IBM Applied Data Science Capstone Project

Recommending a location to open an Indian Restaurant in Toronto, Canada

Introduction/Business Problem:

In this Capstone Project, I am trying to suggest or recommend a suitable location for an entrepreneur to open an Indian restaurant in Toronto, Canada. The idea behind this project is that there is a large number of Indian people who are residing in Canada. Taking their taste and liking for Indian food into consideration, an entrepreneur is interested in opening his restaurant which focusses primarily on Indian cuisine. However there are many neighbourhoods in the area of Toronto and many of them already have restaurants and cafes in these areas. Also there are a lot of Asian people residing in these areas. The Indian cuisine being somewhat similar to what Asian people eat in terms of spices and condiments used in preparation would definitely attract them. With this in mind, the entrepreneur would need required help in choosing a suitable location for his Indian restaurant. The location can be an important factor in terms of popularity and gaining profit.

Data:

We need the following data to solve the problem:

1. The list of neighbourhoods which are in Toronto, Canada :
   This data will be extracted from Wikipedia page of Neighbourhoods in Canada using appropriate Python libraries.
2. The geospatial data (Latitudes and Longitudes) of the neighbourhoods :
   The latitudes and longitudes of the neighbourhoods can be listed down using Geocoder package in Python
3. Data about the different venues that are in the neighbourhoods of Toronto:
   This can be accessed using API services such as Foursquare API. Using this, API calls are made to access the data about the venues around a particular location by passing the coordinates of that location.

Methodology:

The first step which had to be done was collect the necessary data through different sources. For this, the data containing information about the neighbourhoods, boroughs along with postal codes of each borough in Canada was scraped from the Wikipedia page. This data was in the form of a table and was scraped using “read_html()” function of pandas library. The data was present in the table in the following format.

The below table shows the first 10 entries of neighbourhoods and their boroughs.

Table 1

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
7	M8A	Not assigned	Not assigned
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge

However in the above table there are a few entries wherein the neighbourhood and borough details are not available and instead their value is “Not assigned”. Every value which is ‘Not assigned’ is changed to ‘NaN’ value i.e. null value. This has been done so we can identify null values and remove the entries having null values if wanted. This process of eliminating ‘Not assigned’ values is very simple.

The below table shows the replaced NaN values in the place of ‘Not assigned’ values which were present in the borough and neighbourhood columns for the first 10 entries.

Table 2

	Postal Code	Borough	Neighbourhood
0	M1A	NaN	NaN
1	M2A	NaN	NaN
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
7	M8A	NaN	NaN
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge

Then the entries for which borough has NaN values are removed from the table and the neighbourhoods which have a NaN value is replaced with the value in their respective borough if the borough has a non-NaN value.

The resulting table after the changes for the first 10 entries is shown below.

Table 3

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

We also need the coordinates for each borough for this project. This is done with the help of csv file which was made available by IBM. It contains the latitudes and longitudes for each postal code in Canada.

The below table shows the first 10 entries of coordinates for the postal codes in Canada.

Table 4

	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
5	M1J	43.744734	-79.239476
6	M1K	43.727929	-79.262029
7	M1L	43.711112	-79.284577
8	M1M	43.716316	-79.239476
9	M1N	43.692657	-79.264848

Now we have the basic data for beginning the project. The Tables 3 and 4 are now merged together on the basis of postal code to get the below table.

Table 5

	Postal Code	Borough	Neighbourhood	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	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills	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

For this project, our focus is only on the neighbourhoods situated in Toronto. Hence we can remove the entries which do not belong to Toronto.

Thus we get a table consisting of only entries for Toronto. Below are the first 10 entries of this table.

Table 6

	Postal Code	Borough	Neighbourhood	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	India Bazaar, The Beaches West	43.668999	-79.315572
3	M4M	East Toronto	Studio District	43.659526	-79.340923
4	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
5	M4P	Central Toronto	Davisville North	43.712751	-79.390197
6	M4R	Central Toronto	North Toronto West, Lawrence Park	43.715383	-79.405678
7	M4S	Central Toronto	Davisville	43.704324	-79.388790
8	M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160
9	M4V	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049

Now that we have the table containing the neighbourhoods in Toronto along with their coordinates, we will use these coordinates to get nearby venues in these neighbourhoods. For this we will have to make API calls to retrieve this data over internet. We will use Foursquare API which itself is a RESTful set of addresses to which you can send requests. The output which is retrieved is in JSON format. The size of this JSON data depends on the number of requests you make for each neighbourhood and the number of venues returned for that neighbourhood. For our dataset, the JSON data was very large as each neighbourhood had almost more than 15-20 venues located there. The data for each venue contains many attributes. However we need only few of them for our problem. We will need the name of that venue, the category under which this venue falls e.g. ‘Coffee Shop’ and the coordinates where that venue is located in the neighbourhood. After we get this data, we will map it with the data that we collected in Table 6. We will also drop the ‘Postal Code’ column as we would not require it hereafter.

After mapping the venue data, we would get a table as shown below.

Table 7

	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

Now that we have all the data we required for our problem we can perform some exploratory data analysis. We can group our data on the basis of neighbourhood and check the number of venues under each neighbourhood using the count() function. Also we check the number of unique venue categories in our data and which venue category occurs the most number of times in a neighbourhood. For our data there are about 230 unique venue categories. We can use any clustering algorithm here to group neighbourhoods based on their venue categories. We will use K-Means clustering algorithm which is relatively simple clustering technique.

The main idea of K-Means clustering algorithm is to define k centers, one for each cluster. These centers should be placed in an appropriate way because different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

We will use the ‘scikit-learn’ library of python to implement K-Means clustering. However we cannot train our data directly on the K-Means model. Our data has string values for venue categories. We have to convert these string values to integer values otherwise the model cannot process our data. We will use ‘One-Hot Encoding’ technique to do this. Using this we will create a new dataframe consisting of neighbourhoods and columns of each venue category. We had found out earlier that there are more than 230 venue categories, so we get a column for each unique venue category. These columns have values 0 or 1. The value is 0 if a venue of that category is not present in that neighbourhood and the value is 1 if the venue of that category is present in that neighbourhood.

The below table is the resulting table showing first 10 entries.

Table 8

	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	Video Game Store
0	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0
1	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
4	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
5	The Danforth West, Riverdale	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
6	The Danforth West, Riverdale	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
7	The Danforth West, Riverdale	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
8	The Danforth West, Riverdale	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
9	The Danforth West, Riverdale	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

Now the mean number of occurrences of each category of venue for respective neighbourhood is calculated. The venue categories are first grouped together based on the neighbourhoods.

Below is the table for first 10 neighbourhoods.

Table 9

	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

There are many columns in this dataframe, but for our problem we need to check those columns which are related to restaurants and cafes. There are around 34 different types of restaurants in Toronto. We need to check for Indian restaurants and Thai restaurants. We have considered Thai restaurants for this problem in place of Asian restaurants as there were more Thai restaurants than Asian and Indian food being similar to Thai in terms of spice and flavour.

We will need the data of the mean number of occurrences of Indian and Thai restaurant in the nieghbourhoods. We will use this data to train our K-Means model and form clusters.

Below is the table having first 10 entries for mean occurrences of Indian and Thai restaurants.

Table 10

	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

Now we can train our K-Means model on this data to cluster the neighbourhoods. However we also need to decide the number of clusters needed for our data. To find the number of clusters, we will use ‘Elbow Method’.

For this we will iterate the process of training our data and forming cluster labels for a range of values of k i.e. number of clusters. We will find the squared errors for each value of k and plot it on a graph.

Below is the graph for 10 values of k and their squared errors.

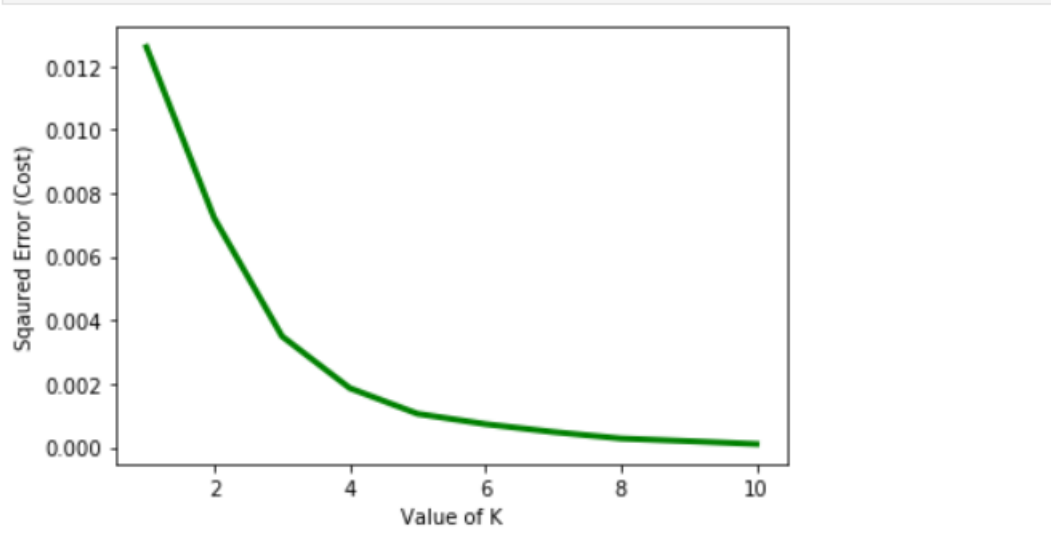


Figure 1

To determine the optimal number of clusters, we have to select the value of k at the "elbow" i.e. the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is **3**.

So we cluster the neighbourhoods into 3 clusters. We can now assign the cluster labels to each neighbourhood and check the occurrence of Indian restaurants in the cluster and check with number of Thai restaurants in that cluster. Accordingly we can be able to recommend suitable places for opening an Indian restaurant. The place will be such that there are not too many Indian restaurants because it opening in such area would prove to be competitive from the beginning. The place will also depend whether there any Thai restaurants along with few Indian restaurants. Thus we can recommend a place on the basis of our analysis.

Results:

After training our model to our data, we get cluster labels for our neighbourhoods. The neighbourhoods are mapped to our dataframe from Table 7.

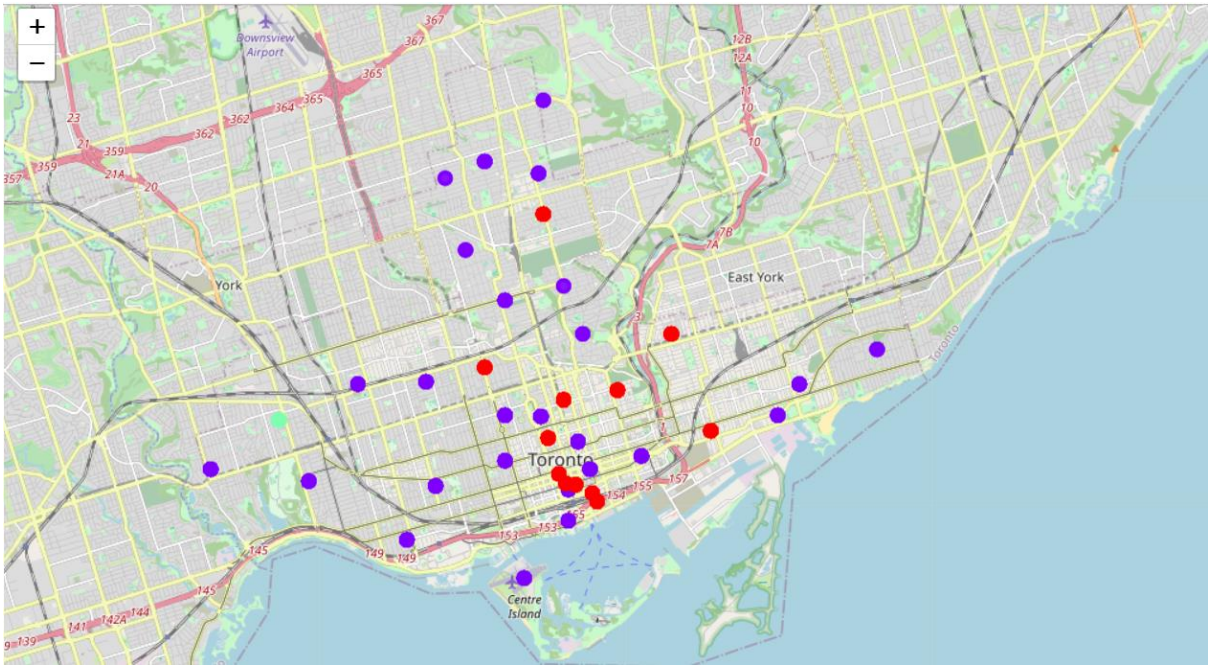
Below is the table of neighbourhoods with cluster labels.

Table 11

	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

The above table also contains the mean occurrences of Indian and Thai restaurants in a neighbourhood for reference.

The neighbourhoods in Toronto are plotted on the map of Canada as per the clusters as shown below



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

The 3 clusters are :

Cluster 0. They are represented by Red colour in the above map. This cluster contains neighbourhoods which have a very high number of Indian and Thai restaurants. The below table shows the neighbourhoods which come under this cluster along with the mean number of Indian and Thai restaurants in these neighbourhoods.

Table 12

	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. They are represented by Blue colour in the above map. This cluster contains neighbourhoods which don't have as many number of Indian and Thai restaurants as compared to Cluster 0. The below table shows the neighbourhoods which come under this cluster along with the mean number of Indian and Thai restaurants in these neighbourhoods

Table 13

	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. They are represented by Green colour in the above map. This cluster contains neighbourhoods which don't have any Indian restaurants, but have the highest number of Thai restaurants as compared to both Cluster 0 and Cluster 1. The below table shows the neighbourhoods which come under this cluster along with the mean number of Indian and Thai restaurants in these neighbourhoods

Table 14

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

Discussion:

Based on the results which we got after training our data and clustering the neighbourhoods, we can say that we got a better understanding of the neighbourhoods and the preferences of people in those neighbourhoods which reflected in the types of restaurants in these neighbourhoods. After analysing these results, we have 2-3 neighbourhoods which can prove beneficial for opening an Indian restaurant.

It is not recommended to open the restaurant in any neighbourhoods which are in Cluster 0. That cluster already has more than enough Indian restaurants. It also has a high number of Thai restaurants which further proves the point of the tastes being similar between these cuisines. Opening the restaurant in this cluster would result in the restaurant facing very difficult competition right from the beginning. Now coming to the Cluster 1, we notice that it 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". In this cluster there are 2 neighbourhoods which have some Thai restaurants. There will be no competition as there are no Indian restaurants in that area. It would give that restaurant the opportunity and appropriate time to grow and flourish its business in that area. These 2 neighbourhoods are "St. James Town" and "Garden District, Ryerson". Lastly for Cluster 2, there is a very high number of Thai restaurants in this area. But this number is very high and it would be preferable to observe this neighbourhood for some time if other cuisines are also preferred by people staying there. At the moment it seems that most of the people prefer this cuisine and there are not many restaurants of other cuisines or categories. So in future if some new restaurants are opened here, then at that time an Indian restaurant can be opened here. However it is not recommended to open a restaurant here at the moment.

So 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.

Limitations/Suggestions:

After completing this project, I realised that some more data about the neighbourhoods and venues would have proven to be beneficial. We could have used data about the population of the neighbourhoods, the age group and ethnicity of people living there. Also the annual income of people and the prices of land in that neighbourhood would have been helpful. We would have been able to cluster the neighbourhoods even more accurately. These things would be taken into consideration next time and I would continue working on this to improve the existing results of my work.

Conclusion:

Hence, in this project I have 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 which is an unsupervised learning algorithm to cluster the neighbourhoods based on venues in these neighbourhoods. On doing so, I successfully segregated the neighbourhoods into different clusters and successfully identified potential places or areas for opening a new Indian restaurant.

References:

- List of neighbourhoods in Canada:
  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- List of coordinates of the neighbourhoods:
  http://cocl.us/Geospatial_data

- Foursquare API documentation:
  https://developer.foursquare.com/docs