As a part of the IBM Data Science professional program Capstone Project, we worked on the real datasets to get an experience of what a data scientist goes through in real life. Main objectives of this project were to define a business problem, look for data in the web and, use Foursquare location data to compare different neighborhoods of Toronto to figure out which neighborhood is suitable for starting a new restaurant business. In this project, we will go through all the process in a step by step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

1. Description of the Business Problem & Discussion of the Background :

Problem Statement : Prospects of opening an Indian Restaurant in Toronto, Canada.

Toronto, the capital of the province of Ontario, is the most populous Canadian city. Its diversity is reflected in Toronto's ethnic neighborhoods such as Chinatown, Corso Italia, Greektown, Kensington Market, Koreatown, Little India, Little Italy, Little Jamaica, Little Portugal & Roncesvalles. One of the most immigrant-friendly cities in North America with more than half of the entire Indian Canadian population residing in Toronto it is one of the best places to start an Indian restaurant.

In this project we will go through step by step process to make a decision whether it is a good idea to open an Indian restaurant. We analyze the neighborhoods in Toronto to identify the most profitable area since the success of the restaurant depends on the people and ambience. Since we already know that Toronto shelter a greater number of Indians than any other city in Canada, it is a good idea to start the restaurant here, but we just need to make sure whether it is a profitable idea or not. If so, where we can place it, so it yields more profit to the owner.

Target Audience

Who will be more interested in this project ? What type of clients or a group of people would be benefitted ?

1. Business personnel who want to invest or open an Indian restaurant in Toronto. This analysis will be a comprehensive guide to start or expand restaurants targeting the Indian crowd.
2. Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.
3. Indian crowd who wants to find neighborhoods with lots of options for Indian restaurants.
4. Business Analyst or Data Scientists, who wish to analyze the neighborhoods of Toronto using Exploratory Data Analysis and other statistical & machine learning techniques to obtain all the necessary data, perform some operations on it and, finally be able to tell a story out of it.

2. Data acquisition and cleaning :

2.1 Data Sources

a) I'm using "List of Postal code of Canada: M" (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) wiki page to get all the information about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.

b) Then I'm using "https://cocl.us/Geospatial_data" csv file to get all the geographical coordinates of the neighborhoods.

c) To get information about the distribution of population by their ethnicity I'm using "Demographics of Toronto" (https://en.m.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity) wiki page. Using this page I'm going to identify the neighborhoods which are densely populated with Indians as it might be helpful in identifying the suitable neighborhood to open a new Indian restaurant.

d) To get location and other information about various venues in Toronto I'm using Foursquare's explore API. Using the Foursquare's explore API (which gives venues recommendations), I'm fetching details about the venues up present in Toronto and collected their names, categories and locations (latitude and longitude).

From Foursquare API (https://developer.foursquare.com/docs), I retrieved the following for each venue:

- Name: The name of the venue.
- Category: The category type as defined by the API.
- Latitude: The latitude value of the venue.
- Longitude: The longitude value of the venue.

2.2 Data Cleaning

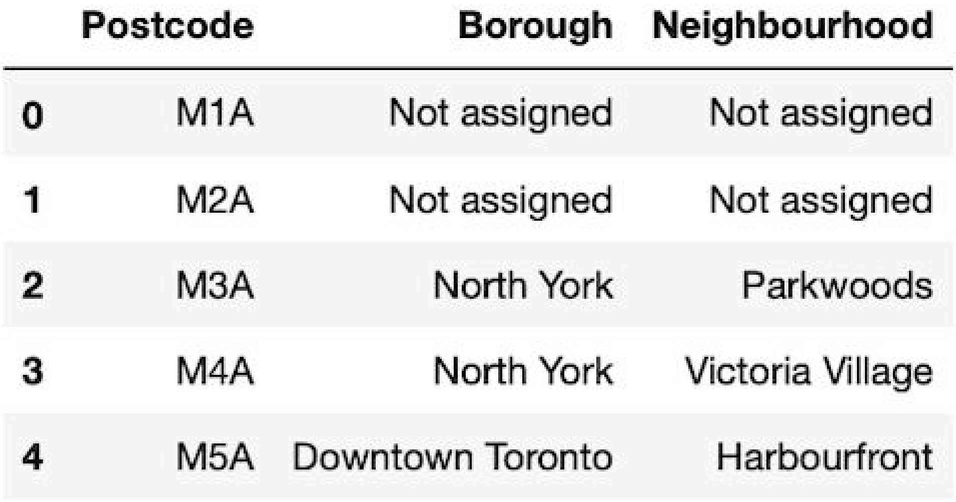
a) Scraping Toronto Neighborhoods Table from Wikipedia

Scraped the following Wikipedia page, "*List of Postal code of Canada: M*" in order to obtain the data about the Toronto & the Neighborhoods in it.

Assumptions made to attain the below DataFrame:

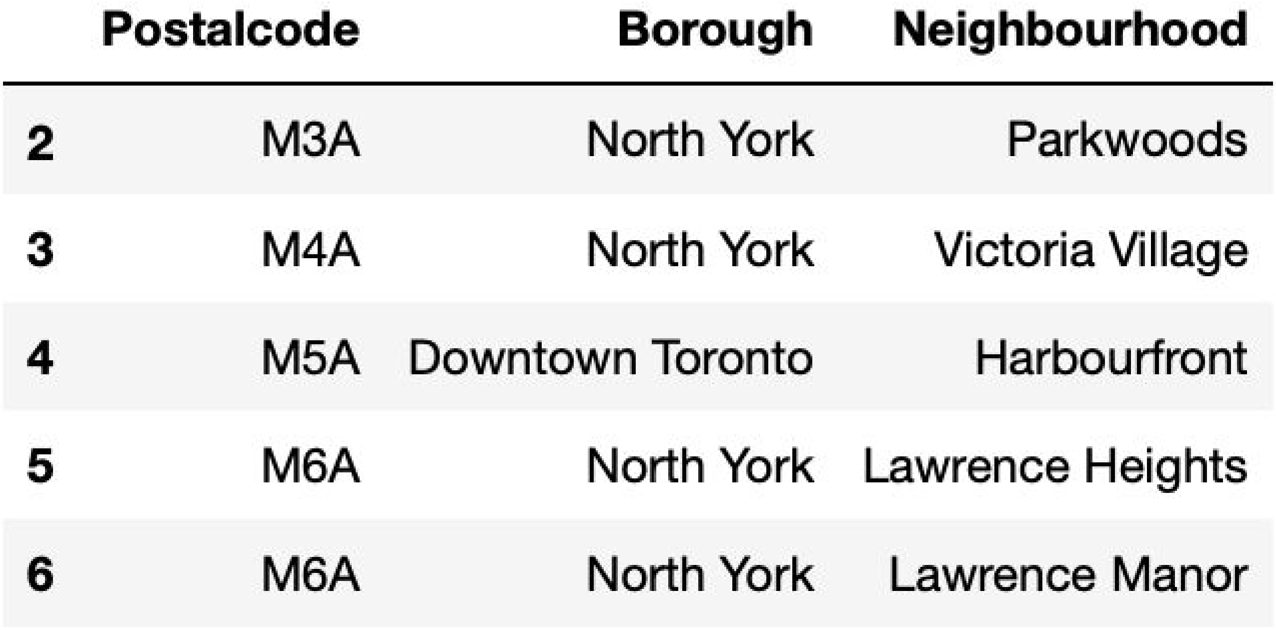
- Dataframe will consist of three columns: PostalCode, Borough, and Neighborhood
- Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.
- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 11 in the above table.

- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.



	Postcode	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	Harbourfront

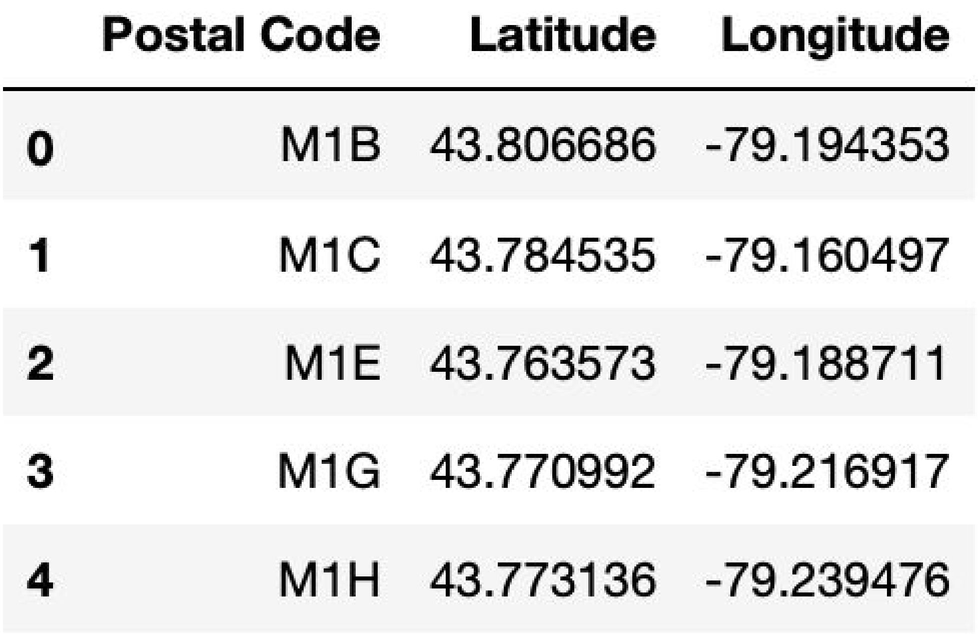
After some cleaning we got the proper dataframe with the Postal code, Borough & Neighborhood information.



	Postalcode	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M6A	North York	Lawrence Heights
6	M6A	North York	Lawrence Manor

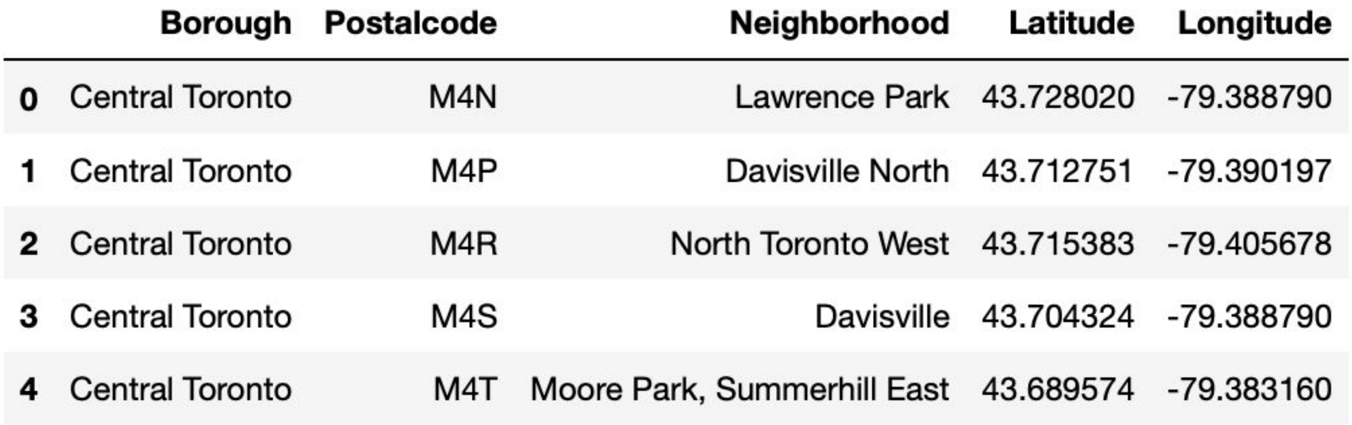
b) Adding geographical coordinates to the neighborhoods

Next important step is adding the geographical coordinates to these neighborhoods. To do so I'm extracting the data present in the Geospatial Data csv file and I'm combining it with the existing neighborhood dataframe by merging them both based on the postal code.



	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

I'm renaming the columns to match the existing dataframe formed from '*List of Postal code of Canada: M' wiki page.* After that I'm merging both the dataframes into one by merging on the postal code.



	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	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

c) Scrap the distribution of population from Wikipedia

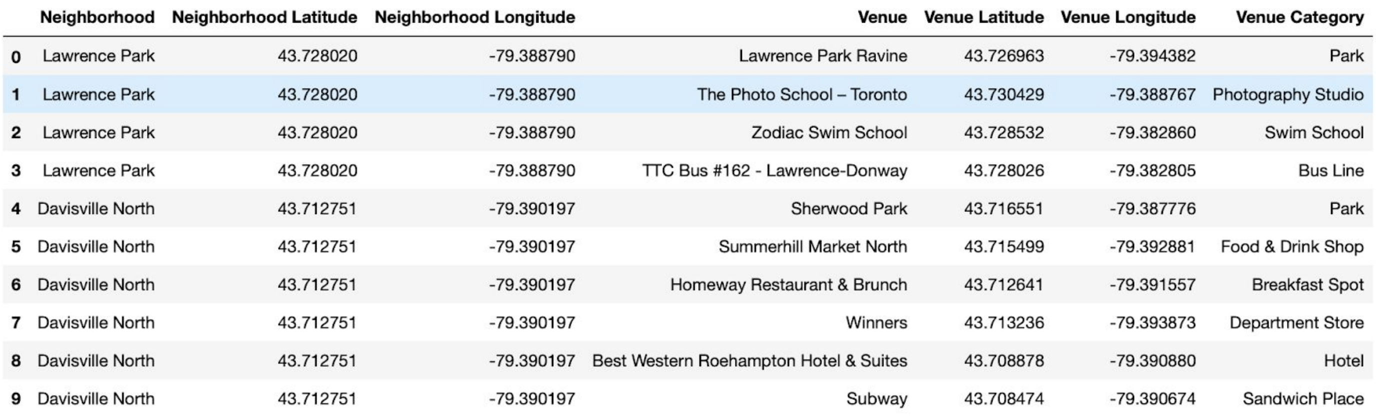
Another factor that can help us in deciding which neighborhood would be best option to open a restaurant is, the distribution of population based on the ethnic diversity for each neighborhood. As this helps us in identifying the neighborhoods which are densely populated with Indian crowd since that neighborhood would be an ideal place to open an Indian restaurant.

Scraped the following Wikipedia page, “Demographics of Toronto” in order to obtain the data about the Toronto & the Neighborhoods in it. Compared to all the neighborhoods in Toronto below given neighborhoods only had considerable amount of Indian crowd. We are examining those neighborhood’s population to identify the densely populated neighborhoods with Indian population.

There were only six neighborhoods in Toronto which Indian population spread across so we are gathering the population, it’s percentage in each riding in those neighborhoods.

d) Get location data using Foursquare

Foursquare API is very useful online application used my many developers & other applications like Uber etc. In this project I have used it to retrieve information about the places present in the neighborhoods of Toronto. The API returns a JSON file and we need to turn that into a data-frame. Here I’ve chosen 100 popular spots for each neighborhood within a radius of 1km.

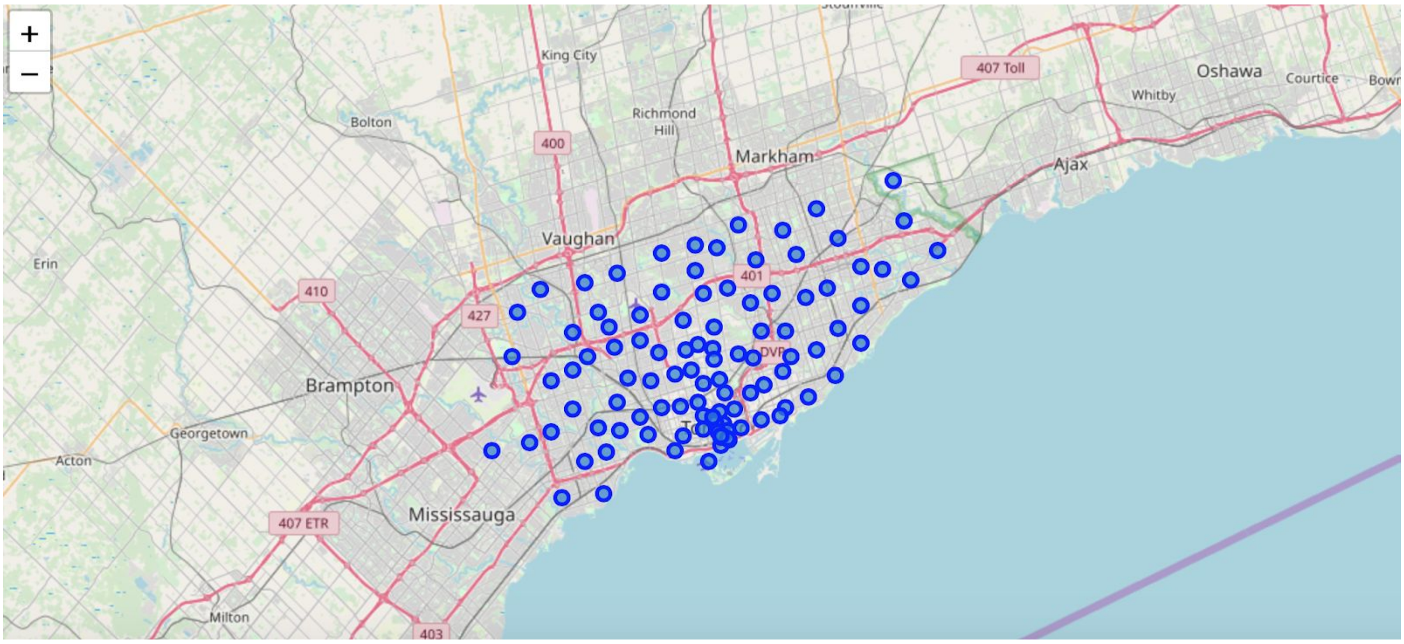


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lawrence Park	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
1	Lawrence Park	43.728020	-79.388790	The Photo School – Toronto	43.730429	-79.388767	Photography Studio
2	Lawrence Park	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
3	Lawrence Park	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
4	Davisville North	43.712751	-79.390197	Sherwood Park	43.716551	-79.387776	Park
5	Davisville North	43.712751	-79.390197	Summerhill Market North	43.715499	-79.392881	Food & Drink Shop
6	Davisville North	43.712751	-79.390197	Homeway Restaurant & Brunch	43.712641	-79.391557	Breakfast Spot
7	Davisville North	43.712751	-79.390197	Winners	43.713236	-79.393873	Department Store
8	Davisville North	43.712751	-79.390197	Best Western Roehampton Hotel & Suites	43.708878	-79.390880	Hotel
9	Davisville North	43.712751	-79.390197	Subway	43.708474	-79.390674	Sandwich Place

3. Exploratory Data Analysis :

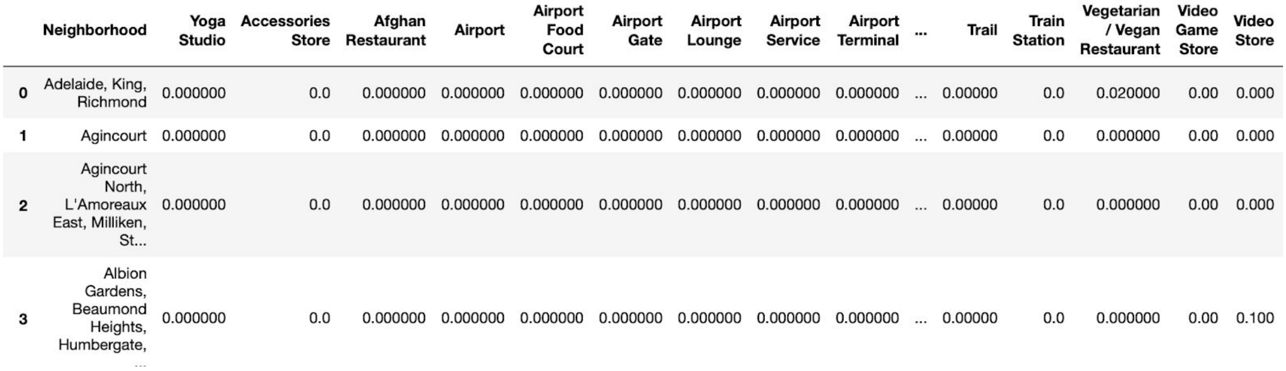
3.1 Folium Library and Leaflet Map

Folium is a python library, I’m using it to draw an interactive leaflet map using coordinate data.



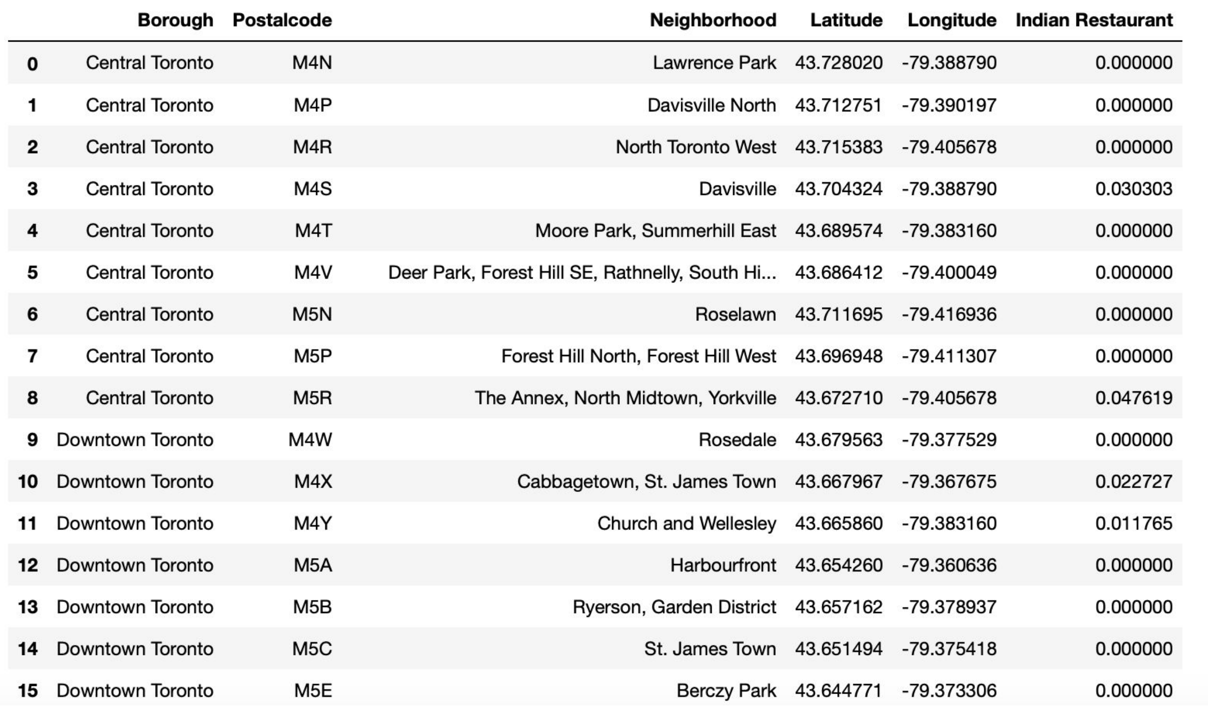
3.2 Relationship between neighborhood and Indian Restaurant

First we will extract the Neighborhood and Indian Restaurant column from the above toronto dataframe for further analysis:



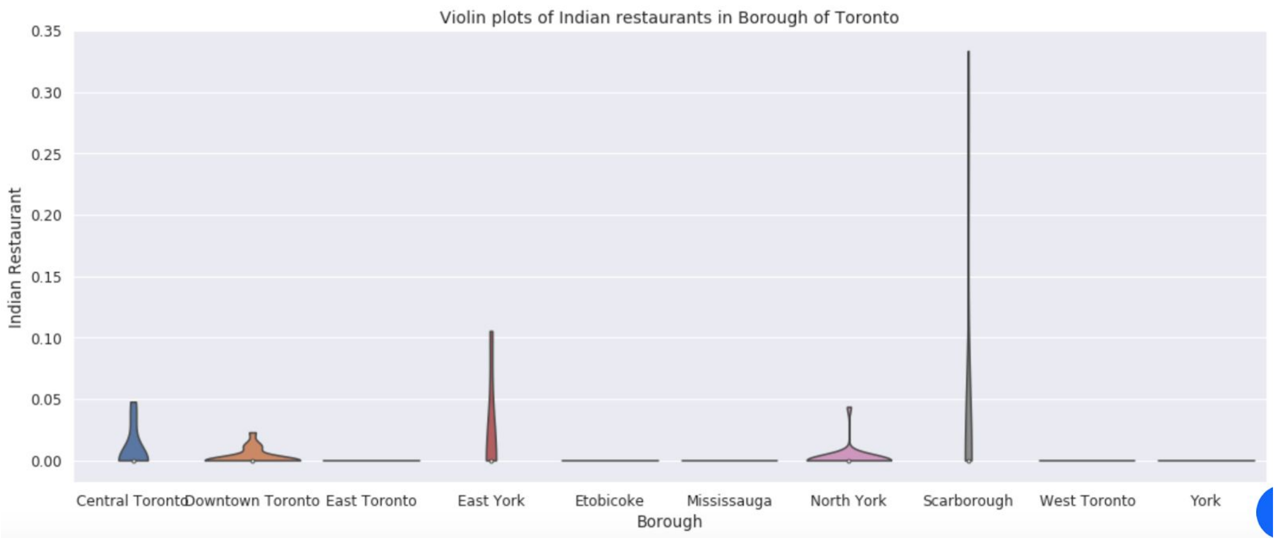
	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	...	Trail	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Video Store
0	Adelaide, King, Richmond	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00000	0.0	0.020000	0.00	0.000
1	Agincourt	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00000	0.0	0.000000	0.00	0.000
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00000	0.0	0.000000	0.00	0.000
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.00000	0.0	0.000000	0.00	0.100

After performing pandas one hot encoding for the venue categories, let us merge this dataframe with the Toronto DataFrame with latitude & longitude information on neighborhood. Finally extract just the Indian restaurant values along with neighborhood information.



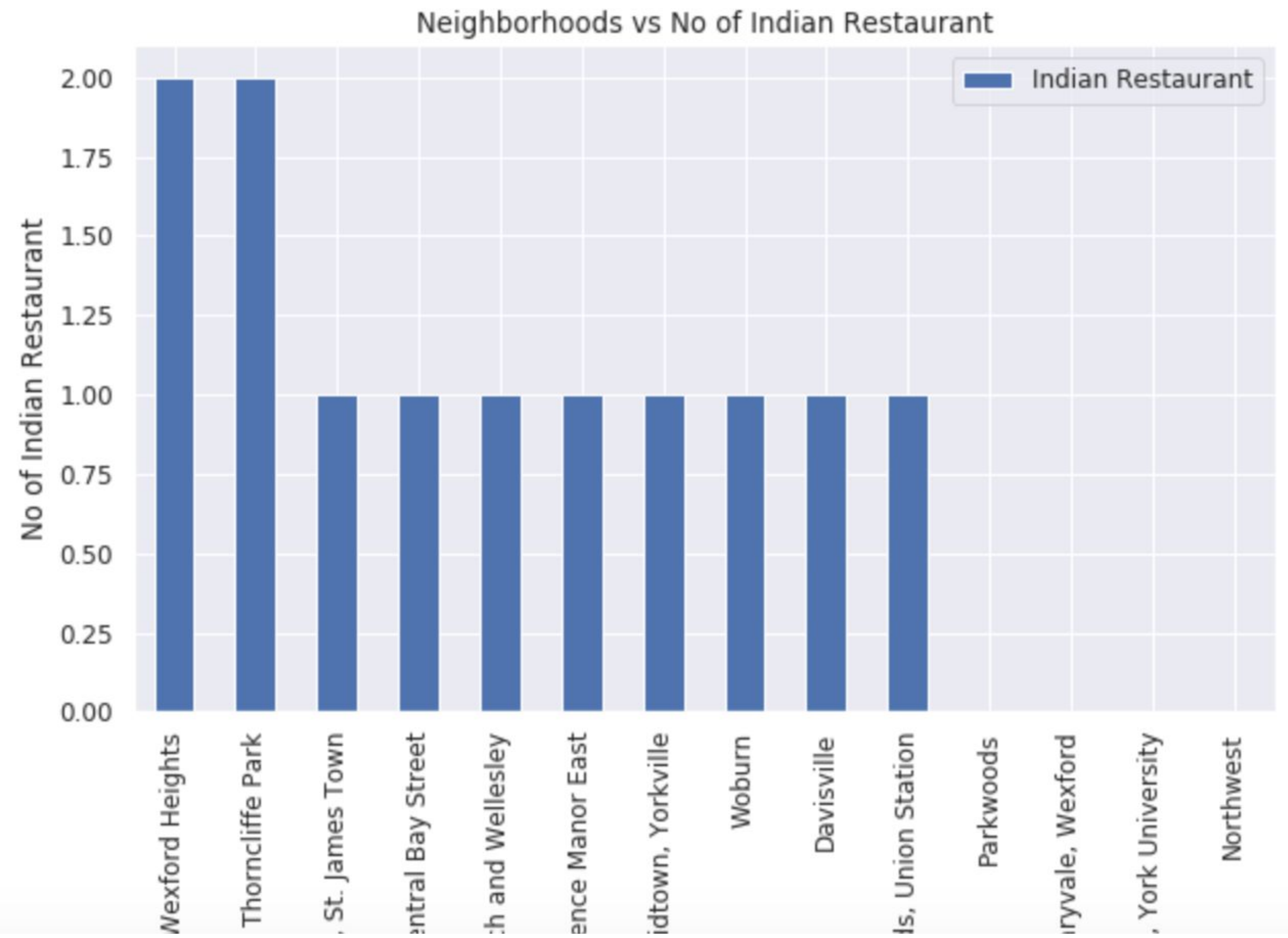
	Borough	Postalcode	Neighborhood	Latitude	Longitude	Indian Restaurant
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790	0.000000
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197	0.000000
2	Central Toronto	M4R	North Toronto West	43.715383	-79.405678	0.000000
3	Central Toronto	M4S	Davisville	43.704324	-79.388790	0.030303
4	Central Toronto	M4T	Moore Park, Summerhill East	43.689574	-79.383160	0.000000
5	Central Toronto	M4V	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686412	-79.400049	0.000000
6	Central Toronto	M5N	Roselawn	43.711695	-79.416936	0.000000
7	Central Toronto	M5P	Forest Hill North, Forest Hill West	43.696948	-79.411307	0.000000
8	Central Toronto	M5R	The Annex, North Midtown, Yorkville	43.672710	-79.405678	0.047619
9	Downtown Toronto	M4W	Rosedale	43.679563	-79.377529	0.000000
10	Downtown Toronto	M4X	Cabbagetown, St. James Town	43.667967	-79.367675	0.022727
11	Downtown Toronto	M4Y	Church and Wellesley	43.665860	-79.383160	0.011765
12	Downtown Toronto	M5A	Harbourfront	43.654260	-79.360636	0.000000
13	Downtown Toronto	M5B	Ryerson, Garden District	43.657162	-79.378937	0.000000
14	Downtown Toronto	M5C	St. James Town	43.651494	-79.375418	0.000000
15	Downtown Toronto	M5E	Berczy Park	43.644771	-79.373306	0.000000

Let’s try to draw some plot using the above dataframe:



With the help of this violin plots we can identify the boroughs with densely populated Indian restaurants. It is drawn using seaborn library to show the distribution of Indian restaurants in different boroughs.

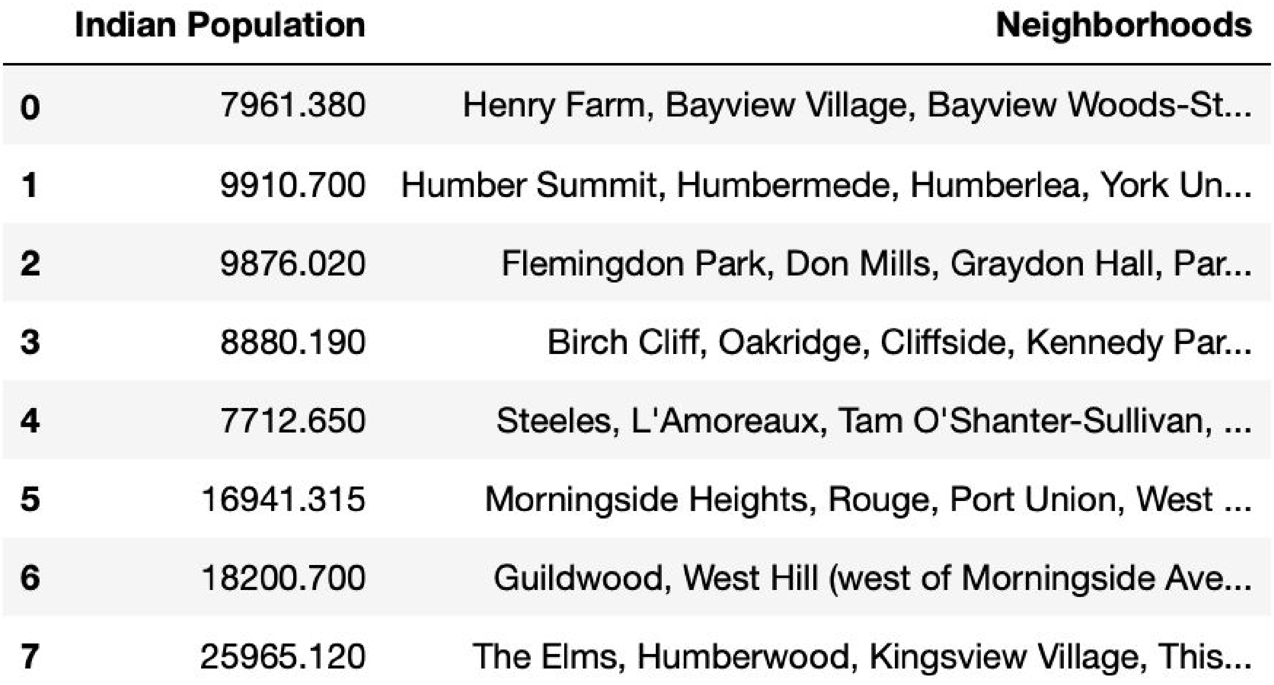
Let’s also visualize the neighborhood with Indian Restaurants:



3.3 Relationship between neighborhood and Indian population

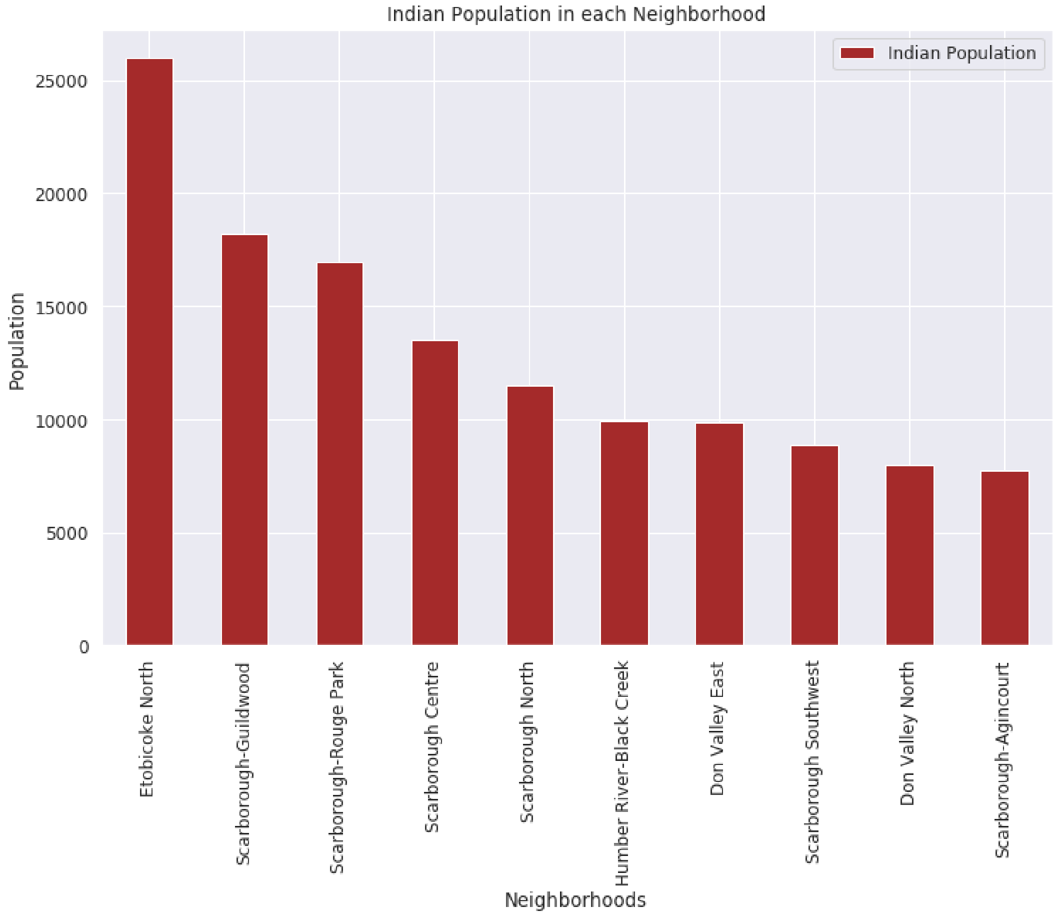
Another key feature is the distribution of Indian crowd in each neighborhoods. Let us analyze the neighborhoods and identify the neighborhoods with highest number of Indian population.

To achieve that we are joining all the neighborhood’s dataframe from using the wiki page with ethnic population and in that we are extracting just the Indian population for each neighborhood.



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

Let’s draw a graph to visualize the population spread in neighborhoods :



This analysis & visualization of the relationship between neighborhoods & Indian population present in those neighborhoods helps us in identifying the highly populated Indian neighborhoods. Once we identify those neighborhoods it helps us in deciding where to place the new Indian restaurant. Indian restaurant placed in an densely populated Indian neighborhood is more likely to get more Indian customers than a restaurant placed in a neighborhood with less or no Indian population. Thus this analysis helps in the determining the success of the new Indian restaurant.

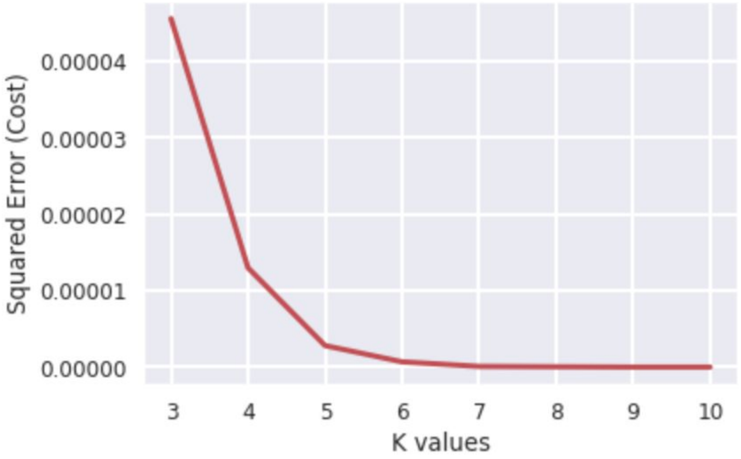
3.4 Relationship between Indian population and Indian restaurant

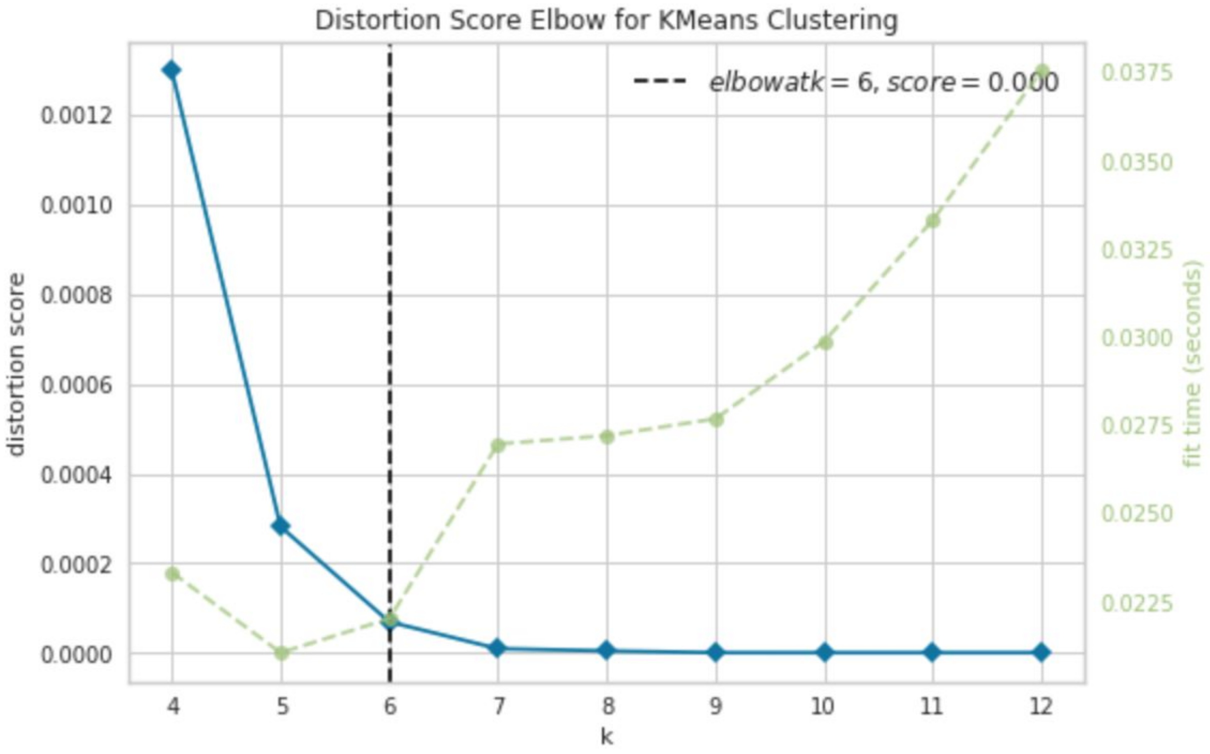
After performing the data cleaning & data analysis we couldn't identify a big relationship established between densely populated Indian neighborhoods & number of Indian restaurants. This might be because of the missing in data as this an area which can improved in future analysis to get a more insight about the business problem.

4. Predictive Modelling:

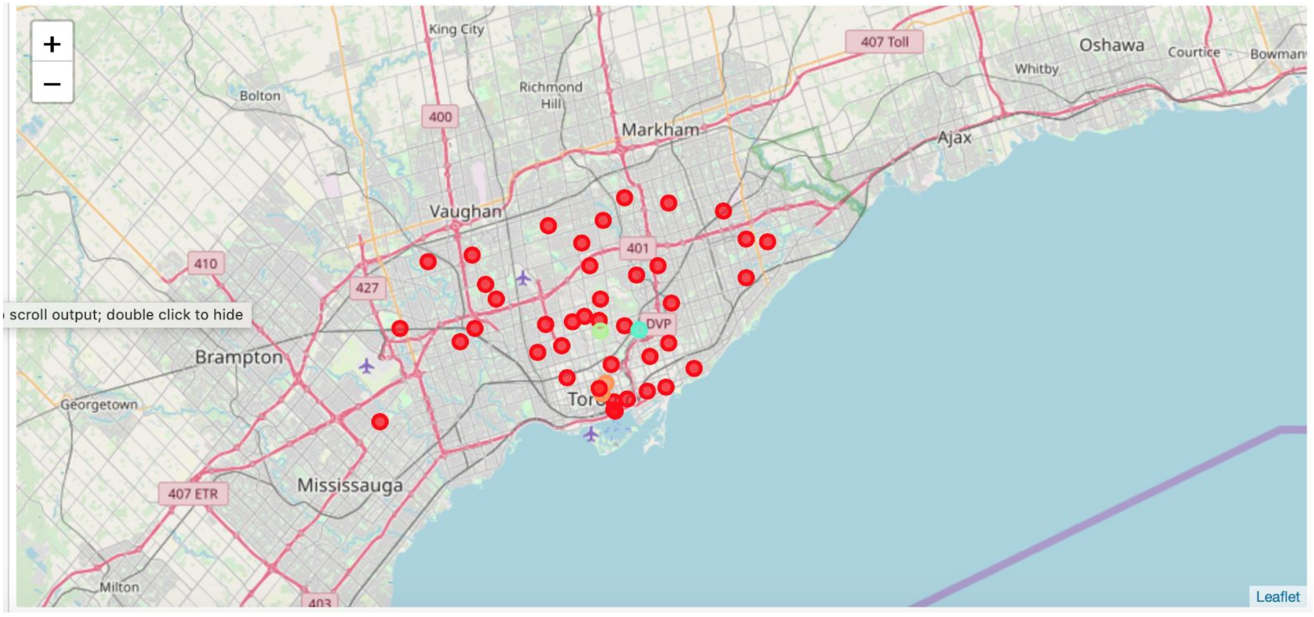
4.1 Clustering Neighborhoods of Toronto:

First step in K-means clustering is to identify best K value meaning the number of clusters in a given dataset. To do so we are going to use the elbow method on the Toronto dataset with Indian restaurant percentage.



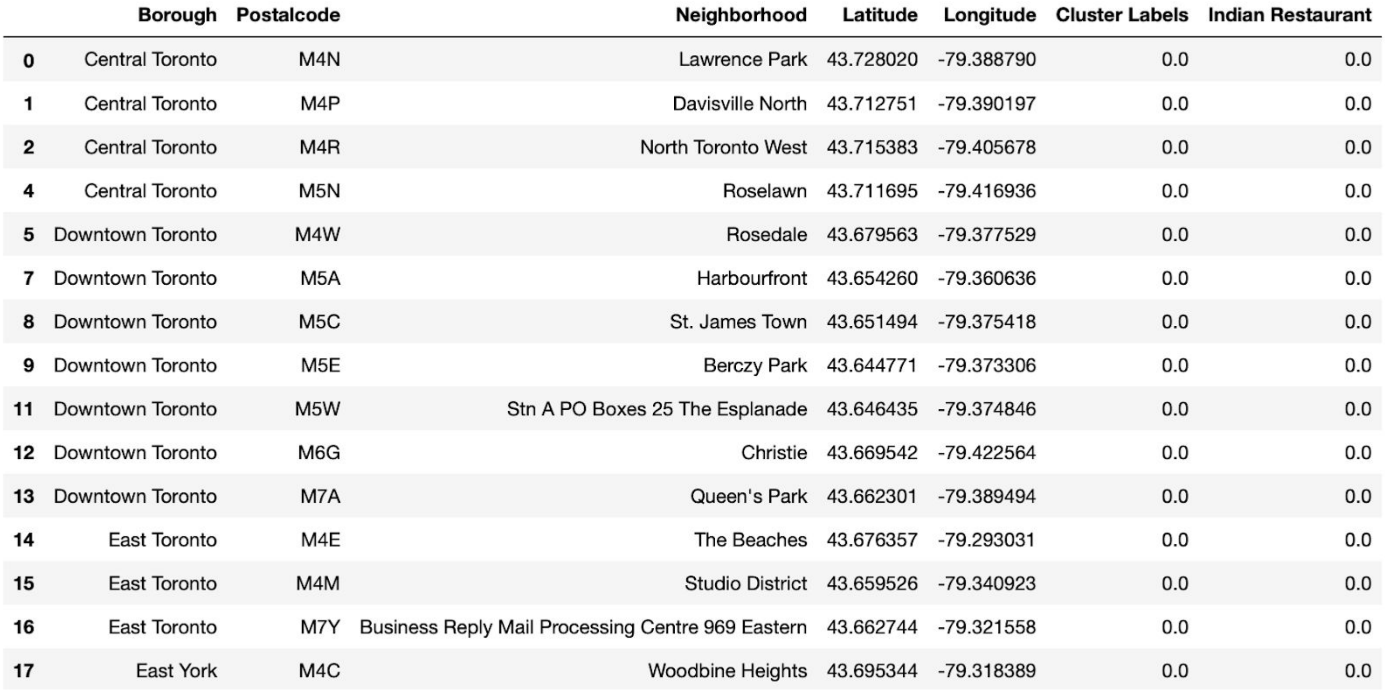


After analysing using elbow method using distortion score & Squared error for each K value, looks like K = 6 is the best value.



4.2 Examine the Clusters:

We have total of 6 clusters such as 0,1,2,3,4,5. Let us examine one after the other.

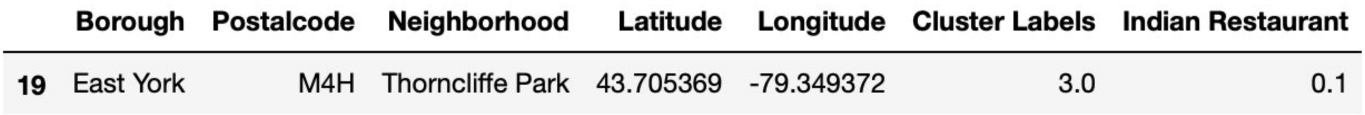


	Borough	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790	0.0	0.0
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197	0.0	0.0
2	Central Toronto	M4R	North Toronto West	43.715383	-79.405678	0.0	0.0
4	Central Toronto	M5N	Roselawn	43.711695	-79.416936	0.0	0.0
5	Downtown Toronto	M4W	Rosedale	43.679563	-79.377529	0.0	0.0
7	Downtown Toronto	M5A	Harbourfront	43.654260	-79.360636	0.0	0.0
8	Downtown Toronto	M5C	St. James Town	43.651494	-79.375418	0.0	0.0
9	Downtown Toronto	M5E	Berczy Park	43.644771	-79.373306	0.0	0.0
11	Downtown Toronto	M5W	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	0.0	0.0
12	Downtown Toronto	M6G	Christie	43.669542	-79.422564	0.0	0.0
13	Downtown Toronto	M7A	Queen's Park	43.662301	-79.389494	0.0	0.0
14	East Toronto	M4E	The Beaches	43.676357	-79.293031	0.0	0.0
15	East Toronto	M4M	Studio District	43.659526	-79.340923	0.0	0.0
16	East Toronto	M7Y	Business Reply Mail Processing Centre 969 Eastern	43.662744	-79.321558	0.0	0.0
17	East York	M4C	Woodbine Heights	43.695344	-79.318389	0.0	0.0

Cluster 0 contains all the neighborhoods which has least number of Indian restaurants. It is shown in red color in the map.

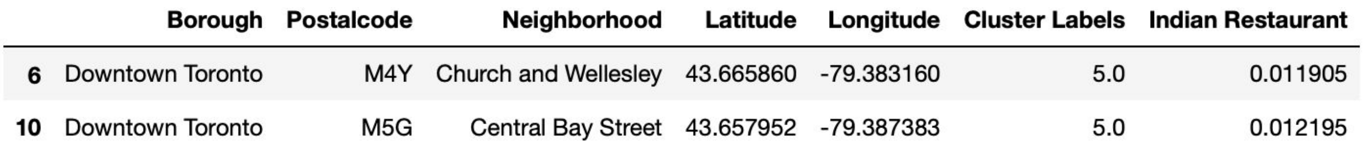
Cluster 1 contains the neighborhoods which is sparsely populated with Indian restaurants. It is shown in purple color in the map.

Cluster 2 & 4 has no rows meaning no data points or no neighborhood was near to these centroids.



	Borough	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
19	East York	M4H	Thorncliffe Park	43.705369	-79.349372	3.0	0.1

Cluster 3 contains all the neighborhoods which is medium populated with Indian restaurants. It is shown in blue color in the map.



	Borough	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
6	Downtown Toronto	M4Y	Church and Wellesley	43.665860	-79.383160	5.0	0.011905
10	Downtown Toronto	M5G	Central Bay Street	43.657952	-79.387383	5.0	0.012195

Cluster 5 contains all the neighborhoods which is densely populated with Indian restaurants. It is shown in Orange color in the map.

5. Results and Discussion:

5.1 Results

We have reached the end of the analysis, in this section we will document all the findings from above clustering & visualization of the dataset. In this project, we started off with the business problem of identifying a good neighborhood to open a new Indian restaurant. To achieve that we looked into all the neighborhoods in Toronto, analysed the Indian population in each neighborhood & number of Indian restaurants in those neighborhoods to come to conclusion about which neighborhood would be a better spot. We have used variety of data sources to set up a very realistic data-analysis scenario. We have found out that —

- In those 11 boroughs we identified that only Central Toronto, Downtown Toronto, East Toronto, East York, North York & Scarborough boroughs have high amount of Indian restaurants with the help of Violin plots between Number of Indian restaurants in Borough of Toronto.
- In all the ridings, Scarborough-Guildwood, Scarborough-Rouge Park, Scarborough Centre, Scarborough North, Humber River-Black Creek, Don Valley East, Scarborough Southwest, Don Valley North & Scarborough-Agincourt are the densely populated with Indian crowd ridings.

- With the help of clusters examining & violin plots looks like Downtown Toronto, Central Toronto, East York are already densely populated with Indian restaurants. So it is better idea to leave those boroughs out and consider only Scarborough, East Toronto & North York for the new restaurant's location.
- After careful consideration it is a good idea to open a new Indian restaurant in Scarborough borough since it has high number of Indian population which gives a higher number of customers possibility and lower competition since very less Indian restaurants in the neighborhoods.

5.2 Discussion

According to this analysis, Scarborough borough will provide the least competition for the new upcoming Indian restaurant as there is very little Indian restaurants spread or no Indian restaurants in few neighborhoods. Also looking at the population distribution looks like it is densely populated with Indian crowd which helps the new restaurant by providing high customer visit possibility. So, definitely this region could potentially be a perfect place for starting a quality Indian restaurants. Some of the drawbacks of this analysis are — the clustering is completely based only on data obtained from Foursquare API and the data about the Indian population distribution in each neighborhood is also based on the 2016 census which is not up-to date. Thus there is huge gap of 3 years in the population distribution data. Even Though there are lots of areas where it can be improved yet this analysis has certainly provided us with some good insights, preliminary information on possibilities & a head start into this business problem by setting the step stones properly.

6. Conclusion :

Finally to conclude this project, We have got a chance to on a business problem like how a real like data scientists would do. We have used many python libraries to fetch the data , to manipulate the contents & to analyze and visualize those datasets. We have made use of Foursquare API to explore the venues in neighborhoods of Toronto, then get good amount of data from Wikipedia which we scraped with help of Wikipedia python library and visualized using various plots present in seaborn & matplotlib. We also applied machine learning technique to predict the output given the data and used Folium to visualize it on a map.

Some of the drawbacks or areas of improvements shows us that this analysis can be further improved with the help of more data and different machine learning technique. Similarly we can use this project to analysis any scenario such as opening a different cuisine restaurant or opening of a new gym and etc. Hopefully, this project helps acts as initial guidance to take more complex real-life challenges using data-science.