Coursera IBM Data Science Capstone Project

Analyzing location for an Indian restaurant in Toronto, Canada

Prepared by:
Ruchir Palkar

1.Introduction & Business Problem :

Problem Background:
Toronto is the densely populated city in Canada. It provides a lot of business opportunities and a business-friendly environment. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. Its economy is highly diversified with strengths in technology, design, financial services, life sciences, education, arts, fashion, business services, environmental innovation, food services, and tourism.

Toronto has attracted big business players from all around the world. It means the Toronto market is highly competitive. Also, the cost of doing business in such a market is high. Hence, any business venture or expansion must be analyzed very carefully. The insights derived from the analysis will give a good understanding of the business environment which help in strategically targeting the market. It also reduces risk and helps in getting a better Return on Investment.

Problem Description

A restaurant is a business establishment which prepares and serves food and drink to customers in return for money. Toronto is famous for its excellent cuisine. Its food culture includes an array of international cuisines influenced by the city's immigrant history.

In the restaurant business, LOCATION is a very important factor for its success. When a restaurant is located in a good and posh location, restaurant owner don't have to push his marketing efforts to generate more footfalls. Due to the location, his business will get more walk-ins than any others. But choosing the location sometimes can pose a big challenge.

Also opening a restaurant in the area with no competition is challenging as well.

Hence, in this project, I will help ABC foods LLC to find the most suitable location for an Indian Restaurant in Toronto with no or little competition with high foot traffic.

Target Audience
Anyone who wants to open a restaurant in Toronto, Canada.

2.Data Section

In this project, my job as Data Scientist was to locate best neighborhoods in Toronto city for starting new Indian restaurant. I analyzed neighborhoods and venue data of Toronto city for this project.

1. First of all, I applied **K means clustering** for finding Cluster that has no or little number of Indian restaurants in its neighbourhoods (In this case- CLUSTER 0)
2. After that I analyzed that "CLUSTER 0" data for finding neighborhoods that has large number of restaurants & hotels. This step will help to find neighborhoods that are **famous for Hotels & Restaurants** in" CLUSTER 0".
3. **More venues in a neighborhood means more foot traffic**. Foot traffic is must for restaurant business.In this step I found total number of venues in each neighborhood by analyzing venue data of each neighbourhood of CLUSTER 0.

With help of above steps, I was able to find neigbhorhoods with no Indian restaurant which are **famous for hotels & restaurants & has lots of foot traffic.**

I will explain it in details in this report.

To solve this problem, I used following data:

1. List of neighborhoods in Toronto, Canada.
2. Latitude and Longitude of these neighborhoods.
3. Indian restaurants Venue data.
4. Data of Hotels & Restaurants in each neighborhood.
5. Data of Venues in each neighborhood

Data 1- List of neighborhoods in Toronto, Canada.

This Wikipedia page ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M") contains a list of neighborhoods in Toronto, with a total of 130 neighborhoods. I used web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages.

Data 2- Latitude and Longitude of these neighborhoods.

I was supposed get the geographical coordinates of the neighborhoods using Python Geocoder package but due to its unstable nature I used csv file which has latitude and longitude coordinates of the neighborhoods.

Data 3- Indian restaurants Venue data of Toronto city

This data was fetched using Foursquare API. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.

Data 4 – Data of Hotels & Restaurants in each Neighborhood

This data was also fetched using Foursquare API.

Data 5- Data of venues in each Neighborhood

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail	4bd461bc77b29c74a07d9282
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store	4ad4c062f964a52011f820e3
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub	4b8daea1f964a520480833e3
3	The Beaches	43.676357	-79.293031	Glen Stewart Ravine	43.676300	-79.294784	Other Great Outdoors	56afcad6498e05333bf42031
4	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood	4df91c4bae60f95f82229ad5

This data was also fetched using Foursquare API.

3. Methodology-

First of all, I did web scraping for getting the list of neighborhoods in Toronto, Canada from Wikipedia page (" https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M").

I used Python requests and beautifulsoup packages for scraping data. However, it only had a list of neighborhood names, Borough and postal codes. I also needed their coordinates for using Foursquare API to fetch the list of venues in these neighborhoods. For this purpose, I tried using Geocoder package but due to its unstable nature, I didn't use it. I used the csv file which contains the coordinates of Toronto neighborhoods. After that I created dataframe containing Postalcode,Borough,Neighborhood,Latitude & Longitude data by merging them. Also did data cleaning and data wrangling before getting final dataframe.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M4E	East Toronto	The Beaches	43.676357	-79.293031
1	M4K	East Toronto	The Danforth West,Riverdale	43.679557	-79.352188
2	M4L	East Toronto	The Beaches West,India Bazaar	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790

After gathering all these coordinates, I visualized the map of Toronto using Folium package for verifying whether these are correct coordinates are correct or not.

Next, I used the Foursquare API to fetch venues data within 500 meters radius. From Foursquare, I fetched the names, categories, latitude and longitude of the venues. With this data, I also checked how many unique categories that I can get from these venues. Then, I analyzed each neighborhood by grouping the rows by neighborhood and took the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

	Neighborhoods	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Mu
0	Adelaide,King,Richmond	0.0	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.03	0.0	0.0	0.010000	
1	Berczy Park	0.0	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.00	0.0	0.0	0.017544	
2	Brockton,Exhibition Place,Parkdale Village	0.0	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.00	0.0	0.0	0.000000	
3	Business Reply Mail Processing Centre 969 Eastern	0.0	0.0000	0.0000	0.0000	0.000	0.0000	0.000	0.00	0.0	0.0	0.000000	
4	CN Tower,Bathurst Quay,Island airport,Harbourf...	0.0	0.0625	0.0625	0.0625	0.125	0.1875	0.125	0.00	0.0	0.0	0.000000	

After that, I fetched data of Indian restaurants for analysis.

	Neighborhoods	Indian Restaurant
34	The Annex,North Midtown,Yorkville	0.043478
11	Davisville	0.028571
6	Central Bay Street	0.024096
37	The Danforth West,Riverdale	0.023810
5	Cabbagetown,St. James Town	0.022222

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, 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 and it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Indian restaurant”.

```
: # set number of clusters
clusters = 3

to_clustering = to_indian.drop(["Neighborhoods"], 1)

# run k-means clustering
kmeans = KMeans(n_clusters=clusters, random_state=0).fit(to_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

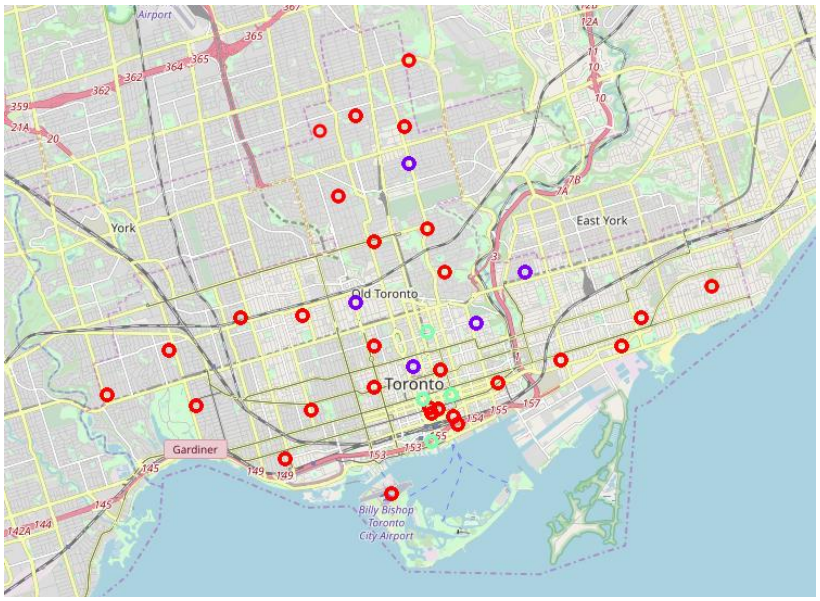
```
: array([1, 1, 1, 1, 1, 2, 2, 2, 2, 0], dtype=int32)
```

4. Results-

PART-1 **= K Means Clustering**

I used K means clustering for segmentation and clustering of Toronto neighborhoods based on how many Indian restaurants are in each neighborhood and got following result.

- Cluster 0: Neighborhoods with no Indian restaurants
- Cluster 1: Neighborhoods with little number of Indian restaurants
- Cluster 2: Neighborhoods with high number of Indian restaurants

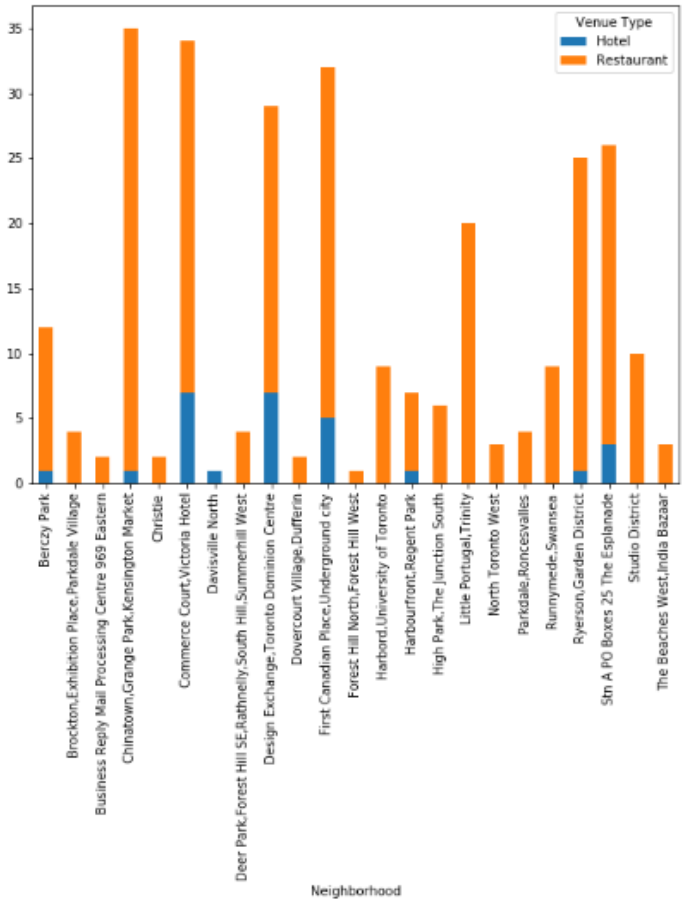


The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple colorand Cluster 2 in light green color. We can say opening restaurant in neighborhoods of Cluster 0 will best option but I did further analysis for selecting best neighborhoods among neighborhoods in Cluster 0 .

PART -2 **= Finding neighborhoods with lots of restaurants & hotels in Cluster 0**

All neighborhoods in Cluster 0 have no Indian restaurant. I went one step further and did analysis on Cluster 0 for finding total number of restaurants & hotels in each neighborhood of Cluster0 and arranging them in descending order for getting neighborhoods that are famous for hotels & restaurants.

Venue Type	Hotel	Restaurant	Total H & R
Neighborhood			
Chinatown,Grange Park,Kensington Market	1	34	35
Commerce Court,Victoria Hotel	7	27	34
First Canadian Place,Underground city	5	27	32
Design Exchange,Toronto Dominion Centre	7	22	29
Stn A PO Boxes 25 The Esplanade	3	23	26



Above table & graph are results of further analysis.

This analysis gave neighborhoods which are famous for restaurants & hotels in Cluster 0.

PART 3= **Finding neighborhoods that has maximum number of Venues.**

More venues in neighborhood mean more foot traffic. In this step,I calculated total venues in each neighborhood of Cluster 0 and merged with total restaurants & hotels table.
The following are results.

:

	Hotel	Restaurant	Total H & R	Total Venues
Neighborhood				
Chinatown,Grange Park,Kensington Market	1	34	35	100
Commerce Court,Victoria Hotel	7	27	34	100
First Canadian Place,Underground city	5	27	32	100
Design Exchange,Toronto Dominion Centre	7	22	29	100
Stn A PO Boxes 25 The Esplanade	3	23	26	96

These are top 5 neighborhoods that I can suggest my client for opening an Indian restaurant.
As I am using free version of Foursquare API, maximum venues I can fetch is 100. If I had paid version, results would be better.

5. Discussion-

In this project, I took only few considerations for finding ideal location for an Indian restaurant. There are other factors as well that can help to find better location like real estate prices, population density,etc. These factors would have definitely improved my analysis.

However, this project gives a nice start to the process and narrowed down a very long list (from 103 to 5 choices)

6. Conclusion-

In this project, I have gone through the process of identifying the business problem, extracting and preparing the data and performing the machine learning by utilizing k-means clustering. Also I did further analysis on clustering data for recommending top 5 neighborhoods for opening an Indian restaurant having no competition and high foot traffic.

7.References

List of neighborhoods in Toronto:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
Foursquare Developer Documentation: https://developer.foursquare.com/docs