

# Exploring Toronto Neighborhoods - To open a new Indian Restaurant

Rajasekar Manokaran

June 1, 2020

## Introduction

As a part of IBM Data Science professional certification Capstone Project, we have to work on the real data to get an experience of what a data scientist go through in real life. I would like to explore the city Toronto and the objectives of this assignments is to define a business problem, find 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. I go through all the process in a step by step manner from Business problem, 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 Problem

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

Toronto, the capital of the province of Ontario, is the most widespread 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 Canada with more than half of the entire Indian Canadian population be located in Toronto. it might be one of the best places to start an Indian restaurant. In this project we will go through step by step process to decide 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. Let us analyze and explore whether it is a profitable idea or not. If profitable in which place, we can start the restaurant, so that it gets 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](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](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](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:

- a. Dataframe will consist of three columns: PostalCode, Borough, and Neighborhood
- b. Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.
- c. 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.

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

Note: Wikipedia — package is used to scrape the data from wiki.

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

#### b) Adding geographical coordinates to the neighbourhoods

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

```
[6] ▶ ML 8/8
lat_long_df = lat_long_df.rename(columns={'Postal Code': 'Postalcode'})
lat_long_df.head()
```

	Postalcode	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 dataframe 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, Lawrence Park	43.715383	-79.405678
3	Central Toronto	M4S	Davisville	43.704324	-79.388790
4	Central Toronto	M4T	Moore Park, Summerhill East	43.689574	-79.383160

```

[8] In [8]: MI
print('The dataframe has {} boroughs and {} neighborhoods.'.format(
    len(toronto_DF['Borough'].unique()),
    toronto_DF.shape[0]
))
The dataframe has 10 boroughs and 103 neighborhoods.

```

### 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 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.

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

```

[9] In [9]: MI
html = wp.page("Demographics of Toronto").html().encode("UTF-8")

```

```

[36] In [36]: MI
TEY_population_df.head(1)

```


	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 %
0	Willowdale	117405	Chinese	25.9	Iranian	12.1	Korean	10.6	NaN	NaN	NaN	NaN	NaN	NaN

```

[37] In [37]: MI
North_population_df.head(1)


```

	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 %
0	Scarborough Centre	110450	Filipino	13.1	East Indian	12.2	Canadian	11.2	Chinese	10.7	English	7.8	Sri Lankan	

[38] ▶ MI 

```
Scar_population_df.head(1)
```

	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	Ethn Orig 6 in %
0	Etobicoke-Lakeshore	127520	English	17.1	Canadian	15.9	Irish	14.4	Scottish	13.5	Polish	9.2	Italian	9


[39] ▶ MI 

```
ETY_population_df.head(1)
```

	Ethnic groups in the Toronto CMA (2016)Source: Focus on Geography Series, 2016 Census; Toronto, (CMA) - Ontario	Ethnic groups in the Toronto CMA (2016)Source: Focus on Geography Series, 2016 Census; Toronto, (CMA) - Ontario.1	Population	Ethnic Origin 1 in %
0	Ethnic group	White	2804630	47.8

#### d) Get location data using Foursquare

Foursquare API is very useful online application used by many developers & other applications like Uber etc. In this project I have used it to retrieve information about the places present in the neighbourhoods 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 neighbourhood within a radius of 1km.

[28] ▶ MI 

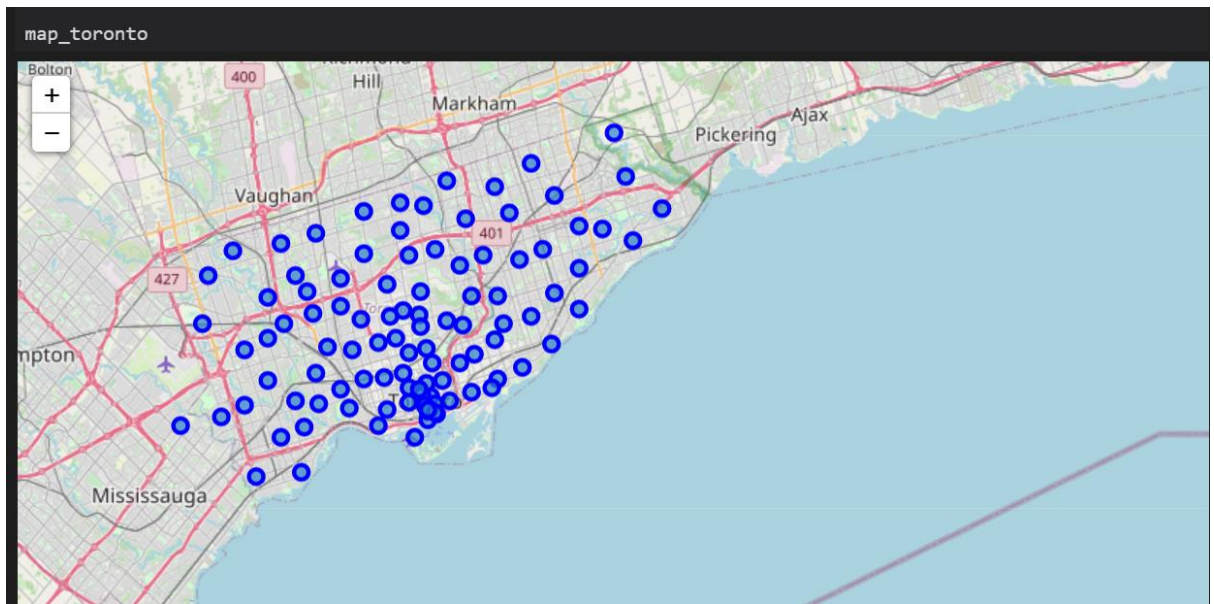
```
toronto_venues.head()
```

	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

### 3. Exploratory Data Analysis:

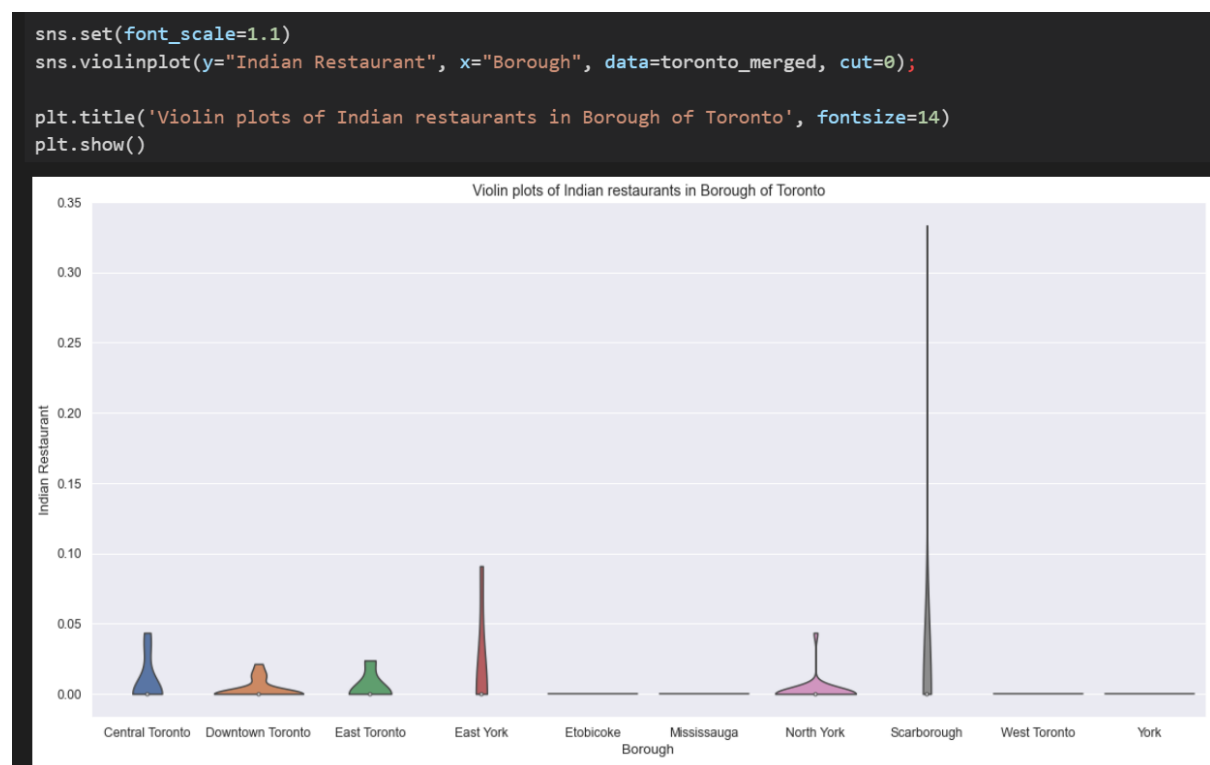
#### 3.1 Folium Library and Leaflet Map

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



### 3.2 Relationship between neighbourhood and Indian Restaurant

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



### 3.3 Relationship between Indian population and Indian Restaurant

First get the list of neighborhoods present in the riding using the wikipedia geography section for each riding. Altering the riding names to match the wikipedia page so we can retrieve the neighborhoods present in those ridings

```

toronto_part['split_neighborhoods'] = toronto_part['Neighborhood'].str.split(',')
toronto_part.drop(columns=['Neighborhood'],inplace=True,axis=1)
toronto_part = toronto_part.split_neighborhoods.apply(pd.Series).merge(toronto_part, left_index=True, right_index = True).drop(["split_neighborhoods"], axis = 1)\
    .melt(id_vars = ['Indian Restaurant'], value_name = "Neighborhood").drop(
("variable", axis = 1).dropna()

toronto_part.reset_index()
toronto_part

```

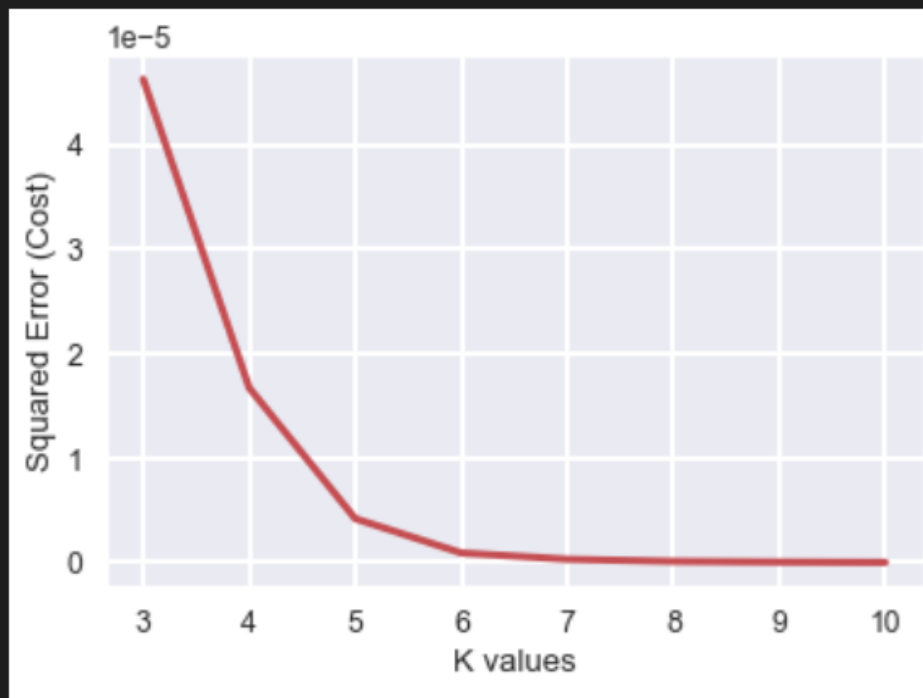
	Indian Restaurant	Neighborhood
0	0.000000	Agincourt
1	0.000000	Alderwood
2	0.000000	Bathurst Manor
3	0.000000	Bayview Village
4	0.043478	Bedford Park
...	...	...
579	0.000000	Island airport
628	0.000000	Royal York South East
641	0.000000	Thistletown
723	0.000000	Kingsway Park South East
736	0.000000	Albion Gardens

#### 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.

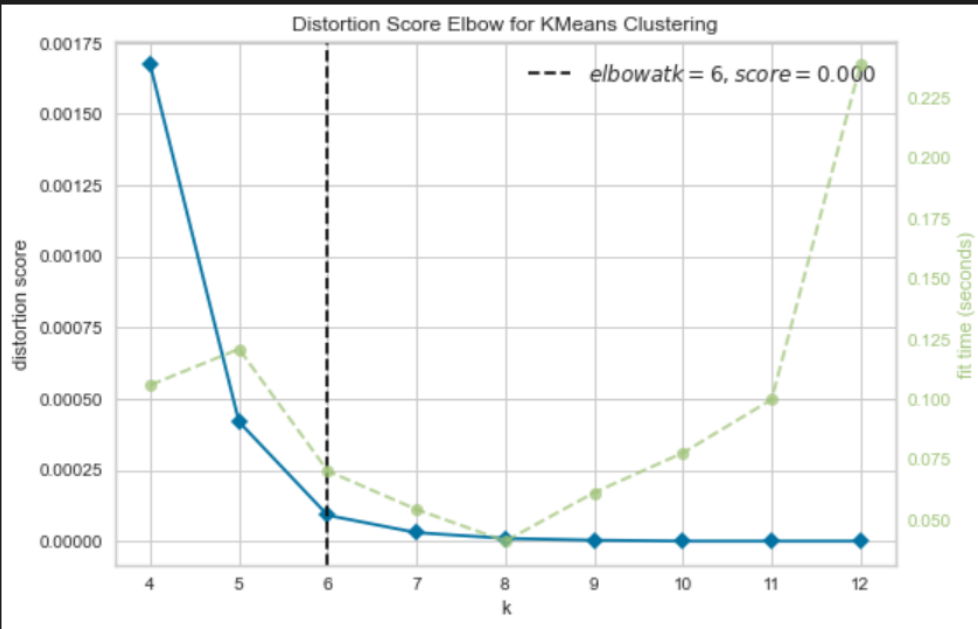
```
plt.plot(range(3,11), error_cost, color='r', linewidth='3')
plt.xlabel('K values')
plt.ylabel('Squared Error (Cost)')
plt.grid(color='white', linestyle='-', linewidth=2)
plt.show()
```





```
# 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 0x13d7d5...
```

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

[50]  M↓

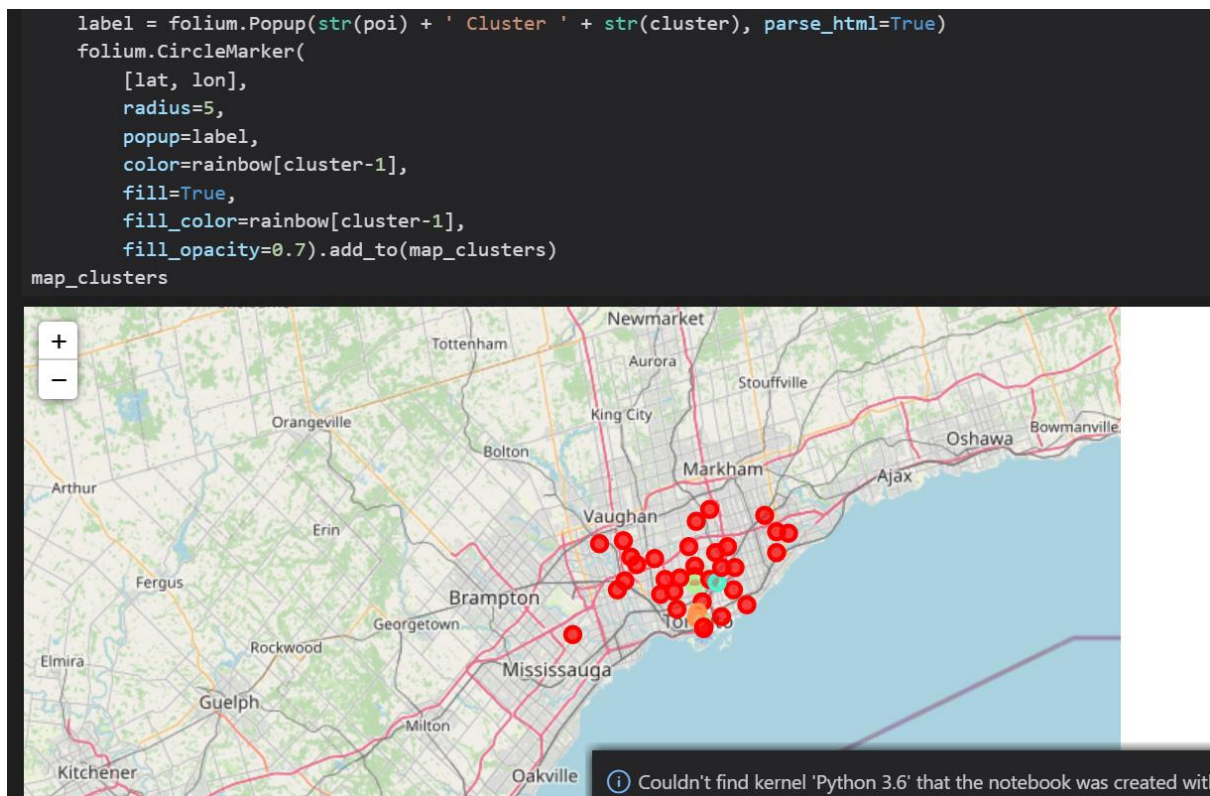
```
kclusters = 6
```

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

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

```
kmeans.labels_
```

[illegible]



## 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 neighborhoods which has least number of Indian restaurants. It is shown in red color in the map

```

[53] In: MI
#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
3	Central Toronto	M5N	Roselawn	43.711695	-79.416936	0.0	0.0

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

```

[56] In: MI
#Cluster 3
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3]

```

	Borough	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
16	East York	M4H	Thorncliffe Park	43.705369	-79.349372	3.0	0.090909

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

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

	Borough	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
5	Downtown Toronto	M4Y	Church and Wellesley	43.665860	-79.383160	5.0	0.012821
9	Downtown Toronto	M5G	Central Bay Street	43.657952	-79.387383	5.0	0.015385

## 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.