Predicting to open an Indian Restaurant

Akhilesh T

JULY 2, 2020

1. Introduction: Business Problem

1.1 Problem Statement: To open 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, and 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, I will go through step by step process to make a decision whether it is a good idea to open an Indian restaurant. We analyse 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.

1.2 Target Audience

1) Business personnel who want to invest or open an Indian restaurant in Toronto. This analysis will be a comprehensive guide to start or expand restaurants targeting the Indian crowd.

2) Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.

3) Indian crowd who wants to find neighborhoods with lots of option for Indian restaurants.

4) Business Analyst or Data Scientists, who wish to analyse 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.

This idea of analysis is good for someone looking to open a restaurant in the Indian dense populated city. And also it is helpful for any contractor to take up this idea and do the business.

2. Data

2.1 Data sources

a) I'm using "List of Postal code 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.
(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

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

c) To get information about the distribution of population by their ethnicity I’m using “Demographics of Toronto” 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.
(https://en.m.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity)

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).

e) 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.

The following Libraries are used:

- pandas
- requests
- folium
- numpy
- matplotlib
- json
- KMeans

2.2 Data Cleaning

a) Scraping Toronto Neighborhoods Table from Wikipedia

Assumptions made to attain the below DataFrame:

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

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

Table.1 Dataframe from ‘List of Postal code of Canada: M’ Wikipedia Table.

Postal Code	Borough	Neighborhood
M3A	North York	Parkwoods
M4A	North York	Victoria Village

M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

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. The following table represents the DataFrame with latitudes & longitudes of Postal codes in Toronto.

Table.2 The DataFrame merged with geospatial data on Postal code.

Postal code	Borough	Neighborhood	Latitude	Longitude
M5G	Downtown Toronto	Central Bay Street	43.65795	-79.3874
M2H	North York	Hillcrest Village	43.80376	-79.3635
M4B	East York	Parkview Hill, Woodbine Gardens	43.7064	-79.3099
M1J	Scarborough	Scarborough Village	43.74473	-79.2395
M4G	East York	Leaside	43.70906	-79.3635
M4M	East Toronto	Studio District	43.65953	-79.3409
M1R	Scarborough	Wexford, Maryvale	43.75007	-79.2958
M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamestown, Mount Olive, Beaumond Heights, Thistletown, Albion Gardens	43.73942	-79.5884
M9L	North York	Humber Summit	43.7563	-79.566
M5V	Downtown Toronto	CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	43.62895	-79.3944
M1B	Scarborough	Malvern, Rouge	43.80669	-79.1944
M5A	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.3606

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 neighborhood's population to identify the densely populated neighborhoods with Indian population.

There were only six neighborhoods in Toronto which Indian population spread across and their corresponding boroughs are Toronto & East York, North York, Scarborough & Etobicoke & York

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 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.

3. Methodology

3.1 Exploratory Data Analysis:

3.1.1 Folium Library and Leaflet Map

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

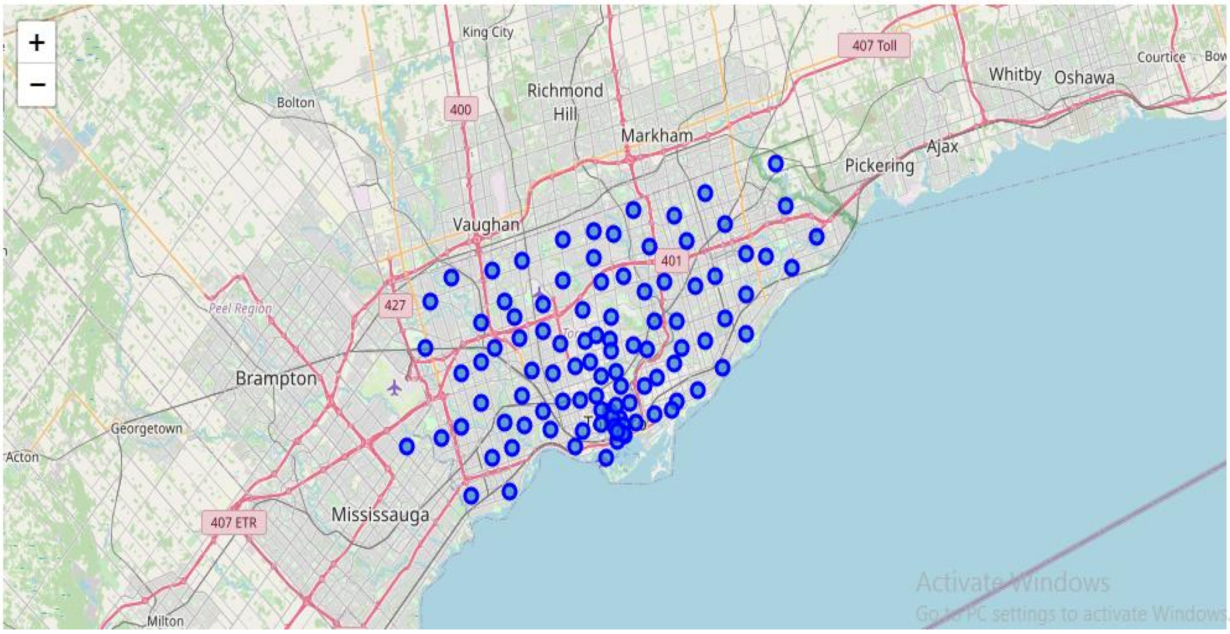

Figure.1 The map of Toronto city with neighborhoods.

3.1.2 Relationship between Borough and Indian Restaurant

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.

In general, violin plots are a method of plotting numeric data and can be considered a combination of the box plot with a kernel density plot. In the violin plot, we can find the same information as in the box plots.

Let's visualize the Boroughs with Indian Restaurants graph with the help of violin plots.

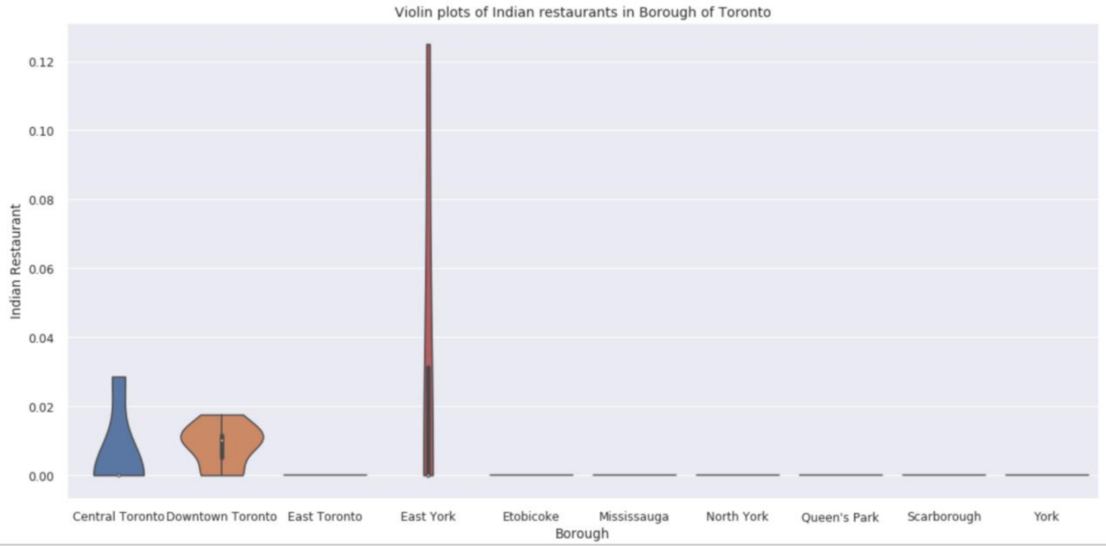

Figure.2 Violin plot with Borough on x-axis and Indian Restaurants on y-axis.

3.1.3 Relationship between Neighborhood and Indian Restaurant.

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

Let’s also visualize the neighborhood with Indian Restaurants:

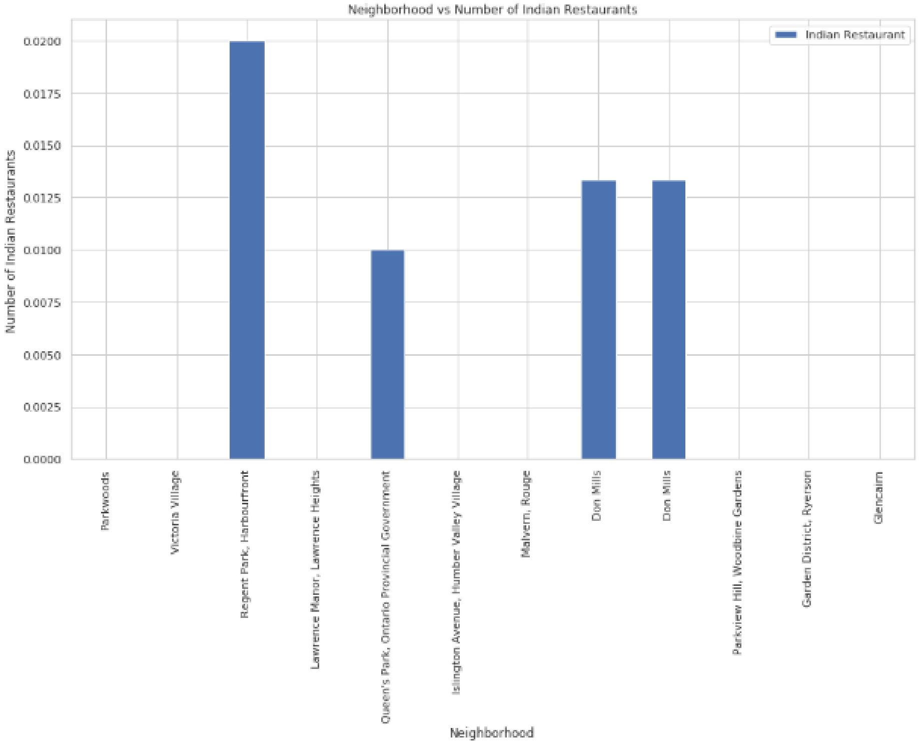

Figure.3 Bar chart with Neighborhood on x-axis and No. of Indian Restaurants on y-axis

3.1.4 Relationship between neighborhood and Indian population

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

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

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

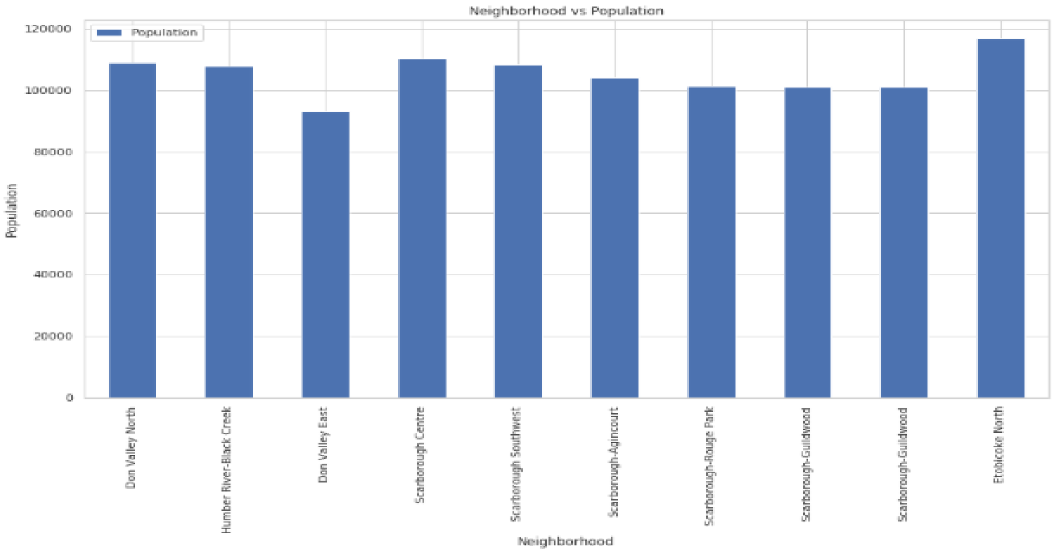

Figure.4 Bar chart with Neighborhood on x-axis and Indian Population on y-axis

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

3.2 Predictive Modelling

3.2.1 Clustering Neighborhoods of Toronto:

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

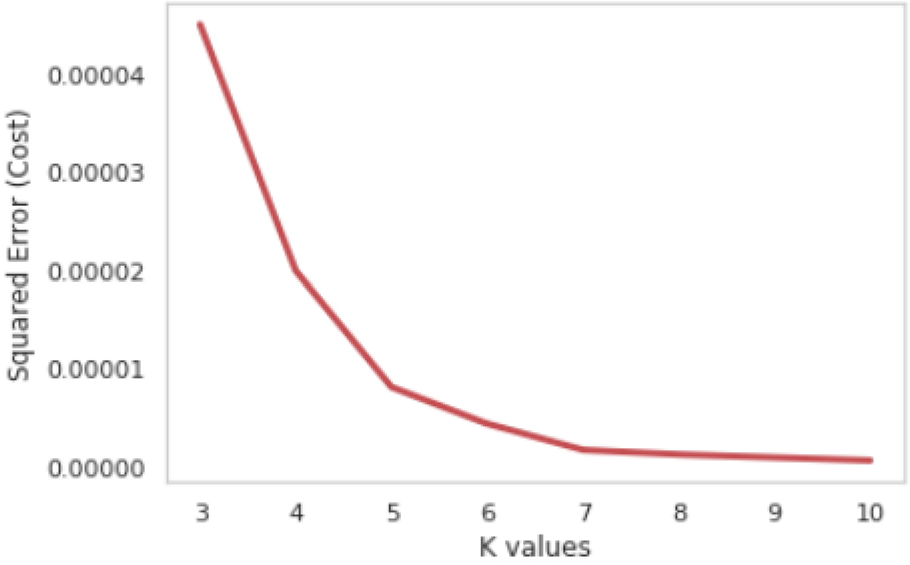

Figure.5 Elbow method to identify best K value

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.

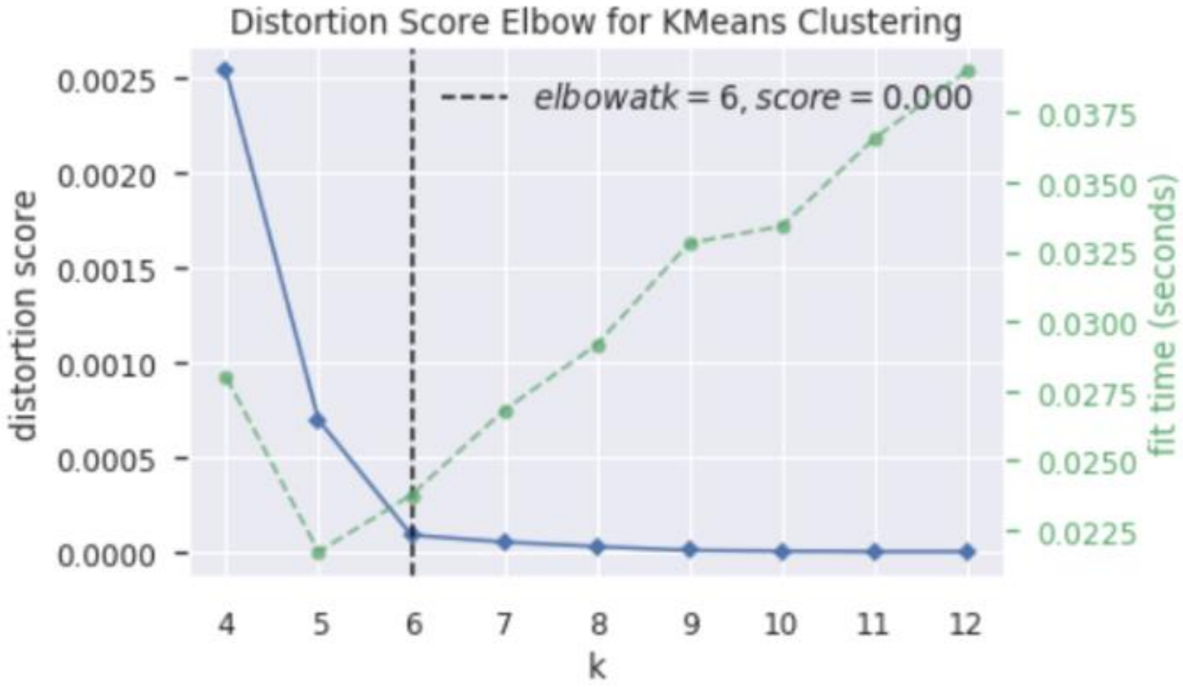

Figure.6 Elbow visualizer to identify the K value

With the help of modelling six clusters are created in the Toronto city. A Clustering Algorithm tries to analyse natural groups of data on the basis of some similarity. It locates the centroid of the group of data points. To carry out effective clustering, the algorithm evaluates the distance between each point from the centroid of the cluster.

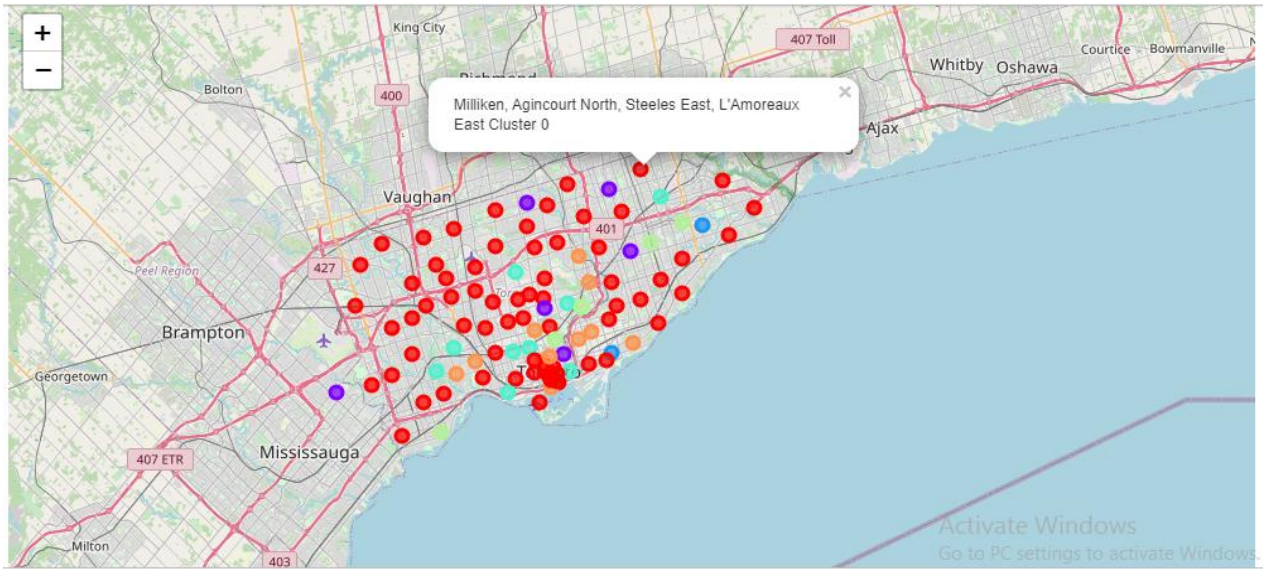
Milliken, Agincourt North, Steeles East, L'Amoreaux East Cluster 0

Figure.7 The following map consists of Clusters of neighborhoods in Toronto city.

3.2.2 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 have least number of Indian restaurants. It is shown in red color in the map.

Cluster 1 contains the neighborhoods which are sparsely populated with Indian restaurants. It is shown in purple color in the map.
Cluster 2 contains least neighborhood which is medium populated with Indian restaurants. It is shown in dark blue in the map.
Cluster 3 contains neighborhood which is medium populated with Indian restaurants. It is shown in light blue in the map.
Cluster 4 contains 2nd least neighborhood which is medium populated with Indian restaurants.
Cluster 5 contains 2nd highest neighborhood which is least populated with Indian restaurants. It is shown in orange colour in the map.

4. Results

We have reached the end of the analysis, in this section we will document all the findings from above clustering & visualization of the dataset. In this project, we started off with the business problem of identifying a good neighborhood to open a new Indian restaurant.

To achieve that we looked into all the neighborhoods in Toronto, analysed the Indian population in each neighborhood & number of Indian restaurants in those neighborhoods to come to conclusion about which neighborhood would be a better spot. We have used variety of data sources to set up a very realistic data-analysis scenario. We have found out that —

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

5. Discussion

According to this analysis, Scarborough borough will provide the least competition for the new upcoming Indian restaurant as there is very little Indian restaurants spread or no Indian restaurants in few neighborhoods. Also looking at the population distribution looks like it is densely populated with Indian crowd which helps the new restaurant by providing high customer visit possibility.
So, definitely this region could potentially be a perfect place for starting quality Indian restaurants. Some of the drawbacks of this analysis are — the clustering is completely based only on data obtained from Foursquare API. 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 neighborhoods of Toronto, then get good amount of data from Wikipedia which we scraped with help of Wikipedia python library and visualized using various plots present in seaborn & matplotlib. We also applied machine learning technique to predict the output given the data and used Folium to visualize it on a map.

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

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

7. Future directions

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 improve in future analysis to get a more insight about the business problem.