IBM Data science Professional Certificate Capstone Project

Best location to open an Indian Restaurant in Toronto

Tanmay Bhise

July 2021

1. Introduction

1.1 Problem statement

Toronto is the most populous and multicultural city in Canada, occupying more than 7.5 percent of the Canadian population [1]. In a report by CIC news, Toronto is ranked third among new immigrants' preferred cities [2]. The Census of 2016 shows that most immigrants in Toronto are from India [3]. Therefore, This project hypothesizes a client who wants to open an Indian restaurant in Toronto and analyzes the best possible location to do so using machine learning. Target audience for this project is anyone who is planning to open an Indian restaurant in Toronto.

1.2 Objective

The overall objective of this project is to discover the locations of Indian restaurants in the neighborhoods of Toronto and ultimately recommend the best location to open a new Indian restaurant to a hypothetical client using machine learning.

1.3 Report structure

First, This project starts with discussion on hypotheses propped in this project along with data acquisition and data cleaning process in section 2. Secondly, the exploratory analysis regarding collected data is detailed in section 3. In addition, the same section also describes the machine learning algorithms utilized to make better recommendations regarding the best location to open a new Indian restaurant in Toronto. Next, Section 4 discusses the outcome of the machine learning algorithms. Section 5 states the recommendations made from the analysis of the results. Ultimately, section 6 concludes this project with final remarks.

2. Hypothesis, Data acquisition and cleaning

2.1 Hypothesis

It is obvious that a Toronto neighborhood with a large population would be among the best candidate locations, to be considered for opening a new restaurant. More the population, the more will be the customers. Furthermore, this neighborhood should have less or no Indian restaurants within its surroundings. Neighborhoods with more Indian restaurants would create a competitive environment and might not be a good sign for a new restaurant. Consequently, a populous neighborhood with less or no Indian restaurants will be among the best locations to open a new Indian restaurant.

2.2 Data sources

This project makes use of a total of four datasets. The first one is Toronto neighborhood data scraped from Wikipedia. This data contains information regarding postal codes, regions and neighborhoods of Toronto. Secondly, Geospatial coordinates of neighborhoods of Toronto are read from the data provided by coursera (IBM) on the 3rd week of the capstone project. The third dataset is the postal codes and population of Toronto as per 2016 census obtained from statcan. These three data sets are used to explore neighborhoods with a large population. Lastly, the fourth dataset is about Indian restaurants in Toronto which is retrieved from Foursquare via RESTful API calls. This dataset in combination with the first three are used to make recommendations regarding the best location to open an Indian restaurant as per hypothesis stated in section 2.1.

2.3 Data cleaning

The main goal of the data cleaning process is to prepare data for exploratory analysis and run machine learning algorithm on it. For some of the neighborhood data scraped from wikipedia, regions were not assigned. Such neighborhoods are not considered for further study and removed during the data cleaning process. Furthermore, wikipedia neighborhood data is combined with longitudes and latitudes as per their postal codes. Any region names which are too long have been shortened. For instance, "MississaugaCanada Post Gateway Processing Centre" has been replaced by simply "Mississauga Canada". The ultimate result of this process is a single table

containing 103 neighborhoods and their respective regions, postal codes, and their longitudes and latitudes as shown in the figure below.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park / Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor / Lawrence Heights	43.718518	-79.464763
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494

Fig 1. First five rows of Toronto neighborhoods dataset

The Toronto Indian restaurants data retrieved via Foursquare API is shown in Fig 2. It consists of all the columns of Toronto neighborhoods data along with an additional "Name" column which corresponds to the name of Indian restaurants in Toronto.

	Name	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	Aroma Fine Indian Restaurant	M5V	Downtown Toronto	CN Tower / King and Spadina / Railway Lands / ...	43.628947	-79.394420
1	309 Dhaba Indian Excellence	M5V	Downtown Toronto	CN Tower / King and Spadina / Railway Lands / ...	43.628947	-79.394420
2	Indian Biriyani House	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
3	Indian Biriyani House	M5H	Downtown Toronto	Richmond / Adelaide / King	43.650571	-79.384568
4	Chaska	M5H	Downtown Toronto	Richmond / Adelaide / King	43.650571	-79.384568

Fig 2. First five rows of Toronto Indian Restaurants dataset

The neighborhood population dataset obtained from statcan website consists of different postal codes of Canada and their respective population.

	PostalCode	Population
0	A0A	46587.0
1	A0B	19792.0
2	A0C	12587.0
3	A0E	22294.0
4	A0G	35266.0

Fig 3. First five rows of Canadian population dataset

All the three datasets have been combined to ultimately create a table consisting of the name of Toronto neighborhoods, number of Indian restaurants per neighborhood and neighborhood population as illustrated in Fig 4.

	Neighborhood	Indian_Restaurants	Population
0	Harbourfront East / Union Station / Toronto Is...	3	14545.0
1	CN Tower / King and Spadina / Railway Lands / ...	2	49195.0
2	Studio District	2	24689.0
3	Richmond / Adelaide / King	2	2005.0
4	The Annex / North Midtown / Yorkville	2	26496.0

Fig 4. First five rows of merged dataset

After normalizing the number of Indian restaurants and population using MinMaxScaler function built in sklearn,the final dataframe is now ready to be fit into a machine learning algorithm.

	Neighborhood	Indian_Restaurants	Population
0	Harbourfront East / Union Station / Toronto Is...	1.000000	0.191641
1	CN Tower / King and Spadina / Railway Lands / ...	0.666667	0.648181
2	Studio District	0.666667	0.325296
3	Richmond / Adelaide / King	0.666667	0.026417
4	The Annex / North Midtown / Yorkville	0.666667	0.349105

Fig 5. First five rows of merged and normalized dataset

Before moving to the machine learning algorithm, exploratory analysis has been carried out to gain some descriptive information regarding the data. Following section discusses both exploratory analysis as well as a Machine learning algorithm.

3. Methodology

3.1 Exploratory data analysis

3.2.1 Neighborhoods in Toronto

The figure below shows all the neighborhoods of Toronto included in the dataframe. Clearly, North York is the region in Toronto with the most number of neighborhoods with count reaching upto 24. On the other hand, Downtown Toronto and Scarregion each have 17 neighborhoods, both ranking second.

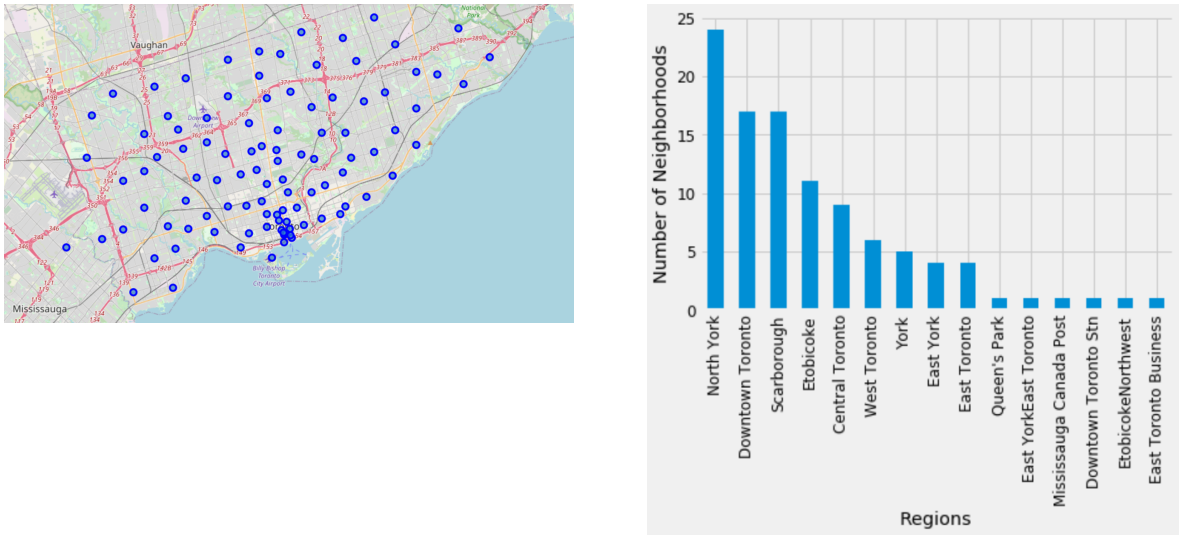

Fig 6. Neighborhoods in Toronto

3.2.2 Population distribution in Toronto

Although, it is not necessary that the most number of neighborhoods corresponds to the most population. This can be clearly seen in the figure below. Even Though, North York is the region with the most number of neighborhoods, Scarborough is the most populated one in Toronto with count reaching upto 714,699. North York and Downtown Toronto rank second and third respectively. The population is distributed normally across Toronto as can be seen in the distribution plot with the population of most of the regions in Toronto is between 200,000 to

300,000. It should be noted that some of the regions scraped from wikipedia with population less than 100 have not been considered for further processing.

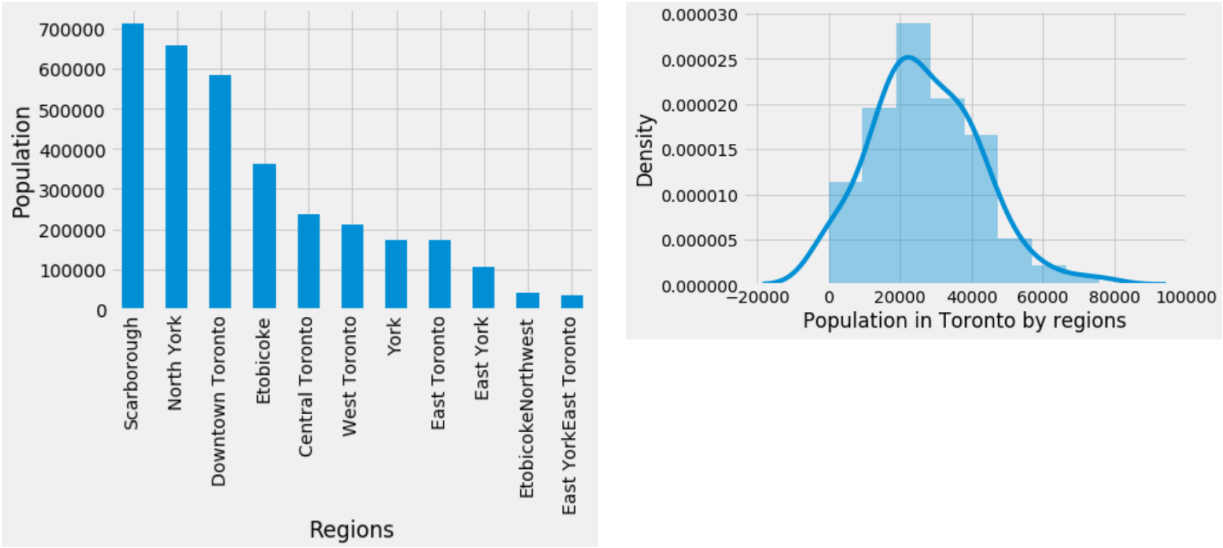

Fig 7. Population distribution in Toronto

3.2.3 Indian restaurants in Toronto

In the figure below, a pie chart shows the distribution of Indian restaurants in Toronto. Pie chart only includes regions with Indian restaurants, and regions with no Indian restaurants have not been included. From the pie chart, it is evident that the majority of Indian restaurants are located downtown.

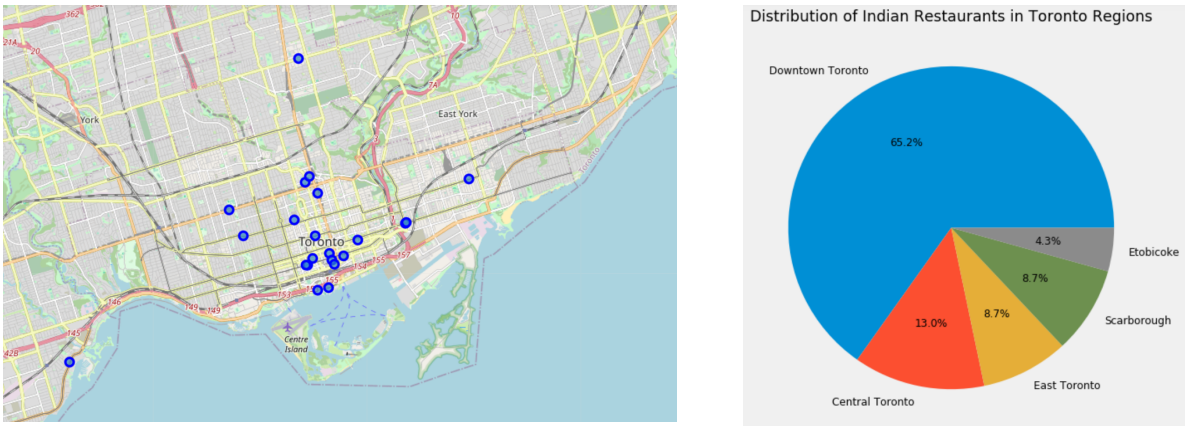

Fig 8. Indian Restaurants in Toronto

3.2.4 Concluding remark on exploratory analysis

Overall, from explorative analysis of data, following observations can be made:

1. North York region includes most number of neighborhoods
2. Scarborough is the most populated region in Toronto. North York and Downtown Toronto regions rank second and third respectively.
3. The population of most of the neighborhoods in Toronto is between 200,000 and 300,000.
4. Most of the Indian restaurants are located in Downtown Toronto

From the explorative analysis, with the most number of Indian restaurants, Downtown Toronto is one of the competitive regions to open a new one. Most populated regions like Scarborough and North York could be the preferable locations to open new Indian restaurant. But before making any conclusion and recommendation regarding the best neighborhood to open an Indian restaurant, an unsupervised machine learning algorithm called K Means clustering has been utilised to cluster neighborhoods as per number of restaurants and their population. This study is discussed in the following section.

3.2 Machine learning

3.2.1 K-means algorithm

K means clustering is the unsupervised machine learning clustering algorithm which minimizes the euclidean distance between data points within the clusters and maximizes the distance between two clusters. This algorithm needs to specify the value of k which is the number of clusters. The optimal value for k is obtained with the help of the elbow method. In this method, the inertia which is within the cluster sum of square error of data points is plotted against the value of k. The number of k where, plot forms an elbow is considered to be an optimal value of the k. This is because after this elbow point, the inertial minimization is relatively slow. The elbow plot for the dataset used in this project is shown in the figure below. It can be seen that the elbow point is with a value of k equal to 3. Therefore, given data which is neighborhoods in our case, has been divided into three clusters.

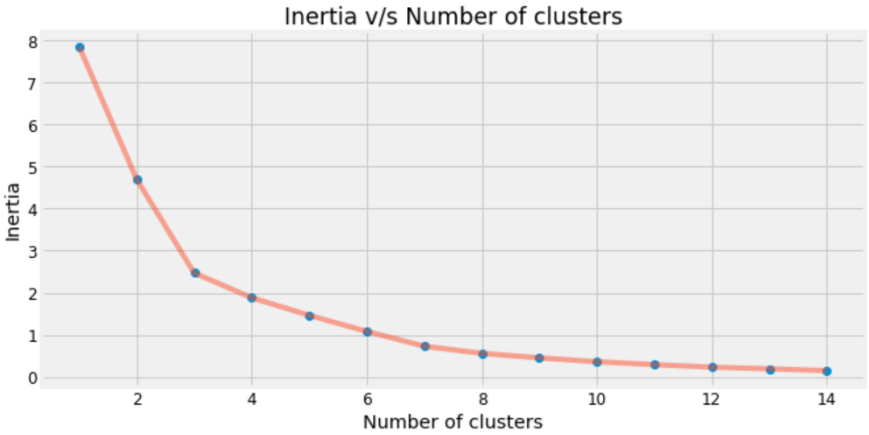

Fig 9. K-means clustering elbow plot

Results of the k-means algorithm are detailed in the next section.

4. Results

The three clusters of neighborhoods of Toronto are shown in the figure below with three different colors of markers. It can be observed that most of the neighborhoods fall into clusters colored with red markers. This is followed by a cluster denoted by a purple marker which consists of the second highest number of neighborhoods while a cluster with green color marker contains least neighborhoods. These three clusters are analyzed further in the following paragraphs.

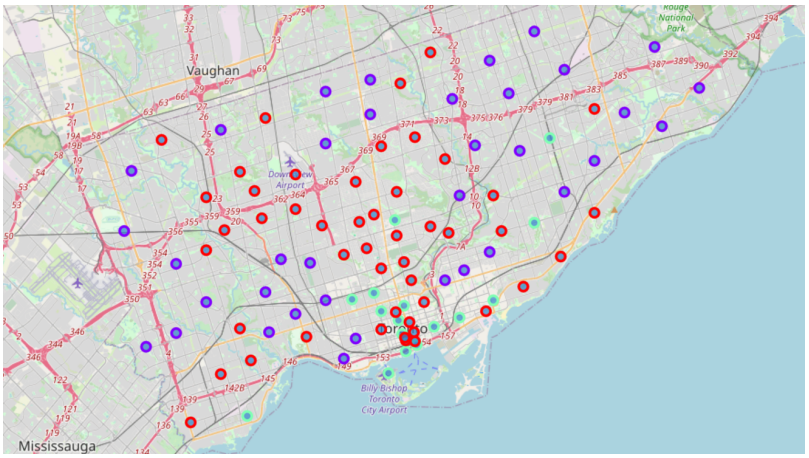

Fig 10.Toronto neighborhoods clusters

5. Discussion

5.1 Cluster 0

Cluster 0 includes a total of 36 neighborhoods. These include neighborhoods with a **large population and no Indian restaurants**. The average population of neighborhoods in this cluster is 41,511. This cluster is the highly recommended cluster since according to available data there are no Indian restaurants within these neighborhoods so less competition. Also, these neighborhoods have a large population and hence more customers can be expected.

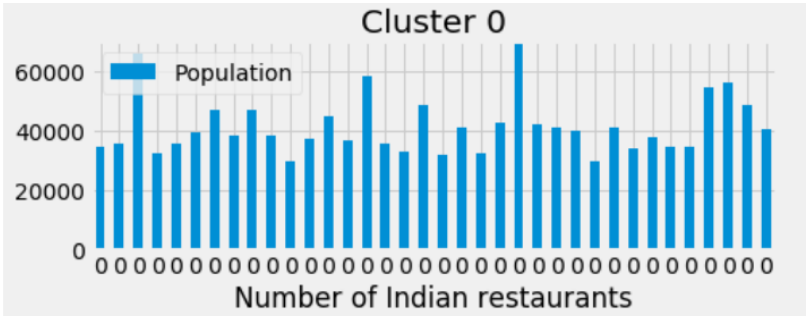

Fig 11. Cluster0: Population vs Number of restaurants

Within cluster 0, with the largest population, Widowwale south in North York, is the highly recommended location to open an Indian restaurant. Table gives five highly recommended locations within cluster 0 to open an indian restaurant.

Neighborhood	Indian_Restaurants	PostalCode	Regions	Latitude	Longitude	Population	cluster
WillowdaleSouth	0	M2N	North York	43.770120	-79.408493	75897.0	0
Malvern / Rouge	0	M1B	Scarborough	43.806686	-79.194353	66108.0	0
Fairview / Henry Farm / Oriole	0	M2J	North York	43.778517	-79.346556	58293.0	0
South Steeles / Silverstone / Humbergate / Jam...	0	M9V	Etobicoke	43.739416	-79.588437	55959.0	0
Milliken / Agincourt North / Steeles East / L'...	0	M1V	Scarborough	43.815252	-79.284577	54680.0	0

Fig 12. Cluster0: 5 highest recommended neighborhoods to open an Indian restaurant

5.2 Cluster 1

The final cluster, cluster 2, includes 12 neighborhoods. Specifically, this cluster contains a **low to medium population and with at least one Indian restaurant**. The average population in neighborhoods of this cluster is 24,550 with the population of some neighborhoods exceeding the 30,000 mark. This is the second highest recommended cluster.

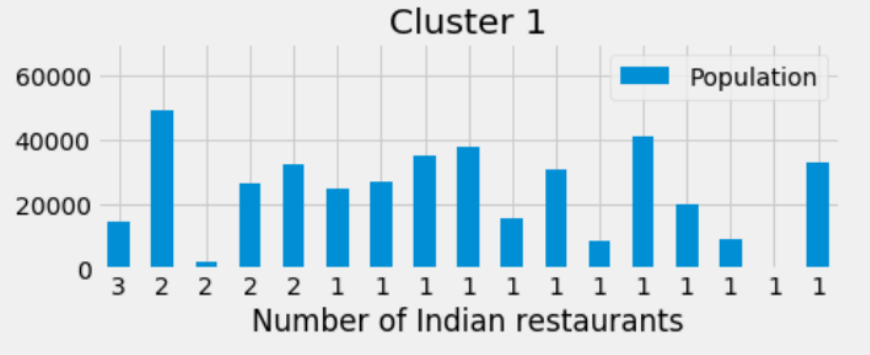

Fig 13. Cluster1: Population vs Number of restaurants

Table below demonstrates the top 5 highly recommended neighborhoods to open an Indian restaurant. The highest recommendation is Regent Park / Harbourfront in Downtown Toronto with a population of 41,078 and only one Indian restaurant.

Neighborhood	Indian_Restaurants	PostalCode	Regions	Latitude	Longitude	Population	cluster
Regent Park / Harbourfront	1	M5A	Downtown Toronto	43.654260	-79.360636	41078.0	1
New Toronto / Mimico South / Humber Bay Shores	1	M8V	Etobicoke	43.605647	-79.501321	37975.0	1
Golden Mile / Clairlea / Oakridge	1	M1L	Scarborough	43.711112	-79.284577	35081.0	1
India Bazaar / The Beaches West	1	M4L	East Toronto	43.668999	-79.315572	32640.0	1
Church and Wellesley	1	M4Y	Downtown Toronto	43.665860	-79.383160	30472.0	1

Fig 14. Cluster1: 5 highest recommended neighborhoods to open an Indian restaurant

5.3 Cluster 2

Cluster 1 includes a total of 49 neighborhoods which is the most number of neighborhoods within any cluster. Neighborhoods in this cluster have **low population with no Indian restaurants**. The average population in this cluster is 17,013. Even though these neighborhoods do not have Indian restaurants, this is the least recommended cluster. This is because with less population in these neighborhoods, relatively few customers are expected.

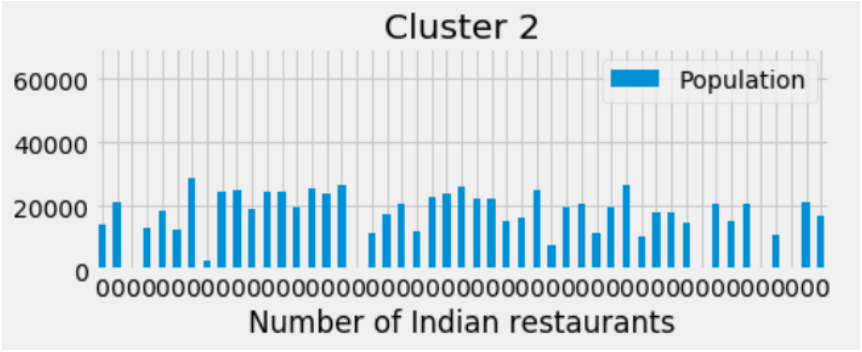

Fig 15. Cluster2: Population vs Number of restaurants

For any other reasons, If someone wants to open an Indian restaurant in neighborhoods within cluster 1, with the largest population, Glencairn in North York, is the highly recommended location. Table gives five highly recommended locations within cluster 1 to open an indian restaurant.

Neighborhood	Indian_Restaurants	PostalCode	Regions	Latitude	Longitude	Population	cluster
Glencairn	0	M6B	North York	43.709577	-79.445073	28522.000000	1
DownsviewEast	0	M3K	North York	43.737473	-79.464763	26728.281142	1
Davisville	0	M4S	Central Toronto	43.704324	-79.388790	26506.000000	1
Bedford Park / Lawrence Manor East	0	M5M	North York	43.733283	-79.419750	25975.000000	1
Northwood Park / York University	0	M3J	North York	43.767980	-79.487262	25473.000000	1

Fig 16. Cluster1: 5 highest recommended neighborhoods to open an Indian restaurant

6. Conclusion

The best neighborhood in Toronto to open a new Indian restaurant was discovered using an unsupervised machine learning clustering algorithm called K-means clustering. Three clusters of neighborhoods were formed. The cluster0 is the cluster of highest recommended neighborhoods to open a new Indian restaurant. Willowdale south in North York is the highest recommended neighborhood in Toronto.

7. Reference

1. https://en.wikipedia.org/wiki/List_of_the_largest_municipalities_in_Canada_by_population
2.https://www.cicnews.com/2020/02/which-cities-in-canada-attract-the-most-immigrants-0213741.html#gs.6um8t1
3.https://www12.statcan.gc.ca/census-recensement/2016/as-sa/fogs-spg/Facts-cma-eng.cfm?LANG=Eng&GK=CMA&GC=535&TOPIC=7