Applied Data Science Capstone by IBM/Coursera

Exploring Toronto's Neighbourhoods to open a new Indian Restaurant

Sherbulandkhan Babi

29 April, 2020

Introduction : Business Problem & Discussion of the background

As a part of the IBM Data Science Professional Course, this Capstone Project is based on 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 neighbourhoods of Toronto to figure out which neighbourhood is a suitable option for starting a new Indian restaurant. 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.

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 neighbourhoods such as Chinatown, Corso Italia, Greek town, Kensington Market, Korea town, 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 try to find an **optimal location for a restaurant** and make a decision whether it is a good idea to open an **Indian** restaurant in **Toronto**, Canada.

We already know that Toronto shelters a greater number of Indians than any other city in Canada and it may be 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 that it yields more profit to the owner.

Since there are lots of restaurants in Toronto we will try to detect **most profitable area since the success of any restaurant depends on the people, ambience and its location**. We are also particularly interested in **areas with Indians living in vicinity**.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Target Audience

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

- Business personnel who wants 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.

- 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 venture.

- Indian crowd who wants to find neighbourhoods with lots of option for Indian restaurants.

- Business Analyst or Data Scientists, who wish to analyze the neighbourhoods 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.

Data (Sources and Cleaning)

Data Sources

- **"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 neighbourhoods present in Toronto. This page has the **postal code**, **borough** & the name of all the **neighbourhoods** present in Toronto.

- In order to get all the **geographical coordinates** of the neighbourhoods, we will be using "https://cocl.us/Geospatial_data" csv file.

- To get info about the distribution of population on the basis of their ethnicity, we will be using **"Demographics of Toronto"** (https://en.m.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity) wiki page. Using this page, I will be easily able to identify the neighbourhoods which are densely populated with Indians as it might be helpful in identifying the suitable neighbourhood to open a new Indian restaurant.

- To get location and other information about various venues in Toronto, we will be using **Foursquare's explore API**. Using the Foursquare's explore API (which gives venues recommendations), we'll be fetching details about the venues up present in Toronto and collect their **names**, **categories** and **locations** (latitude and longitude).

From **Foursquare API** (https://developer.foursquare.com/docs), we get 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.

Data Cleaning

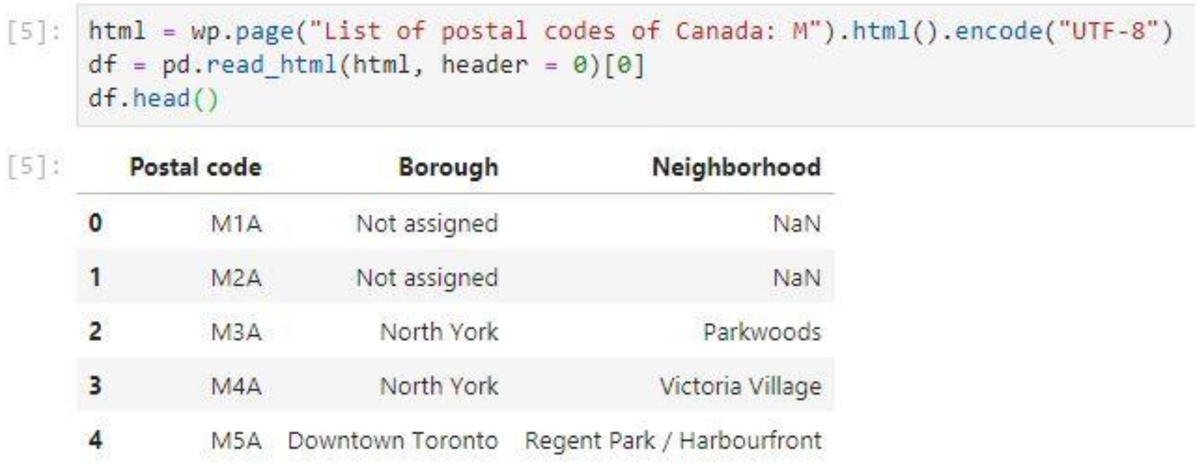
a) Scraping Toronto Neighbourhoods Table from Wikipedia

Assumptions made to attain the below Data Frame:

- Data frame will consist of three columns: **Postal Code**, **Borough**, and **Neighbourhood**
- Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.

- More than one neighbourhood 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 neighbourhoods: **Harbourfront** and **Regent Park**. These two rows will be **combined into one row with the neighbourhoods separated with a comma as shown in row 11 in the above table.
- If a cell has a borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough.

Wikipedia - Package is used to scrape the data from wiki.



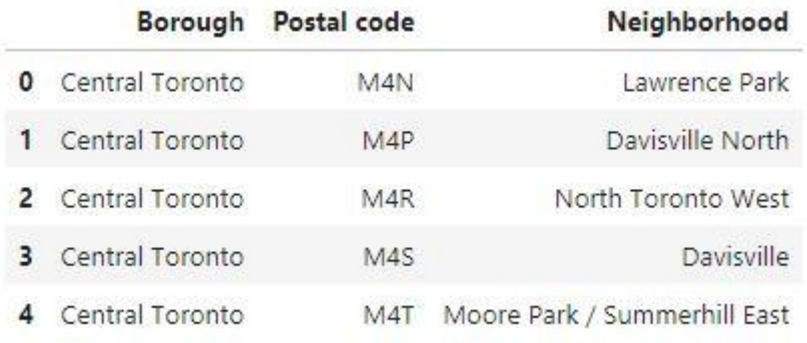
```
[5]: html = wp.page("List of postal codes of Canada: M").html().encode("UTF-8")
     df = pd.read_html(html, header = 0)[0]
     df.head()
```

[5]:

	Postal code	Borough	Neighborhood
0	M1A	Not assigned	NaN
1	M2A	Not assigned	NaN
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park / Harbourfront

Table 1 Dataframe formed from the scraped wikipedia page

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

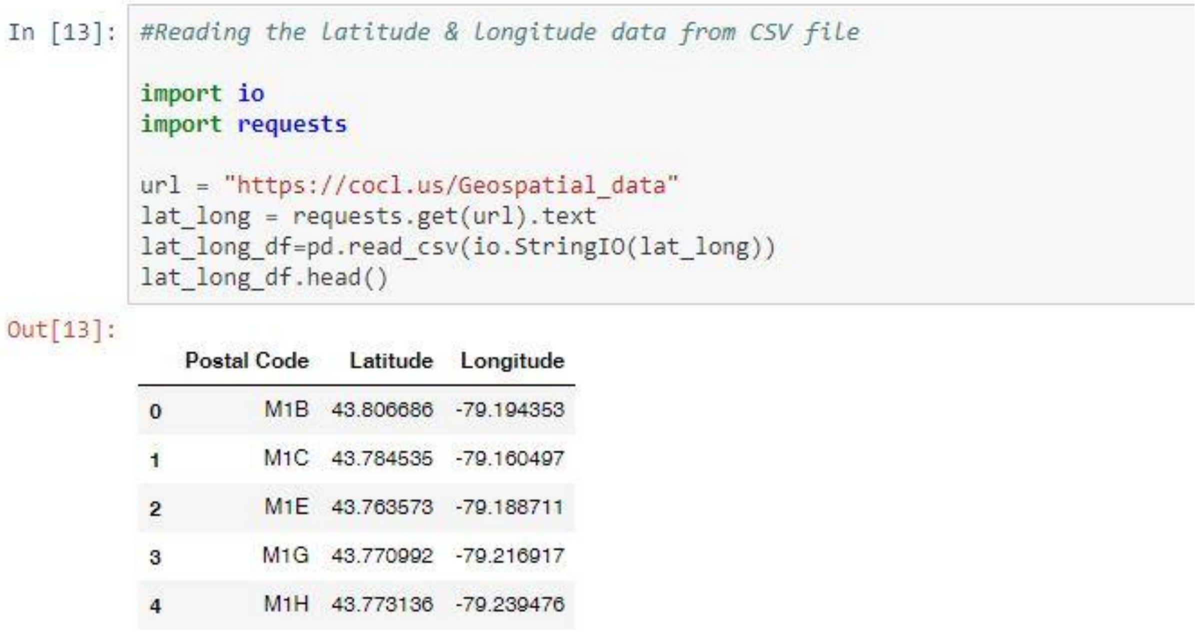


	Borough	Postal code	Neighborhood
0	Central Toronto	M4N	Lawrence Park
1	Central Toronto	M4P	Davisville North
2	Central Toronto	M4R	North Toronto West
3	Central Toronto	M4S	Davisville
4	Central Toronto	M4T	Moore Park / Summerhill East

Table 2 Dataframe from 'List of Postal code of Canada: M' Wikipedia Table

b) Adding geographical coordinates to the neighbourhoods

Next important step is adding the geographical coordinates to these neighborhoods. To do so, we will extract the data present in the Geospatial Data .csv file and combine it with the existing neighbourhood dataframe by merging them both based on the postal code.



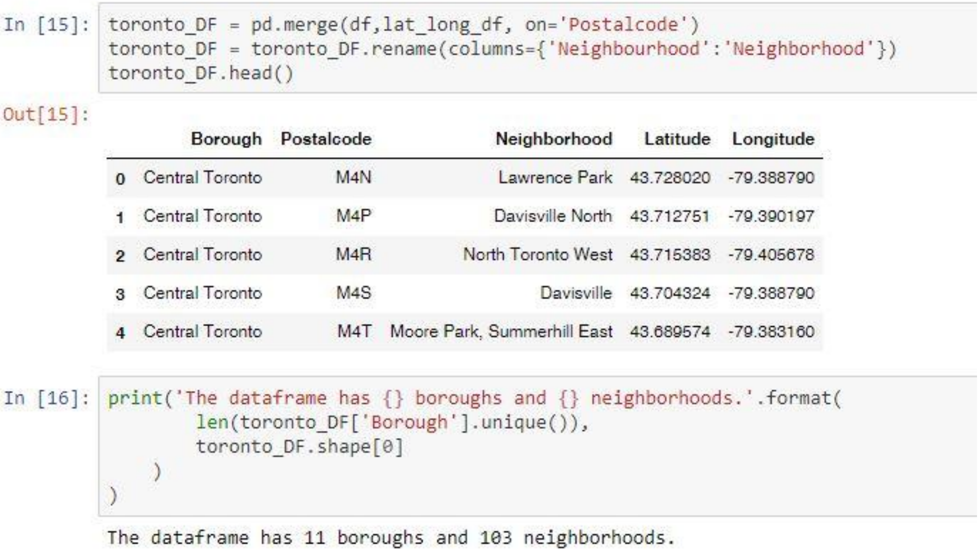
```
In [13]: #Reading the latitude & longitude data from CSV file

         import io
         import requests

         url = "https://cocl.us/Geospatial_data"
         lat_long = requests.get(url).text
         lat_long_df=pd.read_csv(io.StringIO(lat_long))
         lat_long_df.head()
```

Out[13]:

	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

Table 3 DataFrame with latitude & longitude of Postal codes in Toronto

Renaming the columns to match the existing dataframe and then merging both the dataframe into one by matching on the postal code.



```
In [15]: toronto_DF = pd.merge(df,lat_long_df, on='Postalcode')
         toronto_DF = toronto_DF.rename(columns={'Neighbourhood':'Neighborhood'})
         toronto_DF.head()
```

Out[15]:

	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

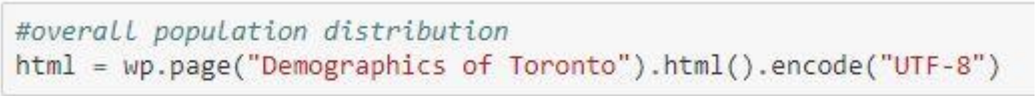
```
In [16]: print('The dataframe has {} boroughs and {} neighborhoods.'.format(
                len(toronto_DF['Borough'].unique()),
                toronto_DF.shape[0]
             )
         )

         The dataframe has 11 boroughs and 103 neighborhoods.
```

Table 4 Merged new dataframe with info about Neighbourhoods, borough, postalcode, latitude & longitude in Toronto

c) Scrap the distribution of population from Wikipedia

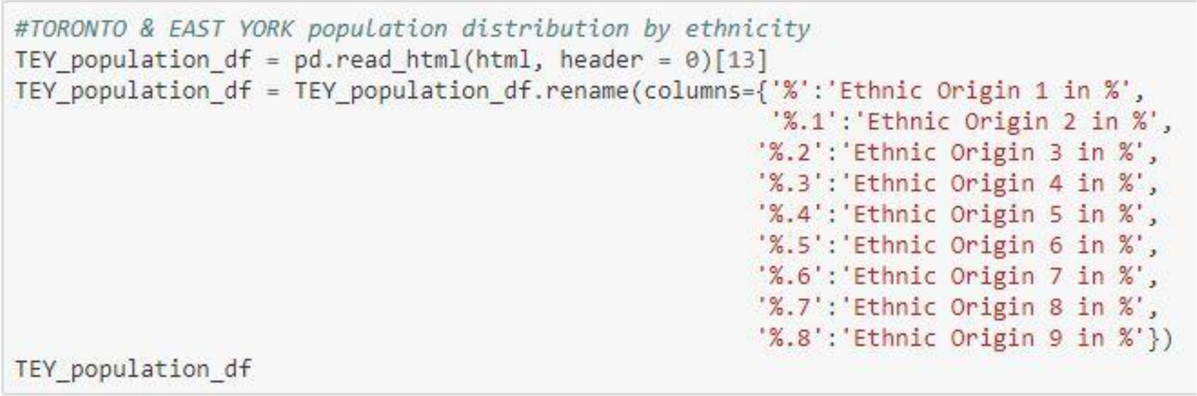
Another factor that can help us in deciding which neighbourhood would be best option to open a restaurant is, the distribution of population based on the ethnic diversity for each neighbourhood. As this helps us in identifying the neighbourhoods which are densely populated 5 with Indian crowd since that neighbourhood 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 Neighbourhoods in it. Compared to all the neighbourhoods in Toronto below given neighbourhoods only had considerable amount of Indian crowd. We are examining those neighbourhood's population to identify the densely populated neighbourhoods with Indian population.



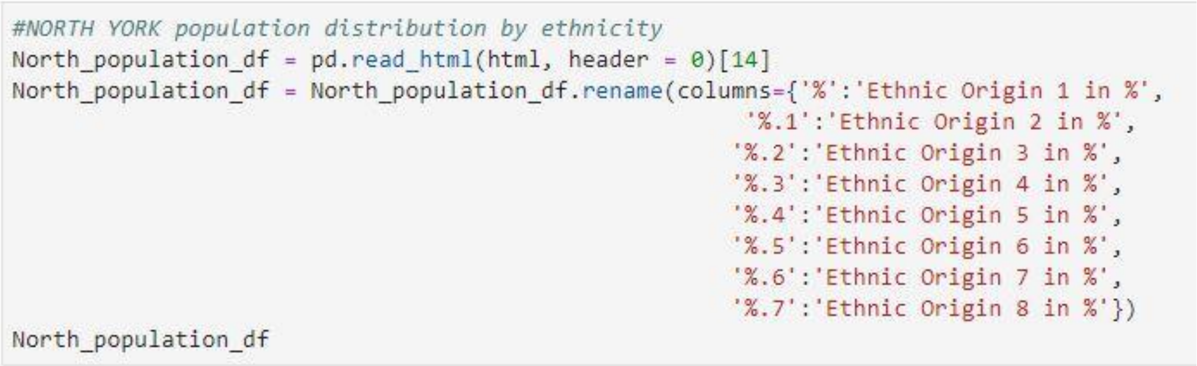
```
#overall population distribution
html = wp.page("Demographics of Toronto").html().encode("UTF-8")
```

Code snippet 1 Scraping the wiki page



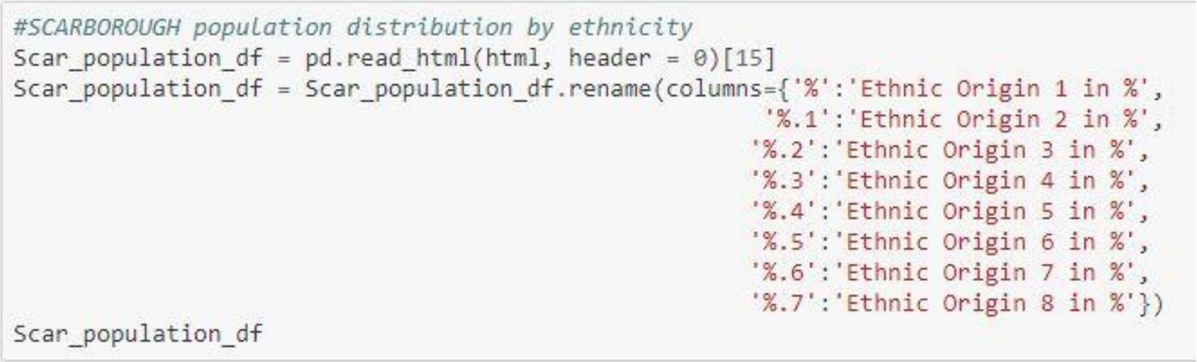
```
#TORONTO & EAST YORK population distribution by ethnicity
TEY_population_df = pd.read_html(html, header = 0)[13]
TEY_population_df = TEY_population_df.rename(columns={'%':'Ethnic Origin 1 in %',
                                                       '%.1':'Ethnic Origin 2 in %',
                                                      '%.2':'Ethnic Origin 3 in %',
                                                      '%.3':'Ethnic Origin 4 in %',
                                                      '%.4':'Ethnic Origin 5 in %',
                                                      '%.5':'Ethnic Origin 6 in %',
                                                      '%.6':'Ethnic Origin 7 in %',
                                                      '%.7':'Ethnic Origin 8 in %',
                                                      '%.8':'Ethnic Origin 9 in %'})
TEY_population_df
```

Table 5 TORONTO & EAST YORK population distribution by ethnicity



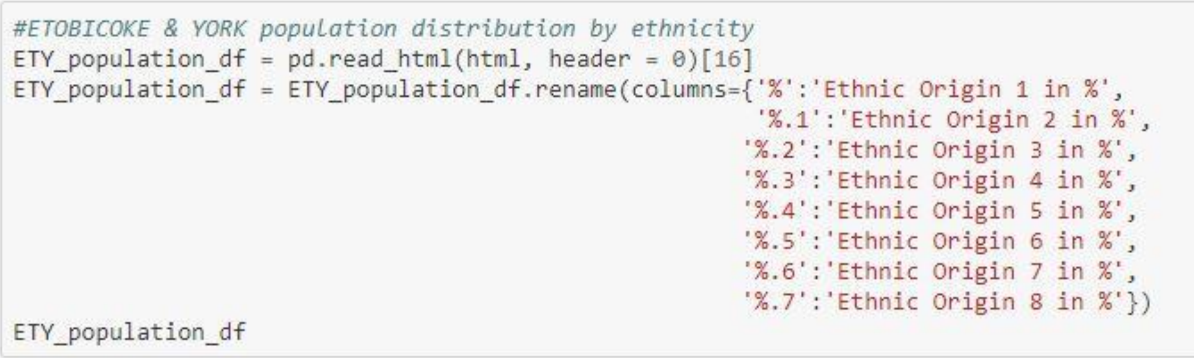
```
#NORTH YORK population distribution by ethnicity
North_population_df = pd.read_html(html, header = 0)[14]
North_population_df = North_population_df.rename(columns={'%':'Ethnic Origin 1 in %',
                                                           '%.1':'Ethnic Origin 2 in %',
                                                          '%.2':'Ethnic Origin 3 in %',
                                                          '%.3':'Ethnic Origin 4 in %',
                                                          '%.4':'Ethnic Origin 5 in %',
                                                          '%.5':'Ethnic Origin 6 in %',
                                                          '%.6':'Ethnic Origin 7 in %',
                                                          '%.7':'Ethnic Origin 8 in %'})
North_population_df
```

Table 6 NORTH YORK population distribution by ethnicity



```
#SCARBOROUGH population distribution by ethnicity
Scar_population_df = pd.read_html(html, header = 0)[15]
Scar_population_df = Scar_population_df.rename(columns={'%':'Ethnic Origin 1 in %',
                                                         '%.1':'Ethnic Origin 2 in %',
                                                        '%.2':'Ethnic Origin 3 in %',
                                                        '%.3':'Ethnic Origin 4 in %',
                                                        '%.4':'Ethnic Origin 5 in %',
                                                        '%.5':'Ethnic Origin 6 in %',
                                                        '%.6':'Ethnic Origin 7 in %',
                                                        '%.7':'Ethnic Origin 8 in %'})
Scar_population_df
```

Table 7 SCARBOROUGH population distribution by ethnicity

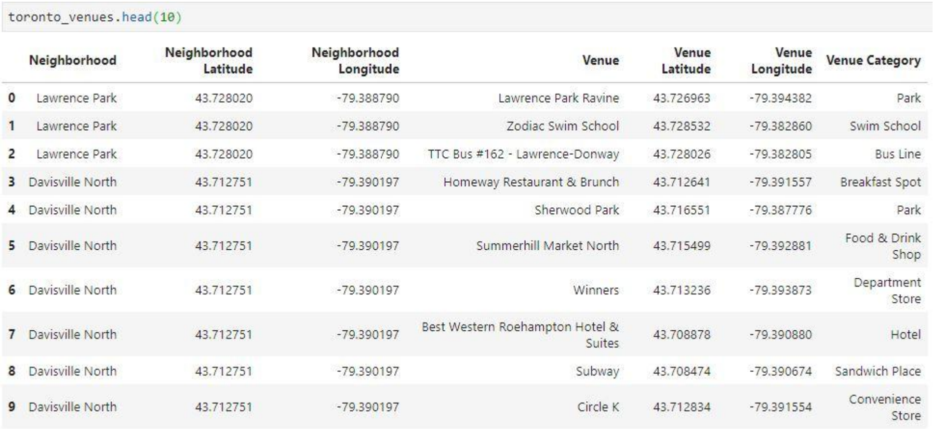


```
#ETOBICOKE & YORK population distribution by ethnicity
ETY_population_df = pd.read_html(html, header = 0)[16]
ETY_population_df = ETY_population_df.rename(columns={'%':'Ethnic Origin 1 in %',
                                                       '%.1':'Ethnic Origin 2 in %',
                                                      '%.2':'Ethnic Origin 3 in %',
                                                      '%.3':'Ethnic Origin 4 in %',
                                                      '%.4':'Ethnic Origin 5 in %',
                                                      '%.5':'Ethnic Origin 6 in %',
                                                      '%.6':'Ethnic Origin 7 in %',
                                                      '%.7':'Ethnic Origin 8 in %'})
ETY_population_df
```

Table 8 ETOBICOKE & YORK population distribution by ethnicity

d) Get location data using Foursquare

Foursquare API is very usefule online application used my many developers & other application like Uber etc. In this project, we 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 dataframe. **Note:- Choosing 100 popular spots for each neighborhood within a radius of 1km.**



```
toronto_venues.head(10)
```

	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	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Lawrence Park	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
3	Davisville North	43.712751	-79.390197	Homeway Restaurant & Brunch	43.712641	-79.391557	Breakfast Spot
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	Winners	43.713236	-79.393873	Department Store
7	Davisville North	43.712751	-79.390197	Best Western Roehampton Hotel & Suites	43.708878	-79.390880	Hotel
8	Davisville North	43.712751	-79.390197	Subway	43.708474	-79.390674	Sandwich Place
9	Davisville North	43.712751	-79.390197	Circle K	43.712834	-79.391554	Convenience Store

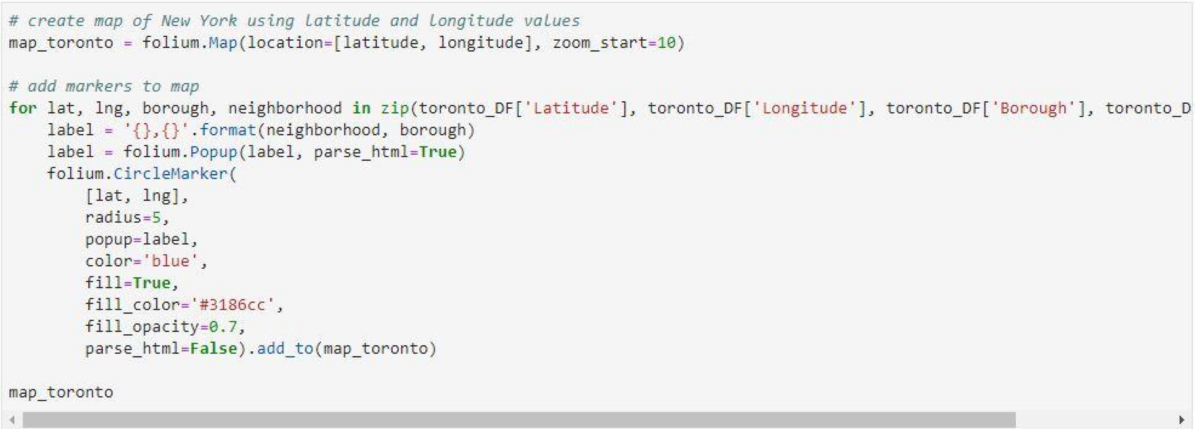
Table 9 Dataframe with venues in each neighbourhood along with the category info of the venues

Methodology (Exploratory Data Analysis, Predictive Modelling)

Exploratory Data Analysis

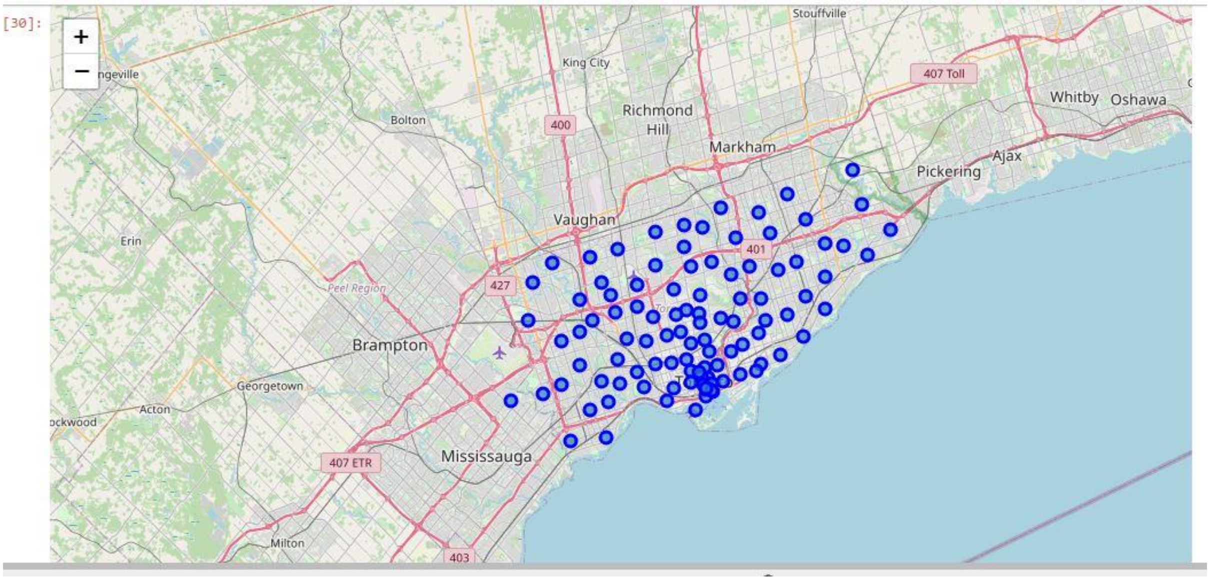
a) Folium Library and Leaflet Map

Folium is a python library which we will be using to draw an interactive leaflet map using coordinate data.



```
# create map of New York using latitude and longitude values
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)

# add markers to map
for lat, lng, borough, neighborhood in zip(toronto_DF['Latitude'], toronto_DF['Longitude'], toronto_DF['Borough'], toronto_D
    label = '{},{}'.format(neighborhood, borough)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_toronto)

map_toronto
```

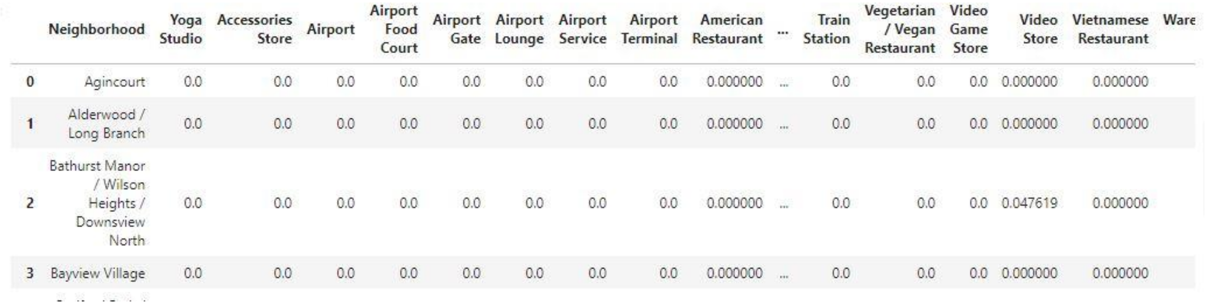
Code snippet 2 To draw the folium map



Map 1 Folium map of Toronto

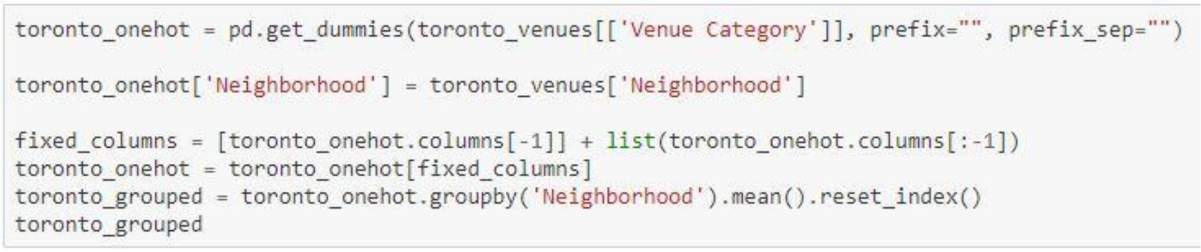
b) Relationship between neighbourhood and Indian Restaurant

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



	Neighborhood	Yoga Studio	Accessories Store	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Train Station	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Ware
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.000000	0.000000	
1	Alderwood / Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.000000	0.000000	
2	Bathurst Manor / Wilson Heights / Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.047619	0.000000	
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.000000	0.000000	

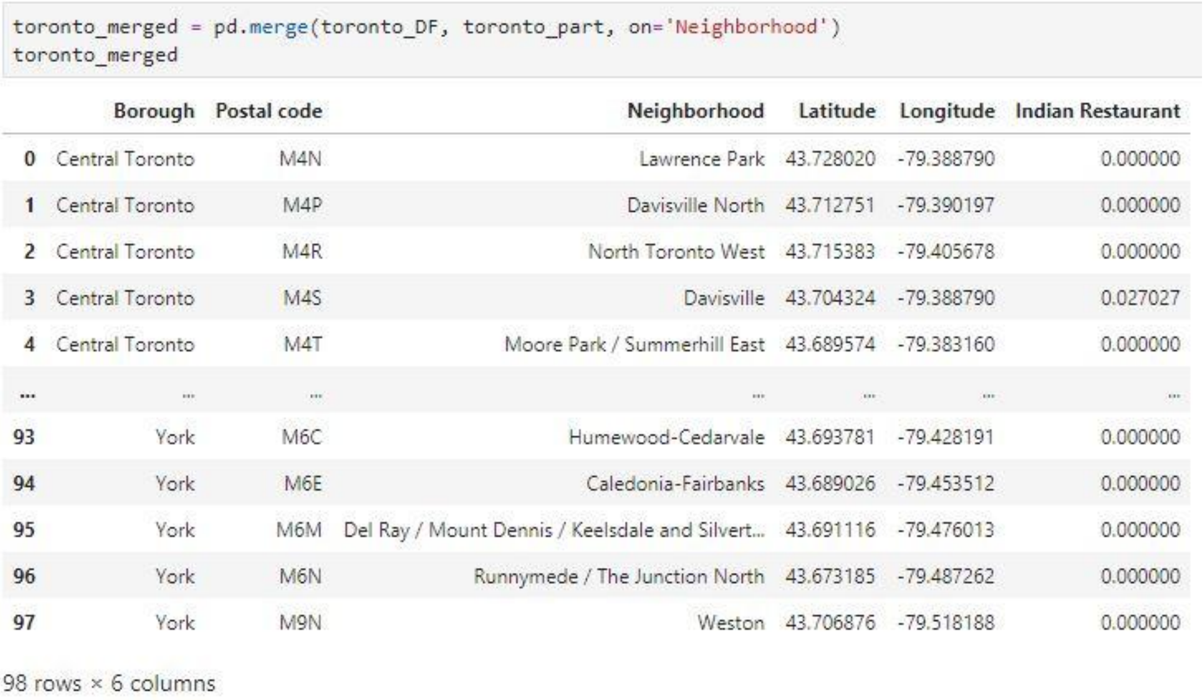
Table 10 Dataframe formed using Foursquare API information about venues in each neighbourhood



```
toronto_onehot = pd.get_dummies(toronto_venues[['Venue Category']], prefix="", prefix_sep="")

toronto_onehot['Neighborhood'] = toronto_venues['Neighborhood']

fixed_columns = [toronto_onehot.columns[-1]] + list(toronto_onehot.columns[:-1])
toronto_onehot = toronto_onehot[fixed_columns]
toronto_grouped = toronto_onehot.groupby('Neighborhood').mean().reset_index()
toronto_grouped
```

Code snippet 3 Manipulating the data to make the analysis easy

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



```
toronto_merged = pd.merge(toronto_DF, toronto_part, on='Neighborhood')
toronto_merged
```

	Borough	Postal code	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	0.027027
4	Central Toronto	M4T	Moore Park / Summerhill East	43.689574	-79.383160	0.000000
...	...	...	...	...	...	...
93	York	M6C	Humewood-Cedarvale	43.693781	-79.428191	0.000000
94	York	M6E	Caledonia-Fairbanks	43.689026	-79.453512	0.000000
95	York	M6M	Del Ray / Mount Dennis / Keelsdale and Silvert...	43.691116	-79.476013	0.000000
96	York	M6N	Runnymede / The Junction North	43.673185	-79.487262	0.000000
97	York	M9N	Weston	43.706876	-79.518188	0.000000

98 rows × 6 columns

Table 11 Toronto Dataframe for Indian restaurants count in each neighbourhood

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

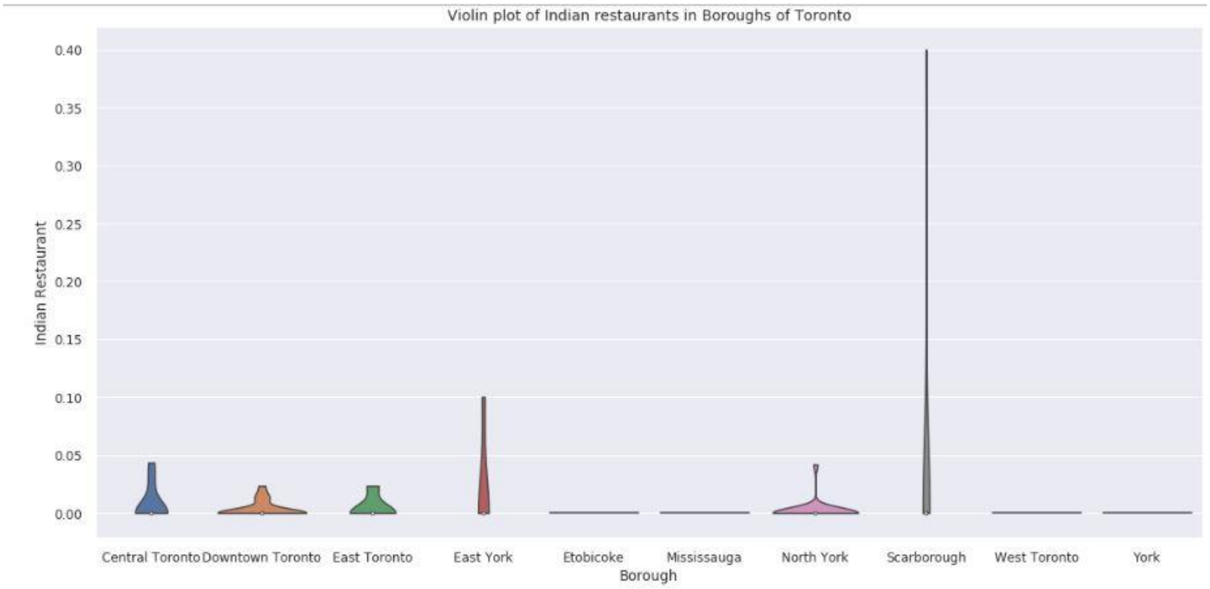


Figure 1 Violin plot

This plot helps in identifying the boroughs with densely populated Indian restaurants.

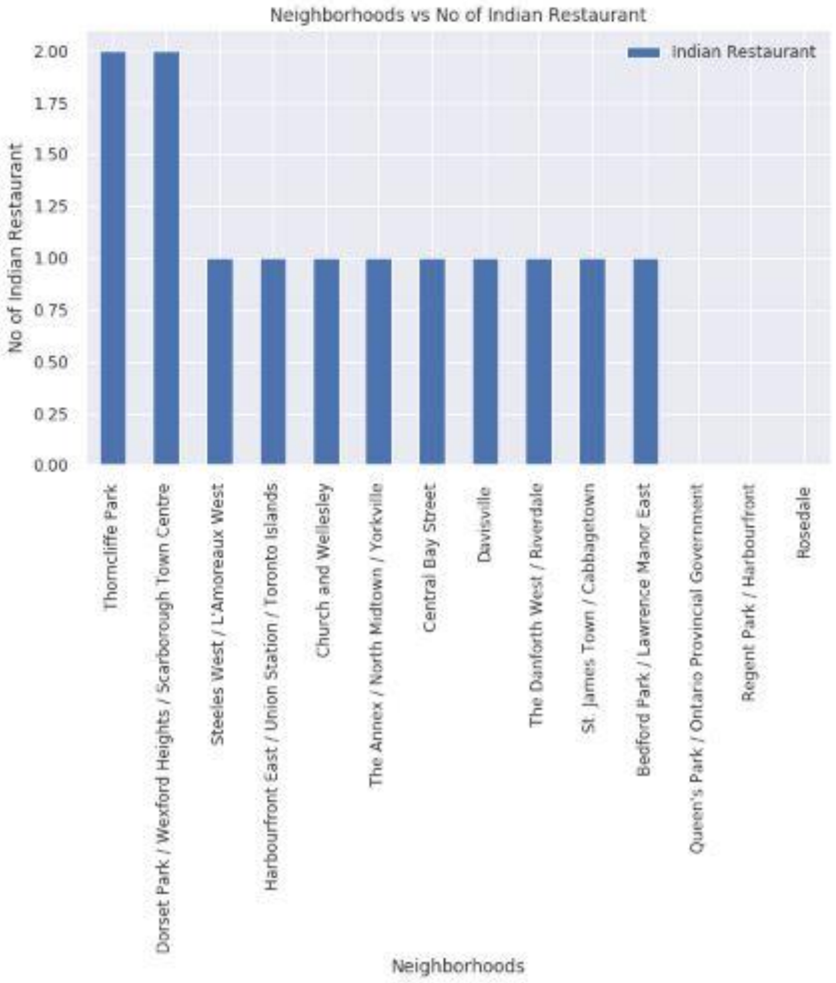
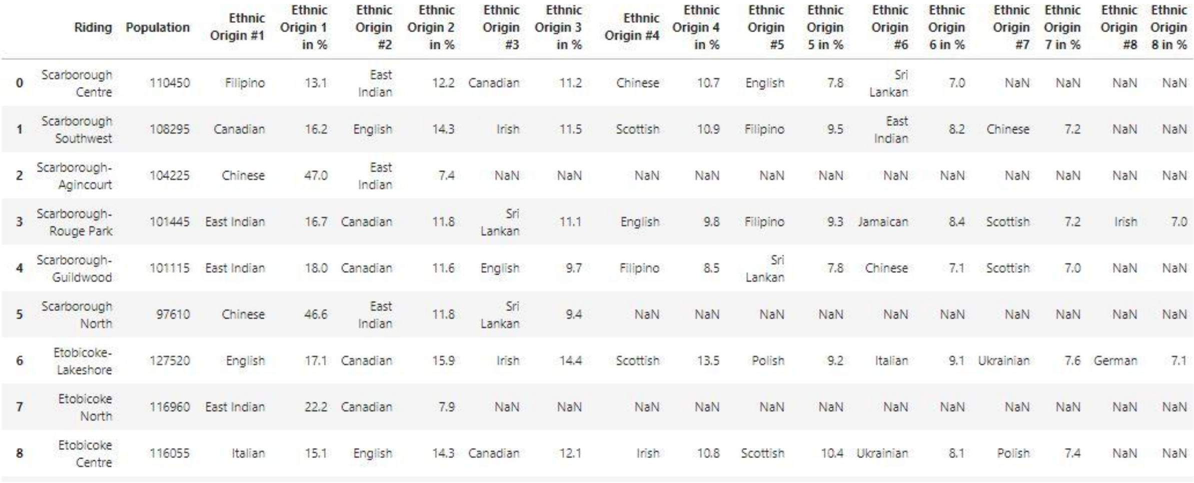


Figure 2Bar Plot

c) Relationship between Neighborhoods and Indian population

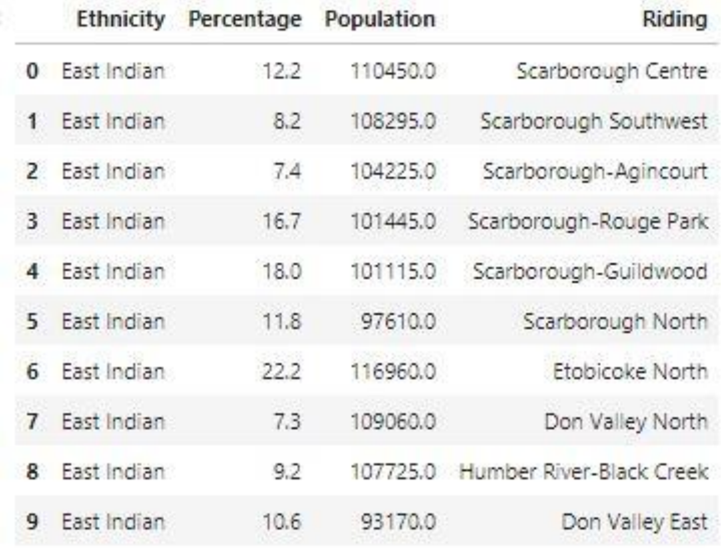
Another key feature is the distribution of Indian crowd in each neighbourhoods. Let us analyse the neighbourhoods and identify the neighbourhoods with highest number of Indian population.

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



	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 %	Ethnic Origin #8	Ethnic Origin 8 in %
0	Scarborough Centre	110450	Filipino	13.1	East Indian	12.2	Canadian	11.2	Chinese	10.7	English	7.8	Sri Lankan	7.0	NaN	NaN	NaN	NaN
1	Scarborough Southwest	108295	Canadian	16.2	English	14.3	Irish	11.5	Scottish	10.9	Filipino	9.5	East Indian	8.2	Chinese	7.2	NaN	NaN
2	Scarborough-Agincourt	104225	Chinese	47.0	East Indian	7.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Scarborough-Rouge Park	101445	East Indian	16.7	Canadian	11.8	Sri Lankan	11.1	English	9.8	Filipino	9.3	Jamaican	8.4	Scottish	7.2	Irish	7.0
4	Scarborough-Guildwood	101115	East Indian	18.0	Canadian	11.6	English	9.7	Filipino	8.5	Sri Lankan	7.8	Chinese	7.1	Scottish	7.0	NaN	NaN
5	Scarborough North	97610	Chinese	46.6	East Indian	11.8	Sri Lankan	9.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	Etobicoke-Lakeshore	127520	English	17.1	Canadian	15.9	Irish	14.4	Scottish	13.5	Polish	9.2	Italian	9.1	Ukrainian	7.6	German	7.1
7	Etobicoke North	116960	East Indian	22.2	Canadian	7.9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	Etobicoke Centre	116055	Italian	15.1	English	14.3	Canadian	12.1	Irish	10.8	Scottish	10.4	Ukrainian	8.1	Polish	7.4	NaN	NaN

Table 12 Dataframe with neighbourhoods & their population distribution



	Ethnicity	Percentage	Population	Riding
0	East Indian	12.2	110450.0	Scarborough Centre
1	East Indian	8.2	108295.0	Scarborough Southwest
2	East Indian	7.4	104225.0	Scarborough-Agincourt
3	East Indian	16.7	101445.0	Scarborough-Rouge Park
4	East Indian	18.0	101115.0	Scarborough-Guildwood
5	East Indian	11.8	97610.0	Scarborough North
6	East Indian	22.2	116960.0	Etobicoke North
7	East Indian	7.3	109060.0	Don Valley North
8	East Indian	9.2	107725.0	Humber River-Black Creek
9	East Indian	10.6	93170.0	Don Valley East

Table 13 Extracted dataframe with just Indian population information

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

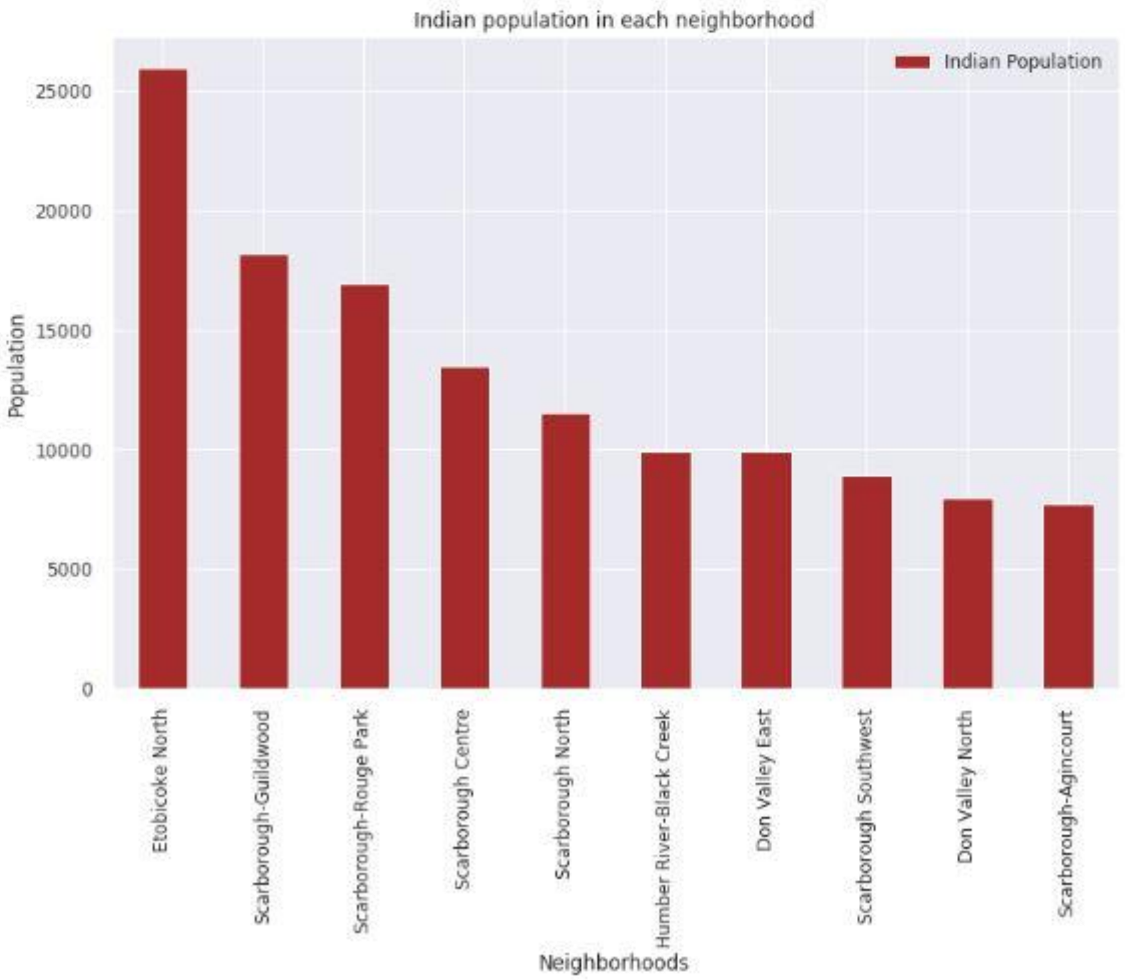


Figure 3Bar graph to show the population in each riding in Toronto

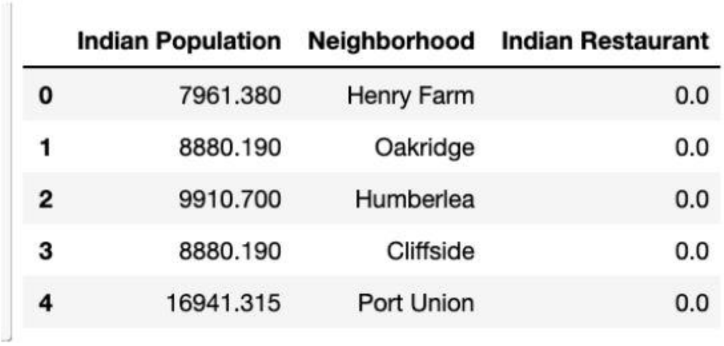
This analysis & visualization of the relationship between neighbourhoods & Indian population present in those neighbourhoods helps us in identifying the our perfect locations for Indian restaurants.

Once we identify those neighbourhoods, it will help us in deciding where to place the new Indian restaurant. Indian restaurant placed in an densely populated Indian neighbourhood is more likely to get more Indian customers than a restaurant placed in a neighbourhood with less or no Indian population.

Thus this analysis helps in determining the success of a new Indian restaurant in Toronto's neighbourhoods.

d) Relationship between Indian population and Indian restaurant

First get the list of neighbourhoods present in the riding using the Wikipedia geography section for each riding. Altering the riding names to match the Wikipedia page in order to retrieve the neighbourhoods present in those ridings.



	Indian Population	Neighborhood	Indian Restaurant
0	7961.380	Henry Farm	0.0
1	8880.190	Oakridge	0.0
2	9910.700	Humberlea	0.0
3	8880.190	Cliffside	0.0
4	16941.315	Port Union	0.0

Table 14 Dataframe of densely populated neighbourhoods with number of Indian restaurants

Predictive Modelling

a) Clustering Neighbourhoods of Toronto:

First step in K-means clustering is to identify the best K value i.e 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 **(i.e. toronto_merged dataframe)**

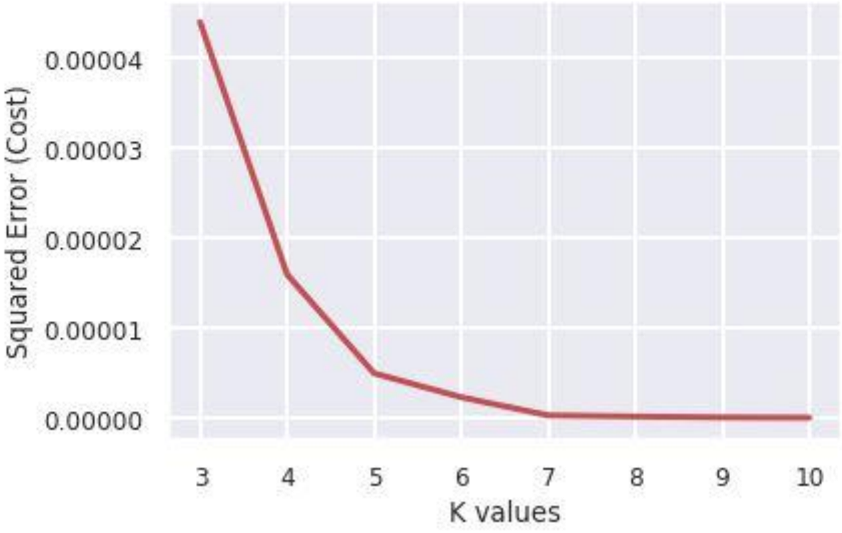
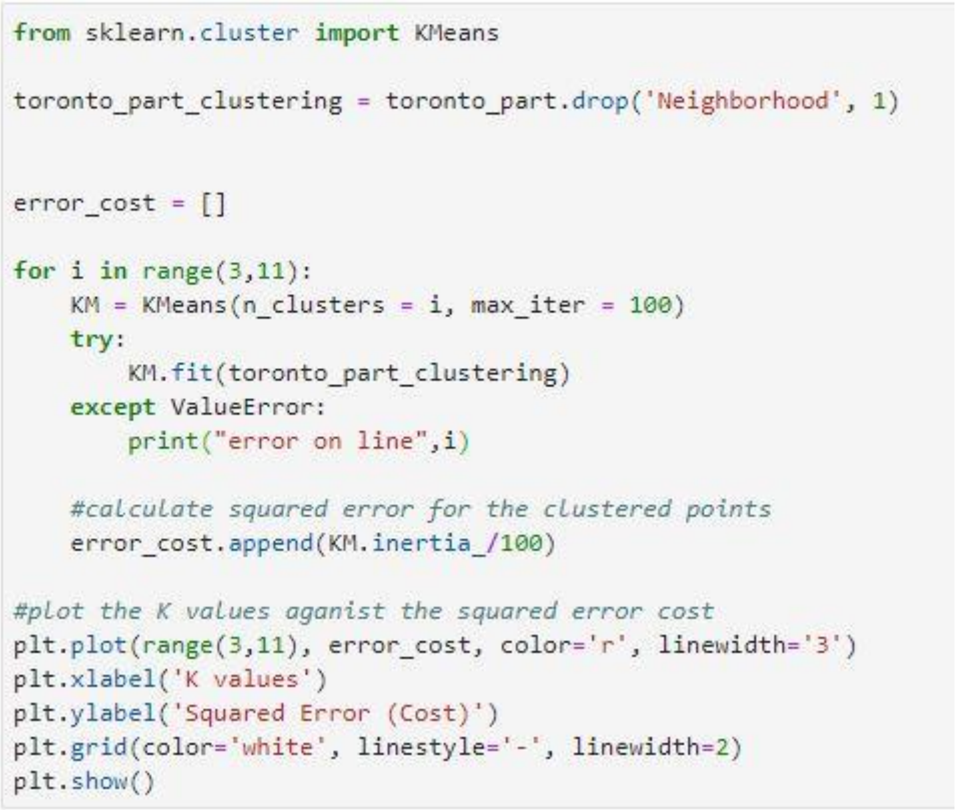
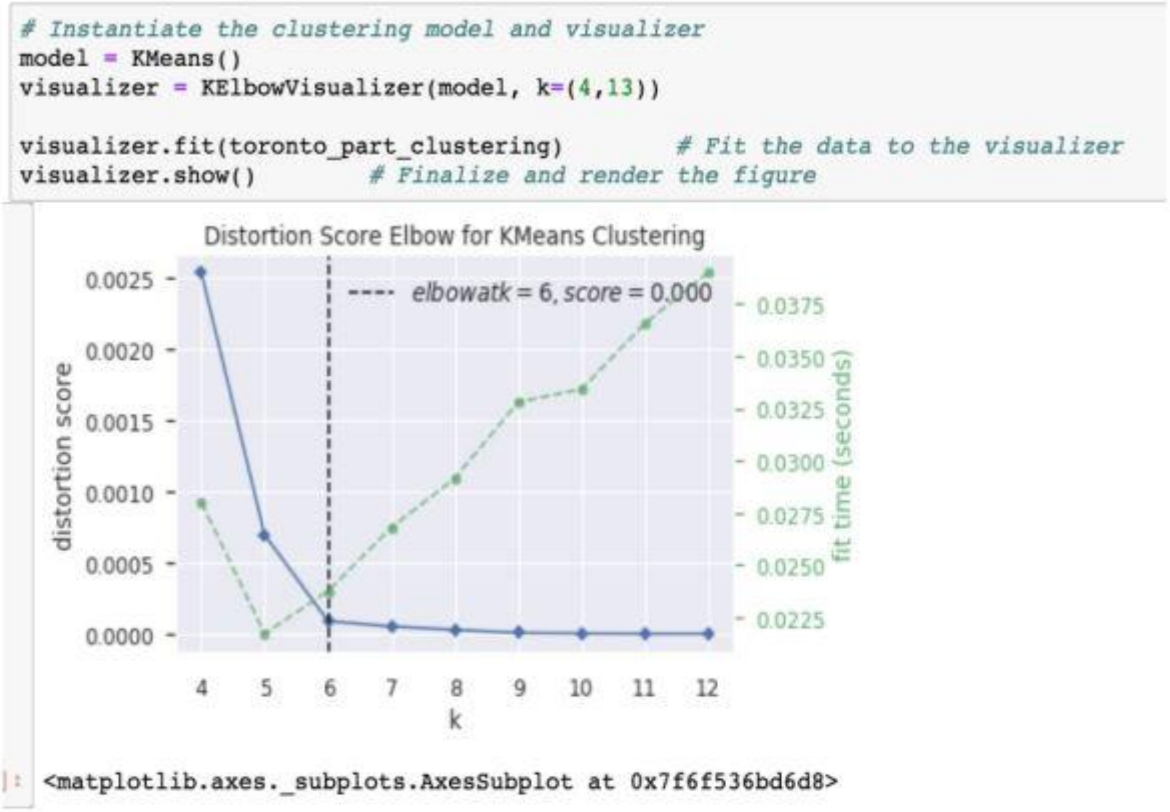


Figure 4 Elbow method to identify best k value



```
from sklearn.cluster import KMeans

toronto_part_clustering = toronto_part.drop('Neighborhood', 1)


error_cost = []

for i in range(3,11):
    KM = KMeans(n_clusters = i, max_iter = 100)
    try:
        KM.fit(toronto_part_clustering)
    except ValueError:
        print("error on line",i)

    #calculate squared error for the clustered points
    error_cost.append(KM.inertia_/100)

#plot the K values aganist the squared error cost
plt.plot(range(3,11), error_cost, color='r', linewidth='3')
plt.xlabel('K values')
plt.ylabel('Squared Error (Cost)')
plt.grid(color='white', linestyle='-', linewidth=2)
plt.show()
```



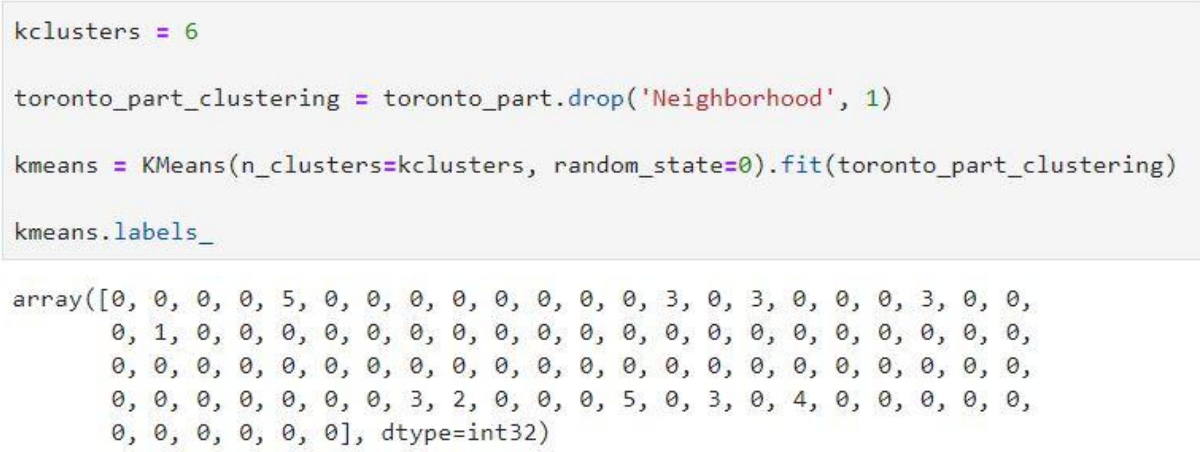
```
# Instantiate the clustering model and visualizer
model = KMeans()
visualizer = KElbowVisualizer(model, k=(4,13))

visualizer.fit(toronto_part_clustering)        # Fit the data to the visualizer
visualizer.show()        # Finalize and render the figure
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6f536bd6d8>
```

Table 15 Elbow visualizer to identify the K value

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

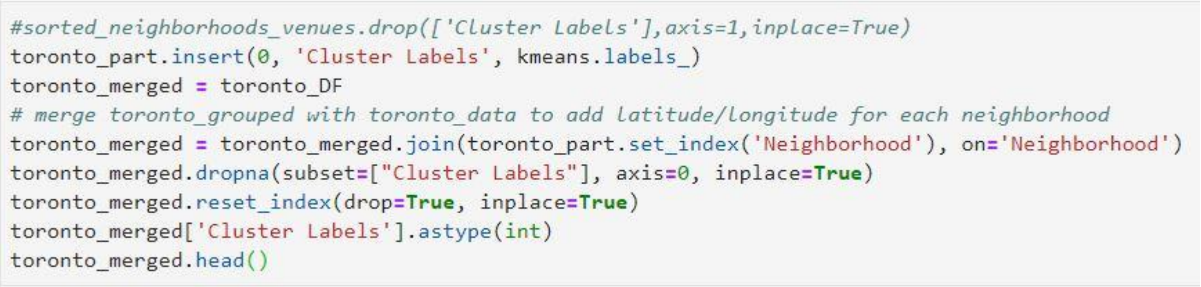
Clustering the Toronto Neighborhood Using K-Means with K = 6



```
kclusters = 6

toronto_part_clustering = toronto_part.drop('Neighborhood', 1)

kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_part_clustering)

kmeans.labels_
```

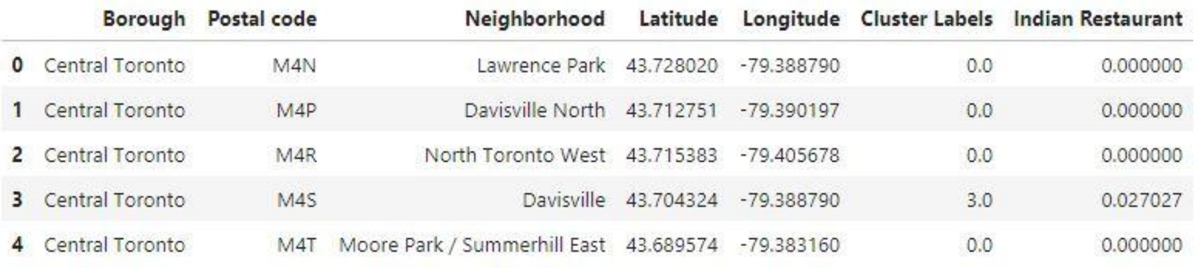
```
array([0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 0, 0, 0, 3, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 3, 2, 0, 0, 0, 5, 0, 3, 0, 4, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0], dtype=int32)
```

Code snippet 4 6 clusters & its labels



```
#sorted_neighborhoods_venues.drop(['Cluster Labels'],axis=1,inplace=True)
toronto_part.insert(0, 'Cluster Labels', kmeans.labels_)
toronto_merged = toronto_DF
# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
toronto_merged = toronto_merged.join(toronto_part.set_index('Neighborhood'), on='Neighborhood')
toronto_merged.dropna(subset=["Cluster Labels"], axis=0, inplace=True)
toronto_merged.reset_index(drop=True, inplace=True)
toronto_merged['Cluster Labels'].astype(int)
toronto_merged.head()
```

Code snippet 5 Clustering the Toronto dataframe



	Borough	Postal code	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790	0.0	0.000000
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197	0.0	0.000000
2	Central Toronto	M4R	North Toronto West	43.715383	-79.405678	0.0	0.000000
3	Central Toronto	M4S	Davisville	43.704324	-79.388790	3.0	0.027027
4	Central Toronto	M4T	Moore Park / Summerhill East	43.689574	-79.383160	0.0	0.000000

Table 16 Dataframe with cluster labels for neighbourhood

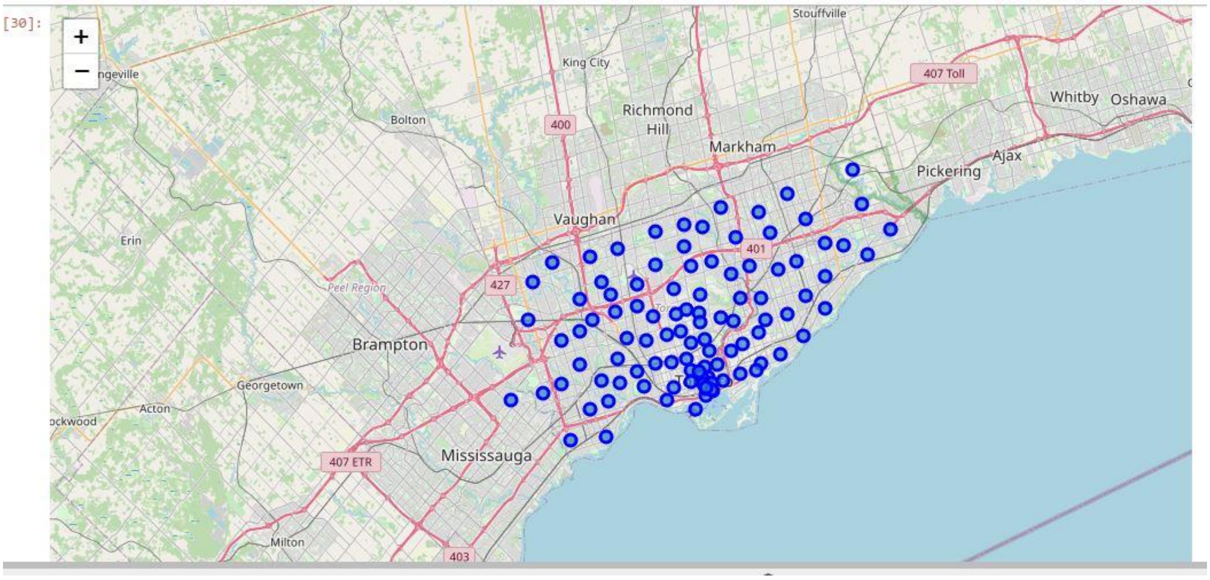
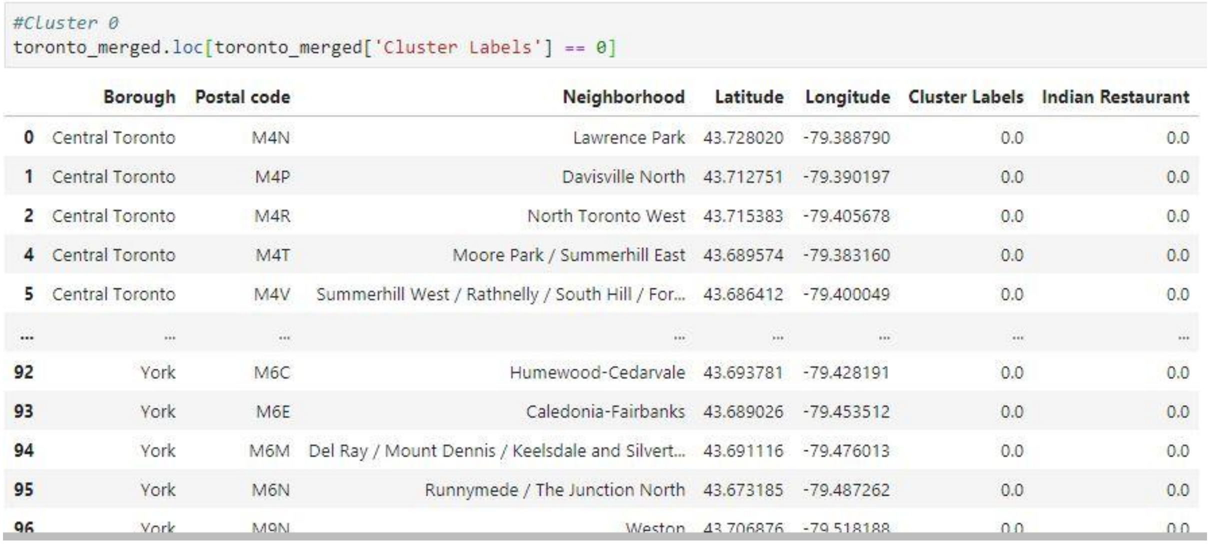


Figure 5 Folium map for the clusters of different neighbourhoods

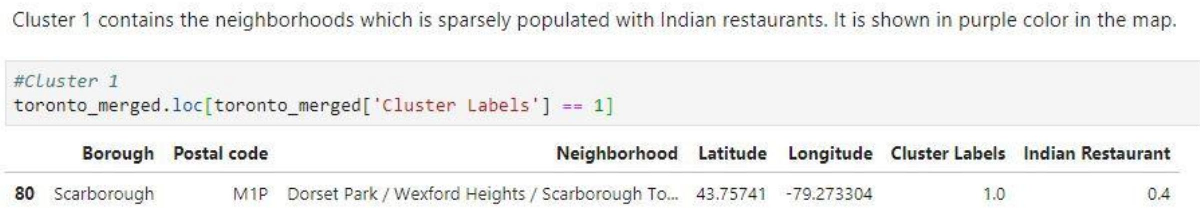
b) Examining the Clusters:

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



```
#Cluster 0
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0]
```

	Borough	Postal code	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	M4T	Moore Park / Summerhill East	43.689574	-79.383160	0.0	0.0
5	Central Toronto	M4V	Summerhill West / Rathnelly / South Hill / For...	43.686412	-79.400049	0.0	0.0
...	...	...	...	...	...	...	...
92	York	M6C	Humewood-Cedarvale	43.693781	-79.428191	0.0	0.0
93	York	M6E	Caledonia-Fairbanks	43.689026	-79.453512	0.0	0.0
94	York	M6M	Del Ray / Mount Dennis / Keelsdale and Silvert...	43.691116	-79.476013	0.0	0.0
95	York	M6N	Runnymede / The Junction North	43.673185	-79.487262	0.0	0.0
96	York	M9N	Weston	43.706876	-79.518188	0.0	0.0

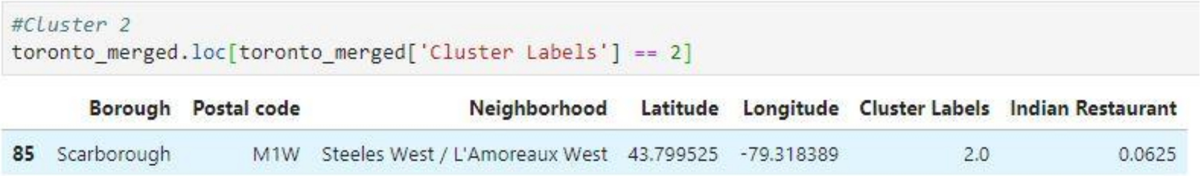


Cluster 1 contains the neighborhoods which is sparsely populated with Indian restaurants. It is shown in purple color in the map.

```
#Cluster 1
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1]
```

	Borough	Postal code	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
80	Scarborough	M1P	Dorset Park / Wexford Heights / Scarborough To...	43.75741	-79.273304	1.0	0.4

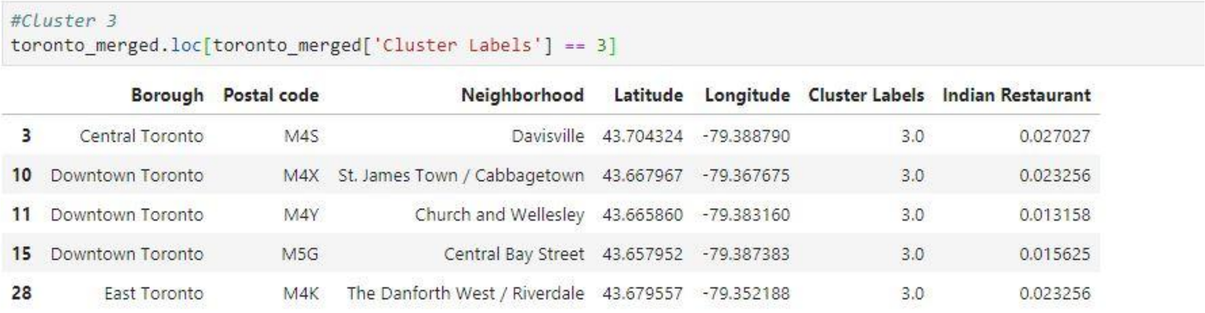
Cluster 2 has no rows meaning no data points or neighborhood was near to this centroid.



```
#Cluster 2
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2]
```

	Borough	Postal code	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
85	Scarborough	M1W	Steeles West / L'Amoreaux West	43.799525	-79.318389	2.0	0.0625

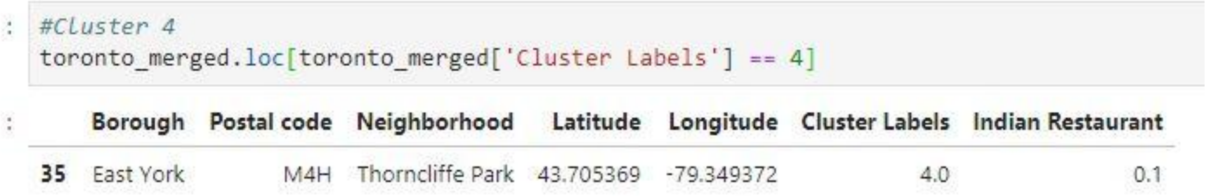
Cluster 3 contains all the neighborhoods which is medium populated with Indian restaurants. It is shown in blue color in the map.



```
#Cluster 3
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3]
```

	Borough	Postal code	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
3	Central Toronto	M4S	Davisville	43.704324	-79.388790	3.0	0.027027
10	Downtown Toronto	M4X	St. James Town / Cabbagetown	43.667967	-79.367675	3.0	0.023256
11	Downtown Toronto	M4Y	Church and Wellesley	43.665860	-79.383160	3.0	0.013158
15	Downtown Toronto	M5G	Central Bay Street	43.657952	-79.387383	3.0	0.015625
28	East Toronto	M4K	The Danforth West / Riverdale	43.679557	-79.352188	3.0	0.023256

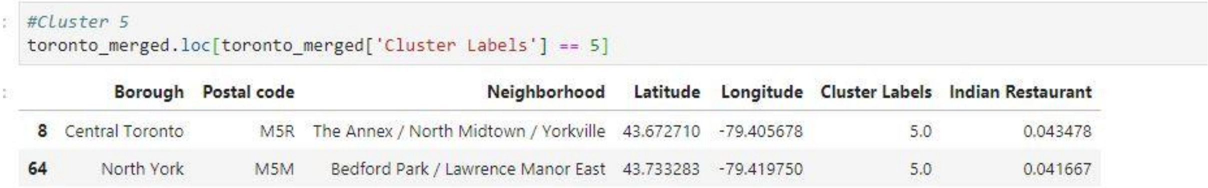
Cluster 4 has one row meaning only one data point or neighborhood was near to this centroid.



```
#Cluster 4
toronto_merged.loc[toronto_merged['Cluster Labels'] == 4]
```

	Borough	Postal code	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
35	East York	M4H	Thorncliffe Park	43.705369	-79.349372	4.0	0.1

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



```
#Cluster 5
toronto_merged.loc[toronto_merged['Cluster Labels'] == 5]
```

	Borough	Postal code	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
8	Central Toronto	M5R	The Annex / North Midtown / Yorkville	43.672710	-79.405678	5.0	0.043478
64	North York	M5M	Bedford Park / Lawrence Manor East	43.733283	-79.419750	5.0	0.041667

Results and Discussion

Results

We have reached the end of the analysis, in the result section we can document all the findings from above clustering & visualization of the data(s). In this project, as the business problem started with identifying a good neighbourhood to open a new Indian restaurant, we looked into all the neighbourhoods in Toronto, analysed the Indian population in each neighbourhood & spread of Indian restaurants in those neighbourhoods to come to conclusion about which neighbourhood would be a better spot for opening a new Indian restaurant. I have used data from web resources like Wikipedia, geospatial coordinates of Toronto neighbourhoods, and Foursquare API, 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-Guild wood, 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 neighbourhoods.

Discussion

According to this analysis, Scarborough borough will provide least competition for the new upcoming Indian restaurant as there are very little Indian restaurants spread or no Indian restaurants in neighborhoods. Also looking at the population distribution, it looks like it is densely populated with Indian crowd which gives the new restaurant high possibility of customers.

So, definitely this region could potentially be a perfect place for starting a quality Indian restaurant.

Some of the drawbacks of this analysis are —

- The clustering is completely based only on data obtained from Foursquare API.
- The Indian population distribution in each neighbourhood is also based on the 2016 census which is not up-to date. Thus population distribution would have definitely changed by 2020 given 3 years gap in the data.
- Since population distribution of Indian crowd in each neighbourhood & number of Indian restaurants are the major features in this analysis, it is not fully up-to date data, this analysis is definitely not far from being conclusive & it has lot of areas where it can be improved.

However, it certainly provides us with some good insights, preliminary information on possibilities & a head start into this business problem by setting the step stones properly. Furthermore, this may also potentially vary depending on the type of clustering techniques that we use to examine the data.

Conclusion

Finally to conclude this project, we got a chance to solve a business problem like real data scientists would do. We have made use of Foursquare API to explore the venues in neighbourhoods of Toronto, then got a good amount of data from Wikipedia which we scraped with the help of Wikipedia Python Library and Visualized it using various plots present in seaborne & matplotlib.

We also applied **machine learning techniques** to predict the output given by the data as well as used Folium to visualize it on a map.

Also, some of the drawbacks or areas of improvements shows us that this analysis can further be improved with the help of more data and different machine learning techniques.

Similarly we can use this project to analysis any scenario such as opening a different cuisine or success of opening a new gym etc. Hopefully, this project helps acts as initial guidance to take more complex real-life challenges using data-science.