Analysis and Selection of Neighborhoods in Toronto for Indian Restaurant

Applied Data Science Capstone Project

1.Introduction

1.1 Background

Toronto is the capital city of the province of Ontario Canada. The city is famous for business, finance, technology, quality education, and so on. This city is one of the largest and multicultural and cosmopolitan cities in North America. It has diverse demography with people from all over the world and one of the popular destinations for immigrants from Asia. Southeast Asia's culture and cuisine, especially that of the Indian community, can be found here due to their dominant population.

1.2 Problem

Indian cuisine is very popular in the city which has created good business for an Indian restaurant. As the population of the South Asian community around the metropolitan area is increasing, the demand for a restaurant with authentic south Asian recipes is also increasing. But due to various reason investor are unable to identify the proper location to open the restaurant. Restaurant of similar type is clustered within the specific area which has not only increased competition within small customer number but also has greatly hindered the profitability.

1.2 Interest

This project is intended to provide a valuable answer to those stakeholders who are thinking of doing business-related in this sector. Moreover, this project will focus on choosing the appropriate location to open Indian restaurants which are not crowded with these types. Also, this project will determine the neighborhoods where there is a higher demand for Indian cuisines.

The goal is to use FourSquare API to extract the geographical information for the neighborhoods of Toronto and identify the venues with a lesser number of Indian Restaurant within the area.

2. Data Acquisition

2.1 Data Sources

As per our problem, we required geographical information and all the neighborhood around the Toronto Metropolitan city. For this purpose, I extracted all the neighborhood data from Wikipedia which includes Postal code, Brough, and Neighborhoods of Toronto city.

After scraping all the aforementioned information from the web, I was required geospatial information for these neighborhoods which was acquired using a CSV file obtained from Kaggle. The file contains the postal code for each neighborhood with its respective longitude and latitude.

Also, we need to find out all the restaurants and related venues in the Toronto Downtown neighborhoods. For this purpose, I utilized FourSquare API to extract all the venues as per their latitude and longitude and later filtered data for restaurants and related venues.

2.2 Data Cleaning

Now that we have scraped neighborhood data from Wikipedia using request library and beautiful soup, we must clean data to move ahead. Also, the HTML file is converted into a pandas dataframe which will make it easier in the cleaning, analysis, and visualization process.
For this process, I simplified names in the Brough column using replace function and put the data into a dataframe. Then, using the shape function the total number of rows and columns of the dataframe was identified.

2.3 Adding Features

Now that we have seen that our dataset consists of 103 rows and 3 columns, let’s add latitude and longitude to these neighborhoods using the dataset obtained from Kaggle. Since I am using IBM Watson Studio for this project I will import the CSV file to the notebook and merge it with our dataframe. First, I extracted the CSV file into the notebook and change it into dataframe, and renamed some of the columns of the geodata_df which makes it easier to merge with the previous data set. After completing all the processes, the final dataframe named "neighborhoods" was created which is shown in the figure below.

```
In [15]: neighborhoods.head()
```

Out[15]:

	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:Neighborhoods data

2.4 Neighborhood Candidates

Since we are interested in determining the neighborhoods in Downtown Toronto, I took the neighborhoods which are located here. First, I found the geographical information for Toronto using the "geolocator" function. Then, I filtered the dataset for the borough of Downtown Toronto. I used a Folium map to visualize these neighborhoods which is shown in the figure below.

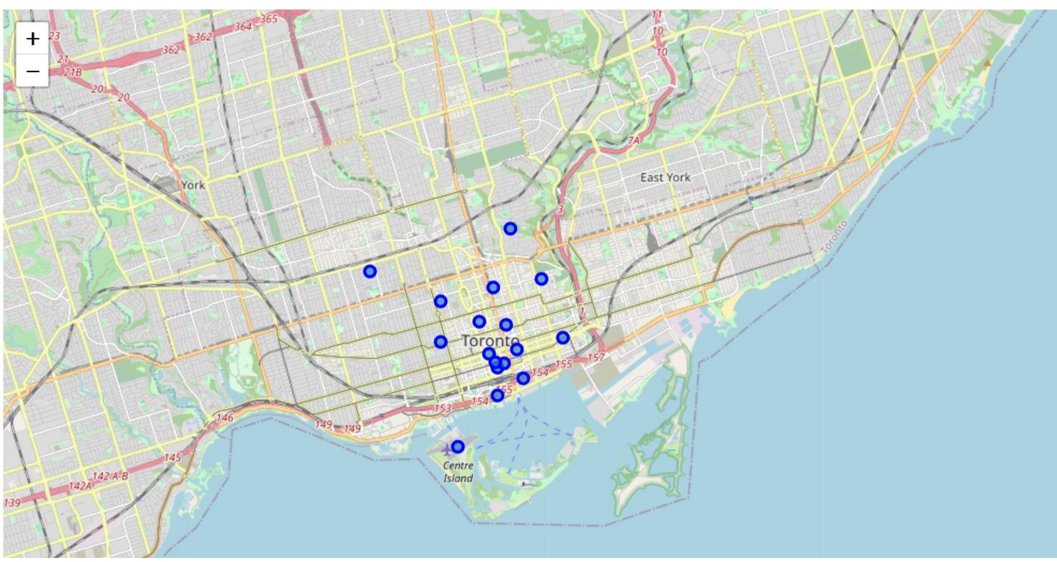

Fig 2: Neighborhoods in downtown Toronto

2.5 FourSquare API

Now that we have selected our candidate's Brough and extracted all the neighborhoods in the area. I used four square API to extract all the venues which deal with the business of restaurant or Indian restaurant or similar category. Thereafter, I created dataframe named "downtown_venues" for all the venues around the Downtown. As our area of concern is only the restaurants in these neighborhoods, I filtered the venue category which has a restaurant.

Out[26]:

	Neighborhood	Neighborhood_Latitude	Neighborhood_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Impact Kitchen	43.65636850543279	-79.356980	Restaurant
1	Regent Park, Harbourfront	43.65426	-79.360636	Souvlaki Express	43.65558391537734	-79.364438	Greek Restaurant
2	Regent Park, Harbourfront	43.65426	-79.360636	Izumi	43.6499697935016	-79.360153	Asian Restaurant
3	Regent Park, Harbourfront	43.65426	-79.360636	Cluny Bistro & Boulangerie	43.650565116074695	-79.357843	French Restaurant
4	Regent Park, Harbourfront	43.65426	-79.360636	El Catrin	43.650600737116996	-79.358920	Mexican Restaurant

Fig 3:Dataset with restaurant category

Furthermore, I identified the Indian restaurants in the neighborhoods from the venue category.

Out[27]:

	Neighborhood	Neighborhood_Latitude	Neighborhood_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	Berczy Park	43.644771	-79.373306	Bindia Indian Bistro	43.64855916613238	-79.371816	Indian Restaurant
1	Central Bay Street	43.657952	-79.387383	Colaba Junction	43.66094	-79.385635	Indian Restaurant
2	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	Indian Roti House	43.63906038875002	-79.385422	Indian Restaurant
3	St. James Town, Cabbagetown	43.667967	-79.367675	Butter Chicken Factory	43.66707247004843	-79.369184	Indian Restaurant
4	Church and Wellesley	43.665860	-79.383160	Kothur Indian Cuisine	43.66787229558206	-79.385659	Indian Restaurant

Fig 4:Dataset with an Indian Restaurant

3. Methodology

In this project, we will focus on detecting the areas near Toronto that have a lower restaurant density, particularly that of Indian restaurants.

In our Data Acquisition step, we have collected the required data with their geographical information for the neighborhoods of Toronto. Afterward, we used FourSquare API to collect all the venues in the data frame. Then, we filtered our venues on the basis of Restaurant which is our main area of concern. Further, we created a separate data frame for Indian Restaurants..

In the Analysis section of this project, we will group every restaurant according to the neighborhood. We will create a separate data frame for each neighborhood restaurant and Indian restaurant. Then, we will visualize our result to check the proximity of these restaurants and check their density within the neighborhoods.

The object of this project is to identify the appropriate location to open an Indian restaurant around Downtown, Toronto. For this purpose, we are required to identify the neighborhood with a higher restaurant density. Also, we need to identify which types of restaurants are distributed around a specific area. Hence, utilizing cluster techniques we will be able to identify the types of restaurant and their density within the neighborhoods. This machine learning approach will allow us to develop a solution for our problem and recommend a suitable venue to the stakeholder.

For this project, we used K-mean clustering techniques to distribute neighborhoods as per our input variables. We planed to cluster our data into 5 different clusters i.e k=5.

3.1 Data Analysis

In this section, I grouped every restaurant according to its neighborhoods and created a separate data frame for each neighborhood restaurant and Indian restaurant. Then, I visualized the result to check the proximity of these restaurants and their density within the neighborhoods.

First, I grouped all the restaurants with their particular neighborhood and found 41 unique categories of restaurants. I saw that the density of restaurant is higher within the central downtown area but that of Indian are very low. Thereafter, I visualize the distribution of these restaurants within Downtown, Toronto using a folium map which is shown in the figure below.

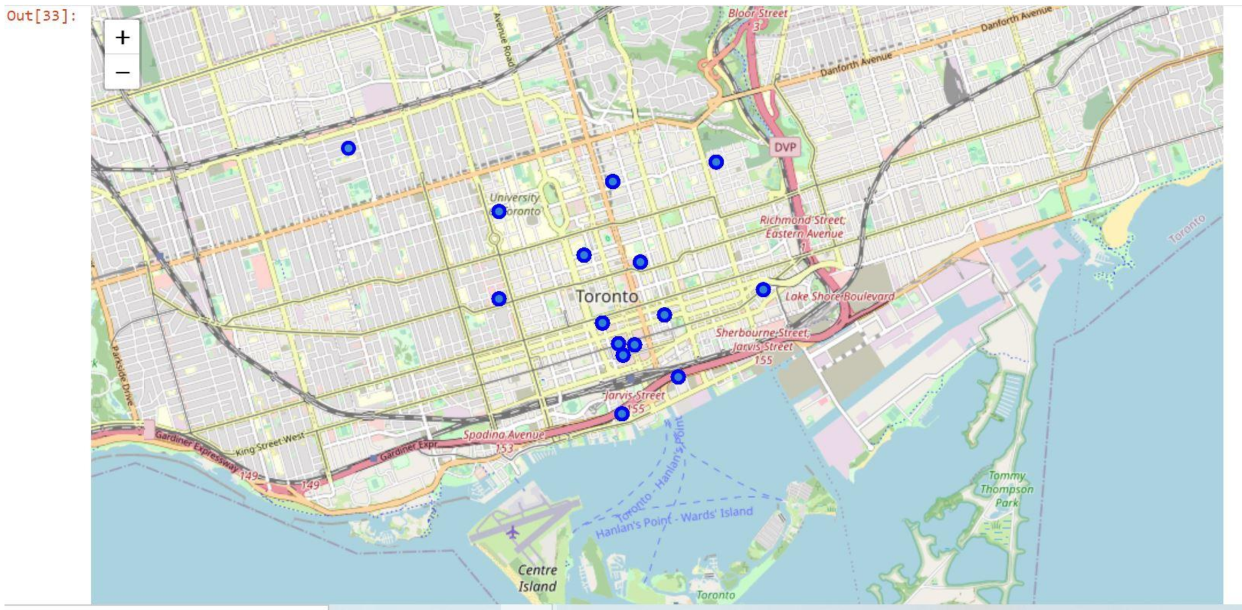

Fig 5:Restaurant distribution in Toronto

3.2 Clustering

The objective of this project was to identify the appropriate location to open an Indian restaurant around Downtown, Toronto. For this purpose, we are required to identify the neighborhood with a higher restaurant density. Also, we need to identify which types of restaurants are distributed around a specific area. Hence, utilizing cluster techniques we will be able to identify the types of restaurant and their density within the neighborhoods. This machine learning approach will allow us to develop a solution for our problem and recommend a suitable venue to the stakeholder.

Before clustering our dataset, I did some statistical testing and data wrangling to prepare our dataset for clustering. First, I change all of our categorical data set into numeric data by assigning 0 or 1 value as per the location of the restaurant within the neighborhood. Then, I normalized the result and group them as per their location. The figure below shows the normalized value for the grouped dataset.

Out[35]:

	Neighborhood	Afghan Restaurant	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Caribbean Restaurant	Chinese Restaurant	Colombian Restaurant	Comfort Food Restaurant	Doner Restaurant	Ethi Resta
0	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.00	0.083333	0.000000	0.000
1	Central Bay Street	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.00	0.000000	0.000000	0.000
2	Christie	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.00	0.000000	0.000000	0.000
3	Church and Wellesley	0.037037	0.037037	0.000000	0.000000	0.000000	0.0	0.000000	0.00	0.000000	0.000000	0.037
4	Commerce Court, Victoria Hotel	0.000000	0.074074	0.111111	0.000000	0.000000	0.0	0.000000	0.00	0.000000	0.000000	0.000
5	First Canadian Place, Underground city	0.000000	0.086957	0.130435	0.000000	0.043478	0.0	0.043478	0.00	0.000000	0.000000	0.000

Fig 6: Normalized dataset as per location

Similarly, I selected the top restaurant in those neighborhoods which helped me to segment the neighborhoods according to the dominant restaurant type.

For this project, I used K-mean clustering techniques to distribute neighborhoods as per our input variables. I planed to cluster our data into 5 different clusters which means by selecting K=5. I generated the array of K-means label and assigned a label to each cluster by merging it with a dataset which contains top 10 restaurants. Finally, I used a Folium map to visualize the result of our cluster. The map is shown in the figure below:

Out[40]:

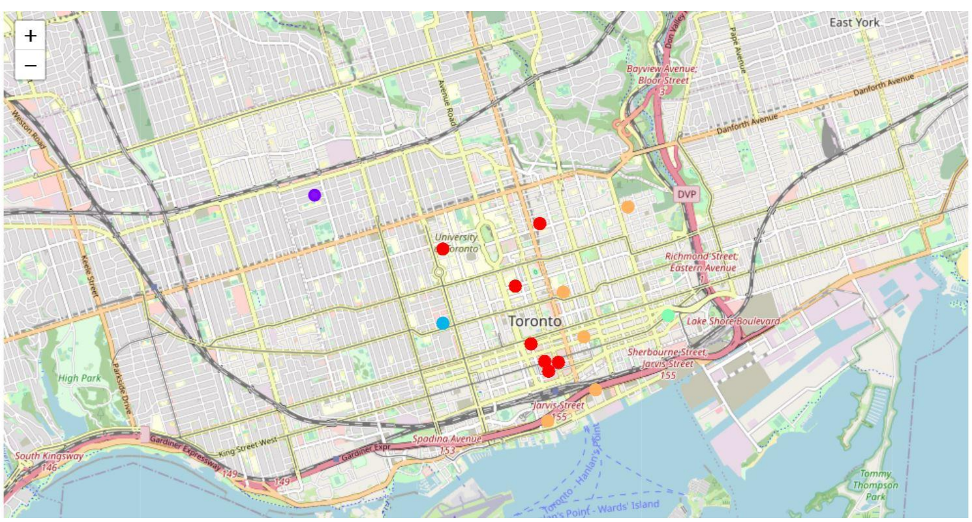

Fig 7: Clustered restaurant within Toronto’s neighborhood

The plot shows that most of the restaurants were clustered within a central downtown area which includes Indian restaurants too. But, most of the popular restaurant was Japanese restaurant which was not a good sign for the person planning to open Indian restaurant within the area. So to determine the best location which has lower restaurant density and suitability for an Indian restaurant, I created a dataset for each cluster. After analyzing each cluster, I found that clusters 2 and 3 have lower restaurant densities with mixed types of restaurants which could be feasible locations. Further analysis and the final decision are shown in the result section. The figure below shows clusters 1 and 3.

Out[42]:

tude	Venue_Category	Cluster Labels	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
	Italian Restaurant	1	Italian Restaurant	Restaurant	Afghan Restaurant	Portuguese Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Molecular Gastronomy Restaurant	Moroccan Restaurant	New American Restaurant
	Restaurant	1	Italian Restaurant	Restaurant	Afghan Restaurant	Portuguese Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Molecular Gastronomy Restaurant	Moroccan Restaurant	New American Restaurant

Fig 8: Cluster 1 with lowest restaurant density

Out[44]:

tude	Venue_Category	Cluster Labels	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
	Restaurant	3	Mexican Restaurant	Asian Restaurant	Greek Restaurant	Restaurant	French Restaurant	Portuguese Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Molecular Gastronomy Restaurant	Moroccan Restaurant
	Greek Restaurant	3	Mexican Restaurant	Asian Restaurant	Greek Restaurant	Restaurant	French Restaurant	Portuguese Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Molecular Gastronomy Restaurant	Moroccan Restaurant
	Asian Restaurant	3	Mexican Restaurant	Asian Restaurant	Greek Restaurant	Restaurant	French Restaurant	Portuguese Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Molecular Gastronomy Restaurant	Moroccan Restaurant
	French Restaurant	3	Mexican Restaurant	Asian Restaurant	Greek Restaurant	Restaurant	French Restaurant	Portuguese Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Molecular Gastronomy Restaurant	Moroccan Restaurant
	Mexican Restaurant	3	Mexican Restaurant	Asian Restaurant	Greek Restaurant	Restaurant	French Restaurant	Portuguese Restaurant	Middle Eastern Restaurant	Modern European Restaurant	Molecular Gastronomy Restaurant	Moroccan Restaurant

Fig 9: Cluster 3 with lower and mixed restaurants

4. Result and Discussion

From our analysis, we found that although there is a great number of restaurants in downtown Toronto, the cluster of restaurants is fairly low if we move farther from downtown. From our initial analysis, we found that there were more than 250 restaurants available within the area of interest considering the whole of Toronto downtown.

Higher concentrations of restaurants were found in Custer 0 i.e within Old Toronto, Commerce Court. Similarly, a small number of restaurants were discovered within Cluster 1, which includes neighborhoods like Christie.

From our precious data analysis, there were only a few Indian restaurants clustered around the neighborhood like Central Bay Street, St.James Town, and Berczy Park. But after K-means clustering we came to realized that although the density of Indian restaurants was higher in this area, they were not so popular. Segmentation of restaurants in this cluster shows that the choice of Indian restaurants was very low while the Japanese restaurant was very popular in the neighborhood.

On the other hand, the density of restaurants in other clusters like 1 and 3 was low which shows the possibilities of Indian restaurants in these neighborhoods. Since we have taken very few variables for this analysis we cannot predict to the whole the best location just from this result alone. Although with fewer assumptions and input variables we could suggest Cluster 2 and 3 will be the best location to open a new restaurant as a mixed type of restaurant are found here with low restaurant density.

5. Conclusion

The purpose of this project was to identify the best location around the Toronto downtown for Indian Restaurant in order to solve the problem of the stakeholder who is interested in investing in this sector. First, we identified the restaurant’s location within the neighborhood of Toronto using FourSquare API and further narrow our search by identifying the Indian restaurants within the area. After that, we performed some data analysis and clustered the locations as per the popularity of restaurants within the neighborhoods. Using the K-means cluster; 5 clusters were identified and visualized on the map. Afterward, a data frame for each cluster was created to make our result more understandable.

Although we have recommended the possible location to open an Indian restaurant within the city, the final decision to select the location will be solely done by stakeholders. For this purpose, he may consider the factor like the density of South Asian community within the area, proximity to major roads, prices and so on.