EXPLORING TORONTO NEIGHBORHOODS TO IDENTIFY A SUITABLE LOCATION FOR A NEW INDIAN RESTAURANT

By Amy Reidy, November 2020

CAPSTONE PROJECT FOR IBM'S DATA SCIENCE PROFESSIONAL CERTIFICATE

Table of contents

1. Introduction
2. Data
3. Methodology
4. Results
5. Discussion
6. Conclusion

---

1. Introduction

1.1 Background

This project is the final assignment of the Capstone Project Course for IBM's Data Science Professional Certificate. Throughout the previous 8 courses, myself and other aspiring data scientists have learned the fundamentals of Data Science, including data science methodology, data analysis and visualisation with Python, and machine learning models, such as K-means. This project is therefore an opportunity to apply all our new knowledge to a hands-on project.

In the previous assignment in this course, we learned how to use the Foursquare API to explore Toronto neighborhoods and to cluster them into different groups based on the most common venues. Building on this exercise, I have decided to continue to explore Toronto – a city that I lived in for two years – to investigate where would be a suitable location for a new Indian restaurant, based on the density of existing Indian restaurants and the number of Indian residents in each neighborhood.

1.2 Problem Statement

"Exploring Toronto Neighborhoods to Identify a Suitable Location for a New Indian Restaurant"

1.3 Problem Description

Toronto is the most populous city in Canada, and in 2016, 51% of the Toronto residents belonged to a visible minority group (including 32% South Asian residents), making it one of the most multicultural and multiracial cities in the world. Unsurprisingly, it also boasts a remarkably diverse food scene and has many different ethnic neighborhoods where you can sample cuisines from around the world, including Little Portugal, Little Italy, Koreatown, Little India, and multiple Chinatowns.

I have decided to investigate and predict where would be the best neighborhood to open a new Indian restaurant not only because it was my favourite takeout food to indulge in when I lived in Toronto, but also, as I believe this could be a very promising business venture as the number of Indian residents in Canada is rapidly growing, and in 2016, almost 51% of the entire Indian Canadian community resided in the Greater Toronto Area (GTA).

There has been a reported 105% increase from 2016 to 2019 in the number of Indians being admitted to Canada as permanent residents , with also a 127% increase of international students from India from 2016 to 2018. This rise could be partly due to large numbers of young Indian tech workers moving to Canada, as it is

becoming increasingly difficult to obtain permanent residence in the United States. Furthermore, Indians accounted for a quarter of new permanent residents permitted to Canada in 2019 , and there is now an estimated 1 million Indians living in Canada. So now more than ever would be a fantastic time to open a new Indian restaurant and profit from this recent surge in the Indian population.

To investigate which neighborhoods could be the optimal location for a new Indian restaurant, I will first explore the population distributions in top visible minority groups in neighborhoods and the relationship between venues in the neighborhoods. Then I will identify the relationship between neighborhoods and existing Indian restaurants, and the relationship between East Indian population and Indian restaurants. And finally, I will use clustering analysis to identify suitable location for opening a new Indian restaurant.

1.4 Interest

This project is targeted at the following stakeholders:

- Entrepreneurs who are interested in opening a new Indian restaurant in Toronto.
- Indian restaurant owners who may be interested in relocating or expanding to another location.
- Food truck owners who serve Indian cuisine and are looking for prime locations to target.

Other stakeholders who may be interested in the results of this project:

- Indian individuals and families who are interested in moving to Toronto and would like to know which areas have the highest population of Indian residents and/or greatest density of Indian restaurants.

2. Data

2.1 Data Sources

- I extracted spatial information, such as latitude and longitude of Toronto neighborhoods from "https://cocl.us/Geospatial_data"
- I used data from the Wikipedia page "List of Postal code of Canada: M" (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) to get information about the neighborhoods in Toronto, for example their name, postal code and borough.
- I extracted information about various venues in Toronto and their spatial information using Foursquare's API. More details about the API can be found on https://developer.foursquare.com/docs. I was interested in learning about existing Indian restaurants. Therefore, I extracted the number of venues listed as Indian restaurants in each neighborhood so I could explore the relationship between neighborhoods and Indian restaurants.
- Although I mentioned that there has been an 125% increase between 2016 to 2019 in the number of Indians being admitted to Canada as permanent residents, unfortunately the most recent dataset I could find for the ethnic makeup of Toronto populations is from 2016. I extracted this population information from the Wikipedia page https://en.wikipedia.org/wiki/Demographics_of_Toronto. I used this demographic data to get population information about East Indian ethnic group in Toronto

neighborhoods. This information together with the Foursquare data about existing Indian restaurants allowed me to explore the relationship between the Indian population and existing restaurants.

2.2 Summary of Pre-processing and Data Cleaning

The pre-processing and data cleaning involved scraping Toronto neighborhood data, extracting latitude and longitude of neighborhoods, scraping venue items, extracting Indian population in neighborhoods and number of existing Indian restaurants. All the extracted data was converted into dataframes for further analysis. Each pre-processing and data cleaning step is discussed and documented as markdowns together with code in the Jupyter notebook.

3. Methodology

3.1 Extracting data about Toronto postal codes from Wikipedia.

The first step was to extract a table of postal codes in Toronto from this Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) and convert into a dataframe.

To clean the dataframe, I renamed some of the column names, dropped the rows where the value for 'Borough' was 'Not assigned, combined neighborhoods which have the same postal code and added the borough value to the neighborhoods where the value was 'Not assigned.

This resulted in the below dataframe with 103 rows and 3 columns.

	Borough	PostalCode	Neighborhood
0	Central Toronto	M4N	Lawrence Park
1	Central Toronto	M4P	Davisville North
2	Central Toronto	M4R	North Toronto West, Lawrence Park
3	Central Toronto	M4S	Davisville
4	Central Toronto	M4T	Moore Park, Summerhill East

3.2 Importing geospatial information (latitude and longitude) of neighborhoods and visualizing on a map

Next I imported geospatial data (latitude and longitude) from https://cocl.us/Geospatial_data. I then changed the postal code column name and merged the latitude and longitude of locations to the dataframe.

	Borough	PostalCode	Neighborhood	Latitude	Longitude
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197
2	Central Toronto	M4R	North Toronto West, Lawrence Park	43.715383	-79.405678
3	Central Toronto	M4S	Davisville	43.704324	-79.388790
4	Central Toronto	M4T	Moore Park, Summerhill East	43.689574	-79.383160
...	...	...	...	...	...
98	York	M6C	Humewood-Cedarvale	43.693781	-79.428191
99	York	M6E	Caledonia-Fairbanks	43.689026	-79.453512
100	York	M6M	Del Ray, Mount Dennis, Keelsdale and Silverthorn	43.691116	-79.476013
101	York	M6N	Runnymede, The Junction North	43.673185	-79.487262
102	York	M9N	Weston	43.706876	-79.518188

103 rows × 5 columns

Once I had the latitude and longitude for each neighborhood, I could use Nominatim for the geopy library to find the geographical coordinates for Toronto, and then I used folium to visualize the neighborhoods on a map of Toronto.

3.3 Extracting demographics information from Wikipedia.

My next task was to extract demographics data of Toronto ridings (collection of neighborhoods) from the 'Demographics of Toronto' Wikipedia page ('https://en.wikipedia.org/wiki/Demographics_of_Toronto'). I

extracted four tables which list the top 10 ethnic groups in East York, North York, Scarborough, and Etobicoke and York. The below table shows the top 10 ethnic groups in East York.

	Riding	Population	Ethnic Origin #1	Ethnic Origin 1 in %	Ethnic Origin #2	Ethnic Origin 2 in %	Ethnic Origin #3	Ethnic Origin 3 in %	Ethnic Origin #4	Ethnic Origin 4 in %	Ethnic Origin #5	Ethnic Origin 5 in %	Ethnic Origin #6	Ethnic Origin 6 in %	Ethnic Origin #7	Ethnic Origin 7 in %	Eth Ori
0	Spadina-Fort York	114315	English	16.4	Chinese	16.0	Irish	14.6	Canadian	14.0	Scottish	13.2	French	7.70	German	7.6	I
1	Beaches-East York	108435	English	24.2	Irish	19.9	Canadian	19.7	Scottish	18.9	French	8.7	German	8.40	NaN	NaN	I
2	Davenport	107395	Portuguese	22.7	English	13.6	Canadian	12.8	Irish	11.5	Italian	11.1	Scottish	11.00	NaN	NaN	I
3	Parkdale-High Park	106445	English	22.3	Irish	20.0	Scottish	18.7	Canadian	16.1	German	9.8	French	8.88	Polish	8.5	I
4	Toronto-Danforth	105395	English	22.9	Irish	19.5	Scottish	18.7	Canadian	18.4	Chinese	13.8	French	8.86	German	8.8	Gr
5	Toronto-St. Paul's	104940	English	18.5	Canadian	16.1	Irish	15.2	Scottish	14.8	Polish	10.3	German	7.90	Russian	7.7	Ita
6	University-Rosedale	100520	English	20.6	Irish	16.6	Scottish	16.3	Canadian	15.2	Chinese	14.7	German	8.70	French	7.7	Ita
7	Toronto Centre	99590	English	15.7	Canadian	13.7	Irish	13.4	Scottish	12.6	Chinese	12.5	French	7.20	NaN	NaN	I

3.4 Extracting location and venue information using Foursquare API

The Foursquare API was used to retrieve information about Indian restaurants present in the neighborhoods of Toronto. The API requires credentials such as client_id, client_secret, version and limit. I was interested in Indian restaurants for each neighborhood, so I also used the category Id to specify that I only wanted to retrieve information about Indian restaurants.

```
def getNearbyVenues(names, latitudes, longitudes, radius=500): #start with radius set at 500m

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        #print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&categoryId={}&radius=
            CLIENT_ID,
            CLIENT_SECRET,
            lat,
            lng,
            VERSION,
            CATEGORYID,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

I alternated the radius length between 500m, 750m and 1km to find the best fit. I decided to use 750m for the radius which gave me a dataframe with 238 rows, but only 145 unique restaurants meaning that many of the restaurants were repeated as they are in the vicinity of multiple neighborhoods. When I

limited the radius to 500m, I only retrieved 74 unique restaurants, so I decided to use the larger dataset from the 750m radius (see below).

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Davisville North	43.712751	-79.390197	Banjara Indian Cuisine	43.707810	-79.393296	Indian Restaurant
1	Davisville	43.704324	-79.388790	Marigold Indian Bistro	43.702881	-79.388008	Indian Restaurant
2	Davisville	43.704324	-79.388790	Banjara Indian Cuisine	43.707810	-79.393296	Indian Restaurant
3	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049	Bombay Roti	43.686661	-79.393705	Indian Restaurant
4	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049	Bombay Roti	43.686650	-79.393630	Indian Restaurant
...	...	...	...	...	...	...	...
233	Humewood-Cedarvale	43.693781	-79.428191	Indus Tavern	43.700117	-79.429157	Indian Restaurant
234	Humewood-Cedarvale	43.693781	-79.428191	The Great Maratha	43.700043	-79.429320	Indian Restaurant
235	Humewood-Cedarvale	43.693781	-79.428191	Empire Grill	43.699852	-79.431782	Indian Restaurant
236	Caledonia-Fairbanks	43.689026	-79.453512	House of Indian Roti	43.694887	-79.449981	Indian Restaurant
237	Runnymede, The Junction North	43.673185	-79.487262	Roti Time	43.668211	-79.487152	Indian Restaurant

238 rows × 7 columns

```
print('There are {} uniques categories.'.format(len(toronto_venues3['Venue'].unique())))
```

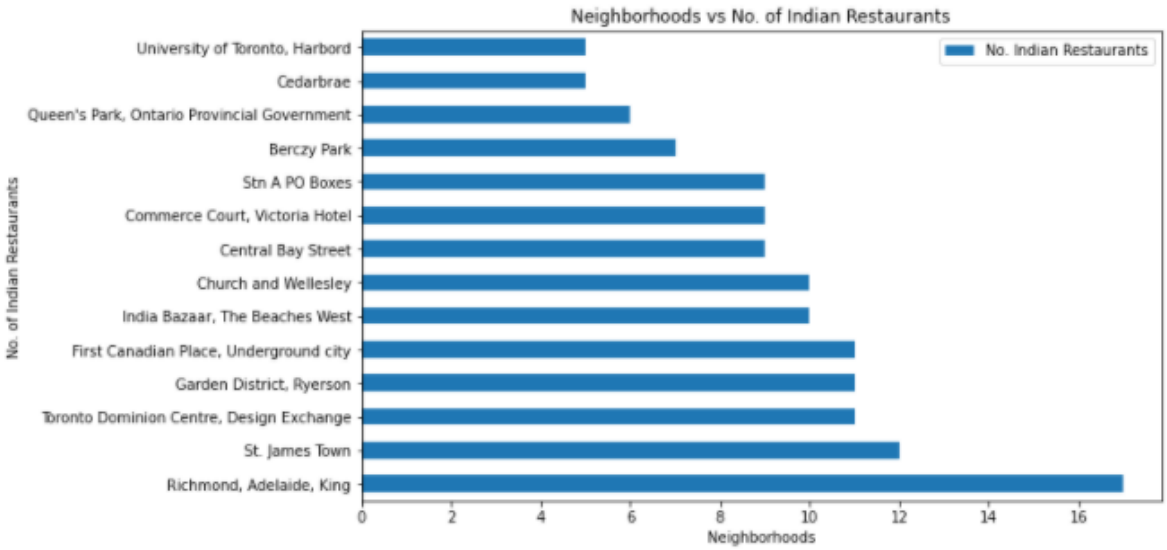
```
There are 145 uniques categories.
```

3.5 Exploring the relationship between neighborhoods and Indian restaurants

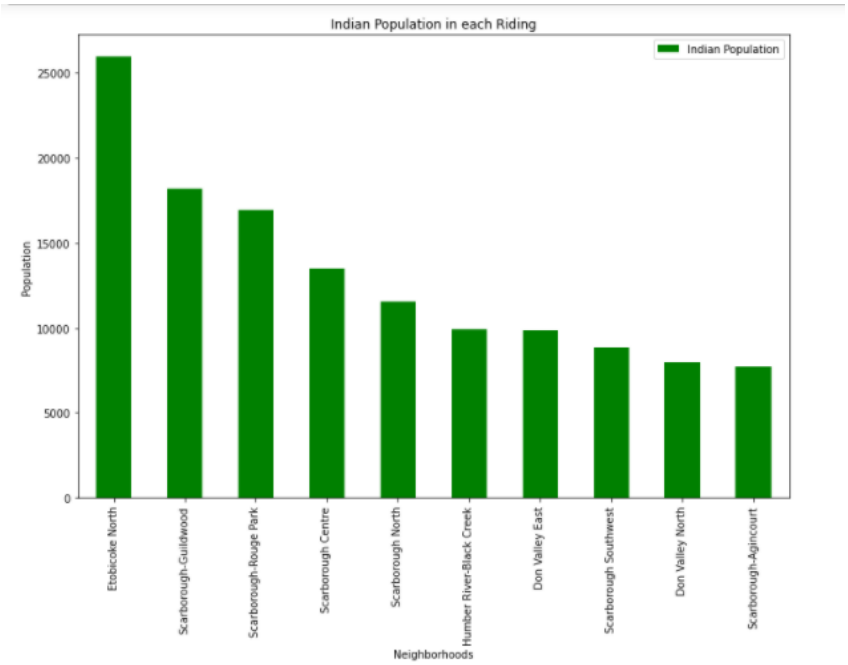
From the dataset I retrieved from Foursquare, I created a new dataframe containing how many Indian restaurants are in each neighborhood.

	Neighborhood	No. Indian Restaurants
0	Agincourt	1
1	Alderwood, Long Branch	1
2	Bedford Park, Lawrence Manor East	4
3	Berczy Park	7
4	Birch Cliff, Cliffside West	1
5	Brockton, Parkdale Village, Exhibition Place	2
6	Business reply mail Processing Centre, South C...	2
7	Caledonia-Fairbanks	1
8	Canada Post Gateway Processing Centre	2
9	Cedarbrae	5
10	Central Bay Street	9
11	Christie	2
12	Church and Wellesley	10
13	Clarks Corners, Tam O'Shanter, Sullivan	3
14	Commerce Court, Victoria Hotel	9
15	Davisville	2
16	Davisville North	1
17	Don Mills	2
18	Dorset Park, Wexford Heights, Scarborough Town...	3
19	East Toronto, Broadview North (Old East York)	2

Then I used matplotlib and seaborn to create a bar chart showing the neighborhoods with the most Indian restaurants nearby.



Next, I merged the dataframes containing the ethnic population distribution for East York, North York, Scarborough, and Etobicoke and York, picked out the ridings with highest Indian population percentage, and visualized them on a bar chart.



After this I got a list of neighborhoods present in the riding using the Wikipedia Geography Section for each riding. I first needed to alter the riding names to match the Wikipedia page so I could retrieve the neighborhoods present in those ridings and merge this to my existing dataframe.

	Riding	Indian Population	Neighborhoods
0	Don Valley North	7961.380	Henry Farm, Bayview Village, Bayview Woods-St...
1	Humber River-Black Creek	9910.700	Humber Summit, Humbermede, Humberlea, York Un...
2	Don Valley East	9876.020	Flemingdon Park, Don Mills, Graydon Hall, Par...
3	Scarborough Southwest	8880.190	Birch Cliff, Oakridge, Cliffside, Kennedy Par...
4	Scarborough-Agincourt	7712.650	Steeles, L'Amoreaux, Tam O'Shanter-Sullivan, ...
5	Scarborough-Rouge Park	16941.315	Morningside Heights, Rouge, Port Union, West ...
6	Scarborough-Guildwood	18200.700	Guildwood, West Hill (west of Morningside Ave...
7	Etobicoke North	25965.120	The Elms, Humberwood, Kingsview Village, This...

Next, I split the neighborhoods in each riding into individual rows and added the number of Indian restaurants that are in each of these neighborhoods.

	Indian Population	Neighborhood	No. Indian Restaurants
4	25965.120	Thistletown	3
3	18200.700	Morningside	1
2	16941.315	Rouge	1
1	8880.190	Oakridge	1
5	8880.190	Clairlea	1
0	7961.380	Henry Farm	2

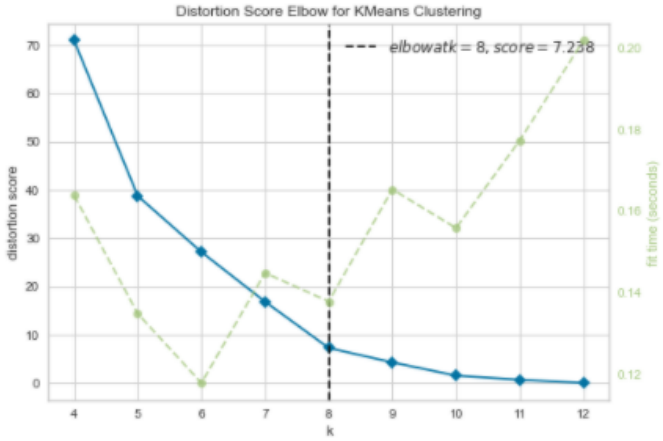
We can see from this dataframe that there does not seem to be a strong relationship between the number of Indian restaurants in a neighborhood and the number of Indian residents in this riding. It is worth noting that Morningside, Rouge, Oakridge and Clairlea are all in the Scarborough area and there is only one Indian restaurant nearby each of these neighborhoods despite the high number of Indians living in this borough.

Also, of the 6 neighborhoods in the Etobicoke North riding, only one neighborhood (Thistletown) has Indian restaurants, even though Etobicoke North is the riding with the greatest numbers of Indian residents.

3.6 Predictive Modelling

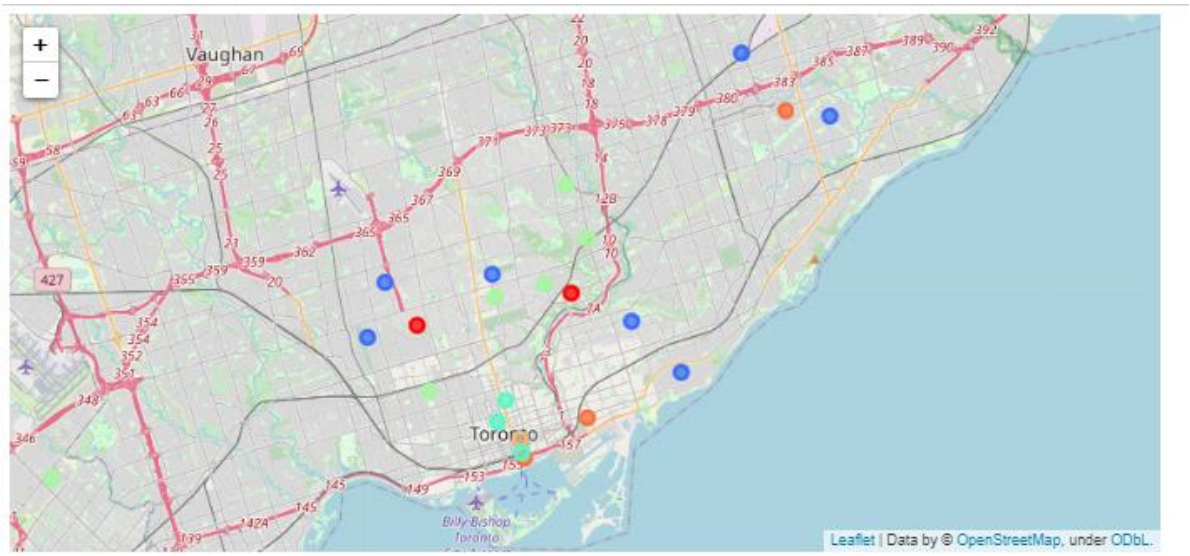
For the final part of my investigation, I used the K-means machine learning model to group neighborhoods into clusters based on the density of Indian restaurants in each neighborhood. Please note that I only used neighborhoods which had at least one Indian restaurants within 750 metres.

The first step in K-means clustering was to identify best K value, meaning the optimal number of clusters in a dataset. To do so, I used the elbow method on the dataset showing the number of Indian restaurants in near each neighborhood.



As you can see from the above chart which shows the distortion score and squared error for each K value, K = 8 is the best value.

So, I clustered the Toronto neighborhoods with K-means K = 8, assigned cluster labels to each neighborhood, added their geographical coordinates, and visualized the results on the map below.



Now let us examine the resulting clusters:

- **Cluster 3** – This cluster has no rows of data, meaning no data points or neighborhoods were near to this centroid.

- **Cluster 1 and 4** contain the neighborhoods which are the most densely populated by Indian restaurants, and we can see all of these are in Downtown Toronto. Cluster 1 is pale orange and Cluster 4 is turquoise color on the map.

	Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	No. Indian Restaurants
3	Downtown Toronto	M5C	St. James Town	43.651494	-79.375418	1.0	12.0

	Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	No. Indian Restaurants
2	Downtown Toronto	M4Y	Church and Wellesley	43.665860	-79.383160	4.0	10.0
6	Downtown Toronto	M5G	Central Bay Street	43.657952	-79.387383	4.0	9.0
7	Downtown Toronto	M5W	Stn A PO Boxes	43.646435	-79.374846	4.0	9.0

- Neighborhoods in **Cluster 7** are medium to highly populated by Indian restaurants. This cluster is orange on the map**.**

	Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	No. Indian Restaurants
5	Downtown Toronto	M5E	Berczy Park	43.644771	-79.373306	7.0	7.0
10	East Toronto	M4M	Studio District	43.659526	-79.340923	7.0	5.0
19	Scarborough	M1H	Cedarbrae	43.773136	-79.239476	7.0	5.0

- Neighborhoods in **Cluster 0** are medium populated by Indian restaurants. Cluster 0 is red on the map.

	Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	No. Indian Restaurants
13	East York	M4H	Thorncliffe Park	43.705369	-79.349372	0.0	4.0
21	York	M6C	Humewood-Cedarvale	43.693781	-79.428191	0.0	4.0

- We can see St. James Town is the only neighborhood in **Cluster 6**, however this neighborhood is also listed in Cluster 1 as having 12 Indian restaurants in the vicinity so we can disregard this cluster.

	Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	No. Indian Restaurants
4	Downtown Toronto	M5C	St. James Town	43.651494	-79.375418	6.0	3.0

- **Cluster 5** contains neighborhoods which are sparsely populated with Indian restaurants. It is shown in lime green color on the map.

	Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	No. Indian Restaurants
1	Central Toronto	M4S	Davisville	43.704324	-79.388790	5.0	2.0
8	Downtown Toronto	M6G	Christie	43.669542	-79.422564	5.0	2.0
12	East York	M4G	Leaside	43.709060	-79.363452	5.0	2.0
14	Mississauga	M7R	Canada Post Gateway Processing Centre	43.636966	-79.615819	5.0	2.0
15	North York	M3B	Don Mills	43.745906	-79.352188	5.0	2.0
16	North York	M3C	Don Mills	43.725900	-79.340923	5.0	2.0

- **Cluster 2** contains the neighborhoods with the least number of Indian restaurants in the vicinity. It is shown in blue color on the map. This cluster shows neighborhoods which may be suitable for opening a new Indian restaurant as there is low competition here.

	Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	No. Indian Restaurants
0	Central Toronto	M4P	Davisville North	43.712751	-79.390197	2.0	1.0
9	East Toronto	M4E	The Beaches	43.676357	-79.293031	2.0	1.0
11	East York	M4C	Woodbine Heights	43.695344	-79.318389	2.0	1.0
17	North York	M6B	Glencairn	43.709577	-79.445073	2.0	1.0
18	Scarborough	M1G	Woburn	43.770992	-79.216917	2.0	1.0
20	Scarborough	M1S	Agincourt	43.794200	-79.262029	2.0	1.0
22	York	M6E	Caledonia-Fairbanks	43.689026	-79.453512	2.0	1.0

4. Results

We have come to the end of our analysis, and in this section, I will discuss my findings. To begin this project, I introduced the business problem of identifying a suitable neighborhood to open a new Indian restaurant in Toronto. To achieve this, I investigated all the neighborhoods in Toronto, analysed the Indian population in each riding and the number of Indian restaurants in each neighborhood to come to a conclusion about which neighborhood would be an ideal location. I used variety of data sources to set up a very realistic data-analysis scenario.

My findings:

- Of all the ridings, Etobicoke North, Scarborough-Guildwood, Scarborough-Rouge Park, Scarborough Centre, Scarborough North are the most densely populated by the East Indian ethnic group.
- By examining the clusters and bar chart, it looks like Downtown Toronto, Central Toronto, East York are already densely populated with Indian restaurants, so I believe it is a better idea to avoid opening another restaurant in these areas and consider only Scarborough, Etobicoke & North York for the new restaurant's location.
- After careful consideration, I think it is a good idea to open a new Indian restaurant in Etobicoke North as it is the riding with the largest Indian population, but there are only 3 Indian restaurants in one of the 6 neighborhoods in Etobicoke North, so there doesn't seem to be too much competition there.
- The Scarborough area would also be an excellent option since it too has a high Indian population so there are more possible interested customers than other areas, and there would be low competition since there is not a high number of Indian restaurants in these neighborhoods.

5. Discussion

According to this analysis, Etobicoke North or Scarborough seem to be very suitable locations to open a new Indian restaurant, as they both have high numbers of Indians living there, which means

there could be a strong customer base. And they also have low numbers of existing Indian restaurants so there should not be a great deal of competition in these areas.

While my analysis showed these areas as possibly the most suitable locations for a new restaurant in Toronto, there were also some limitations to my analysis, for example:

- The data I used for the Indian population distribution is from 2016, and as I mentioned the number of Indians being admitted into Canada as permanent residents has doubled in the past few years, so we cannot assume that the distribution remains the same today, and it's very possible that some areas have become more popular since 2016.
- The clustering is based only on data obtained from Foursquare API. And my search results really depended on the radius I chose. In some neighborhoods, it would have been better to have a radius larger that 750m, but when I applied a 1 km radius to my request, this resulted into many restaurants overlapping into multiple neighborhoods.
- There are many other factors that are important to consider when deciding where to open a new restaurant such as the cost of rent and the amount of foot traffic in the area, however analysing these features are beyond the scope of this capstone project.
- As a beginner to data science, I'm still very new to data visualisation and analysis, and if I had greater skills in these areas, I would have liked to try to visualize the Indian population on a geographical heatmap to better show the spread of the Indian population over Toronto. I would also liked to have tried other forms of machine learning for my analysis, but K-means was the main type of machine learning that was introduced in this course.

Although there are lots of areas where it can be improved, I think overall this analysis has still provided us with some good insights into potential locations which could be ideal for opening a new Indian restaurant.

6. Conclusion

In this capstone project, we were given the opportunity to work on a typical business problem that a data scientist might try to solve. I was able to utilize most of the key learning outcomes from the past eight courses, including using Python libraries to fetch data, and to manipulate, analyse and visualize these dataframes. I have used Foursquare API to search for Indian restaurants in neighborhoods in Toronto, and I have gotten data from Wikipedia by scraping pages with help from the Wikipedia python library. I also applied the K-means machine learning technique to predict the output given the data and used Folium to visualize it on a map. Some areas of improvements for future analysis include using more data from more diverse sources and utilizing different machine learning techniques.

Finally, while this project investigated the most ideal location to open a new Indian restaurant in Toronto, this type of analysis could also be applied to other kinds of business ventures, like where to open a Asian supermarket, and to other cities around the world.