Exploring Locations of Toronto to Open an Indian Restaurant

Applied Data Science Capstone Project
by Rashmi Ranjan Mohanty

1. Introduction

1.1 - Background

As a part of the Applied Data Science Capstone Project, we have to work on data science toolsets on a real-life problem on real world datasets to get an experience of what data scientists do. Objectives of this assignment is to source data available in the internet and use Foursquare Location Data to compare different neighborhoods of Toronto and figure out which neighborhood is the most suitable for starting a new restaurant business. In this project, we'll progress in a step by step manner from problem description, data preparation to final analysis and finally will provide a recommendation that can be leveraged by the stakeholders to make business decisions.

1.2 - Business Problem

Toronto is the capital city of the Canadian province of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. Diversity of the city is reflected in the ethnic neighborhoods such as Chinatown, Corso Italia, Greektown, Kensington Market, Koreatown, Little India, Little Italy, Little Jamaica, Little Portugal & Roncesvalles. Toronto is one of the most immigrant friendly cities in North America. With more than half of the entire Indian Canadian population residing here, Toronto is one of the best cities to start an Indian Restaurant Business.

In this project we will go through a step by step process to take an informed decision on whether it's a good idea to open an Indian Restaurant in Toronto. We will analyze all the neighborhoods in Toronto to identify the most profitable area for the restaurant to operate. Since we already know that Toronto shelters a huge number of Indians than any other city in Canada, it will be a good idea to start an Indian Restaurant here, however we need to establish whether it is a profitable idea or not. If so, in which locality we can open it to maximize returns to the investors.

1.3 - Target Audience

- Investors who want to invest in Indian Restaurant Business in the city of Toronto, Canada. This analysis will provide the investors with a list of suitable localities for consideration to start a restaurant targeting the Indian food lovers
- Food lovers can use this analysis to find neighborhoods with availability of Indian Food
- Data Scientists who wish to analyze the neighborhoods of Toronto using Exploratory Data Analysis and other Statistical & Machine Learning techniques can use this analysis as a guidance

2. Data

2.1 - Data Sources

- Neighborhoods Data - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  This Wikipedia Page contains Postal Code, Borough & Neighborhood Names in Toronto
- Geographical Coordinates - https://cocl.us/Geospatial_data
  This CSV File contains Geographical coordinates of the Neighborhoods
- Demographics Data - https://en.m.wikipedia.org/wiki/Demographics_of_Toronto
  This Wikipedia Page contains information on the distribution of population by ethnicity
- Venues Details - Foursquare Explore API (https://developer.foursquare.com/)
  Foursquare API can be used to retrieve details for each of the venues in Toronto

2.2 - Data Cleaning

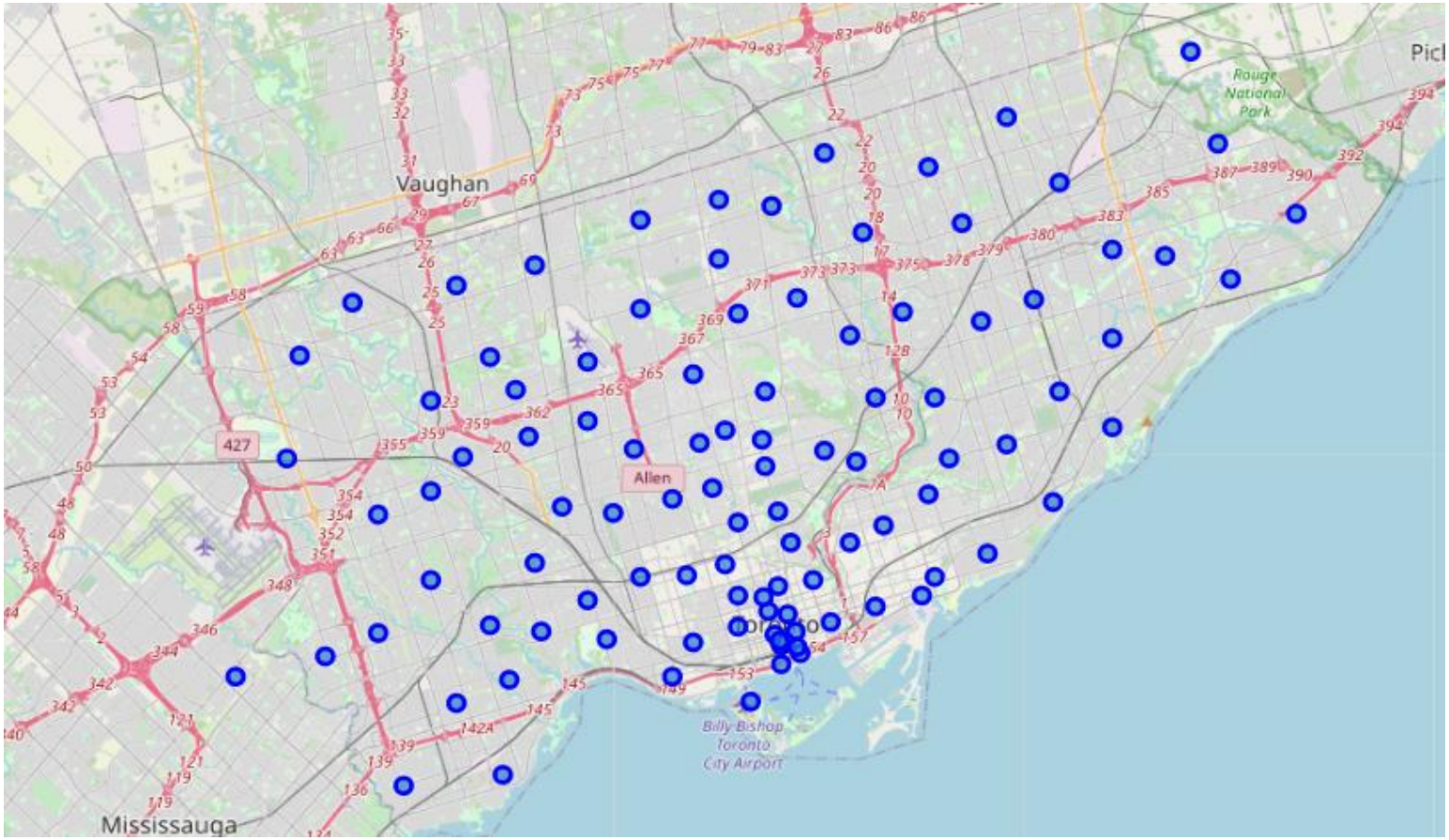
- Neighborhood data scrapped from Wikipedia has three features; Postal Code, Borough and Neighborhood
- The records with no assigned Borough (< 1%) were excluded from the analysis
- It was observed that more than one Neighborhoods can exist in the same Postal Code Area
- These Neighborhoods were combined into one row separated by comma to remove duplicate Postal Codes in the dataset
- For Boroughs with unassigned Neighborhoods, the Neighborhood name has been considered same as the Borough
- Demographics Data extracted from Wikipedia was filtered to keep only the data points related to Indian Population
- For this analysis Venue Details from Foursquare Explored API was filtered to retain only the top 100 venues within 1 KM radius of each neighborhood
- The following four features from the Venue Details were retained for further analysis:

  - Name - Name of the Venue
  - Category - Category Type as defined by the API
  - Latitude - Latitude value of the Venue
  - Longitude - Longitude value of the Venue

3. Exploratory Data Analysis

3.1 - Map of Toronto Neighborhoods

Map of the neighborhoods in Toronto using Geospatial Coordinates and Folium Library in Python

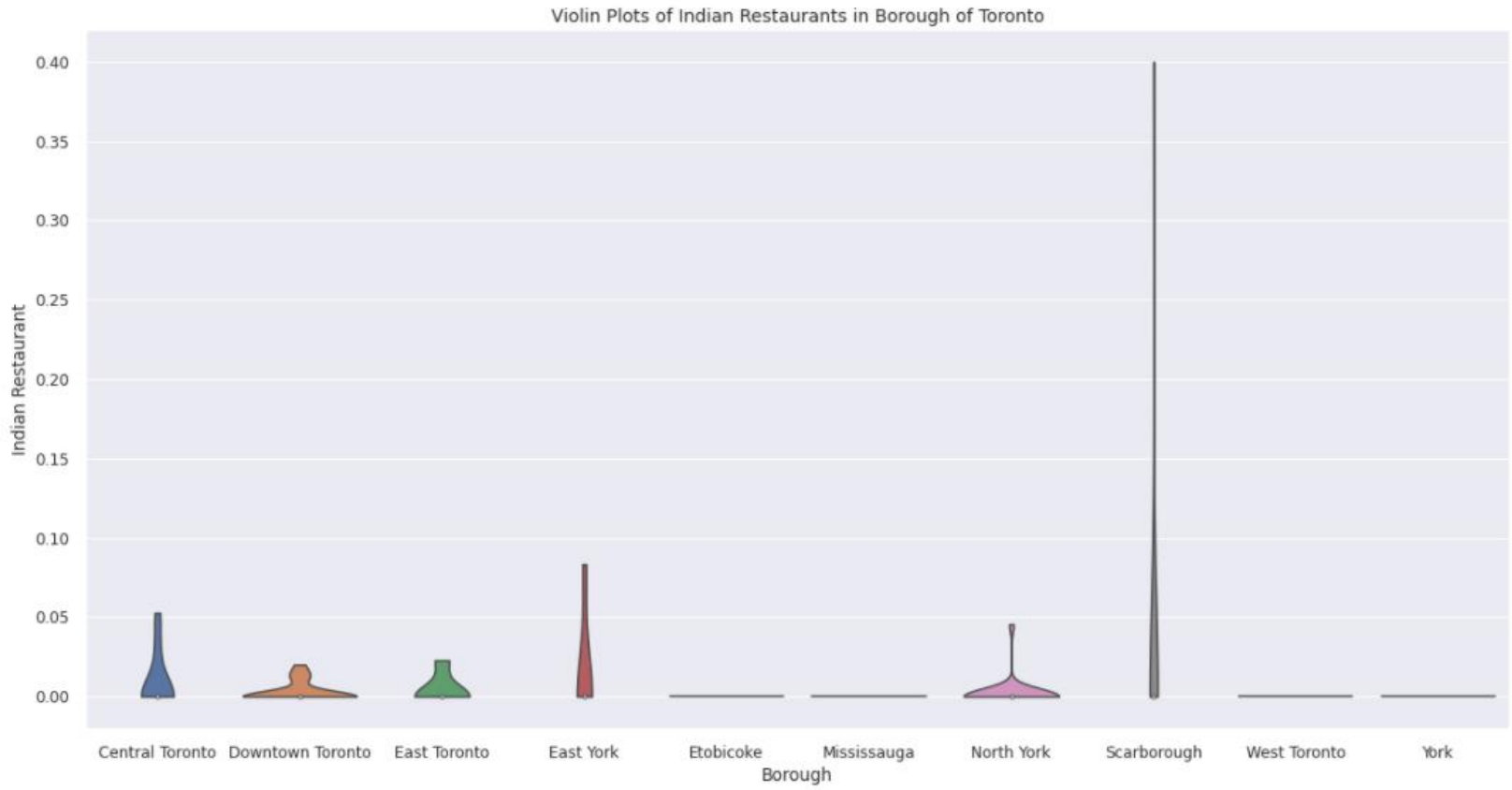


3. Exploratory Data Analysis

3.2 - Relationship between Neighborhoods and Indian Restaurants

Number of existing Indian Restaurants in a Neighborhood is a key feature in deciding a locality for opening a new Indian Restaurant as high density of same business in a locality means high competition and distribution of profitability and market share.

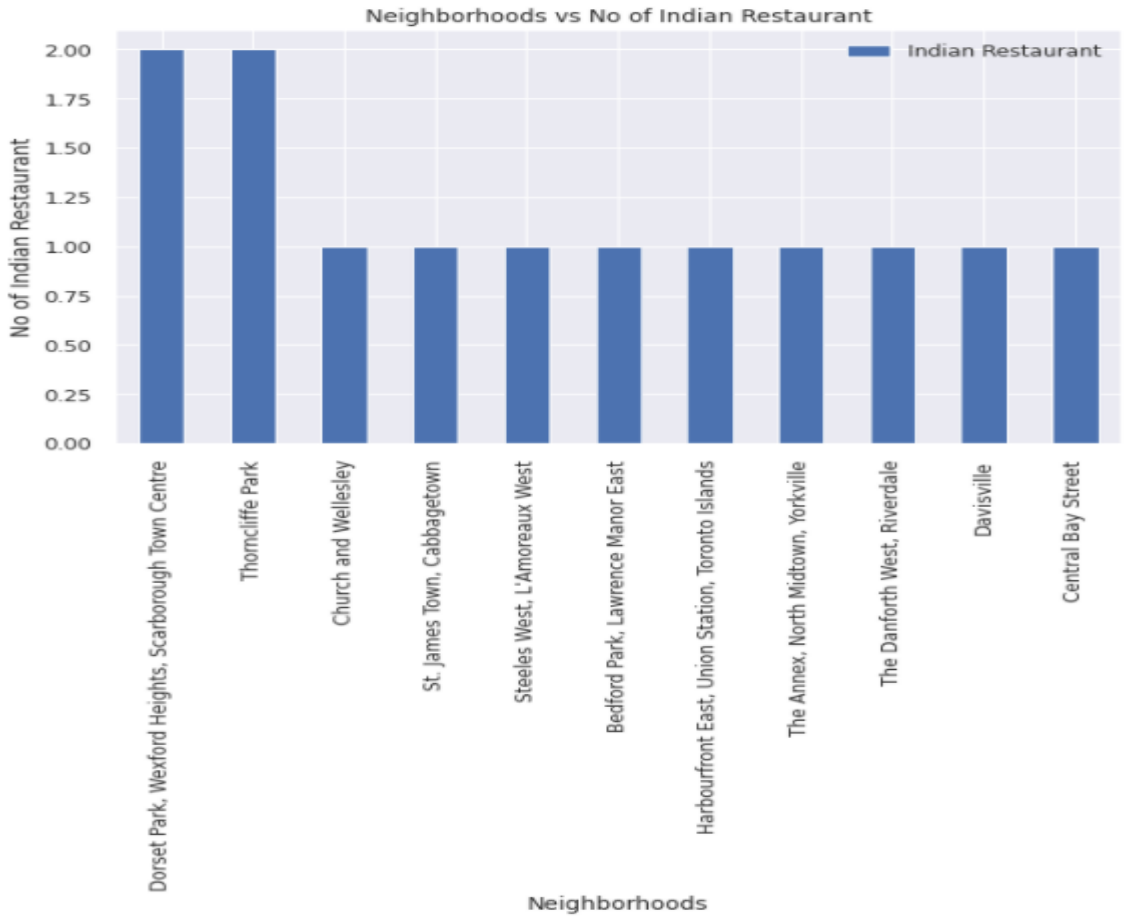
Violin Plot showing density of Indian Restaurants in Neighborhoods of Toronto



3. Exploratory Data Analysis

3.2 - Relationship between Neighborhoods and Indian Restaurants

Bar Chart showing number of Indian Restaurants in Neighborhoods of Toronto

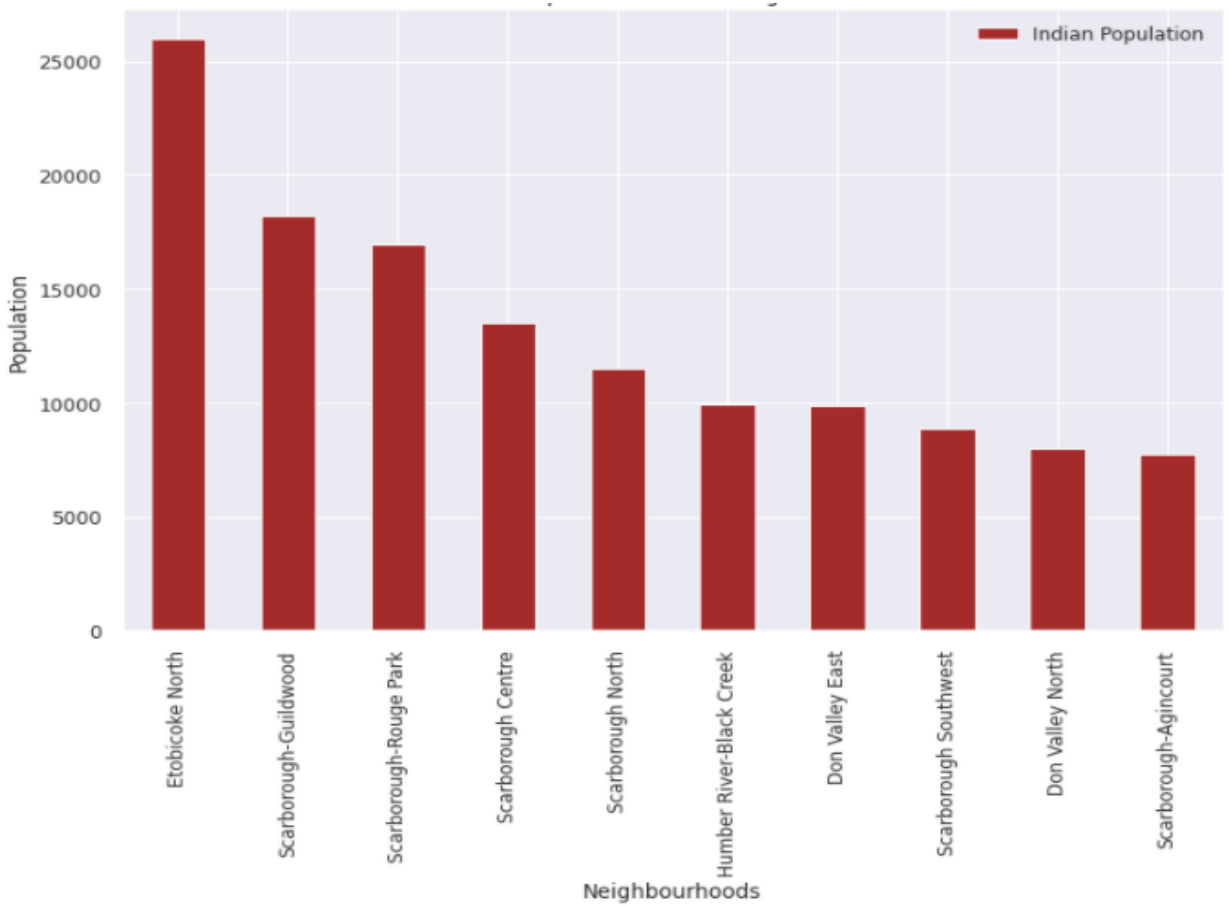


3. Exploratory Data Analysis

3.3 - Relationship between Neighborhoods and Indian Population

Another key feature is the distribution of Indian Population in each of the Neighborhoods.

Bar Chart showing Indian Population in Neighborhoods of Toronto



This analysis & visualization of the relationship between Neighborhoods & Indian Population living in those Neighborhoods will help us in identifying the highly populated Indian Neighborhoods. Once we identify those Neighborhoods it helps us in deciding where to open the new Indian Restaurant. Indian Restaurant placed in an densely populated Indian Neighborhood is more likely to get more customers than a restaurant placed in a Neighborhood with less or no Indian Population. Thus this analysis helps us in determining the success of the new business.

3. Exploratory Data Analysis

3.4 - Relationship between Restaurants and Indian Population

After performing the analysis, we couldn't see a strong positive relationship between densely populated Indian Neighborhoods & number of Indian Restaurants. This might be because of the missing/not up-to-date data as this is an area which can be improved in future analysis to get more insights.

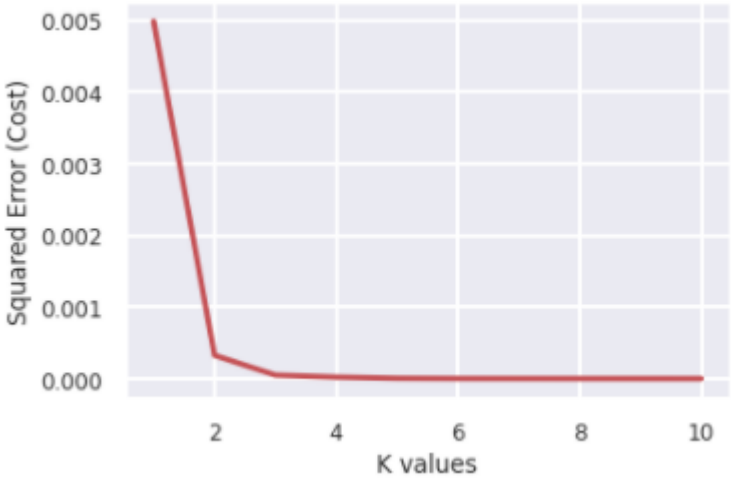
The following table displays the number of Indian Restaurants in densely populated Indian Neighborhoods

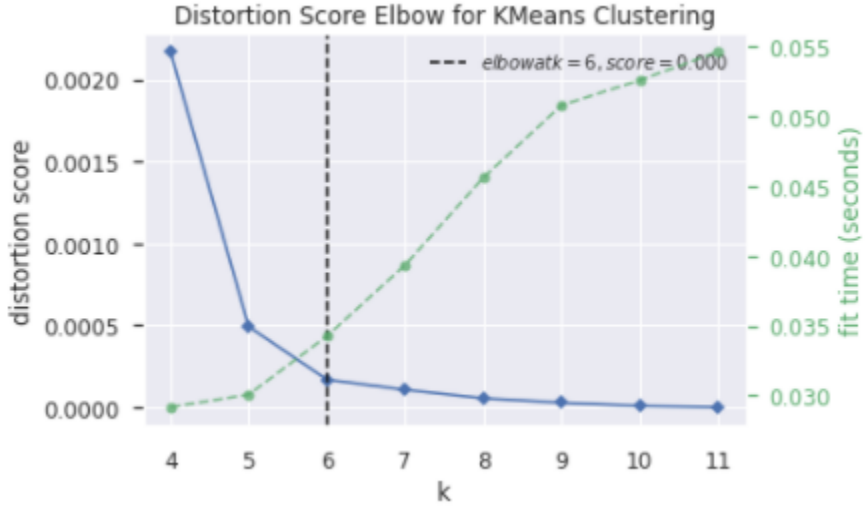
	Indian Population	Neighborhood	Indian Restaurant
0	7961.380	Henry Farm	0.0
1	8880.190	Oakridge	0.0
2	16941.315	Rouge	0.0
3	16941.315	Port Union	0.0
4	18200.700	Morningside	0.0
5	25965.120	Thistletown	0.0
6	8880.190	Clairlea	0.0
7	16941.315	Highland Creek	0.0
8	13474.900	Maryvale	0.0

4. Predictive Modelling

4.1 - Clustering Neighborhoods 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.





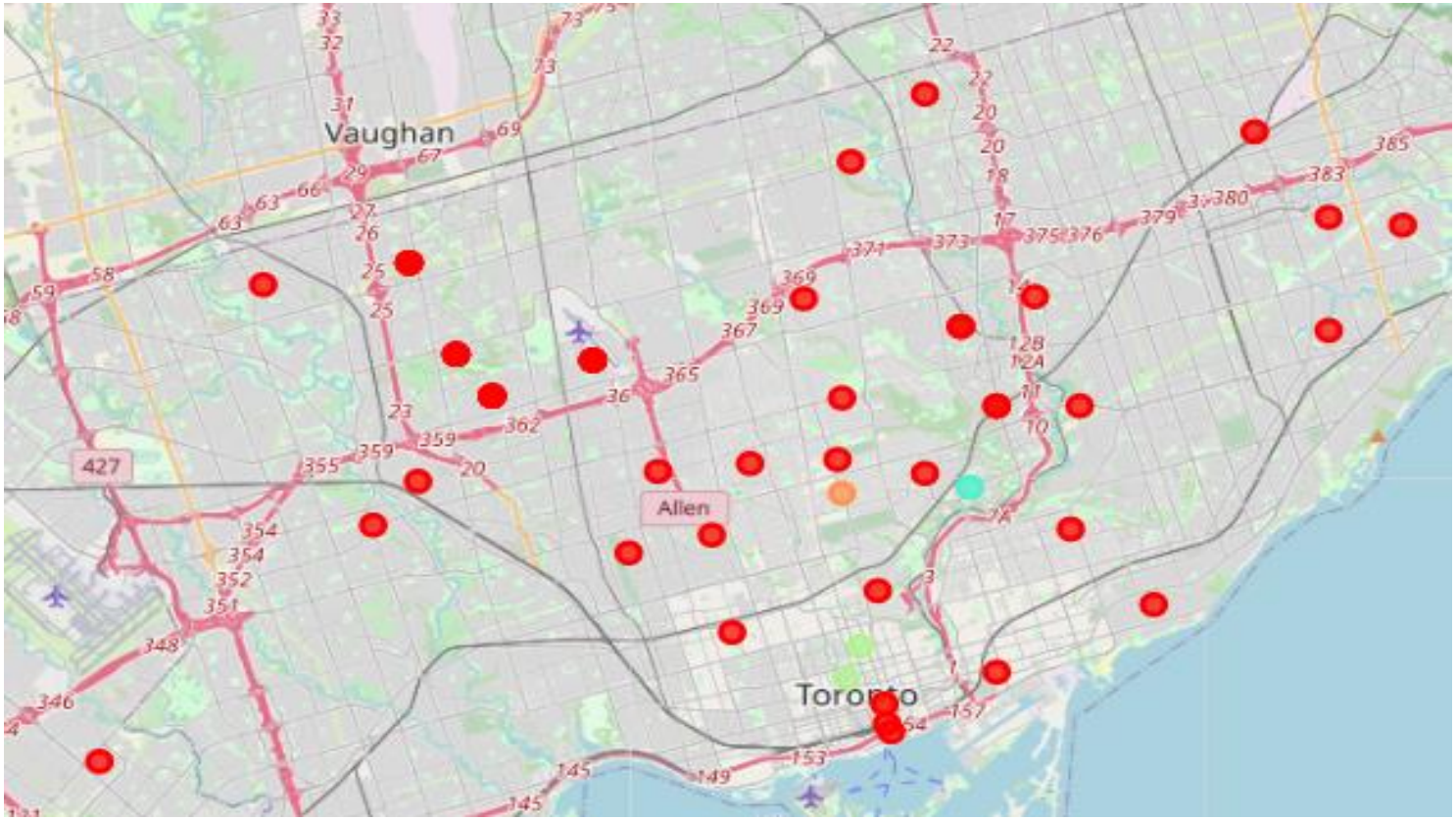
After analyzing the dataset using elbow method using Distortion Score & Squared Error for each K value, looks like K = 6 is the best value.

The following table shows the Toronto Neighborhood dataset with Cluster Labels.

	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	M4S	Davisville	43.704324	-79.388790	5.0	0.029412
3	Central Toronto	M5N	Roselawn	43.711695	-79.416936	0.0	0.000000
4	Downtown Toronto	M4W	Rosedale	43.679563	-79.377529	0.0	0.000000

4. Predictive Modelling

4.2 - Examining the Clusters



Cluster 0 - Red Circles represents Neighborhoods with least number of Indian Restaurants
Cluster 1 - Has no rows meaning no data points or Neighborhood near to this centroid
Cluster 2 - Has no rows meaning no data points or Neighborhood near to this centroid
Cluster 3 - Blue Circles represents Neighborhoods with medium density of Indian Restaurants
Cluster 4 - Green Circles represents Neighborhoods with medium to high density of Indian Restaurants
Cluster 5 - Orange Circles represents Neighborhoods with high density of Indian Restaurants

5. Results and Discussions

5.1 - Results

In this project, the business problem was to identify a good location to open a new Indian Restaurant, we looked into all the Neighborhoods in Toronto, analyzed the Indian Population & density of Indian Restaurants in those Neighborhoods to come to a conclusion about which Neighborhood would be a better location for opening a new Indian Restaurant. Using data from various sources and unsupervised machine learning techniques we have found out that -

- Amongst the Boroughs in Toronto only Central Toronto, Downtown Toronto, East Toronto, East York and North York Boroughs have high density of Indian Restaurants. This was established with help of Violin Plots and Bar Charts.
- Etobicoke North, Scarborough-Guildwood, Scarborough-Rouge Park, Scarborough Centre, Scarborough North, Humber River-Black Creek, Don Valley East, Scarborough Southwest, Don Valley North & Scarborough-Agincourt have dense Indian Population.
- Central Toronto, Downtown Toronto and East York are already densely populated with Indian Restaurants. So it is better to leave those Boroughs out and consider only Scarborough, East Toronto & North York for the new restaurant's location.
- After further consideration it seems to be a good idea to open the new Indian Restaurant in **Scarborough** Borough since it has high number of Indian Population which gives a higher number of prospect customers and lower competition since very less Indian Restaurants are operating in the locality.

5.2 - Discussions

According to this analysis, **Scarborough** Borough will have least competition for the new upcoming Indian Restaurant as there are very less Indian Restaurants in the Neighborhood. Also looking at the population distribution it looks like it is densely populated with Indian Population which will helps the new business as the number of prospect customers will be high. Hence, this location could potentially be a perfect place for starting a new Indian Restaurant Business.

Some of the drawbacks of this analysis:

- The clustering is completely based only on data obtained from Foursquare API
- Indian Population distribution in each Neighborhood is based on 2016 census which is not up-to date

As discussed above, the analysis has some areas of improvement. However, it certainly provides us with some good insights, preliminary information on possibilities & a head start into this business problem by guiding in the right direction.

6. Conclusion

Conclusion

I got a chance to analyze a business problem as a data scientist would do. I have used Python to fetch the data, manipulate the contents & analyze and visualize the datasets. I have made use of Foursquare API to explore the venues in the Neighborhoods of Toronto, then extracted good amount of data from Wikipedia which was scraped with help of Wikipedia Python library and visualized using various plots available in the Seaborn & Matplotlib libraries. I also applied machine learning technique to predict the output given the data and used Folium to visualize it on a map. Also, some of the drawbacks or areas of improvements show us that this analysis can further be improved with help of more data and different machine learning techniques. Similarly, we can use this project to analyze different scenarios. Hopefully, this project will help as an initial guidance to take more complex real-life challenges using data-science.