview Village, Martin Grove Gardens, Richview Gardens, St. Phillips
eights, Humbergate, Jamestown, Mount Olive, Silverstone, South Steeles, Thistletown
Northwest

Following is the ‘Average Housing price Area-wise’ bar graph,

New people would find it helpful in analysing it their comfortabilities:



5. DISCUSSION SECTION

Problem Which Tried to Solve: The major purpose of this project, is to suggest a better neighbourhood in a new city for the person who are shifting there. Social presence in society in terms of like-minded people. Connectivity to the airport, bus stand, city centre, markets and other daily needs things nearby.

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

6. CONCLUSION SECTION

In this Capstone project, using k-means cluster algorithm I separated the neighbourhood into 10(Ten) different clusters and for 103 different latitude and longitude from dataset, which have very-similar neighbourhoods around them. Using the charts above results presented to a particular neighbourhood 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 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 Capstone 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 Developed the Project: Pandas: For creating and manipulating data frames.

- *Folium: Python visualization library would be used to visualize the neighbourhoods 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 Requests: To scrap and library to handle http requests.*
- *Matplotlib: Python Plotting Module.*