Report on
Best Location for Indian Restaurant in Toronto
Coursera-Capstone Project

Table of Contents

Introduction

Toronto is Canada's largest city and a world leader in such areas as business, finance, technology, entertainment and culture. Its large population of immigrants from all over the globe has also made Toronto one of the most multicultural cities in the world. It is the most populous city in Canada and the country’s financial and commercial centre.

While opening a restaurant can be a very lucrative business, a lack of demand or over saturation causes many restaurants to close within the first year of opening. There are many different factors that can account for a restaurant’s success such as location, competition and quality of the food. This is an important question that every business owner must face while choosing whether to open a restaurant or not? or which location is suitable for the business?

In this project will try to find an optimal location to open an Indian restaurant. Specifically, this report will be targeted to stakeholders interested in opening an Indian Restaurant in Toronto. Finding a suitable location for Indian restaurants in major cities like Toronto proves to be a daunting task. Hence, customers can bolster their decisions using the descriptive and predictive capabilities of data science.

We need to find locations(Neighborhood) that have a potentially unfulfilled demand for Indian Restaurant. Also, we need locations that have low competition and are not already crowded. We would also prefer location as close to popular city Neighborhood, assuming the first two conditions are met.

Target Audience

We will use our data science knowledge to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best final location can be chosen by stakeholders.

Business Problem

The objective of this project is to analyze and select the best locations in Toronto 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, if someone is looking to open a new Indian restaurant, where would you recommend that they open it?

Data Acquisition

Based on definition of our problem, factors that will influence our decision are:

- Number of existing restaurants in the neighborhood (any type of restaurant)
- Number of and distance to Indian restaurants in the neighborhood.
- Distance of neighborhood from popular neighborhoods.

In our project we will:

- Acquire the names and boroughs of the neighborhoods by scrapping a Wikipedia page.
- Next, we use the foursquare API to find all types of restaurants within a 1000 meter radius for every neighborhood.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Toronto. The list is available in the Wikipedia page and is acquired from 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'. We will use web scraping with Python requests and the beautifulsoup package to extract a list of the neighborhood's data.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 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 DataFrame and then visualize the neighborhoods on a map using the Folium package. This allows us to perform a stare and compare to make sure that the geographical coordinates returned by the Geocoder package are correctly plotted in the city of Toronto.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. 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 validate how many venues were returned for each neighborhood and examine how many unique categories can be created 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 a venue category for the neighborhoods.

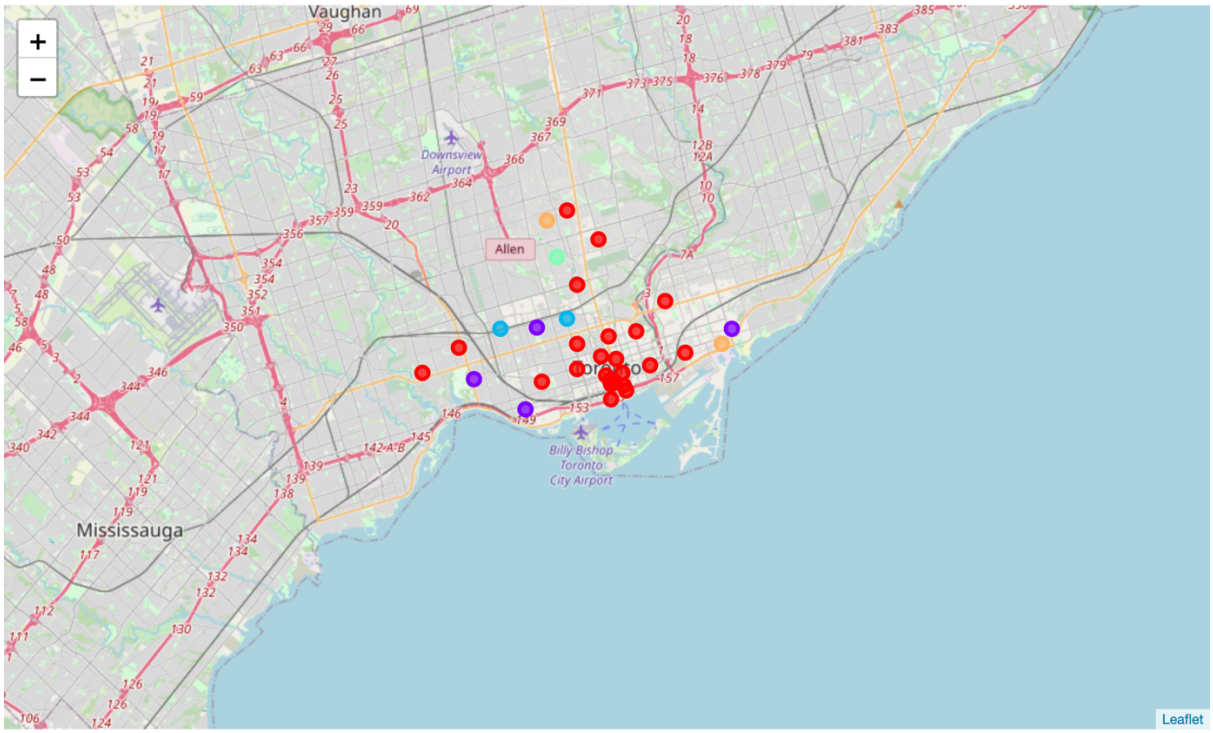
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and most popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

We will cluster the neighborhoods into 5 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 Indian Restaurants. Based on the occurrence of Indian Restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Indian Restaurants.

	index	Neighborhood	Latitude	Longitude
0	21	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576
1	22	Commerce Court, Victoria Hotel	43.648198	-79.379817
2	30	First Canadian Place, Underground city	43.648429	-79.382280

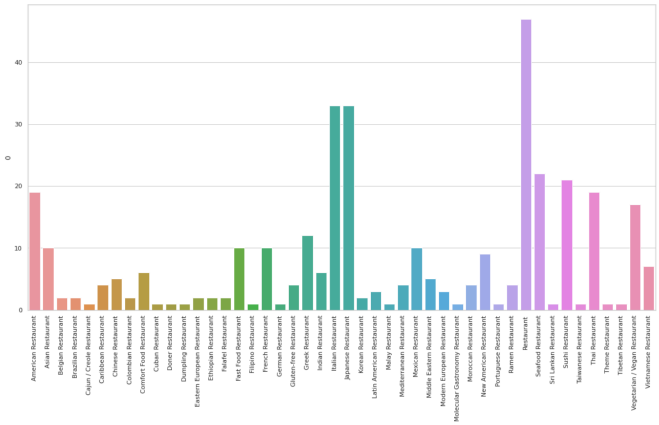
Result

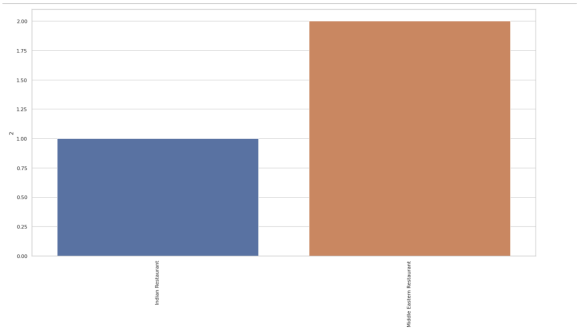
Clusters



The results from k-means clustering shows that we can categorise Toronto neighborhoods into 5 clusters based on how many Indian Restaurants are in each neighborhood.

Now, as clusters 0 and 2 have a maximum number of Indian Restaurants, we will focus our search on neighborhoods within these two clusters. Clusters 0 and 2 may collectively have the highest number of Indian restaurant but there will be some neighborhoods in these clusters which would have a demand for Indian Restaurants, as these neighborhoods are in the same cluster, but would not have enough supply.





Discussion Section

Our Analysis was done on over 39 neighborhoods, containing over 373 restaurants within 2km radius of every neighborhood. We segragated these neighborhoods on the basis of types and amounts of restaurants. Five clusters were obtained, each having a unique collection of restaurants. Since, we were focused on finding optimal neighborhoods for opening Indian restaurants, we selected cluster 0 and 2 which had the highest number of Indian restaurants. The above actions left us with the only those neighborhoods that had a shared characteristics of and that had a high demand for Indian restaurants.

Next, we plotted a heat map for analysing the density of restaurants in the remaining neighborhoods. This allowed us to select neighborhoods that had few or no Indian restaurants and were not overcrowded by other kinds of restaurants. After this, we found out the top three most popular neighborhoods (namely: Toronto Dominion Centre, Design Exchange,Commerce Court, Victoria Hotel, First Canadian Place, Underground city), and the distance of every remaining neighborhoods from all three of them. Then, we extracted top 5 closest neighborhoods from each of three most popular neighborhoods mentioned above.

Taking the union of the resulting three dataset we get 6 neighborhoods that satisfy all three conditions laid out in the business problem by the customer.

Conclusion

The neighborhoods recommendation obtained here are not completely accurate. This is due to the limitations in the dataset used in the project. Due to lack of cross referencing sources, we may have missed a few neighborhoods from our consideration.