Exploring Toronto Neighbourhoods - To open a new Indian Restaurant

Amit Kumar Saha

Sept 9, 2020

Introduction

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 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.

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

b) /en.m.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity) wiki page. Using this page I'm going 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.

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.

c) Longitude: The longitude value of the venue.Then I'm using "https://cocl.us/Geospatial_data" csv file to get all the geographical coordinates of the neighbourhoods.

d) To get information about the distribution of population by their ethnicity I'm using "Demographics of Toronto" (https:/

2.2 Data Cleaning

a) Scraping Toronto Neighbourhoods 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 Neighbourhoods in it.

<u>Assumptions made to attain the below DataFrame</u>:

- Dataframe will consist of three columns: PostalCode, 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.

```
In [10]: html = wp.page("List of postal codes of Canada: M").html().encode("UTF-8")
         df = pd.read_html(html, header = 0)[0]
         df.head()
```

Out[10]:

	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

Table 1: Dataframe formed from the scraped wiki page

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

Out[12]:

	Borough	Postalcode	Neighbourhood
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 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.

In [13]:
```python
#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

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 dataframe into one by merging on the postal code.

In [15]:
```python
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]:
```python
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 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.

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

Code snippet: Scraping the wiki page

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.

	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 %	Ethnic Origin #9	Ethnic Origin 9 in %
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	NaN	NaN	NaN	NaN
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	NaN	NaN	NaN	NaN
2	Davenport	107395	Portuguese	22.7	English	13.6	Canadian	12.8	Irish	11.5	Italian	11.1	Scottish	11.00	NaN	NaN	NaN	NaN	NaN	NaN
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	NaN	NaN	NaN	NaN
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	Greek	7.3	NaN	NaN
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	Italian	7.3	French	7.2
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	Italian	7.4	NaN	NaN
7	Toronto Centre	99590	English	15.7	Canadian	13.7	Irish	13.4	Scottish	12.6	Chinese	12.5	French	7.20	NaN	NaN	NaN	NaN	NaN	NaN

Table 4: TORONTO & EAST YORK population distribution by ethnicity

	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	Willowdale	117405	Chinese	25.9	Iranian	12.1	Korean	10.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Eglinton-Lawrence	112925	Canadian	14.7	English	12.6	Polish	12.0	Filipino	11.0	Scottish	9.7	Italian	9.5	Irish	9.2	Russian	8.4
2	Don Valley North	109060	Chinese	32.4	East Indian	7.3	Iranian	7.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Humber River-Black Creek	107725	Italian	12.8	East Indian	9.2	Jamaican	8.5	Vietnamese	8.0	Canadian	7.4	NaN	NaN	NaN	NaN	NaN	NaN
4	York Centre	103760	Filipino	17.0	Italian	13.4	Russian	9.5	Canadian	8.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	Don Valley West	101790	English	19.2	Canadian	15.1	Scottish	14.9	Irish	14.2	Chinese	11.2	NaN	NaN	NaN	NaN	NaN	NaN
6	Don Valley East	93170	East Indian	10.6	Canadian	10.4	English	10.1	Chinese	8.9	Irish	8.1	Scottish	8.0	Filipino	7.8	NaN	NaN

Table 5: NORTH YORK population distribution by ethnicity

	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

Table 6: SCARBOROUGH population distribution by ethnicity

	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	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
1	Etobicoke North	116960	East Indian	22.2	Canadian	7.9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	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
3	York South-Weston	115130	Portuguese	14.5	Italian	12.8	Canadian	8.7	Jamaican	8.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 7: ETOBICOKE & YORK population distribution by ethnicity

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.

```
In [32]: toronto_venues.head(10)
```

Out[32]:

	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	Summerhill Market North	43.715499	-79.392881	Food & Drink Shop
5	Davisville North	43.712751	-79.390197	Sherwood Park	43.716551	-79.387776	Park
6	Davisville North	43.712751	-79.390197	Winners	43.713236	-79.393873	Clothing 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	43.708378	-79.390473	Sandwich Place
9	Davisville North	43.712751	-79.390197	Gym	43.713126	-79.393537	Gym

Table 8: Dataframe with venues in each neighbourhood along with the category info of the venues.

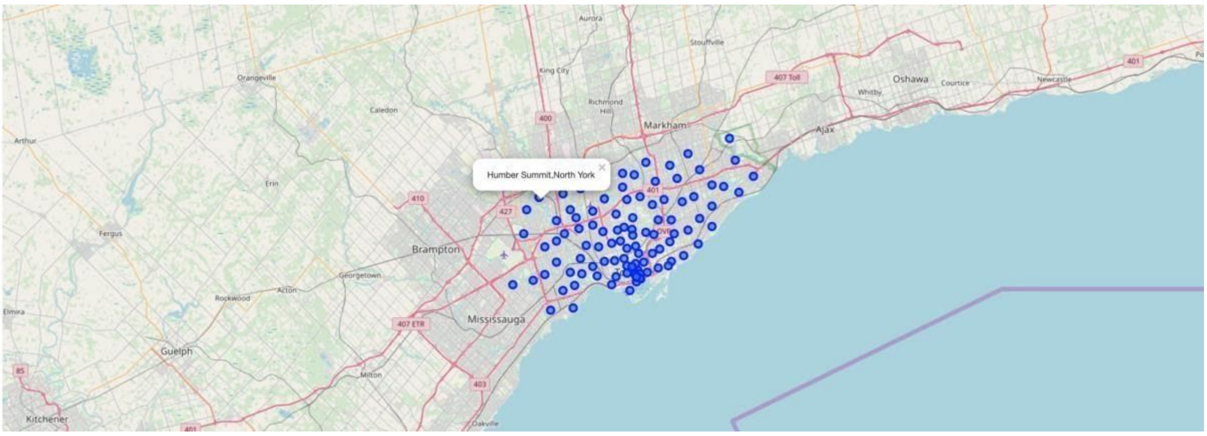
3. Exploratory Data Analysis:

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.

```python
# 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_DF['Neighborhood']):
    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: To draw the folium map



Map 1: Folium map of Toronto Neighbourhood with popup label

3.2 *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	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	W
0	Adelaide, King, Richmond	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.030000	0.000000	0.0	0.010000	0.010000	0.000000	0.03	0.000000	0.0	
1	Agincourt	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00	0.000000	0.0	
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	0.000000	0.0	0.000000	0.000000	0.000000	0.00	0.000000	0.0	
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00	0.000000	0.0	

Table 9: Dataframe formed using Foursquare API information about venues in each neighbourhood

Code snippet

```
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: 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	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790	0	0.000000
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197	0	0.000000
2	Central Toronto	M4R	North Toronto West	43.715383	-79.405678	0	0.000000
3	Central Toronto	M4S	Davisville	43.704324	-79.388790	4	0.028571
4	Central Toronto	M5N	Roselawn	43.711695	-79.416936	0	0.000000
5	Downtown Toronto	M4W	Rosedale	43.679563	-79.377529	0	0.000000
6	Downtown Toronto	M4Y	Church and Wellesley	43.665860	-79.383160	5	0.011765
7	Downtown Toronto	M5C	St. James Town	43.651494	-79.375418	5	0.010000
8	Downtown Toronto	M5E	Berczy Park	43.644771	-79.373306	4	0.017544
9	Downtown Toronto	M5G	Central Bay Street	43.657952	-79.387383	5	0.011364
10	Downtown Toronto	M5W	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	5	0.010101

Table 10: Toronto Dataframe for Indian restaurants count in each neighbourhood

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

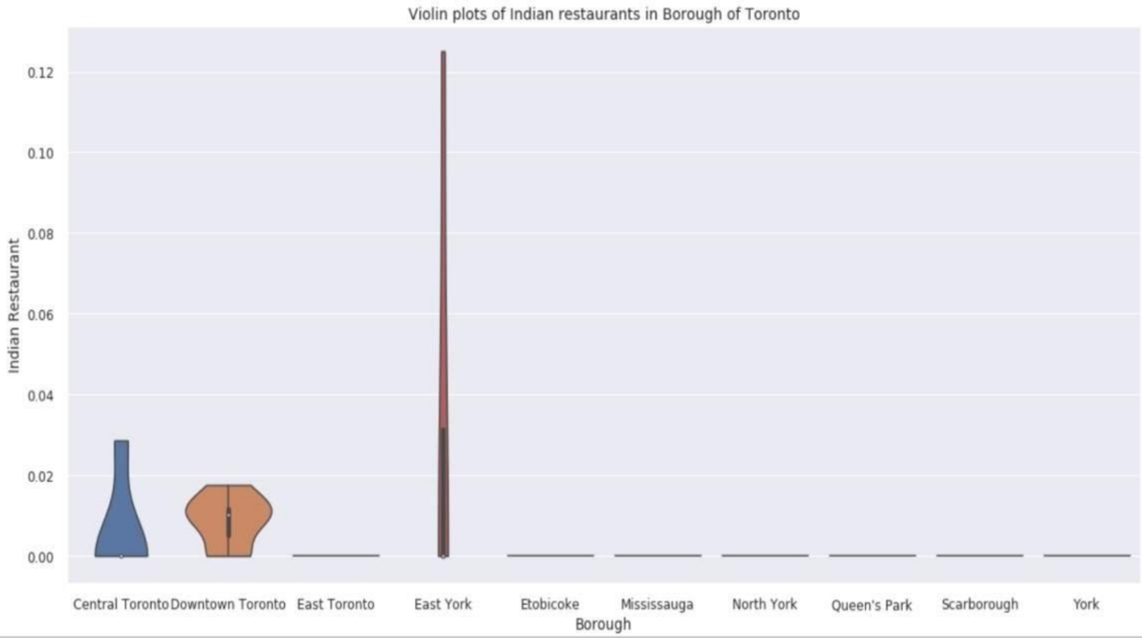


Figure 1: Violin plot

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 neighbourhood with Indian Restaurants:

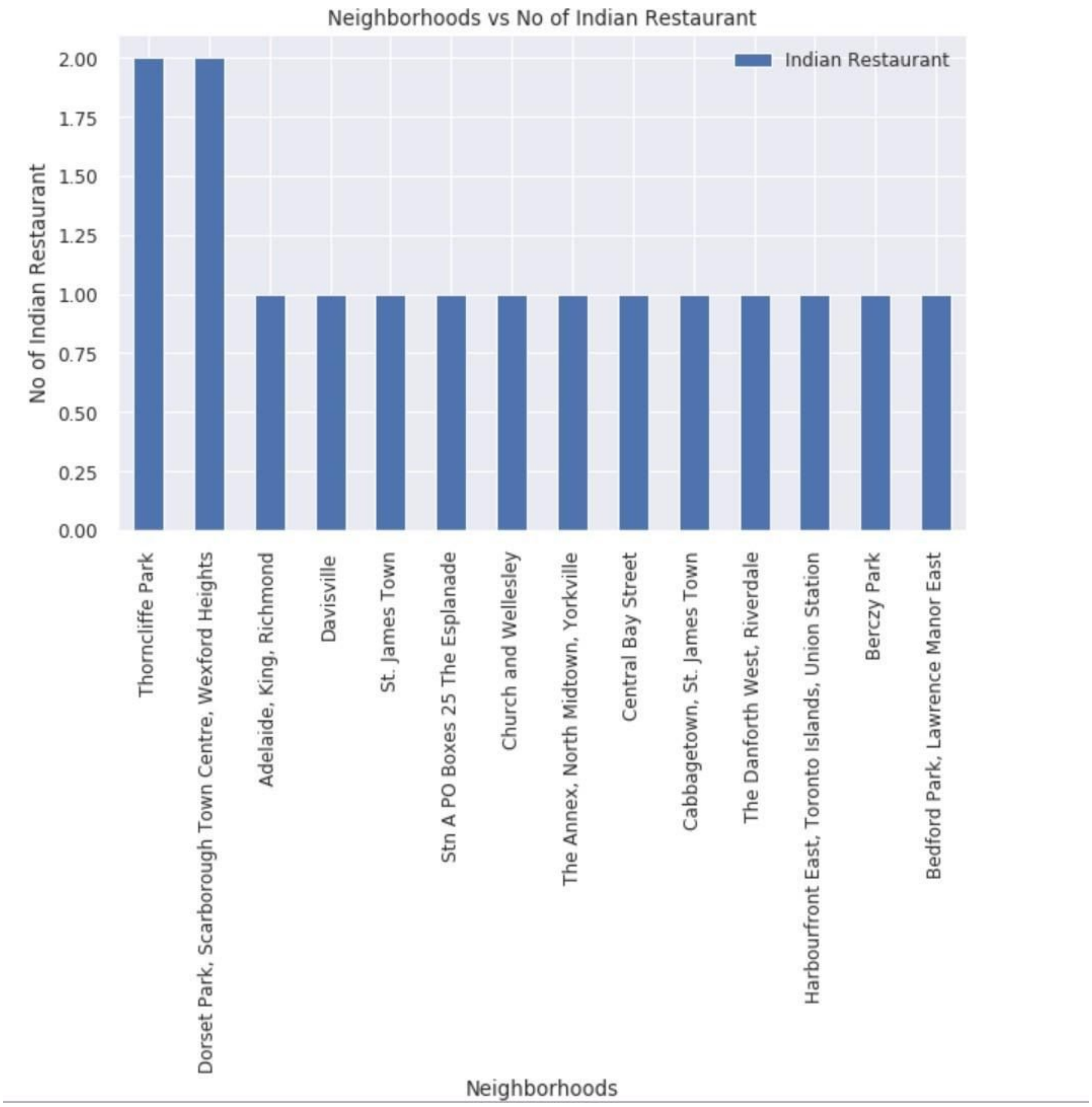


Figure 2: Bar plot

3.3 Relationship between neighbourhood 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 %	Ethnic Origin #9	Ethnic Origin 9 in %
0	Willowdale	117405	Chinese	25.9	Iranian	12.1	Korean	10.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Eglinton-Lawrence	112925	Canadian	14.7	English	12.6	Polish	12.0	Filipino	11.0	Scottish	9.7	Italian	9.50	Irish	9.2	Russian	8.4	NaN	NaN
2	Don Valley North	109060	Chinese	32.4	East Indian	7.3	Iranian	7.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Humber River-Black Creek	107725	Italian	12.8	East Indian	9.2	Jamaican	8.5	Vietnamese	8.0	Canadian	7.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	York Centre	103760	Filipino	17.0	Italian	13.4	Russian	9.5	Canadian	8.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	Don Valley West	101790	English	19.2	Canadian	15.1	Scottish	14.9	Irish	14.2	Chinese	11.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	Don Valley East	93170	East Indian	10.6	Canadian	10.4	English	10.1	Chinese	8.9	Irish	8.1	Scottish	8.00	Filipino	7.8	NaN	NaN	NaN	NaN
7	Scarborough Centre	110450	Filipino	13.1	East Indian	12.2	Canadian	11.2	Chinese	10.7	English	7.8	Sri Lankan	7.00	NaN	NaN	NaN	NaN	NaN	NaN
8	Scarborough Southwest	108295	Canadian	16.2	English	14.3	Irish	11.5	Scottish	10.9	Filipino	9.5	East Indian	8.20	Chinese	7.2	NaN	NaN	NaN	NaN
9	Scarborough-Agincourt	104225	Chinese	47.0	East Indian	7.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
10	Scarborough-Rouge Park	101445	East Indian	16.7	Canadian	11.8	Sri Lankan	11.1	English	9.8	Filipino	9.3	Jamaican	8.40	Scottish	7.2	Irish	7.0	NaN	NaN
11	Scarborough-Guildwood	101115	East Indian	18.0	Canadian	11.6	English	9.7	Filipino	8.5	Sri Lankan	7.8	Chinese	7.10	Scottish	7.0	NaN	NaN	NaN	NaN
12	Scarborough North	97610	Chinese	46.6	East Indian	11.8	Sri Lankan	9.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 11: Dataframe with neighbourhoods & their population distribution

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

Table 12: Extracted dataframe with just Indian population information

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

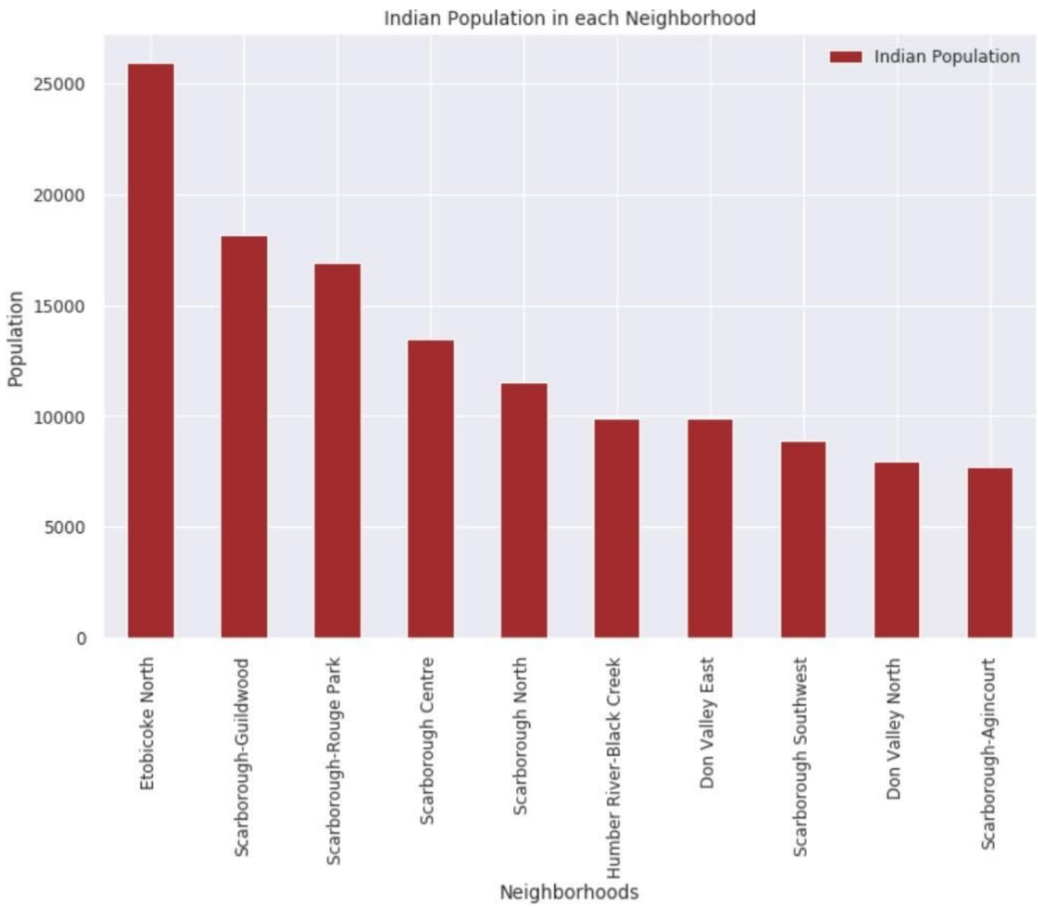


Figure 3: Bar graph to show the population in each riding in Toronto

This analysis & visualization of the relationship between neighbourhoods & Indian population present in those neighbourhoods helps us in identifying the highly populated Indian neighbourhoods. Once we identify those neighbourhoods it helps 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 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.

	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 13: Dataframe of densely populated neighbourhoods with number of Indian restaurants

4. Predictive Modelling:

4.1 Clustering Neighbourhoods 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.

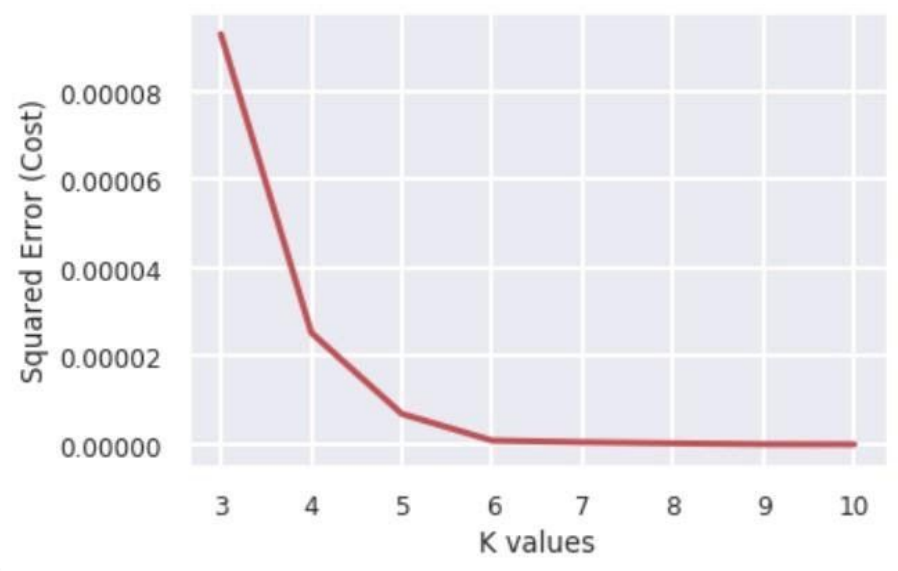
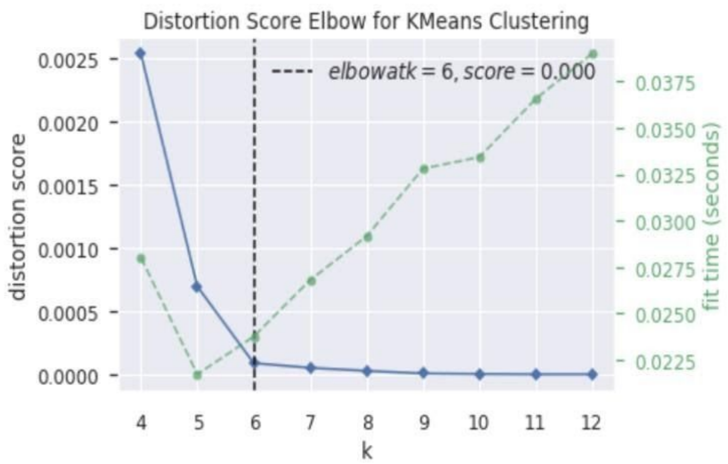


Figure 4: Elbow method to identify best k value

Code snippet —

```
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 14: Elbow visualizer to identify the K value

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

Code snippet: 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: clustering the Toronto dataframe

	Borough	Postalcode	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	4.0	0.028571
4	Central Toronto	M5N	Roselawn	43.711695	-79.416936	0.0	0.000000

Table 14: Dataframe with cluster labels for neighbourhood

Out[213]:

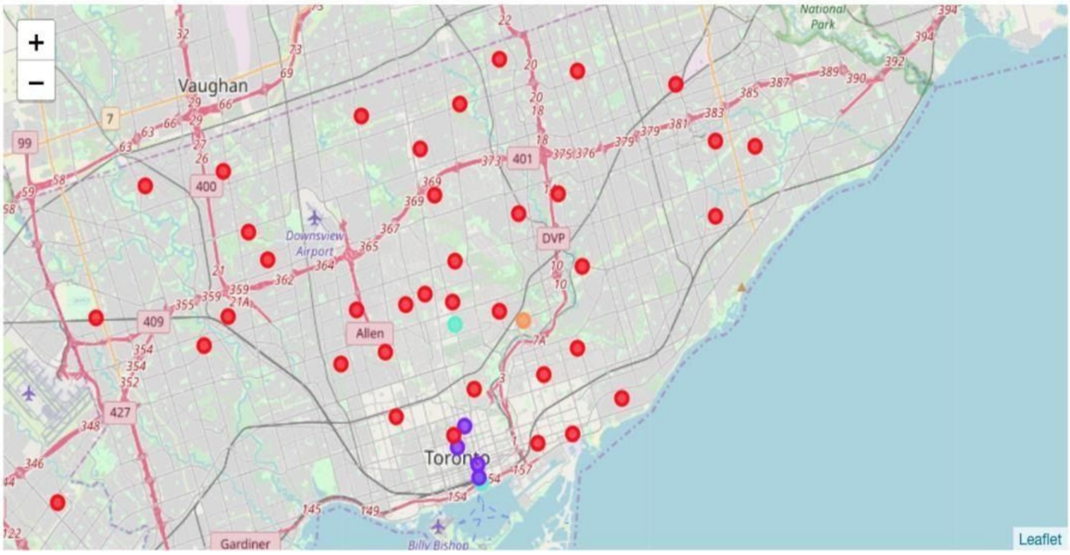


Figure 5: Folium map for the clusters of different neighbourhoods

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.

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

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

	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
11	Downtown Toronto	M6G	Christie	43.669542	-79.422564	0.0	0.0
12	East Toronto	M4E	The Beaches	43.676357	-79.293031	0.0	0.0
13	East Toronto	M4M	Studio District	43.659526	-79.340923	0.0	0.0
14	East Toronto	M7Y	Business Reply Mail Processing Centre 969 Eastern	43.662744	-79.321558	0.0	0.0
15	East York	M4C	Woodbine Heights	43.695344	-79.318389	0.0	0.0
16	East York	M4G	Leaside	43.709060	-79.363452	0.0	0.0
18	East York	M4J	East Toronto	43.685347	-79.338106	0.0	0.0
19	Etobicoke	M9P	Westmount	43.696319	-79.532242	0.0	0.0
20	Etobicoke	M9W	Northwest	43.706748	-79.594054	0.0	0.0
21	Mississauga	M7R	Canada Post Gateway Processing Centre	43.636966	-79.615819	0.0	0.0

Table 15: Cluster 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	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
6	Downtown Toronto	M4Y	Church and Wellesley	43.665860	-79.383160	1.0	0.011765
7	Downtown Toronto	M5C	St. James Town	43.651494	-79.375418	1.0	0.010000
9	Downtown Toronto	M5G	Central Bay Street	43.657952	-79.387383	1.0	0.011364
10	Downtown Toronto	M5W	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	1.0	0.010101

Table 16: Cluster 1

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

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

]:

Borough	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant

Table 17: Cluster 2

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

]:

Borough	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant

Table 18: Cluster 4

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	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
3	Central Toronto	M4S	Davisville	43.704324	-79.388790	3.0	0.028571
8	Downtown Toronto	M5E	Berczy Park	43.644771	-79.373306	3.0	0.017544

Table 19: cluster 3

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

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

]:

	Borough	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
17	East York	M4H	Thorncliffe Park	43.705369	-79.349372	5.0	0.125

Table 10: Cluster 5

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 neighbourhood to open a new Indian restaurant. To achieve that we looked into all the neighbourhoods in Toronto, analysed the Indian population in each neighbourhood & number of Indian restaurants in those neighbourhoods to come to conclusion about which neighbourhood 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 neighbourhoods.

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 neighbourhoods. 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 neighbourhood 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 analyse and visualize those datasets. We have made use of Foursquare API to explore the venues in neighbourhoods 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.