PROJECT REPORT

BY : PRANJAL SHRIVASTAVA

The report is part of the IBM Applied Data Science Specialization Capstone Project. The main objectives of this project were to define a business problem, look for data on the web and use Foursquare location data to compare to different neighborhoods of Toronto to figure out which neighborhood is suitable for starting a new restaurant business.

INTRODUCTION –

Toronto city, capital of the province of Ontario, southeastern Canada. It is the most populous city in Canada and the country's financial and commercial center. The demographics of Toronto make Toronto one of the most multicultural and multiracial cities in the world. More than half of the entire Indian-Canadian population resides in Toronto, people from India love food and I love to eat, thus Toronto is one of the best places to start an INDIAN RESTAURANT.

BUSINESS PROBLEM –

In this capstone project, we will analyze the neighborhoods in Toronto to identify the most profitable neighborhood for opening an Indian Restaurant, by using Web Scraping, Data Pre-processing, Machine learning algorithms like K-Means clustering algorithm, and Foursquare API Service.

TARGET AUDIENCE –

- The business owner who wants to invest or open a start-up company or restaurant.
- The freelancer who loves to have their own small company or restaurant as a side business.
- Indian crowd who wants to find neighborhoods with lots of options for Indian restaurants.
- Tourists who want to eat Indian food.

Data Sources –

- Toronto City Neighbourhoods Data – https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Geographical Coordinates of the Neighbourhoods – https://cocl.us/Geospatial_data
- Location Data of Neighbourhood – Foursquare API Services

Methodology –

- First, web scraping for data using the 'BeautifulSoup' package is done.
- Then a dataframe is created which contains Postal Code, Borough, Neighbourhood.
- Then further processing of the dataframe is done, like removing unassigned values, merging different neighborhoods with the same borough.

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

- The next step is to add geographical coordinates, for that the Geospatial_data.csv is used.

	Postal Code	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

- Then merging is done based on Neighbourhood.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills North	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937

- Finally, the dataframe is modified, in which Borough contains 'Toronto'.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
19	M4E	East Toronto	The Beaches	43.676357	-79.293031
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306
24	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
25	M6G	Downtown Toronto	Christie	43.669542	-79.422564
30	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568
31	M6H	West Toronto	Dufferin, Dovercourt Village	43.669005	-79.442259
35	M4J	East York/East Toronto	The Danforth East	43.685347	-79.338106

- Then using Foursquare API developer services, 100 venues were explored under the radius of 500m. A Foursquare developer account to obtain a Client ID and Client Secret key to pull the data. From Foursquare, the names, categories, latitude, and longitude of the venues were pulled.

- Then one-hot encoding was performed, for each of the neighborhood's venues were turned into the frequency at how many of those Venues were located in each neighborhood. Then, the values in each venue category were grouped by the average of the frequency of each venue category.

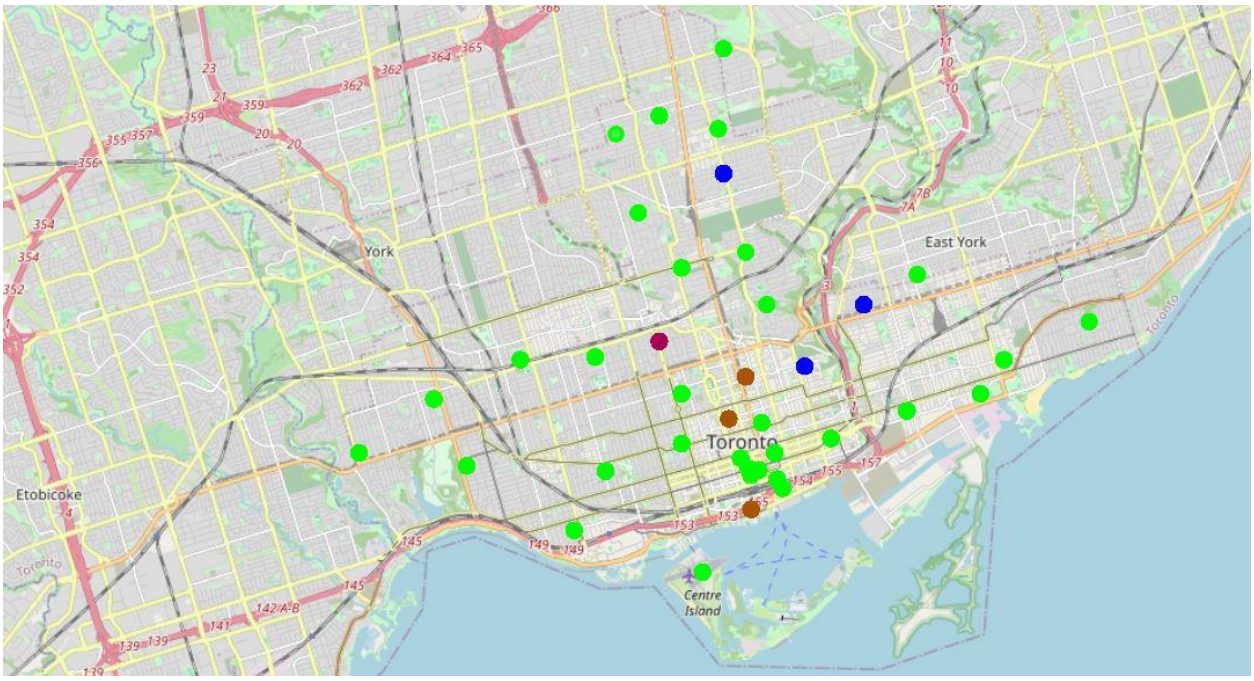
	Neighborhood	Yoga Studio	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Tibetan Restaurant	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Wine Bar	Wine Shop
0	Berczy Park	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.018182	0.0	0.000000	0.0	0.0	0.0
1	Brockton, Parkdale Village, Exhibition Place	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0
2	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.0	0.071429	0.071429	0.071429	0.142857	0.071429	0.071429	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0
3	Central Bay Street	0.015385	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.015385	0.0	0.015385	0.0	0.0	0.0
4	Christie	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0

- Then, a new dataframe was created, which stored Neighbourhood and Indian Restaurant column. It was created to make further operations simpler.

	Neighborhood	Indian Restaurant
0	Berczy Park	0.000000
1	Brockton, Parkdale Village, Exhibition Place	0.000000
2	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
3	Central Bay Street	0.015385
4	Christie	0.000000
5	Church and Wellesley	0.012821
6	Commerce Court, Victoria Hotel	0.000000
7	Davisville	0.027027
8	Davisville North	0.000000
9	Dufferin, Dovercourt Village	0.000000

- At last, clustering of Indian Restaurant was done, it was based on K-Means Clustering Algorithm. Here k-value was taken as 4, then the model was fitted. And finally, the clustered map was displayed.

Results & Discussion–



- Cluster 1 – Green Colour
- Cluster 2 – Blue Colour
- Cluster 3 – Purple Colour
- Cluster 4 – Brown Colour

Cluster 1 - Depicts the least frequency of Indian Restaurants among the neighborhoods.

Cluster 3(The Annex, North Midtown, Yorkville) - Depicts the maximum frequency of Indian Restaurants among the neighborhoods.

After analyzing, it is found that The Annex, North Midtown, Yorkville has the highest frequency amongst all other neighbourhoods, followed by Davisville. Approximately 80 percentage of the neighborhood has no authentic Indian Restaurant, thus it gives a good opportunity for business owner and freelancer to open a new Restaurant. The green cluster(Cluster 1) can be a good option to open an Indian Restaurant, for example The Beaches, St.James Town, India Bazaar, Forest Hill and Parkdale Village are some good options. This concludes the findings for the location and recommends the business owner and freelancer to open an authentic Indian restaurant in these locations.

Conclusion –

Finally, to conclude this project, I have got a glimpse of how data-science project look-like. I have used various libraries like, folium, pandas, sklearn, requests. I have also used BeautifulSoup package for web scraping. Here, I have also used Foursquare API services to explore the neighborhoods. And finally, I have used machine learning algorithm, K-Means Clustering Algorithm, to predict the most profitable neighborhood for opening an Indian Restaurant.