IBM Data Science Capstone – Battle of the Neighbourhoods

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. The city is the anchor of the Golden Horseshoe, an urban agglomeration of 9,245,438 people (as of 2016) surrounding the western end of Lake Ontario, while the Greater Toronto Area (GTA) proper had a 2016 population of 6,417,516. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world.

Source: https://en.wikipedia.org/wiki/Toronto

Toronto is a major Canadian city along Lake Ontario's northwestern shore. It's a dynamic metropolis with a core of soaring skyscrapers, all dwarfed by the iconic, free-standing CN Tower. Toronto also has many green spaces, from the orderly oval of Queen's Park to 400-acre High Park and its trails, sports facilities and zoo.

This means there is a huge opportunity to do business here and the market it highly competitive because it is highly developed and densely populated area. Any new person or a company who are interested in doing business there should understand and analyse the market carefully before venturing.

The insights derived from the project will provide a good understanding of the market there which helps make strategic decision.

Problem Description:

A restaurant is a business which prepares and serves food and drink to customers in return for money, either paid before the meal, after the meal, or with an open account. The City of Toronto is famous for its excellent cuisine and India is one of them. In this project, I am assuming a hypothetical scenario where there might not be enough India restaurants around the Toronto Area. This project will provide good insights to business stake holders who might want to open a new Indian restaurant there.

Target Audience:

This would interest anyone who wants to open a new Indian restaurant in Toronto.

Data description:

Data is available here: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Will use Toronto dataset which we scrapped from Wikipedia will have major attributes like borough, latitude, longitude and zip codes.

Foursquare API:

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. 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
```

Libraries used:

Pandas: For creating and manipulating dataframes.

Folium: Python visualization library would be used to visualize the neighborhoods 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.

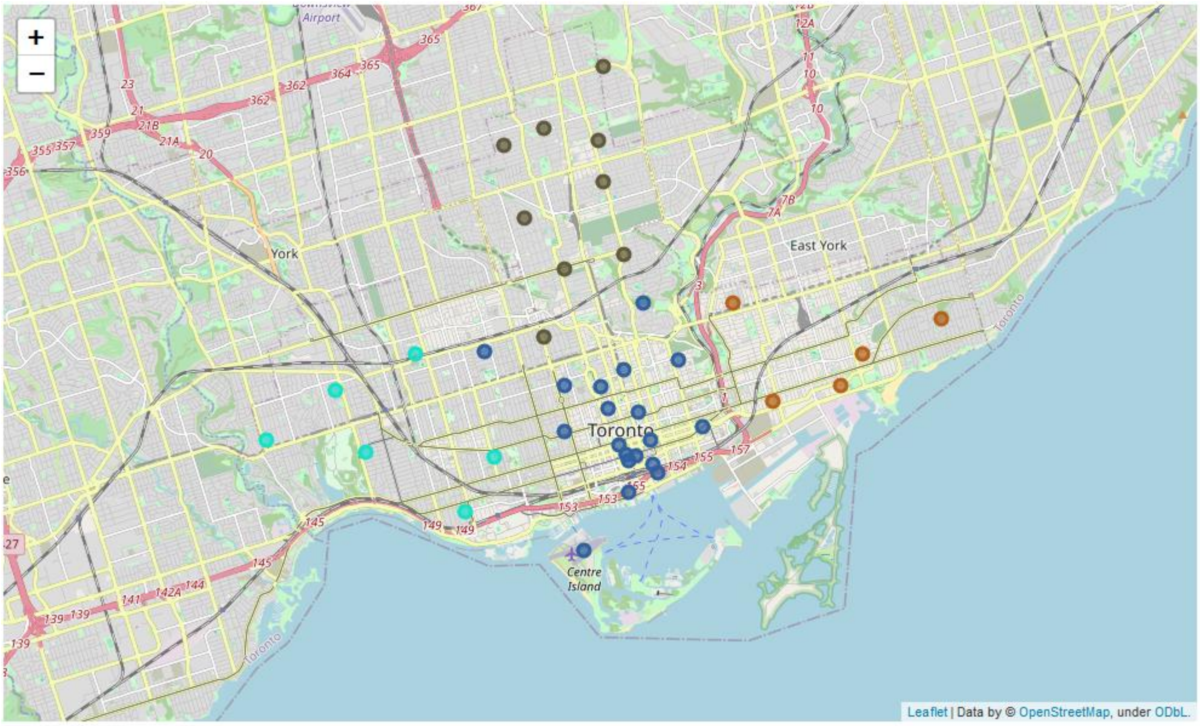
Geocoder: To retrieve Location Data.

Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.

Methodology:

After webscraping, transforming the data into a pandas dataframe and performing data prepossessing it resulted in a simple table containing neighbourhood names and postal codes. To get the coordinates Geocoder was used to get the location data. Once the coordinates are gathered a map of Toronto has been visualized with Folium.

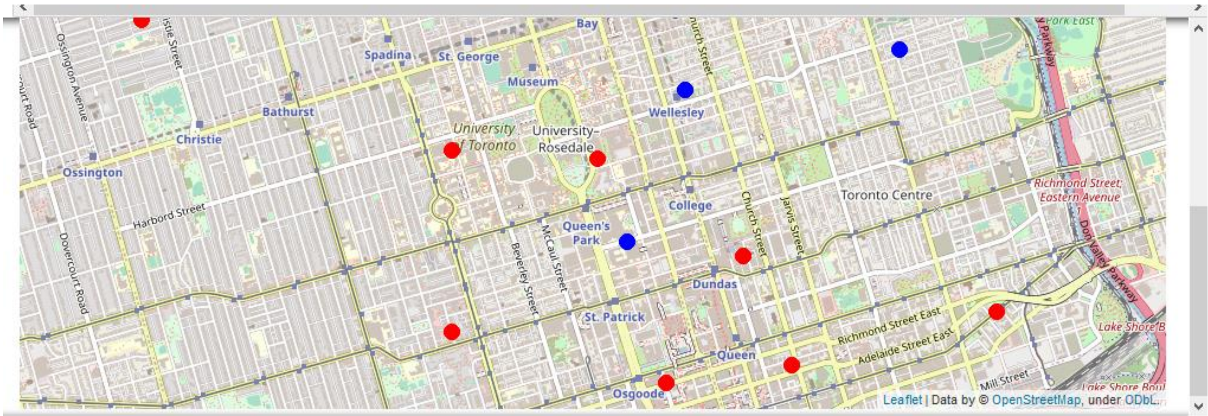


Foursquare API is used to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I was able to pull the names, categories, latitude, and longitude of the venues. With this data, one can also check how many unique categories that one can get from these venues.

Analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare for clustering which will be done later.

Clustering Approach:

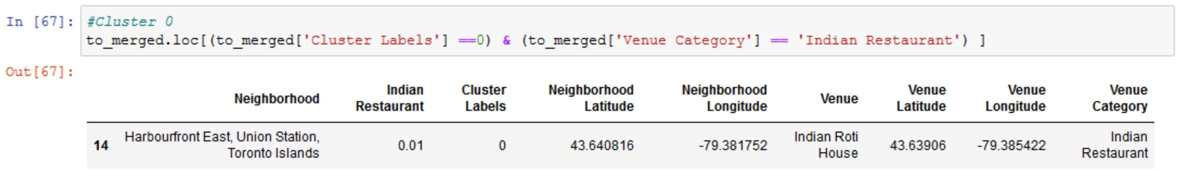
Clustering method by using k-means clustering has been performed on the data. The K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms. I have clustered the neighbourhoods in Toronto into 3 clusters based on their frequency of occurrence for “Indian food”. Based on the concentration of cluster, a recommendation of the ideal location to open the restaurant has been given.



Results:

The results from k-means clustering show that we can categorize Toronto neighbourhoods into 3 clusters based on how many Indian restaurants are in each neighbourhood:

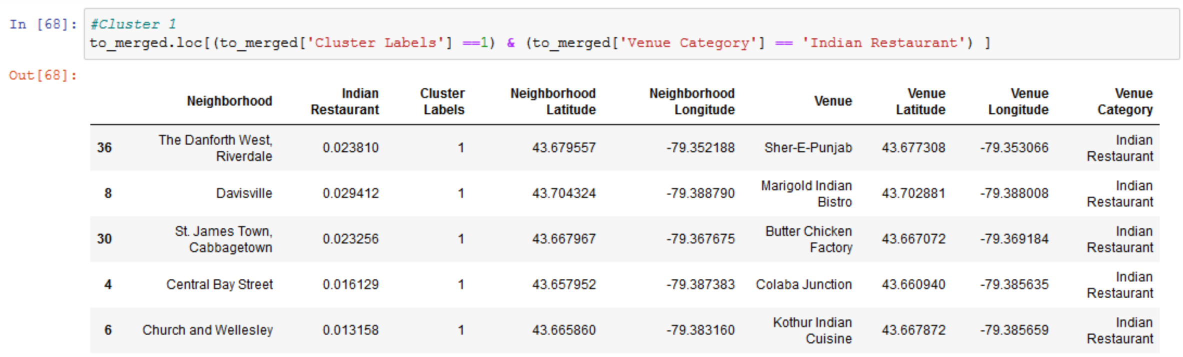
Cluster 1:



```
In [67]: #Cluster 0
         to_merged.loc[(to_merged['Cluster Labels'] ==0) & (to_merged['Venue Category'] == 'Indian Restaurant') ]
Out[67]:
```

	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
14	Harbourfront East, Union Station, Toronto Islands	0.01	0	43.640816	-79.381752	Indian Roti House	43.63906	-79.385422	Indian Restaurant

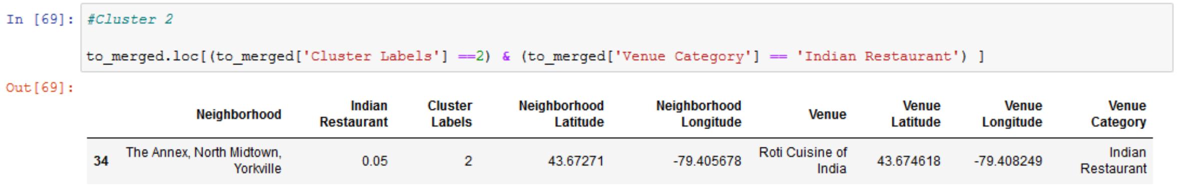
Cluster 2:



```
In [68]: #Cluster 1
         to_merged.loc[(to_merged['Cluster Labels'] ==1) & (to_merged['Venue Category'] == 'Indian Restaurant') ]
Out[68]:
```

	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
36	The Danforth West, Riverdale	0.023810	1	43.679557	-79.352188	Sher-E-Punjab	43.677308	-79.353066	Indian Restaurant
8	Davisville	0.029412	1	43.704324	-79.388790	Marigold Indian Bistro	43.702881	-79.388008	Indian Restaurant
30	St. James Town, Cabbagetown	0.023256	1	43.667967	-79.367675	Butter Chicken Factory	43.667072	-79.369184	Indian Restaurant
4	Central Bay Street	0.016129	1	43.657952	-79.387383	Colaba Junction	43.660940	-79.385635	Indian Restaurant
6	Church and Wellesley	0.013158	1	43.665860	-79.383160	Kothur Indian Cuisine	43.667872	-79.385659	Indian Restaurant

Cluster 3:



```
In [69]: #Cluster 2

         to_merged.loc[(to_merged['Cluster Labels'] ==2) & (to_merged['Venue Category'] == 'Indian Restaurant') ]
Out[69]:
```

	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
34	The Annex, North Midtown, Yorkville	0.05	2	43.67271	-79.405678	Roti Cuisine of India	43.674618	-79.408249	Indian Restaurant

Most of the Indian restaurants are in cluster 2 which is around The Danforth West, Davisville, St.James Town, Central Bay Street, Church and Wellesley

There are only 1 Indian resturant in cluster 1 and cluster 3 which are Harbourfront East, Union Station, Toronto Islands and The Annex, North Midtown, Yorkville. So it looks like a good oppurtunity to open a new resturant here.

Discussion:

Most of the Indian restaurants are in cluster 2 which is around The Danforth West, Davisville, St.James Town, Central Bay Street, Church and Wellesley

There are only 1 Indian resturant in cluster 1 and cluster 3 which are Harbourfront East, Union Station, Toronto Islands and The Annex, North Midtown, Yorkville. So it looks like a good opportunity to open a new restaurant here.

Conclusion:

This project can be modified further to shows clusters of a different city