Capstone Project – Neighborhoods- Toronto

SIDDESH KUMAR IYER

APPLIED DATA SCIENCE CAPSTONE FINAL ASSIGNMENT

INTRODUCTION

SCARBOROUGH, TORONTO

INTRO- PURPOSE OF THIS PROJECT

- Purpose – Help people explore different neighbourhoods around them in an efficient manner out of a number of places in Scarborough, Toronto
- Ease of access to Cafes, Schools, Super markets, pharmacies, grocery shops, malls, theatres, hospitals, etc.
- This Capstone Project aims 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 life

DATA

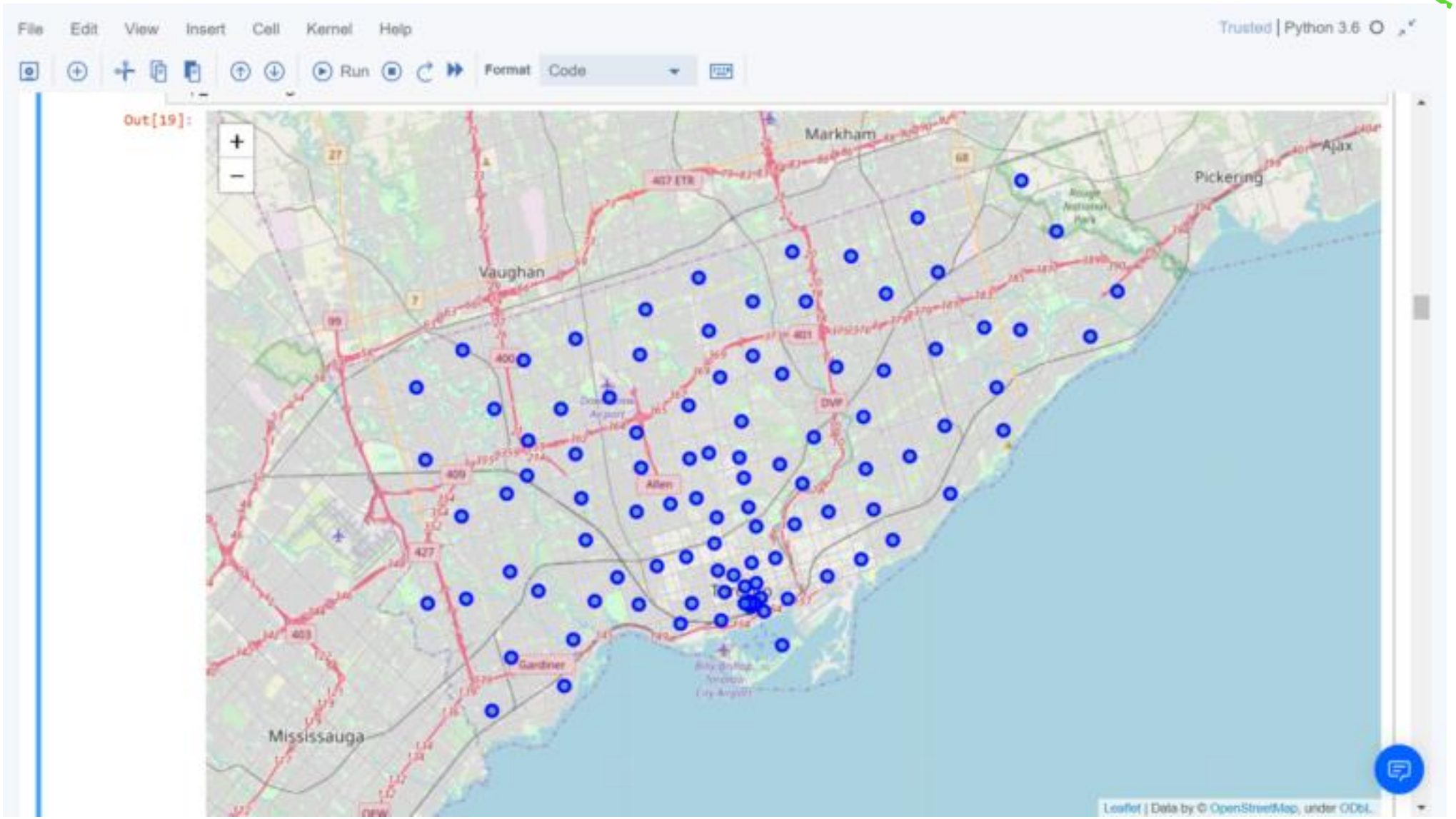
SCARBOROUGH, TORONTO

DATA

- Data Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- The data used will be the one that we scraped from Wikipedia during the Week 3 assignment
- Data includes latitudes, longitudes and zip codes
- Foursquare API shall be used. It is a powerful tool that provides us location data with information about all various venues and events within an area of interest
- The information obtained per venue is as follows:

    - Neighborhood
    - Neighborhood Latitude
    - Neighborhood Longitude
    - Venue
    - Name of the venue e.g. the name of a store or restaurant
    - Venue Latitude
    - Venue Longitude
    - Venue Category

MAP OF SCARBOROUGH



Methodology

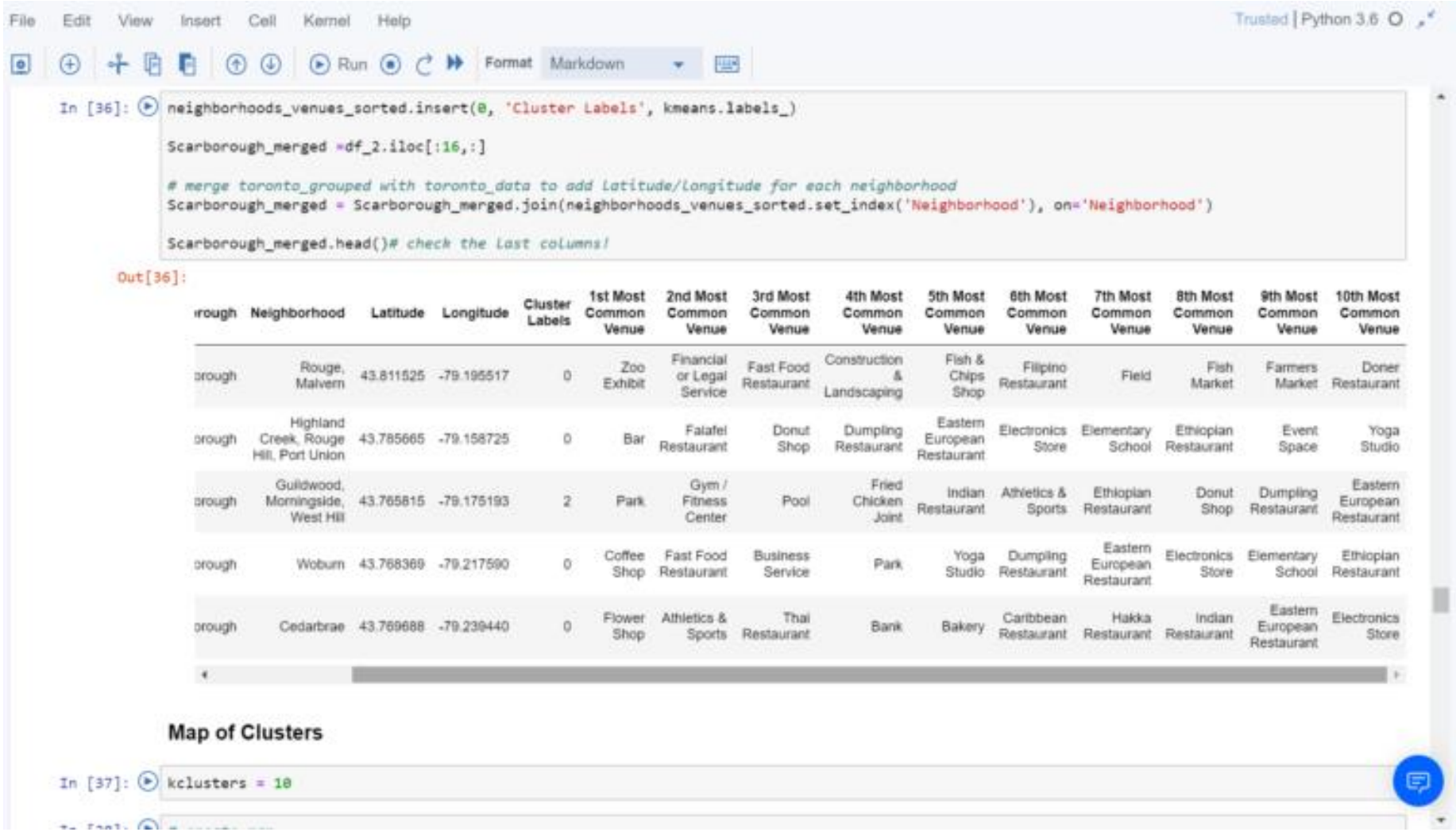
CLUSTERING APPROACH – K MEANS

SCARBOROUGH, TORONTO

CLUSTERING APPROACH

- To compare the similarities of two cities, we decided to explore neighborhoods, and cluster them to find similar neighborhoods in a big city like Toronto
- In order to achieve this, we deployed a form of unsupervised machine learning called the K-Means clustering algorithm

K-MEANS CLUSTERING – MOST COMMON VENUE



File Edit View Insert Cell Kernel Help

Trusted | Python 3.6

Run Format Markdown

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]:

rough	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
orough	Rouge, Malvern	43.811525	-79.195517	0	Zoo Exhibit	Financial or Legal Service	Fast Food Restaurant	Construction & Landscaping	Fish & Chips Shop	Filipino Restaurant	Field	Fish Market	Farmers Market	Doner Restaurant
orough	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	Ethiopian Restaurant	Event Space	Yoga Studio
orough	Guildwood, Morningside, West Hill	43.765815	-79.175193	2	Park	Gym / Fitness Center	Pool	Fried Chicken Joint	Indian Restaurant	Athletics & Sports	Ethiopian Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant
orough	Woburn	43.768369	-79.217590	0	Coffee Shop	Fast Food Restaurant	Business Service	Park	Yoga Studio	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Elementary School	Ethiopian Restaurant
orough	Cedarbrae	43.769688	-79.239440	0	Flower Shop	Athletics & Sports	Thai Restaurant	Bank	Bakery	Caribbean Restaurant	Hakka Restaurant	Indian Restaurant	Eastern European Restaurant	Electronics Store

Map of Clusters

In [37]:
```
kclusters = 10
```

MOST COMMON VENUES NEAR NEIGHBORHOOD – CLUSTERING APPROACH

File Edit View Insert Cell Kernel Help

Trusted | Python 3.6

Run Format Markdown

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
1	Agincourt	Chinese Restaurant	Shopping Mall	Pizza Place	Supermarket	Sushi Restaurant	Breakfast Spot	Print Shop	Mediterranean Restaurant	Coffee Shop	Pool
2	Agincourt North, L'Amoreaux East, Milliken, St...	Pharmacy	Sandwich Place	Sushi Restaurant	Doner Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Elementary School	Ethiopian Restaurant
3	Albion Gardens, Beaumond Heights, Humbergate, ...	Grocery Store	Park	Sandwich Place	Discount Store	Japanese Restaurant	Fried Chicken Joint	Beer Store	Hardware Store	Pizza Place	Fast Food Restaurant
4	Alderwood, Long Branch	Convenience Store	Pub	Sandwich Place	Coffee Shop	Gas Station	Dance Studio	Gym	Pharmacy	Pizza Place	Falafel Restaurant

WORK FLOW – FOURSQUARE API

Step 1

Foursquare API credentials and mining of features of nearby neighbourhoods

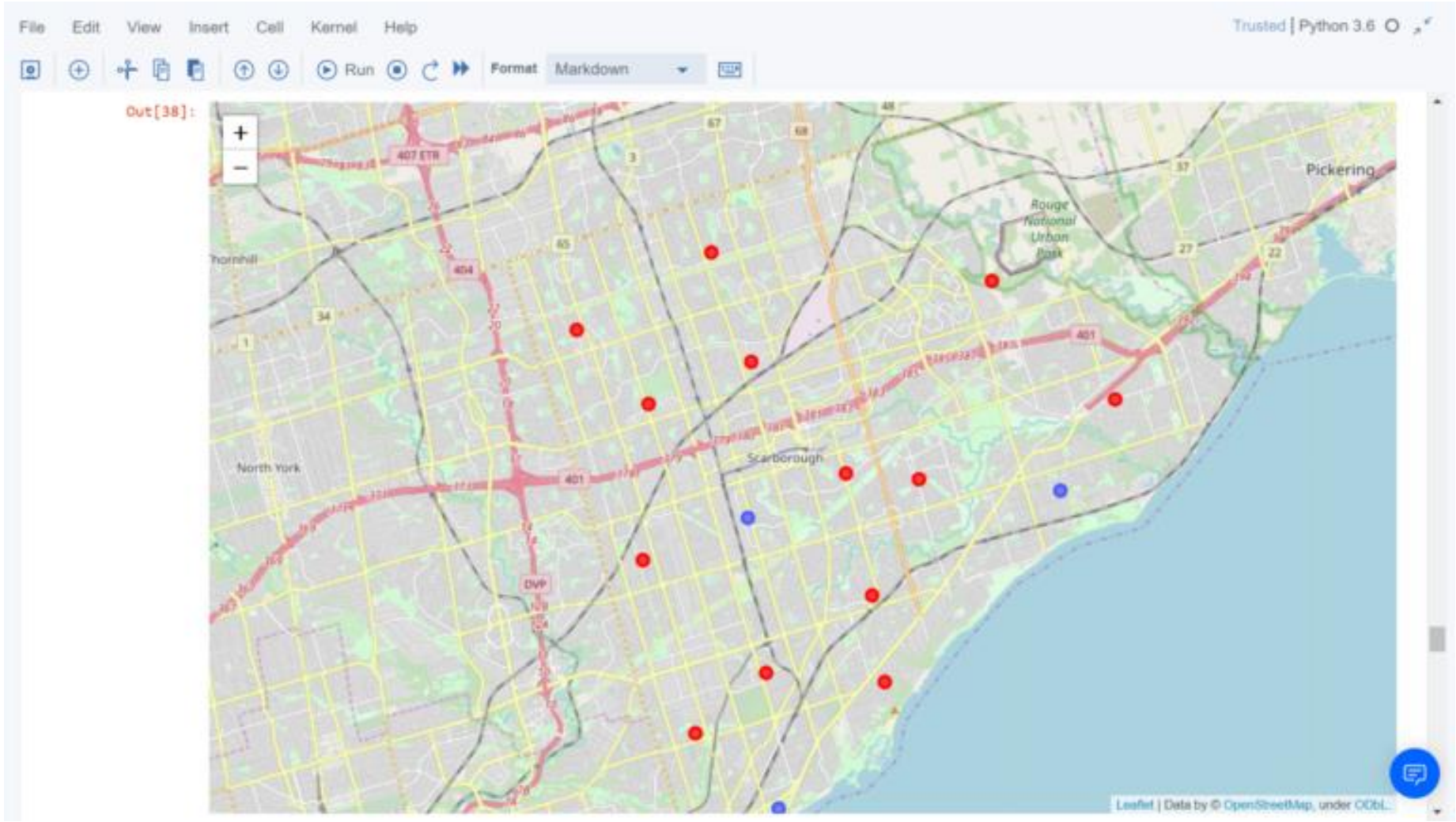
Step 2

HTTP request limitations would restrict us to set the number of places per neighbourhood to 100 & radius to 500

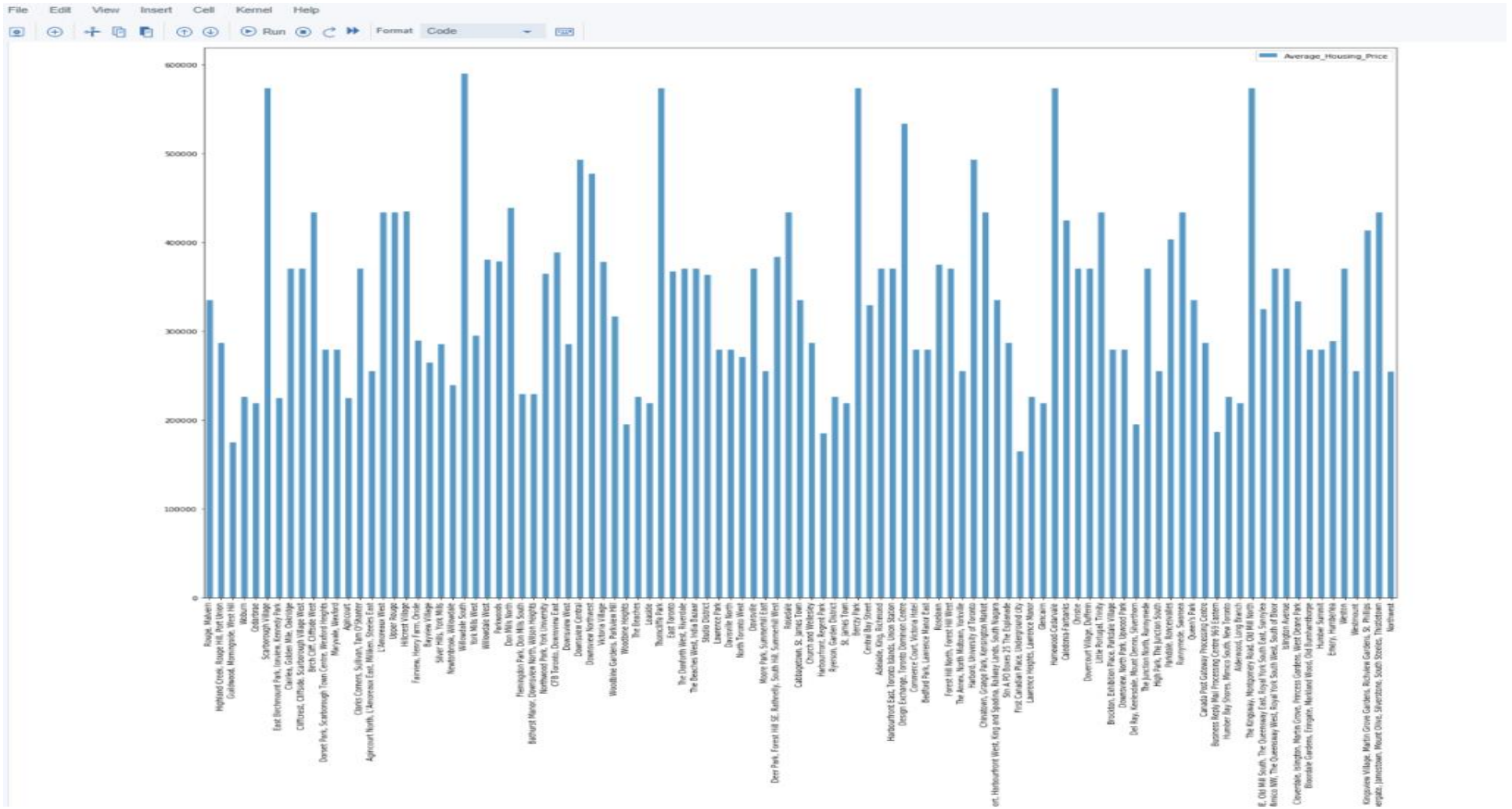
Results

SCARBOROUGH, TORONTO

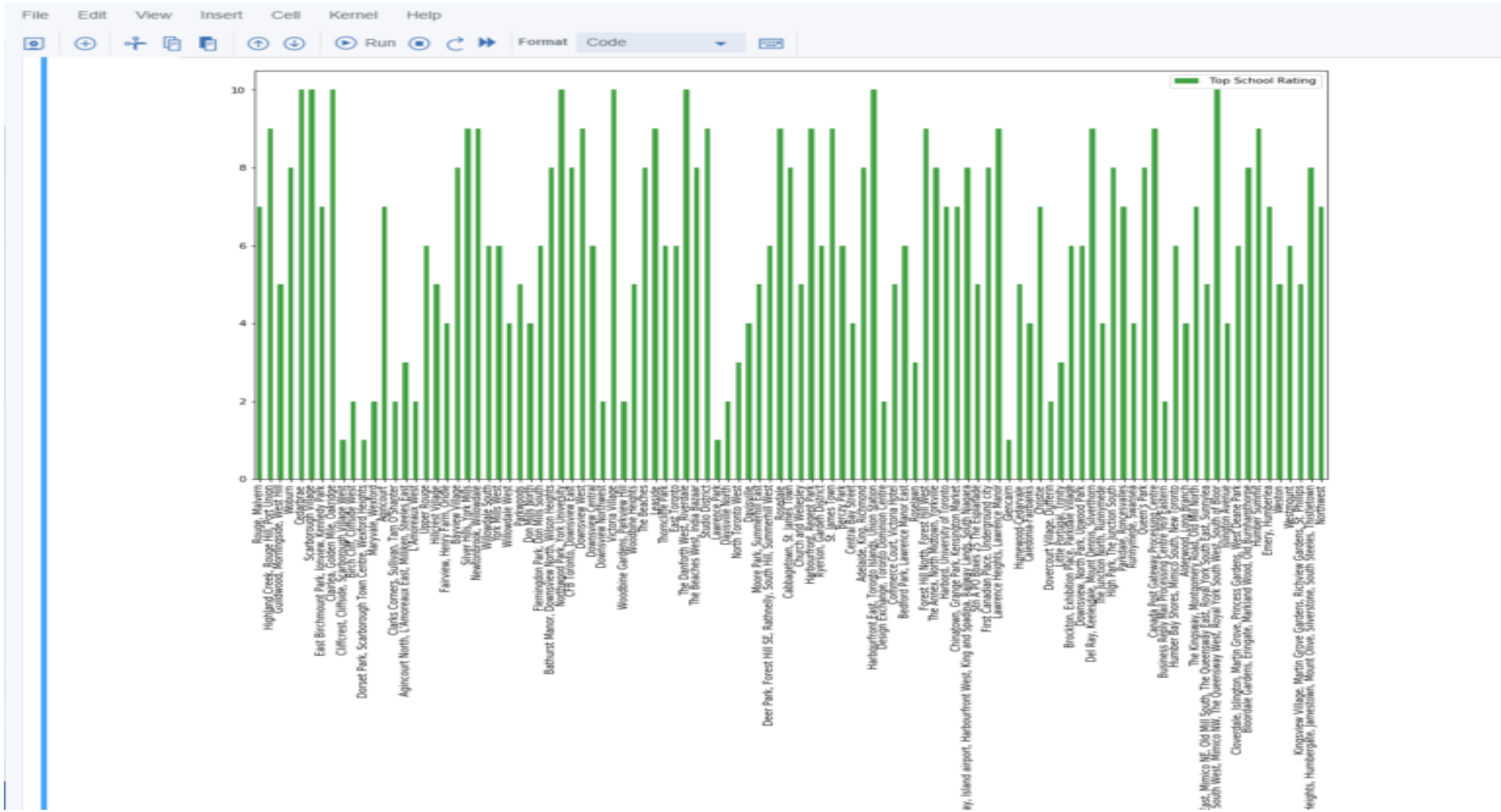
MAP OF CLUSTERS IN SCARBOROUGH



AVERAGE HOUSING PRICE BY CLUSTERS IN SCARBOROUGH



SCHOOL RATINGS BY CLUSTERS IN SCARBOROUGH



THE LOCATION

- Scarborough is a popular destination for immigrants to reside in Canada
- It offers a diverse and multicultural environment in the Greater Toronto Area
- Although restrictions on immigration are trending these days, the immigration trend into Canada has been on the rise in recent times

Discussion

SCARBOROUGH, TORONTO

PROBLEM WE TRIED TO ADDRESS

- The major purpose of this project, is to suggest a better neighborhood in a new city for people moving there. Social presence in society in terms of like minded people. Connectivity to the various transport modes, city center, markets and other daily needs things nearby.
- Two major pointers in this project:
  - 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

SCARBOROUGH, TORONTO

CONCLUSION - BRIEF

- In this project, using k-means cluster algorithm allowed us to separate the neighborhood into 10(Ten) different clusters and for 103 different latitude and longitude combinations from the dataset, which have very-similar neighborhoods around them.
- Using the charts above, results presented to a particular neighborhood are based on average house prices and school rating
- This project has given a practical application to resolve a real situation that has far-reaching impacts on personal and financial fronts 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 - BUILDING UP ON THIS PROJECT

- This project can be continued for making it more precise in terms of finding the best housing areas in Scarborough.
- It can be made more precise in terms of the daily needs of people and optimum economic impacts that it may have on people's budgets.

LIBRARIES USED TO DEVELOP 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.

Thank You