Recommending location for Indian Restaurant In Toronto

Business Problem

- A large number of Indians reside in Canada, especially in the city of Toronto.
- There are also many Asians living in Toronto.
- Indian cuisine is similar to Asian cuisine and hence this restaurant can attract Asians as well.
- An entrepreneur wishes to open an Indian restaurant which caters not only to the tastes of the Indians but also Asians and other locals living in that area.
- The entrepreneur thus wants to know potential places in Toronto for opening this Indian restaurant as a suitable location can be beneficial for the restaurant to be popular and also profitable.

Data Acquisition and Cleaning

- The list of neighbourhoods in Canada which contains the boroughs, their postal codes and neighbourhoods in each borough scraped from the Wikipedia page of postal codes of Canada
- The data about the coordinates of neighbourhoods can be extracted from the csv file provided by IBM
- The data about the venues in all neighbourhoods is accessed through API calls using Foursquare API services.
- The boroughs and neighbourhoods whose value was not available were removed from the dataset.

- The table containing data about neighbourhoods and boroughs in Canada is merged with table containing the coordinates of all neighbourhoods
- The boroughs and neighbourhoods which do not come under Toronto are dropped from the table
- The boroughs column is removed as it is not required hereafter.
- Using Foursquare API data about every venue in each neighbourhood is collected and merged with the existing table.

Data till now…

This is table that is formed after the previous steps were done:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood
4	The Danforth West, Riverdale	43.679557	-79.352188	MenEssentials	43.677820	-79.351265	Cosmetics Shop
5	The Danforth West, Riverdale	43.679557	-79.352188	Pantheon	43.677621	-79.351434	Greek Restaurant
6	The Danforth West, Riverdale	43.679557	-79.352188	La Diperie	43.677702	-79.352265	Ice Cream Shop
7	The Danforth West, Riverdale	43.679557	-79.352188	Dolce Gelato	43.677773	-79.351187	Ice Cream Shop
8	The Danforth West, Riverdale	43.679557	-79.352188	Cafe Fiorentina	43.677743	-79.350115	Italian Restaurant
9	The Danforth West, Riverdale	43.679557	-79.352188	Louis Cifer Brew Works	43.677663	-79.351313	Brewery

Methodology

- Using an appropriate clustering algorithm, the neighbourhoods in Toronto can be grouped together into different clusters.
- These clusters are based on the number of Indian restaurants in the neighbourhoods.
- For this problem Asian restaurants have to be considered as well, but in place of Asian restaurants, Thai restaurants have been selected.
- The number of Asian restaurants was very less and could not have been used for solving this problem.
- Neighbourhoods with occurrences of both Indian and Thai restaurants have to be studied and conclusion has to be made on the basis of competition from other existing restaurants in the area and scope for opening a new Indian restaurant there.

What is a Clustering Algorithm?

- Clustering is a Machine Learning technique that involves the grouping of data points.
- Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group.
- Data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features.
- Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.
- K-Means clustering, Agglomerative Hierarchical clustering, DBSCAN clustering, Mean-Shift clustering are some clustering algorithms.

K-Means Clustering Algorithm

- For this project, we are using K-Means clustering algorithm.
- To begin, we first select a number of classes/groups to use and randomly initialize their respective center points.
- Each data point is classified by computing the distance between that point and each group center, and then classifying the point to be in the group whose center is closest to it.
- Based on these classified points, we recompute the group center by taking the mean of all the vectors in the group.
- Repeat these steps for a set number of iterations or until the group centers don’t change much between iterations.

More changes in the data…

- The available data cannot be trained to the K-Means model as it has the venue categories as String values and not numeric values.
- To convert the string values of venue categories to numeric values, we will use the technique ‘One-Hot Encoding’ to create dummy variables for each category.
- Then the data is grouped according to neighbourhoods and the mean occurrence of each venue category is calculated for each neighbourhood.

	Neighborhoods	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Theme Restaurant	Toy / Game Store	Trail	Train Station	Vegetarian / Vegan Restaurant
0	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.0	0.017241
1	Brockton, Parkdale Village, Exhibition Place	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.0	0.000000
2	Business reply mail Processing Centre, South C...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.0	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.055556	0.055556	0.055556	0.111111	0.166667	0.111111	0.000000	0.0	...	0.000000	0.000000	0.0	0.0	0.000000
4	Central Bay Street	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.0	0.015625
5	Christie	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.0	0.000000
6	Church and Wellesley	0.013333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.013333	0.0	...	0.013333	0.000000	0.0	0.0	0.000000
7	Commerce Court, Victoria Hotel	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.0	...	0.000000	0.000000	0.0	0.0	0.020000
8	Davisville	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.028571	0.0	0.0	0.000000
9	Davisville North	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.000000	0.000000	0.0	0.0	0.000000

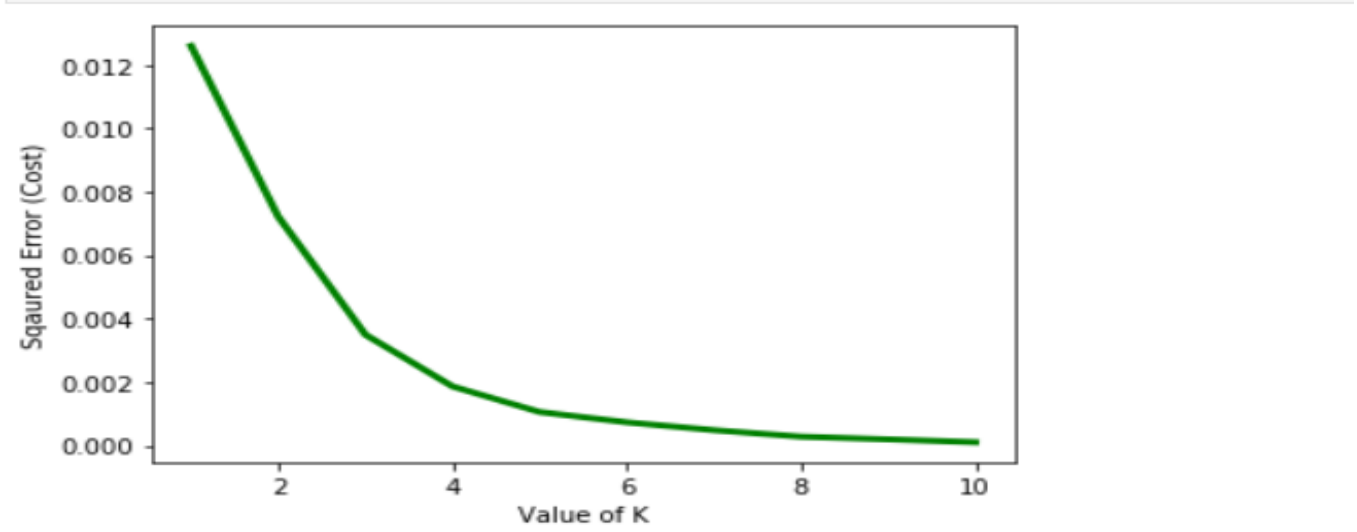
Finalizing the data

- Now only neighbourhoods along with Indian restaurants and Thai restaurants are selected from previous table.
- This final data contains the occurrences of Indian and Thai restaurants in the neighbourhoods of Toronto.

	Neighborhood	Indian Restaurant	Thai Restaurant
0	Berczy Park	0.017241	0.017241
1	Brockton, Parkdale Village, Exhibition Place	0.000000	0.000000
2	Business reply mail Processing Centre, South C...	0.000000	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.000000
4	Central Bay Street	0.015625	0.015625
5	Christie	0.000000	0.000000
6	Church and Wellesley	0.013333	0.013333
7	Commerce Court, Victoria Hotel	0.000000	0.020000
8	Davisville	0.028571	0.028571
9	Davisville North	0.000000	0.000000

Choosing number of clusters

- For initializing the K-Means model, we need to set up appropriate number of clusters.
- This can be done by calculating square errors for values of clusters ‘k’ from 1 to 10 by training the model against the above data for each value of k and plotting the values on graph.
- The elbow method is used to find the correct value of k from this graph.



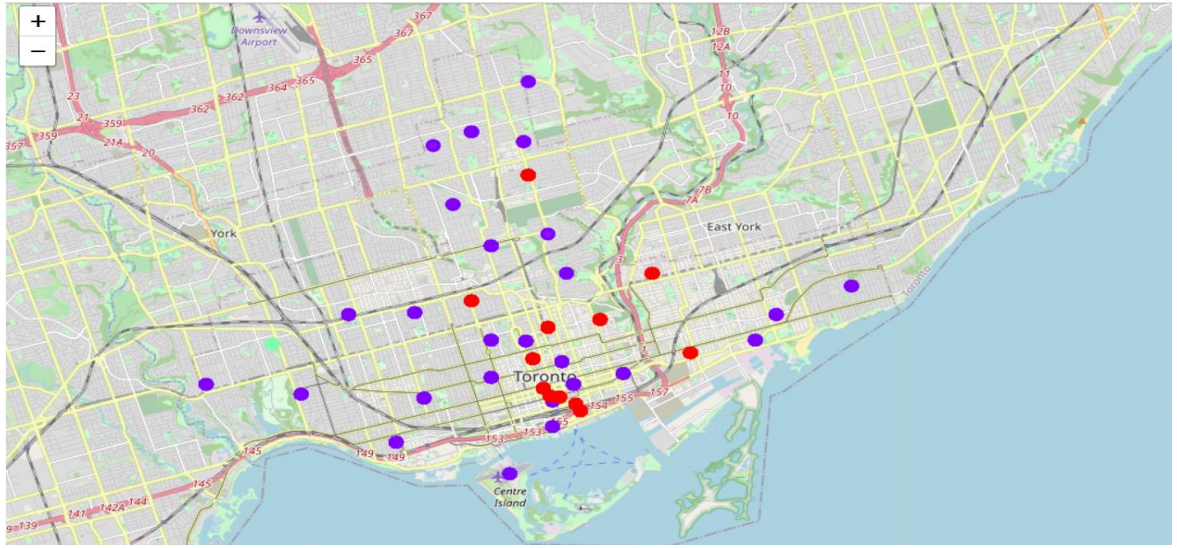
Forming clusters from the data

- After observing the above graph, we select the value of k as 3.
- So after clustering, we separate the neighbourhoods into 3 clusters.
- The cluster labels are predicted for the given data and these cluster labels are added to our dataset.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Indian Restaurant	Thai Restaurant	New Cluster Labels
0	The Beaches	43.676357	-79.293031	Glen Manor Ravine	43.676821	-79.293942	Trail	0.00000	0.0	1
1	The Beaches	43.676357	-79.293031	The Big Carrot Natural Food Market	43.678879	-79.297734	Health Food Store	0.00000	0.0	1
2	The Beaches	43.676357	-79.293031	Grover Pub and Grub	43.679181	-79.297215	Pub	0.00000	0.0	1
3	The Beaches	43.676357	-79.293031	Upper Beaches	43.680563	-79.292869	Neighborhood	0.00000	0.0	1
4	The Beaches	43.676357	-79.293031	Seaspray Restaurant	43.678888	-79.298167	Asian Restaurant	0.00000	0.0	1
5	The Danforth West, Riverdale	43.679557	-79.352188	MenEssentials	43.677820	-79.351265	Cosmetics Shop	0.02381	0.0	0
6	The Danforth West, Riverdale	43.679557	-79.352188	Pantheon	43.677621	-79.351434	Greek Restaurant	0.02381	0.0	0
7	The Danforth West, Riverdale	43.679557	-79.352188	La Diperie	43.677702	-79.352265	Ice Cream Shop	0.02381	0.0	0
8	The Danforth West, Riverdale	43.679557	-79.352188	Dolce Gelato	43.677773	-79.351187	Ice Cream Shop	0.02381	0.0	0
9	The Danforth West, Riverdale	43.679557	-79.352188	Cafe Fiorentina	43.677743	-79.350115	Italian Restaurant	0.02381	0.0	0

Visualizing the clusters on the map

- The 3 clusters are represented by 3 different colors.
- The red color represents cluster 0 which has neighbourhoods with high number of Indian and Thai restaurants.
- The blue color represents cluster 1 which has neighbourhoods with very low number of Indian and Thai restaurants.
- The green color represents cluster 2 which has neighbourhoods with extremely high number of Thai restaurants.



Clusters

- Cluster 0 :

	Neighborhood	Indian Restaurant	Thai Restaurant	New Cluster Labels
0	Berczy Park	0.017241	0.017241	0
4	Central Bay Street	0.015625	0.015625	0
6	Church and Wellesley	0.013158	0.013158	0
7	Commerce Court, Victoria Hotel	0.000000	0.020000	0
8	Davisville	0.029412	0.029412	0
11	First Canadian Place, Underground city	0.000000	0.020000	0
25	Richmond, Adelaide, King	0.000000	0.030000	0
30	St. James Town, Cabbagetown	0.023256	0.023256	0
31	Stn A PO Boxes	0.010204	0.010204	0
32	Studio District	0.000000	0.025000	0
34	The Annex, North Midtown, Yorkville	0.047619	0.000000	0
36	The Danforth West, Riverdale	0.023810	0.000000	0

- Cluster 1 :

	Neighborhood	Indian Restaurant	Thai Restaurant	New Cluster Labels
14	Harbourfront East, Union Station, Toronto Islands	0.01	0.000000	1
29	St. James Town	0.00	0.011628	1
13	Garden District, Ryerson	0.00	0.010000	1
3	CN Tower, King and Spadina, Railway Lands, Har...	0.00	0.000000	1
5	Christie	0.00	0.000000	1
9	Davisville North	0.00	0.000000	1
10	Dufferin, Dovercourt Village	0.00	0.000000	1
12	Forest Hill North & West, Forest Hill Road Park	0.00	0.000000	1
16	India Bazaar, The Beaches West	0.00	0.000000	1
17	Kensington Market, Chinatown, Grange Park	0.00	0.000000	1

- Cluster 2 :

	Neighborhood	Indian Restaurant	Thai Restaurant	New Cluster Labels
15	High Park, The Junction South	0.0	0.08	2

Observations

- It is not recommended to open the restaurant in any neighbourhoods which are in Cluster 0 as that cluster already has more than enough Indian restaurants.
- Opening the restaurant in this cluster would result in the restaurant facing very difficult competition right from the beginning.
- Cluster 1 contains few Thai restaurants in some neighbourhoods, but even few Indian restaurants.
- The Indian restaurants are only situated in the neighbourhood of “Harbourfront East, Union Station, Toronto Islands”.
- “St. James Town” and “Garden District, Ryerson” are 2 neighbourhoods which have some Thai restaurants and there will be no competition as there are no Indian restaurants in that area.

- Cluster 2 has a very high number of Thai restaurants in that area.
- This is a very high number and it would be preferable to observe this neighbourhood for some time if other cuisines are also preferred by people staying there.
- It is not recommended to open a restaurant here at the moment.
- after studying the problem and observing the available data over the internet about neighbourhoods and venues, it is strongly recommended to open the Indian restaurant in “St. James Town” or “Garden District, Ryerson” areas.

Conclusion

- In this project, gone through data from various sources available over the internet and used the API services of Foursquare API to collect additional data about the places in the neighbourhoods of Toronto, Canada.
- Using this collected data, I managed to clean and format this data into a more condensed form in order to process it and used K-Means clustering to cluster the neighbourhoods based on venues in these neighbourhoods.
- Successfully segregated the neighbourhoods into different clusters and successfully identified potential places or areas for opening a new Indian restaurant.