Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

1. INTRODUCTION

This is a capstone project for IBM Data Science Professional Certificate. In this project, I am going to showcase a scenario regarding number of Indian restaurants in Toronto and how it is going to benefit for entrepreneurs to open Indian restaurant in Toronto and its neighborhood. Therefore it might be a great opportunity for Canadian based entrepreneurs. More than a million Indian people resides in Canada. So entrepreneurs might think of opening its business in the areas near to Indian communities. With the purpose, finding the best location to open such a restaurant is one of the most important decisions for these entrepreneurs and this project will help them to find the most suitable location.

2. BUSINESS PROBLEM

"*What is the most suitable location for an entrepreneur to open an Indian Restaurant in Toronto or Canada?"* The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Indian Restaurant in Canada, basically near Toronto. So, I will be leveraging the concept of Data Science Methodology along with Machine Learning Algorithms such as "Clustering" and I will be going to suggest some of the possible location and solution to this business problem.

3. TARGET AUDIENCE

All interested entrepreneurs who want to know the best suitable location to open an authentic Indian restaurant in Canada.

4. DATA

To explore into the data and finding a best possible solution, we will need below data:

- List of Postal Code, borough , Neighborhoods in Toronto, Canada [https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M]
- Latitude and Longitude of these Neighborhoods [http://cocl.us/Geospatial_data]
- Venue data related to Indian restaurants.

This will help us to find prospect neighborhoods/ location that are more suitable to open an Indian Restaurant.

5. DATA EXTRACTION

- Scrapping of Toronto neighborhood details by following Wikipedia page [https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,]
- Getting geographical coordinates of the neighborhoods using the Geocoder package, geographical coordinates of each postal code.
- Leveraging Foursquare API to get venue data related to different neighborhood.

6. METHODOLOGY

- At first, I tried to get a list of boroughs, neighborhoods in Toronto, Canada. I have extracted from Wikipedia link mentioned in the DATA Section -4.

- Performed ***web scraping*** by utilizing html parser with *BeatifulSoup* object and stored the data into a data frame using **pandas** library. This helps me to pull web page data into a tabular format.

  Note: Below screenshots contains sample data. In this process no specific data is discarded.

Web scraping & Data Cleansing

Out[85]:

	Postal code	Borough	Neighborhood
1	M1A	Not assigned	
2	M2A	Not assigned	
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village
5	M5A	Downtown Toronto	Regent Park / Harbourfront

- HTML scraping method is easier and more convenient to pull data directly from a web page into a Data frame. Now, I got a list of list of neighborhood names and borough with postal codes. Then, I performed Data cleaning like splitting neighborhood data separated by "/" into rows and merged with corresponding postal code and borough data. Also, discarded "Non-assigned" Borough data from the table.

Data Splitting

Out[92]:

	PostalCode	Neighborhood
0	M3A	Parkwoods
1	M4A	Victoria Village
2	M5A	Regent Park
3	M5A	Harbourfront
4	M6A	Lawrence Manor
5	M6A	Lawrence Heights
6	M7A	Queen's Park
7	M7A	Ontario Provincial Government
8	M9A	Islington Avenue
9	M1B	Malvern
10	M1B	Rouge
11	M3B	Don Mills

Merged Data

Out[93]:

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park
3	M5A	Downtown Toronto	Harbourfront
4	M6A	North York	Lawrence Manor
5	M6A	North York	Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park
7	M7A	Downtown Toronto	Ontario Provincial Government
8	M9A	Etobicoke	Islington Avenue
9	M1B	Scarborough	Malvern
10	M1B	Scarborough	Rouge
11	M3B	North York	Don Mills
12	M4B	East York	Parkview Hill
13	M4B	East York	Woodbine Gardens
14	M5B	Downtown Toronto	Garden District, Ryerson

Grouped Neighborhood based on Postal Code :

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Malvern , Rouge
1	M1C	Scarborough	Rouge Hill , Port Union , Highland Creek
2	M1E	Scarborough	Guildwood , Morningside , West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
5	M1J	Scarborough	Scarborough Village
6	M1K	Scarborough	Kennedy Park , Ionview , East Birchmount Park
7	M1L	Scarborough	Golden Mile , Clairlea , Oakridge
8	M1M	Scarborough	Cliffside , Cliffcrest , Scarborough Village...
9	M1N	Scarborough	Birch Cliff , Cliffside West
10	M1P	Scarborough	Dorset Park , Wexford Heights , Scarborough ...
11	M1R	Scarborough	Wexford , Maryvale

- To get the coordinate information i.e. latitude and longitude, corresponding to fetched data, I have utilized Geospatial data [http://cocl.us/Geospatial_data] and merged with the main table. Joined these data based on postal code.

Geographical Coordinates :

	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Merging neighborhood data with geospatial coordinates:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern , Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill , Port Union , Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood , Morningside , West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Number of Borough & Neighborhood in the dataframe:

```
The dataframe has 10 Borough and 103 Neighborhood.
```

Number of Borough based on Neighborhood:

```
Borough
Central Toronto      9
Downtown Toronto    19
East Toronto         5
West Toronto         6
Name: Neighborhood, dtype: int64
```

Toronto Coordinate retrieve “geopy.geocoders” package :

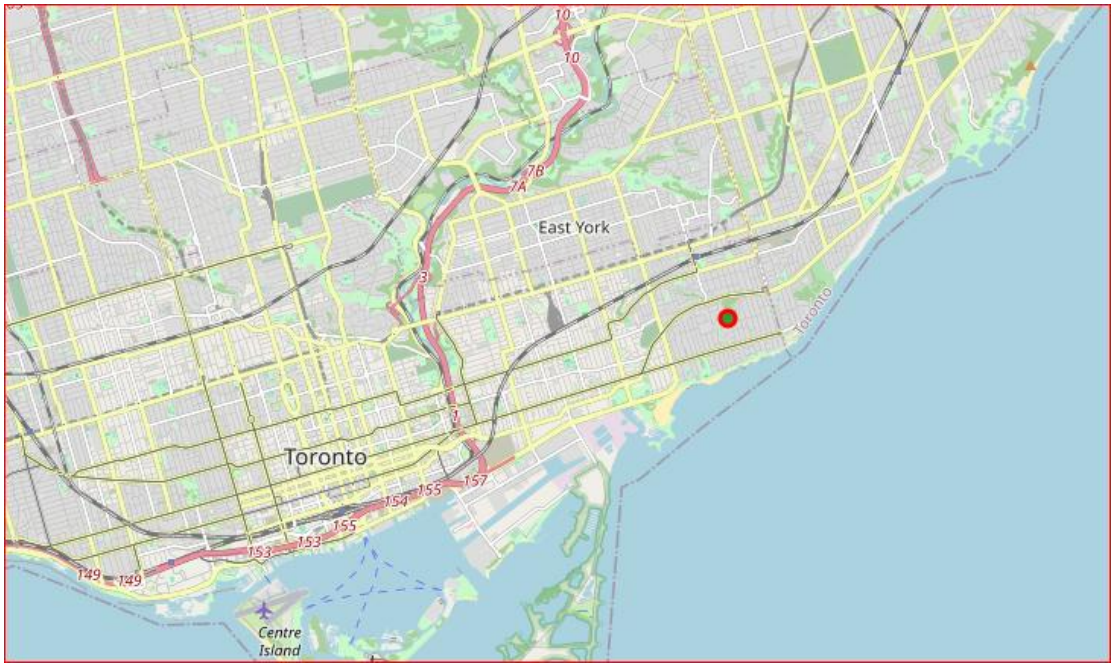
```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

List of Neighborhoods near Toronto :

	PostalCode	Borough	Neighborhood	Latitude	Longitude
37	M4E	East Toronto	The Beaches	43.676357	-79.293031
41	M4K	East Toronto	The Danforth West , Riverdale	43.679557	-79.352188
42	M4L	East Toronto	India Bazaar , The Beaches West	43.668999	-79.315572
43	M4M	East Toronto	Studio District	43.659526	-79.340923
44	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

<u>Map showing Boroughs in Toronto :</u>

Using geospatial data and Folium package, I have created the map of Toronto.



- ***Foursquare API to explore the Neighborhoods –***

  Using *Foursquare API*, I pulled the list of top 100 venues within 500 Meter radius in Toronto. Before using this API, I have created an account in Foursquare Developer and obtained client id and secret key to pull the data.

  Using this API, different venue names, categories, its corresponding coordinates were pulled and populated in a dataframe.

  Sample Data :

	PostalCode	Neighborhood	Neighborhood_Latitude	Neighborhood_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	M4E	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	M4E	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	M4E	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	M4E	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	M4E	The Beaches	43.676357	-79.293031	Dip 'n Sip	43.678897	-79.297745	Coffee Shop

  With this data, I checked number of unique categories. Then, I analyzed each neighborhood by grouping the rows by neighborhood and taking mean on the frequency of occurrence of each venue category. This will be used during Clustering.

  Number of unique Venue Category :

```
There are 228 uniques categories.
```

  Example of unique venue category :

```
array(['Trail', 'Health Food Store', 'Pub', 'Neighborhood', 'Coffee Shop',
       'Cosmetics Shop', 'Greek Restaurant', 'Italian Restaurant',
       'Ice Cream Shop', 'Yoga Studio', 'Brewery',
       'Fruit & Vegetable Store', 'Dessert Shop', 'Restaurant',
       'Pizza Place', 'Juice Bar', 'Bookstore', 'Bubble Tea Shop',
       'Furniture / Home Store', 'Grocery Store', 'Spa', 'Bakery',
       'Caribbean Restaurant', 'Café', 'Indian Restaurant',
       'Japanese Restaurant', 'Lounge', 'Frozen Yogurt Shop',
       'American Restaurant', 'Gym', 'Fish & Chips Shop',
       'Fast Food Restaurant', 'Sushi Restaurant', 'Park', 'Liquor Store',
       'Pet Store', 'Steakhouse', 'Burrito Place', 'Movie Theater',
       'Sandwich Place', 'Light Rail Station', 'Fish Market', 'Gay Bar',
       'Seafood Restaurant', 'Cheese Shop', 'Middle Eastern Restaurant',
       'Comfort Food Restaurant', 'Stationery Store', 'Wine Bar',
       'Thai Restaurant', 'Coworking Space', 'Latin American Restaurant',
       'Gastropub', 'Gym / Fitness Center', 'Bar', 'Convenience Store',
       'Bank', 'Diner', 'Clothing Store', 'Swim School', 'Bus Line',
       'Breakfast Spot', 'Food & Drink Shop', 'Department Store', 'Hotel',
       'Salon / Barbershop', 'Mexican Restaurant', 'Chinese Restaurant',
       'Sporting Goods Shop', 'Rental Car Location', 'Toy / Game Store',
       'Gas Station', 'Farmers Market', 'Gourmet Shop', 'Pharmacy',
       'New American Restaurant', 'Playground', 'Supermarket',
       'Sports Bar', 'Fried Chicken Joint', 'Vietnamese Restaurant',
       'Bagel Shop', 'Health & Beauty Service', 'Jewelry Store',
       'General Entertainment', 'Butcher', 'Taiwanese Restaurant',
```

Here, I have put one constraint as "**Indian Restaurant**" and filtered out the data. Taken mean upon the data :

	Neighborhood	Indian Restaurant
0	Berczy Park	0.000000
1	Brockton , Parkdale Village , Exhibition Place	0.000000
2	Business reply mail Processing CentrE	0.000000
3	CN Tower , King and Spadina , Railway Lands ...	0.000000
4	Central Bay Street	0.016667

Clustering

I have performed clustering method by using *k-means clustering*. Here, I have set k=4 [number of clusters] and then located data points to nearest cluster while keeping centroid as small as possible. It is one of the simplest and popular Unsupervised Machine Learning Algorithms. I have segregated neighborhoods into 4 clusters based on their frequency of occurrence for Indian Food. Based on this observation, recommendation can be done for most suitable location for Indian Restaurant in Toronto.

Clustering:

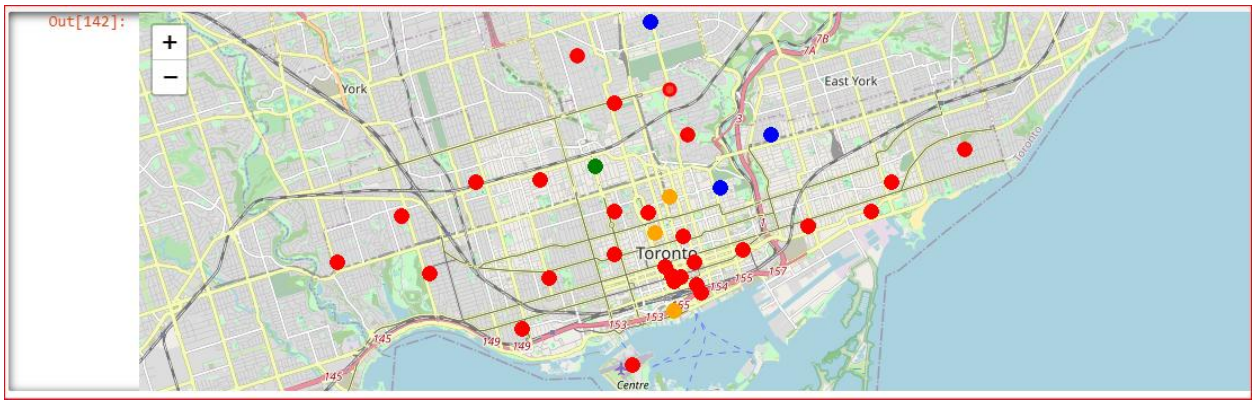
```
Out[118]: array([0, 0, 0, 0, 1, 0, 1, 0, 2, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0],
                dtype=int32)
```

Merged and sorted Data:

	Neighborhood	Indian Restaurant	Cluster_Labels	Neighborhood_Latitude	Neighborhood_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	Berczy Park	0.0	0	43.644771	-79.373306	LCBO	43.642944	-79.372440	Liquor Store
25	Richmond , Adelaide , King	0.0	0	43.650571	-79.384568	Bulldog On The Block	43.650652	-79.384141	Coffee Shop
25	Richmond , Adelaide , King	0.0	0	43.650571	-79.384568	John & Sons Oyster House	43.650656	-79.381613	Seafood Restaurant
25	Richmond , Adelaide , King	0.0	0	43.650571	-79.384568	M Square Coffee Co	43.651218	-79.383555	Coffee Shop
25	Richmond , Adelaide , King	0.0	0	43.650571	-79.384568	Downtown Toronto	43.653232	-79.385296	Neighborhood

7. RESULT & DISCUSSION

Cluster Result:



The result shows distribution of Indian Restaurants in each neighborhood in Toronto from k-means clustering. I have grouped into 4 clusters.

- **Cluster 0** : Neighborhoods with no Indian restaurants.
- **Cluster 1**: Neighborhoods with more number of Indian restaurants.
- **Cluster 2** : Neighborhoods with less number of Indian restaurants
- **Cluster 3** : Neighborhoods with more number of Indian restaurants

Cluster -1

	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	0.030303	1	43.704324	-79.388790	Marigold Indian Bistro	43.702881	-79.388008	Indian Restaurant
30	St. James Town , Cabbagetown	0.022222	1	43.667967	-79.367675	Butter Chicken Factory	43.667072	-79.369184	Indian Restaurant

Cluster -2

	Neighborhood	Indian Restaurant	Cluster_Labels	Neighborhood_Latitude	Neighborhood_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
34	The Annex , North Midtown , Yorkville	0.041667	2	43.67271	-79.405678	Roti Cuisine of India	43.674618	-79.408249	Indian Restaurant

Cluster -3

	Neighborhood	Indian Restaurant	Cluster_Labels	Neighborhood_Latitude	Neighborhood_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
14	Harbourfront East , Union Station , Toronto ...	0.010000	3	43.640816	-79.381752	Indian Roti House	43.639060	-79.385422	Indian Restaurant
6	Church and Wellesley	0.012987	3	43.665860	-79.383160	Kothur Indian Cuisine	43.667872	-79.385659	Indian Restaurant
4	Central Bay Street	0.016667	3	43.657952	-79.387383	Colaba Junction	43.660940	-79.385635	Indian Restaurant

8. Conclusion

Most of the Indian restaurants are in Cluster 2 and 3, around - The Dan forth West, Riverdale, Davisville, St. James Town, Cabbage town, Harbourfront East, Union Station, Church and Wellesley and Central Bay Street. *Lowest* in the Cluster 0 and 1, areas like The Annex, North Midtown, and Yorkville.

So, cluster 0 might be a good location to open Indian Restaurant as there is not any Indian Restaurant in these areas. Therefore, this project recommends the entrepreneur to open any Authentic Indian Restaurant in these locations.