Coursera Capstone

Opening New Indian Restaurant in Toronto City, Canada

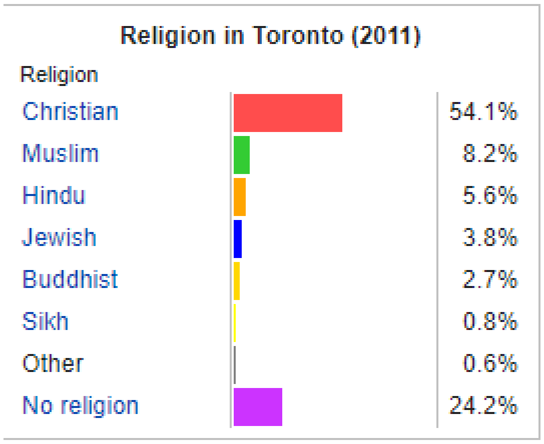
By

Santosh Sawant

06.08.2020

Introduction:

Toronto is the capital city of the Canadian province of Ontario. It is the most populous city in Canada and fourth most populous city in North America. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. If we look at the overall religion status, around 15% population belongs to asian community (Fig. pertaing to 2011 as per Wikipedia). The city is home to the Toronto Stock Exchange, the headquarters of Canada's five largest banksand the headquarters of many large Canadian and multinational corporations. Due to this population density of immigramts os more in Toronto than other Canadian cities. Also statistics indicates that, immigrants to Canda from India are highest after China. So There is high probability of successful running of Indian restaurants in Toronto.



Business Problem:

The objective of this capstone project is to analyse and select the best locations in the city of Toronto, Canada to open a new Indian Restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Toronto, Canada, if a popular restaurant chain owner is looking to open a new Indian restaurant, where would you recommend that they open it?

Target Audience of this project :

This project is particularly useful to Restaurant owners, investors looking to open or invest in new Indian restaurant in the Toronto city in Canada.

Data:

To solve the problem, we will need the following data:

- List of neighborhoods in Toronto. This defines the scope of this project which is confined to the city of Toronto in Canada.

- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to Indian restaurants. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of neighborhoods in Toronto, with a total of 103 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods. After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Indian restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used

Methodology

Firstly, we need to get the list of neighborhoods in the city of Toronto. The list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) provided in week3 assignment on Coursera. We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas Data Frame and then visualize the neighborhoods in a map using Folium package.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Indian Restaurant” data, we will filter the “Indian Restaurant” as venue category for the neighborhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Indian Restaurant”. The results will allow us to identify which neighborhoods have higher concentration of Indian Restaurants while which neighborhoods have fewer number of those. Based on the occurrence of Indian restaurant in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Indian Restaurant.

Results:

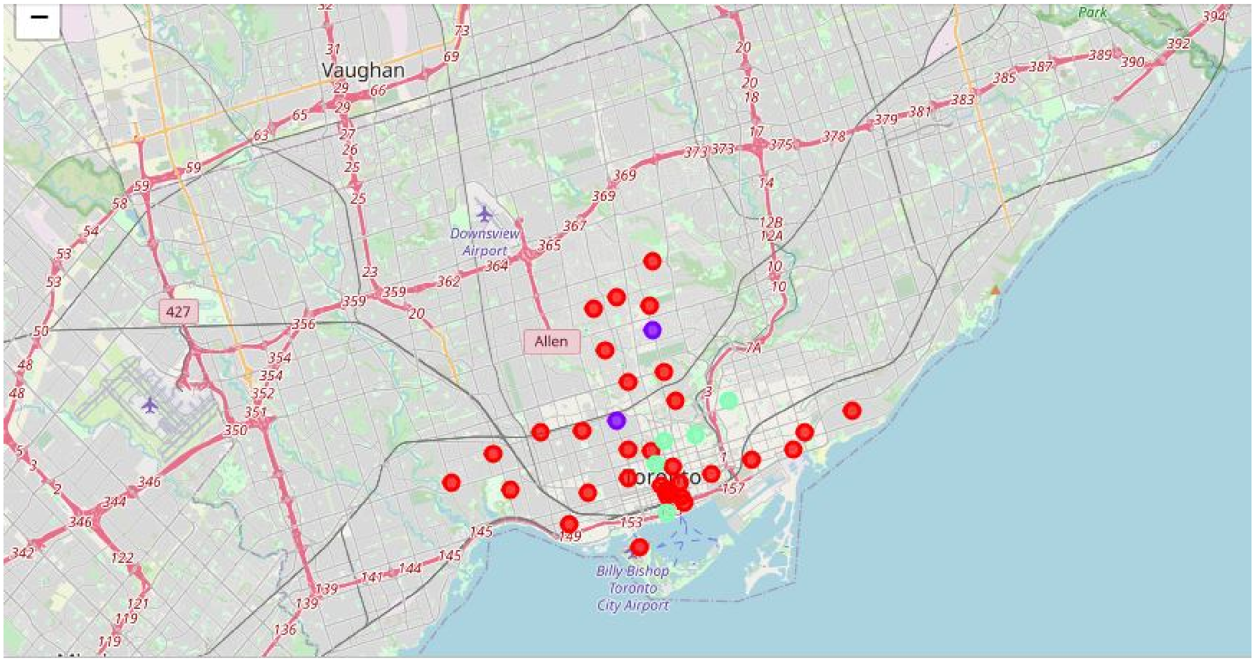
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Indian Restaurant”:

Cluster 0: Neighborhoods with Zero number of Indian restaurants

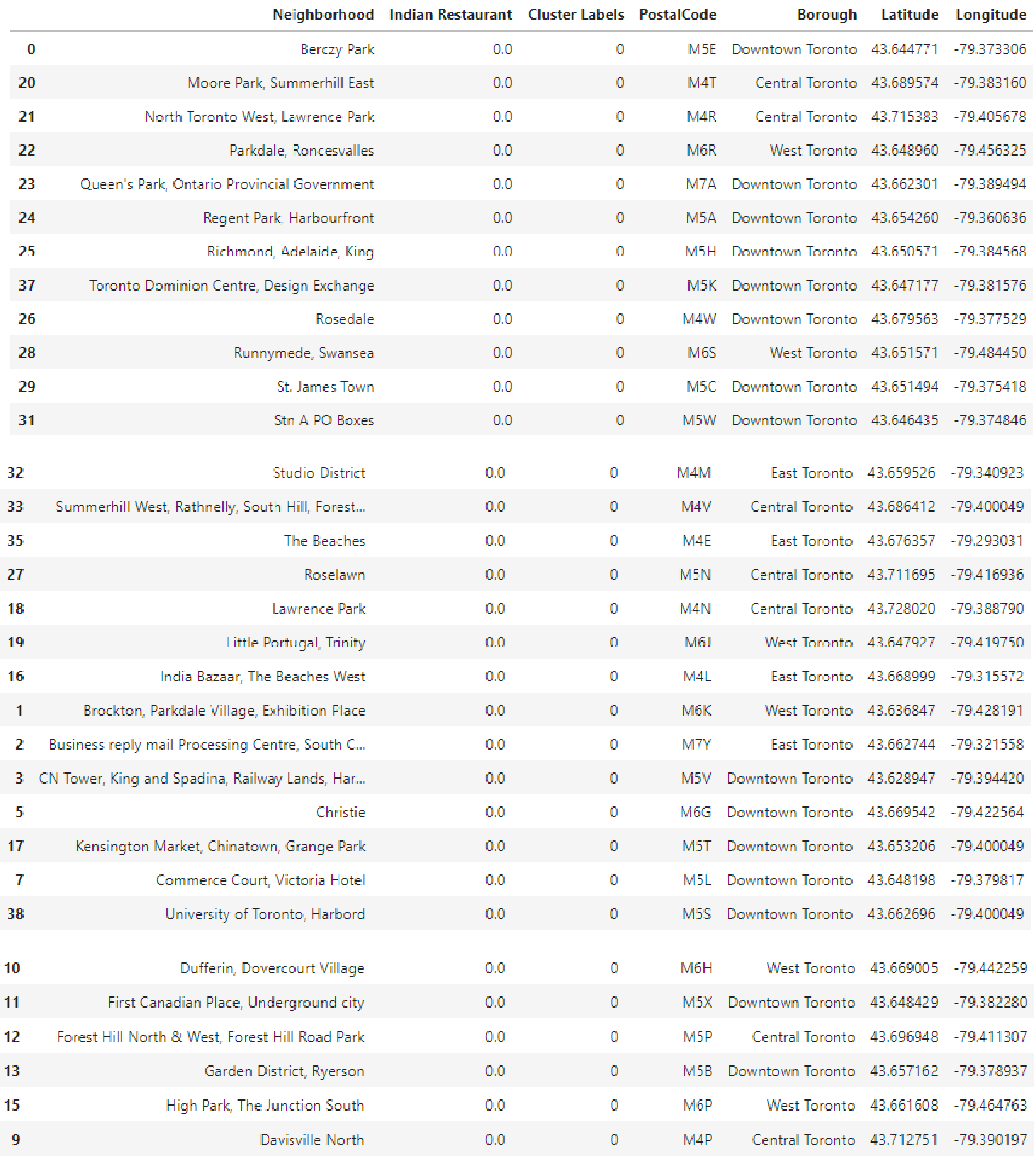
Cluster 1: Neighborhoods with moderate number of Indian restaurants

Cluster 2: Neighborhoods with high concentration of Indian restaurants

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.

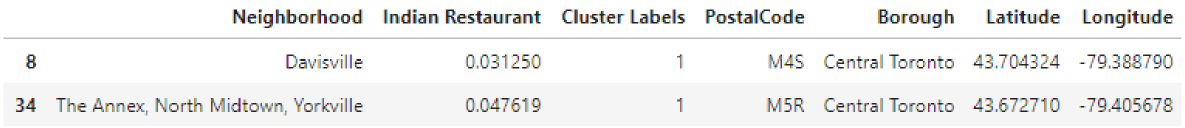


Cluster 0:



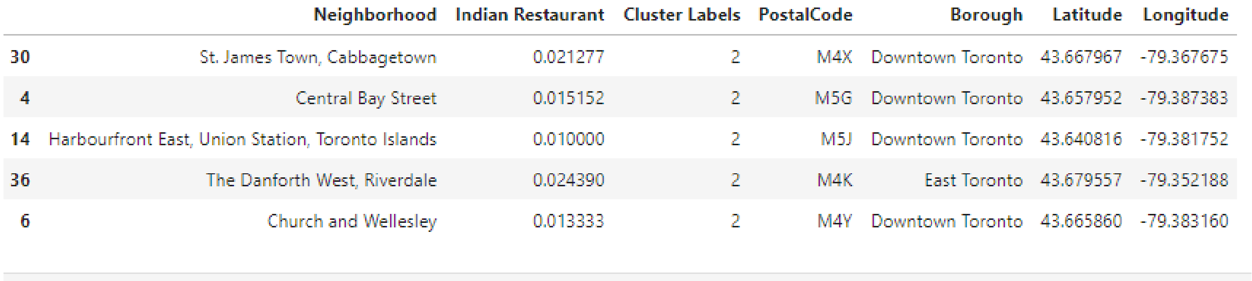
	Neighborhood	Indian Restaurant	Cluster Labels	PostalCode	Borough	Latitude	Longitude
0	Berczy Park	0.0	0	M5E	Downtown Toronto	43.644771	-79.373306
20	Moore Park, Summerhill East	0.0	0	M4T	Central Toronto	43.689574	-79.383160
21	North Toronto West, Lawrence Park	0.0	0	M4R	Central Toronto	43.715383	-79.405678
22	Parkdale, Roncesvalles	0.0	0	M6R	West Toronto	43.648960	-79.456325
23	Queen's Park, Ontario Provincial Government	0.0	0	M7A	Downtown Toronto	43.662301	-79.389494
24	Regent Park, Harbourfront	0.0	0	M5A	Downtown Toronto	43.654260	-79.360636
25	Richmond, Adelaide, King	0.0	0	M5H	Downtown Toronto	43.650571	-79.384568
37	Toronto Dominion Centre, Design Exchange	0.0	0	M5K	Downtown Toronto	43.647177	-79.381576
26	Rosedale	0.0	0	M4W	Downtown Toronto	43.679563	-79.377529
28	Runnymede, Swansea	0.0	0	M6S	West Toronto	43.651571	-79.484450
29	St. James Town	0.0	0	M5C	Downtown Toronto	43.651494	-79.375418
31	Stn A PO Boxes	0.0	0	M5W	Downtown Toronto	43.646435	-79.374846
32	Studio District	0.0	0	M4M	East Toronto	43.659526	-79.340923
33	Summerhill West, Rathnelly, South Hill, Forest...	0.0	0	M4V	Central Toronto	43.686412	-79.400049
35	The Beaches	0.0	0	M4E	East Toronto	43.676357	-79.293031
27	Roselawn	0.0	0	M5N	Central Toronto	43.711695	-79.416936
18	Lawrence Park	0.0	0	M4N	Central Toronto	43.728020	-79.388790
19	Little Portugal, Trinity	0.0	0	M6J	West Toronto	43.647927	-79.419750
16	India Bazaar, The Beaches West	0.0	0	M4L	East Toronto	43.668999	-79.315572
1	Brockton, Parkdale Village, Exhibition Place	0.0	0	M6K	West Toronto	43.636847	-79.428191
2	Business reply mail Processing Centre, South C...	0.0	0	M7Y	East Toronto	43.662744	-79.321558
3	CN Tower, King and Spadina, Railway Lands, Har...	0.0	0	M5V	Downtown Toronto	43.628947	-79.394420
5	Christie	0.0	0	M6G	Downtown Toronto	43.669542	-79.422564
17	Kensington Market, Chinatown, Grange Park	0.0	0	M5T	Downtown Toronto	43.653206	-79.400049
7	Commerce Court, Victoria Hotel	0.0	0	M5L	Downtown Toronto	43.648198	-79.379817
38	University of Toronto, Harbord	0.0	0	M5S	Downtown Toronto	43.662696	-79.400049
10	Dufferin, Dovercourt Village	0.0	0	M6H	West Toronto	43.669005	-79.442259
11	First Canadian Place, Underground city	0.0	0	M5X	Downtown Toronto	43.648429	-79.382280
12	Forest Hill North & West, Forest Hill Road Park	0.0	0	M5P	Central Toronto	43.696948	-79.411307
13	Garden District, Ryerson	0.0	0	M5B	Downtown Toronto	43.657162	-79.378937
15	High Park, The Junction South	0.0	0	M6P	West Toronto	43.661608	-79.464763
9	Davisville North	0.0	0	M4P	Central Toronto	43.712751	-79.390197

Cluster 1:



	Neighborhood	Indian Restaurant	Cluster Labels	PostalCode	Borough	Latitude	Longitude
8	Davisville	0.031250	1	M4S	Central Toronto	43.704324	-79.388790
34	The Annex, North Midtown, Yorkville	0.047619	1	M5R	Central Toronto	43.672710	-79.405678

Cluster 2:



	Neighborhood	Indian Restaurant	Cluster Labels	PostalCode	Borough	Latitude	Longitude
30	St. James Town, Cabbagetown	0.021277	2	M4X	Downtown Toronto	43.667967	-79.367675
4	Central Bay Street	0.015152	2	M5G	Downtown Toronto	43.657952	-79.387383
14	Harbourfront East, Union Station, Toronto Islands	0.010000	2	M5J	Downtown Toronto	43.640816	-79.381752
36	The Danforth West, Riverdale	0.024390	2	M4K	East Toronto	43.679557	-79.352188
6	Church and Wellesley	0.013333	2	M4Y	Downtown Toronto	43.665860	-79.383160

Discussion:

As observations noted from the map in the Results section, most of the Indian restaurants are concentrated in the in cluster 2 and moderate number in cluster 1 . On the other hand, cluster 0 has no occurrence of Indian restaurants in the neighborhoods. This represents a great opportunity and high potential areas to open new Indian restaurant as there is very little to no competition from existing ones. Meanwhile, Indian restaurant in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Indian restaurants. Therefore, this project recommends restaurant owners/ investors to use these findings to open new Indian restaurant in neighborhoods in cluster 0 with no competition. Lastly, investors are advised to avoid neighborhoods in cluster 2 which already have high concentration of Indian restaurants and may be suffering from intense competition.

Limitations and Suggestions for Future Research:

In this project, we only consider one factor i.e. frequency of occurrence of Indian Restaurant, there are other factors such as population (Mainly Asian community) and income of residents that could influence the location decision of opening new Indian restaurant. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new Indian restaurant.

Conclusion:

Answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new Indian Restaurant. Further decision to identify location in cluster 0 neighborhoods may be based on other factors such as population density of Asian community / income levels of population. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding higher competition as in cluster 2.

References:

Wikipedia pages:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

https://en.wikipedia.org/wiki/Toronto