INDIAN RESTAURANT IN TORONTO

A DATA SCIENCE APPROACH

Anshul Sharma
mail.ansh@gmail.com

Table of Contents

INTRODUCTION

This report is intended to provide an insight on "How Data Science can be used to solve a business problem?" The report sections will discuss a specific business problem and a data science approach to solve the same.

BUSINESS SCENARIO BACKGROUND

Canada has been a preferred immigration destination for millions of people over the last decade. As the population has increased exponentially in Toronto, there is a high demand of services catering to the needs of immigrants. One such service is restaurants. The Hotel and restaurant business is on a boom ever since Toronto is flocked with immigrants. Every year, a high number of immigrants settle in the city Toronto who have immigrated from India. Hence there is a high demand for Indian Restaurants in Toronto. However, investing in a restaurant business is make or break life decision for anyone.

With this background, this report will try to address the following question for any entrepreneur who is planning to invest in opening a new Indian restaurant in Toronto city.

1. How many Indian restaurants are already present in Toronto City?
2. Which neighbourhoods in Toronto city has an Indian restaurant?
3. Which neighbourhoods in Toronto city will be ideal to start a new India Restaurant?
4. What are the descriptive highlights of these ideal neighbourhoods?

   With these questions addressed, an entrepreneur will have a fair idea on where to start his business.
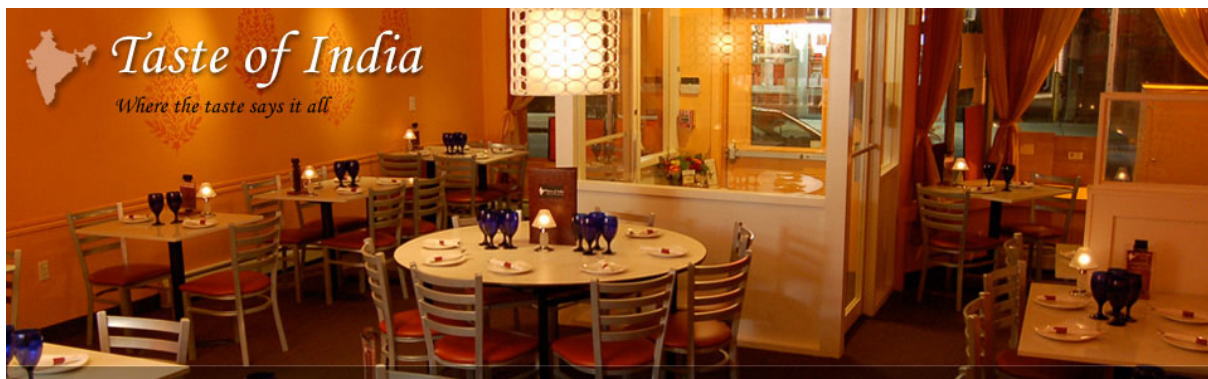

Image 1: Concept image of an Indian Restaurant in Toronto.

DATA OVERVIEW

As the scope of the report is around restaurants business in Toronto, the analysis would require data related to the geography of Toronto city. These may include but not limited to postal codes, neighbourhoods of Toronto city with their geo coordinates. Also, as the main business background is related to the restaurant business, data related to the list of restaurants in Toronto city will be required. Below are the list of datasets used in the report and background study.

1. Postal codes of Canada
2. Geographical coordinates of the postal codes.
3. Venues data of Toronto City.

DATA DESCRIPTION AND ACQUISITION
The above mentioned dataset are explained as follows.

Postal Codes of Canada – Web Scraping
The study for this report starts with exploring the geography of Toronto. For this, we need to understand the defined Postal codes of Toronto. Each postal code is assigned to a Borough. A Borough is defined as an area within a city which is either made of a single or multiple neighbourhoods. A list of all the postal codes with assigned Boroughs and neighbourhoods is available on a wiki page. Below is the link to the same. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Image 2: Wikipedia page detailing "list of postal codes of Canada."

To acquire this data, we can use two approaches.
A. Using pandas.read_html(): Pandas library in python offers support to load the data content of any html onto your working environment. Once the data is loaded, we can extract any tables or texts from the same html page.
   For the purpose of this study, I have used the same to extract postal codes of Canada table from the wiki page. Once the data was extracted, it was sliced to create a data frame with subset of only postal codes, Borough and Neighbourhoods of Toronto city.
B. Using BeautifulSoup package: This package is also used to scrape the data from web for more complex cases. However, for the purpose of this study I have not used this package.

Geographical coordinates of the postal codes – Web Scraping
The geographical coordinates data for all the postal codes and Borough is required as this is the main input for acquiring other venue related data. This data can be extracted using Google Maps Geocoding API or Geocoder Python package or using a direct link to a csv file http://cocl.us/Geospatial_data .
For this study, I have used
1.   Geocoder Python package and
2.   the csv file
to extract geographical coordinates of location.

	Postal Code	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

Image3: The CSV file data

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031

Image 4: CSV data merged with postal code data

Image5: Boroughs of Toronto plotted on Map using Folium library.

Venues data of Toronto City – Foursquare API

As the study requires to investigate restaurant business, we need to acquire the venue data in Toronto. The venue data details all types of venues, their category and geo-coordinates for any desired location. For this study, the venue data was extracted using the Foursquare API. The venue data consists of the following as shown in figure below.

1. Venue Name
2. Venue Category
3. Venue Latitude
4. Venue Longitude

	name	categories	lat	lng
0	Roselle Desserts	Bakery	43.653447	-79.362017
1	Tandem Coffee	Coffee Shop	43.653559	-79.361809
2	Cooper Koo Family YMCA	Distribution Center	43.653249	-79.358008
3	Impact Kitchen	Restaurant	43.656369	-79.356980
4	Body Blitz Spa East	Spa	43.654735	-79.359874

Image 6: Venue data extracted using Foursquare API

Final dataset for analysis

Once all the individual datasets were retrieved from web scraping and Foursquare API, the datasets were merged together to create one data frame.

At first, the arriving data frame had each row defined as each venue in the city of Toronto. However, to the object of the exercise was to identify a suitable neighbourhood. Hence the data frame had to be transformed to represent the venues in terms of the neighbourhoods. To do so,

the rows with same neighbourhood names were grouped together. **One hot encoding** was used for the same. The resulting data frame had each row representing each neighbourhoods and the venues were represented as columns with each data point to show the frequency of occurrence of the venue in that neighbourhood.

	Neighbourhood	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant
0	Berczy Park	0.0	0.0000	0.0000	0.0000	0.000	0.000	0.0000	0.0	0.0	0.0	0.018182	0.000000	0.0	0.0
1	Brockton, Parkdale Village, Exhibition Place	0.0	0.0000	0.0000	0.0000	0.000	0.000	0.0000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0
2	Business reply mail Processing Centre, South C...	0.0	0.0000	0.0000	0.0000	0.000	0.000	0.0000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0
3	CN Tower, King and Spadina, Railway Lands, Har...	0.0	0.0625	0.0625	0.0625	0.125	0.125	0.0625	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0
4	Central Bay Street	0.0	0.0000	0.0000	0.0000	0.000	0.000	0.0000	0.0	0.0	0.0	0.000000	0.014706	0.0	0.0

Image 7: Final data frame used for exploratory data analysis and machine learning

With this the data frame was ready for further data analysis and creating machine learning algorithm which will suggest the best possible neighbourhood to start a new Indian restaurant in Toronto.

METHEDOLOGY

EXPLORATORY DATA ANALYSIS
For the exploratory data analysis the following was done.

TOTAL VENUE CATEGORIES.
Identifying total number of venues in each neighbourhood would give us the idea about the number of 'to-do' options available to individual living in Toronto. The table below shows the list of neighbourhoods in Toronto with total number of venues in them.

Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Berczy Park	55	55	55	55	55	55
Brockton, Parkdale Village, Exhibition Place	23	23	23	23	23	23
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	16	16	16	16	16	16
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	16	16	16	16	16	16
Central Bay Street	68	68	68	68	68	68

Image 8: Total number of venue categories in each Neighbourhood

TOP 10 VENUES
In this section we explored the top 10 venues by their occurrence in the neighbourhood. This would help us know if there are any Indian restaurant in top 10 category in any neighbourhood. If yes, this can be a serious competition. We would try to avoid such neighbourhood to start our restaurant.

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Coffee Shop	Seafood Restaurant	Cocktail Bar	Farmers Market	Beer Bar	Restaurant	Cheese Shop	Bakery	Sandwich Place	Department Store
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Nightclub	Coffee Shop	Climbing Gym	Burrito Place	Restaurant	Italian Restaurant	Intersection	Bar
2	Business reply mail Processing Centre, South C...	Park	Pizza Place	Light Rail Station	Skate Park	Burrito Place	Farmers Market	Fast Food Restaurant	Butcher	Restaurant	Recording Studio
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Lounge	Airport Service	Boutique	Harbor / Marina	Boat or Ferry	Rental Car Location	Bar	Plane	Coffee Shop	Sculpture Garden
4	Central Bay Street	Coffee Shop	Café	Italian Restaurant	Sandwich Place	Salad Place	Bubble Tea Shop	Department Store	Burger Joint	Japanese Restaurant	Thai Restaurant

Image 9: TOP 10 VENUES in each Neighbourhood

INDIAN RESTAURANTS IN TORONTO.

As we are interested to open an Indian Res., its very important to know the total number of Indian restaurants and their frequency of occurrence in each of the neighbourhoods. Again this is to understand the competition and avoid it.

	Neighbourhood	Venue
0	Central Bay Street	1
1	Church and Wellesley	1
2	Davisville	1
3	Harbourfront East, Union Station, Toronto Islands	1
4	St. James Town, Cabbagetown	1
5	The Annex, North Midtown, Yorkville	1
6	The Danforth West, Riverdale	1

Image 10: Indian restaurants in Toronto

INFERENTIAL STATISTICS - MACHINE LEARNING

In order to identify a suitable neighbourhood to start an Indian restaurant, I have clustered the neighbourhoods of Toronto based on the frequency of other Indian restaurant present in them. I have create 4 clusters of neighbourhoods. Furthermore, I have analysed each of these clusters to identify the most suitable one for us.

For clustering the neighbourhoods, I have used **K-mean clustering** method.

The following clusters were created.

1. Cluster 1 – Red

2. Cluster 2 – Purple

3. Cluster 3 – Aquamarine

4. Cluster 4 – Dark khaki'

Below is a sample table depicting the outcome of the clustering. In this each of the neighbourhood is tagged to a cluster.

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0	0.0
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	0	0.0
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	0	0.0
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	0	0.0
4	M4E	East Toronto	The Beaches	43.676357	-79.293031	0	0.0

Image 11: Neighbourhoods tagged to clusters

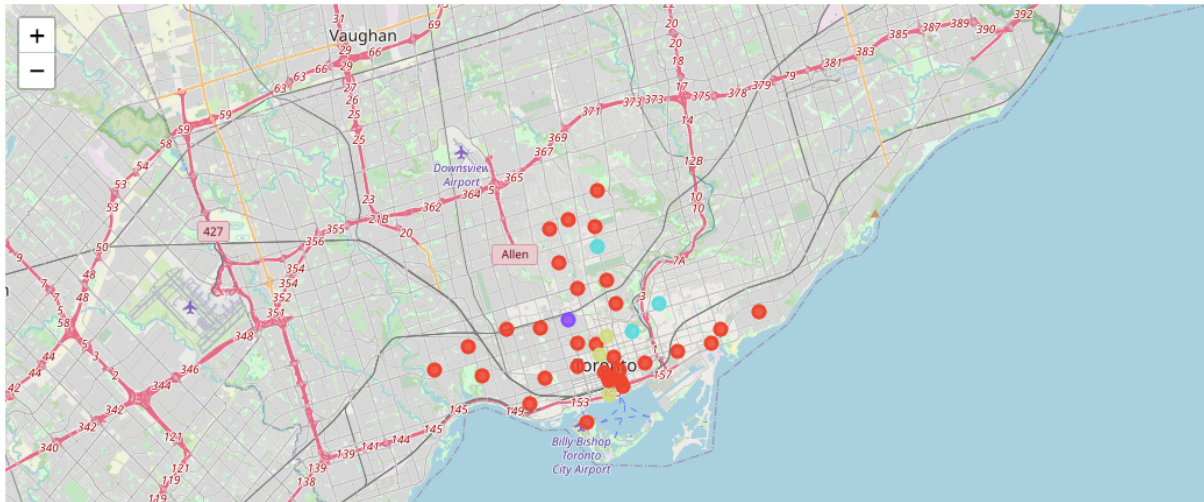Also the resulting clusters were plotted on the map of Toronto to show their spread.

Image 12: Neighbourhoods clusters plotted on a map

RESULTS

Following analysis were done on the clusters to identify the best neighbourhood for opening a new Indian Restaurant.

Neighbourhood in clusters

The image below shows the distribution of neighbourhoods across the 4 clusters identified. Clearly cluster 0 has highest number ie 32 neighbourhood in it. Followed by cluster 2 and 3 with 3 neighbourhoods each in them and finally cluster 1 with just one neighbourhood in it.
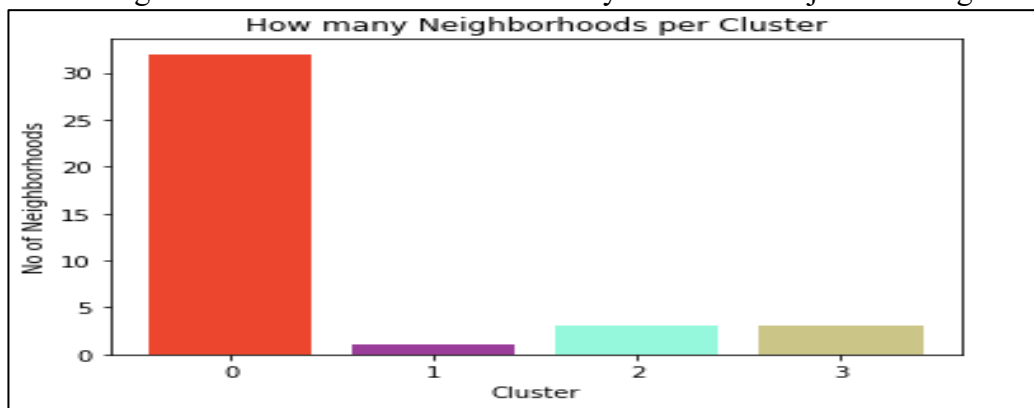

Image 13: Neighbourhoods per clusters

Indian restaurants across clusters

There are no Indian restaurants present in cluster 0. Also there are 3 Indian restaurants present in cluster 2 and 3 each and 1 in cluster 1.
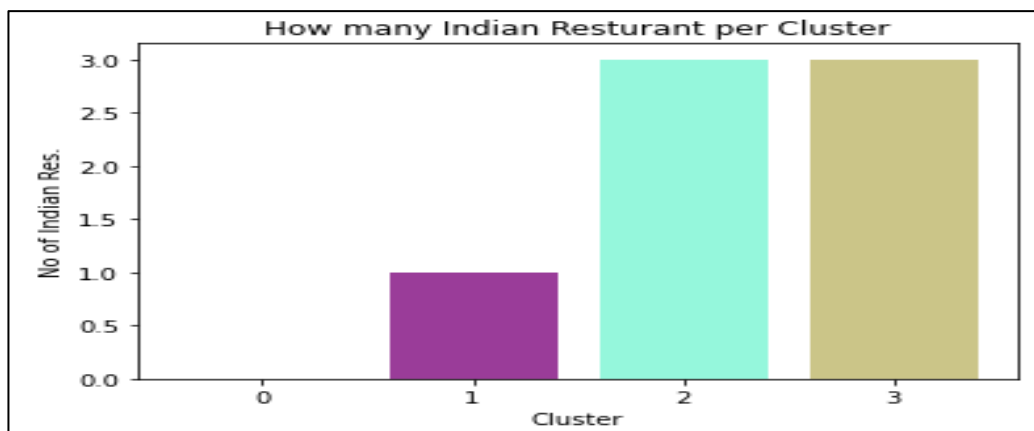

Image 14: Indian restaurants per cluster

Cluster 0 - Red

Cluster 0 has no Indian restaurants in it. It is spread across all 4 Borough ie 'Downtown Toronto', 'East Toronto', 'West Toronto' and 'Central Toronto'. This can be an Ideal cluster for setting up an Indian Restaurant as there is no competition present in the same. Let us try to find an ideal neighbourhood in this cluster for our restaurant.

Cluster 1 - Purple

Cluster 1 is present in only 1 borough ie Central Toronto. Since there is only 2 neighbourhoods in this cluster and with 1 Indian Restaurant already present in it, its not an ideal choice for us to open another Indian Restaurant here.

Cluster 2 - Aquamarine

Cluster 2 is present in only 3 borough and is spread across 5 neighbourhoods. With 3 Indian Restaurant already present in it, its not an ideal choice for us to open another Indian Restaurant here.

Custer 3 - Dark khaki'

Cluster 3 is present in only 3 borough and is spread across 5 neighbourhoods. With 3 Indian Restaurant already present in it, its not an ideal choice for us to open another Indian Restaurant here.

DISCUSSION

Clearly cluster 1, 2, 3 are not our preferred choices to start the Indian restaurant because of the above mentioned reasons. However, cluster 0 seems to be suitable. Also Interestingly, within cluster 0 there are few neighbourhoods which have very common Parks and bars. The same was extracted using the dataset below. This would give us a good sense of which neighbourhood would be a better choice.

	Borough	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
15	East Toronto	India Bazaar, The Beaches West	Park	Gym	Sushi Restaurant	Sandwich Place	Liquor Store	Burrito Place	Restaurant	Italian Restaurant	Fast Food Restaurant	Fish & Chips Shop
18	Central Toronto	Lawrence Park	Park	Swim School	Bus Line	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant	Dog Run
20	Central Toronto	Davisville North	Park	Hotel	Breakfast Spot	Sandwich Place	Food & Drink Shop	Dog Run	Dance Studio	Department Store	Gym / Fitness Center	Cosmetics Shop
21	Central Toronto	Forest Hill North & West, Forest Hill Road Park	Park	Sushi Restaurant	Jewelry Store	Trail	Escape Room	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant
33	Downtown Toronto	Rosedale	Park	Trail	Playground	Deli / Bodega	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant	Dog Run
	Borough	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
11	West Toronto	Little Portugal, Trinity	Bar	Coffee Shop	Vietnamese Restaurant	Restaurant	Café	Asian Restaurant	Men's Store	Cuban Restaurant	Record Shop	Pizza Place

Image 15: clusters0 Neighbourhoods with Parks and bars

CONCLUSION

Hence only cluster 0 spread across 32 neighbourhoods and with no existing Indian Restaurant present in it is the ideal choice to start a new Indian Restaurant. Also, neighbourhoods like India Bazaar, The Beaches West and Little Portugal, Trinity are most suitable neighbourhoods with lots of parks and bars around.

This Analysis can be made more robust by adding demographic data. This will create clusters with more detailed features and will help us making more informed decisions.

Find me on LinkedIn @ https://www.linkedin.com/in/anshul-sharma-14445728/
and the codes for this project on GitHub
https://github.com/Anshul-DS/Coursera_Capstone/blob/main/Final%20Project.ipynb