Final Report | Capstone Project – The Battle of Neighbourhoods Finding a Better Place in Guelph, Toronto

1. Introduction:

The purpose of this Project is to help people in exploring better facilities around their neighbourhood. It will help people make smart and efficient decisions on selecting a great neighbourhood out of numbers of other neighbourhoods in Guelph, Toronto.

Lots of people are migrating to various states of Canada and need 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 access to Cafe, School, Supermarket, medical shops, grocery shops, mall, theatre, hospital, like minded people, etc.

This Project aims to create an analysis of features for people migrating to Guelph 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 wastewater and excrement conveyed in sewers and recreational facilities.

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.

2. Data Section

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

Will use Guelph 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 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.

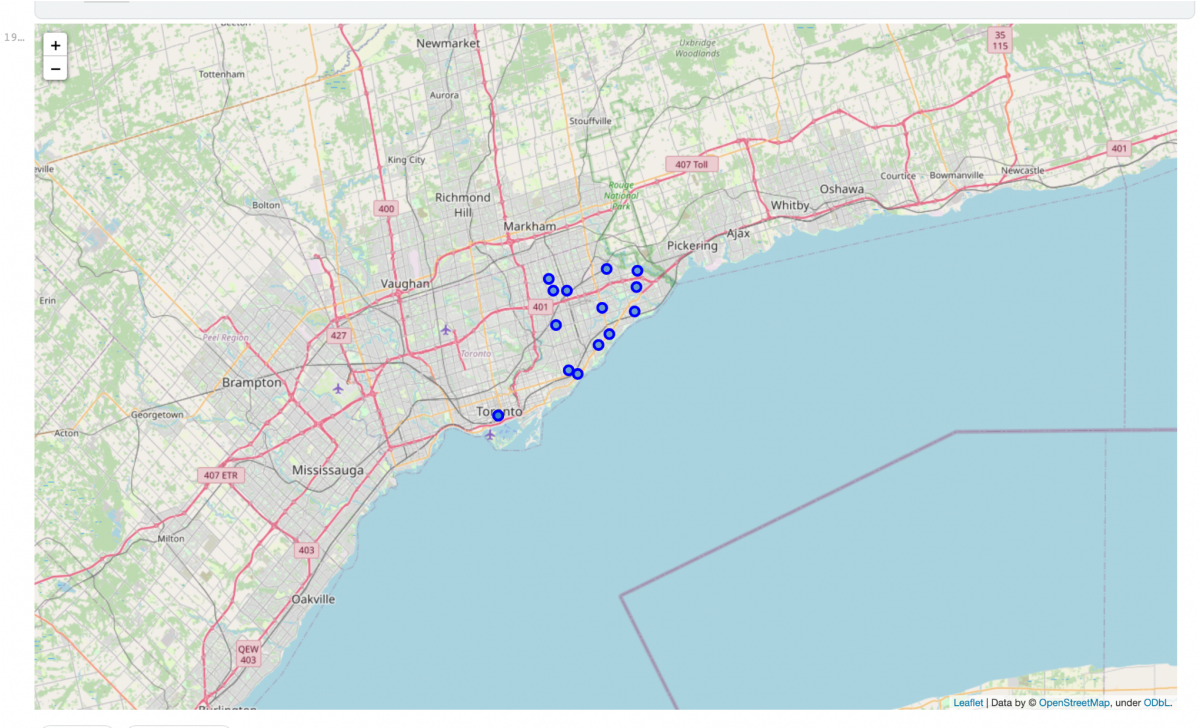
After finding the list of neighbourhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighbourhood. For each neighbourhood, we have chosen the radius to be 100 metres.

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 Guelph



3. Methodology Section

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.

Using K-Means Clustering Approach

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

Guelph_merged =df_2.iloc[:16,:]

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

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

[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	8th Most Common Venue	9th Most Common Venue	10t C
0	\nM1ANot assigned\n\n	\nM2ANot assigned\n\n	M9AEtobicoke(Islington Avenue)	43.64869	-79.38544	1.0	Coffee Shop	Hotel	Café	Beer Bar	American Restaurant	Restaurant	Gym	Japanese Restaurant	Concert Hall	
1	\nM1BScarborough(Malvern / Rouge)\n\n	\nM2BNot assigned\n\n	M9BEtobicoke(West Deane Park / Princess Garden...	43.81023	-79.22038	0.0	Fast Food Restaurant	Pizza Place	Park	Pharmacy	Sandwich Place	Bubble Tea Shop	Supermarket	Skating Rink	Grocery Store	
2	\nM1CScarborough(Rouge Hill / Port Union / Hig...	\nM2CNot assigned\n\n	M9CEtobicoke(Eringate / Bloordale Gardens / Ol...	43.78948	-79.17614	2.0	Fish & Chips Shop	Convenience Store	Yoga Studio	Pharmacy	Park	Office	New American Restaurant	Museum	Movie Theater	Mo La
3	\nM1EScarborough(Guildwood / Morningside / Wes...	\nM2ENot assigned\n\n	M9ENot assigned	43.76343	-79.17820	0.0	Park	Gym / Fitness Center	Restaurant	Convenience Store	Yoga Studio	Mediterranean Restaurant	Pharmacy	Office	New American Restaurant	
4	\nM1GScarborough(Woburn)\n\n	\nM2GNot assigned\n\n	M9GNot assigned	43.76748	-79.22829	1.0	Indian Restaurant	Park	Bank	Coffee Shop	Yoga Studio	Mexican Restaurant	Plaza	Pizza Place	Pharmacy	

Most Common venues near Neighborhood

[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'] = Guelph_grouped['Neighborhood']

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

neighborhoods_venues_sorted.head()
```

[34...

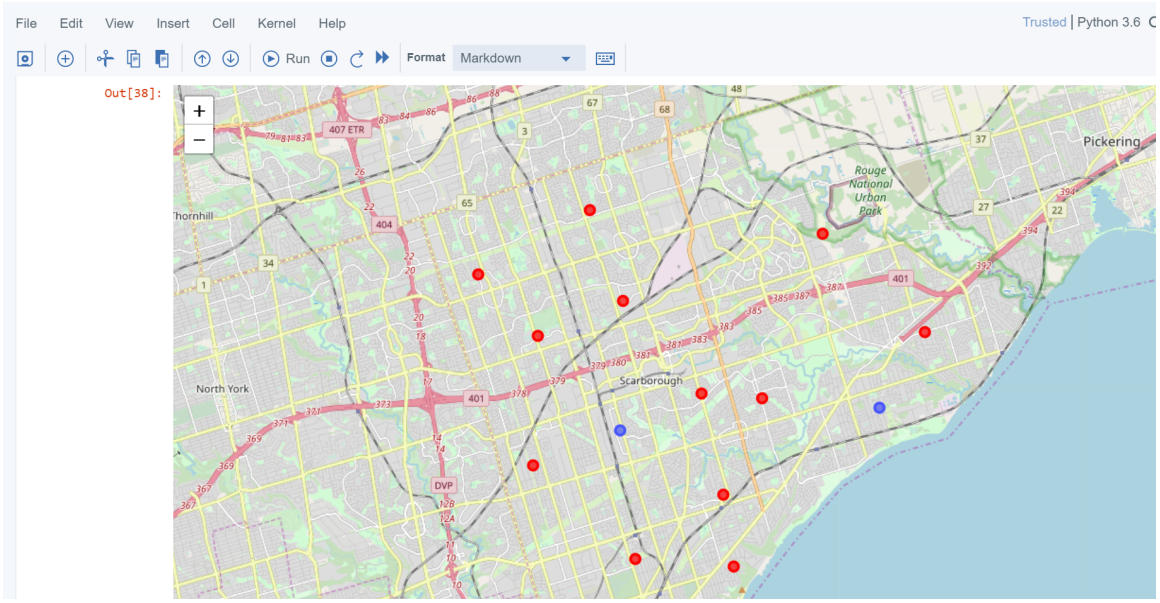
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	M9AEtobicoke(Islington Avenue)	Coffee Shop	Hotel	Café	Beer Bar	American Restaurant	Restaurant	Gym	Japanese Restaurant	Concert Hall	Pizza Place
1	M9BEtobicoke(West Deane Park / Princess Garden...	Fast Food Restaurant	Pizza Place	Park	Pharmacy	Sandwich Place	Bubble Tea Shop	Supermarket	Skating Rink	Grocery Store	Gym / Fitness Center
2	M9CEtobicoke(Eringate / Bloordale Gardens / Ol...	Fish & Chips Shop	Convenience Store	Yoga Studio	Pharmacy	Park	Office	New American Restaurant	Museum	Movie Theater	Monument / Landmark
3	M9ENot assigned	Park	Gym / Fitness Center	Restaurant	Convenience Store	Yoga Studio	Mediterranean Restaurant	Pharmacy	Office	New American Restaurant	Museum
4	M9GNot assigned	Indian Restaurant	Park	Bank	Coffee Shop	Yoga Studio	Mexican Restaurant	Plaza	Pizza Place	Pharmacy	Office

Workflow:

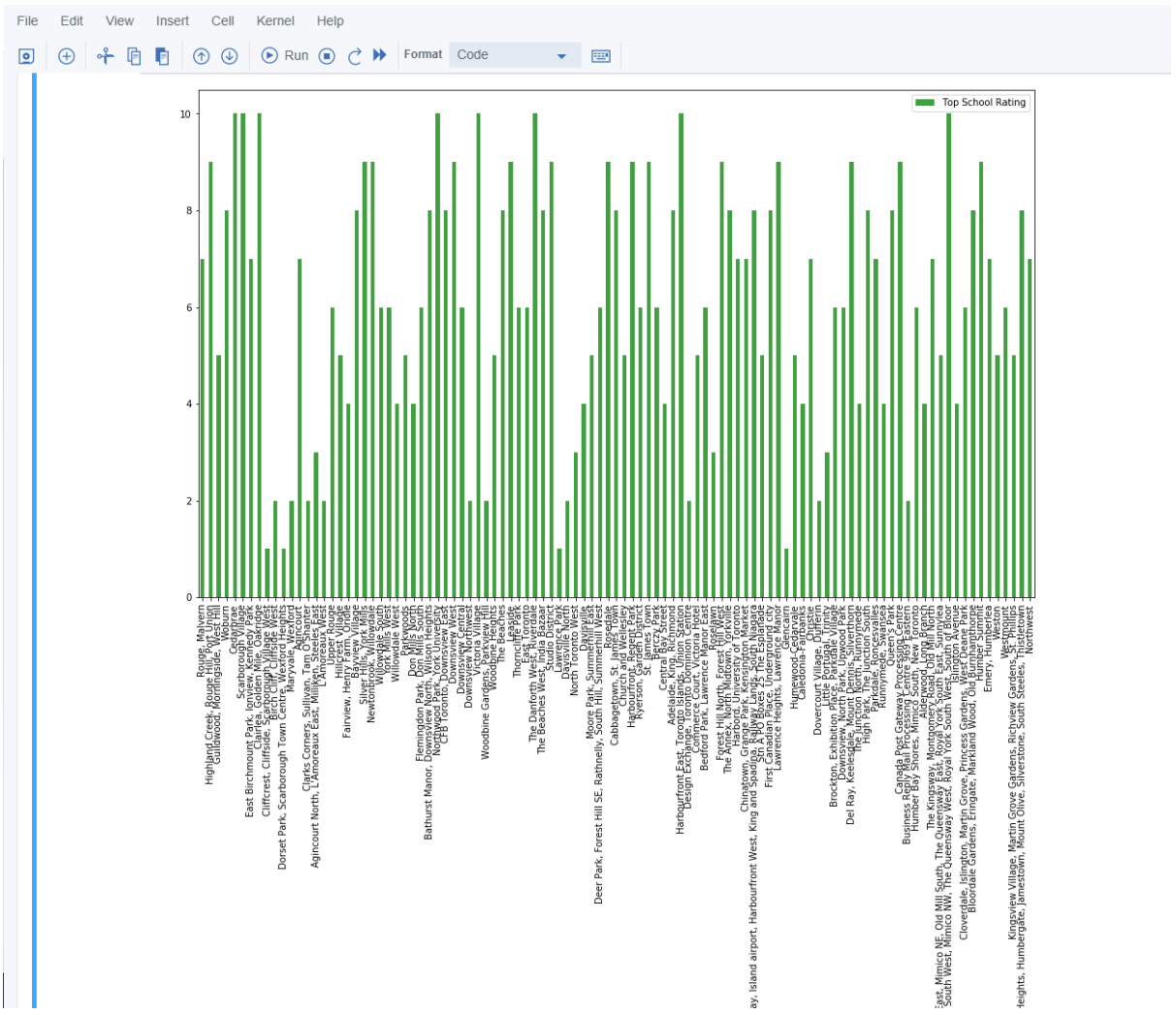
Using credentials of Foursquare API, features of near-by places of the neighbourhoods would be mined. Due to http request limitations the number of places per neighbourhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

4. Results Section

Map of Clusters in Guelph



School Ratings by Clusters in Guelph



The Location:
Guelph 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 has 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.

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 people 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.

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

6. Conclusion Section

In this project, using k-means cluster algorithm I separated the neighbourhood into 10(Ten) different clusters and for 103 different latitude and longitudes from the 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 project can be continued to make it more precise in terms of finding the best house in Guelph. 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 effectiveness.

Libraries Which are Used to Develop the Project:

Pandas: For creating and manipulating dataframes.

Folium: Python visualisation library would be used to visualise the neighbourhoods cluster distribution 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.

Blog Post Link:
https://www.kaggle.com/sarashahin/applied-data-science-capstone-finalproject/edit