Coursera IBM Data Science Professional Certificate

Capstone Project

The Battle of Neighborhoods

Toronto

Potential areas to open a New Indian Restaurant

Introduction

Toronto is a highly multicultural city and has residents from various nationalities. One of the most important time out for the people is eating out with friends and family. People from different nationalities like to visit the restaurant of their ethnicity as it makes them feel closer to home and their culture, apart from satisfying their taste buds. Hence, restaurant owners and want-to-be owners are always on constant lookout for places where they can set up the restaurants. Toronto has a high Indian diaspora and the people are scattered through out the city. Although, some areas cater well to the diaspora in terms of restaurant offerings, others lack them. This report tries to highlight potential areas where such restaurants can be set up.

Objectives

Location is the biggest investment for a restaurant and the choice can define whether a restaurant will be successful or not. Any investment requires careful evaluation of all variables and decide the optimum scenario. The purpose of this report is to highlight potential areas based on available data to aid in such a decision making. It tries to find under-served areas which can be potential markets.

Limitations

A lot of considerations go into the selection of venue and this report focuses on one of the aspects only. This report also highlights the potential areas and further evaluation requires careful evaluation with other variables. Also, only areas are suggested and not their potential.

Data requirement

Following data is required to analyse the business problem:

- List of neighbourhoods in Toronto. The scope of this report is throughout the Greater Toronto Area, which combines neighbour cities and have great inter-connectivity.
- Latitude and longitudes of the neighborhoods. These co-ordinates will be required to plat the data on the map and also analyse them.
- Venue data, which provides different available amenities. Indian restaurants will be selected among them and then will be clustered to analyse the results.

Sources of Data

- Postal Codes, Borough and Neighborhood Data is available on Wikipedia.
- Latitude and Longitude information can be found out using Geoencoder or downloaded from here. In this report, downloaded data is used as Geoencoder is not very reliable.
- Venue information can be gathered using Foursquare API. A personal Foursquare developer account is used. Although, it gives limited access, it is enough to satisfy the requirements of this report.

Data Link

Toronto Neighborhood Data

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

Toronto Co-ordinates data for various Postal Codes

https://cocl.us/Geospatial_data

Foursquare Developer API

https://api.foursquare.com/

Methodology

The first step is to get the list of neighborhoods in Toronto. There is a page available on Wikipedia pertaining to this. The link has been provided earlier. The Wikipedia page is scraped for data on Postal codes, Borough and neighborhood data using Pandas. Now the data is cleaned. All Boroughs which are not assigned are dropped. Then any row with null value is checked. Once confirmed that there is no null value or not assigned value, all the neighborhoods are grouped using postal codes. They are converted to dataframe.

Only neighborhood names are not enough, coordinates are also required. For this part, latitude and longitude data has been downloaded from the link provided. The data is as per postal codes. The previous dataframe is combined with the new dataframe to form a new dataframe with all neighborhoods with latitude and longitude values.

Then the neighborhood data superimposed over Toronto map is prepared using Folium.

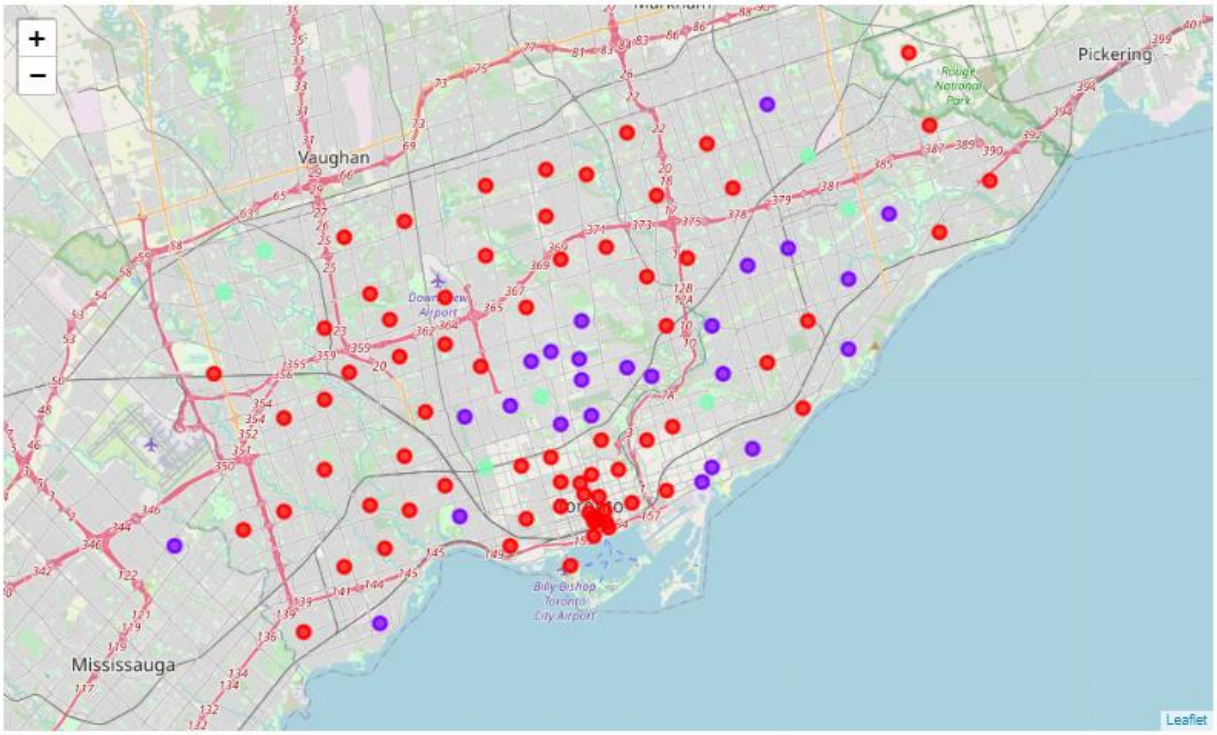
The next part of the analysis involves using Foursquare API to get venues data. Venue data for all the neighborhoods is downloaded with my Client ID and Client Secret. These venues are limited to 100 per neighborhood and within a 3 km radius. It is then converted to a dataframe with neighborhood details and venue details including latitude, longitude and Category. All the different categories of venues are listed to check whether 'Indian Restaurant' is listed or not. Each neighborhood is then grouped with the mean of frequency for each venue. Then data for only 'Indian Restaurant' is populated in a new dataframe.

The last part involves clustering. Clustering was done using K-means algorithm. Initially, a cluster of 5 is defined and all the neighborhoods with different restaurant density if populated. It was found that 2 clusters had similar density in a business sense. Although their density (mean of frequency) was different but the difference was not significant in a practical sense. A cluster of 4 also gave similar results. Finally, a cluster of 3 gave optimum results. The 3 clusters represented areas with high density, medium density and low density. All the 3 clusters were populated. The 3 clusters were then superimposed on the map of Toronto and populated.

Finally, at the end analysis was done as to how the different clusters looked and what inference can be deduced from the clustering of the data.

Results

Below map shows the superimposed clusters on the map of Toronto.



The 3 clusters are:

1. Red – Low density of Indian restaurants
2. Blue – Medium density of Indian restaurants
3. Aqua - High density of Indian restaurants

It can be seen that only a few areas have high density of Indian restaurants. While medium density areas are much higher in number. Low density areas are pretty spread out around Toronto.

Discussion

It can be seen from the results that there are only a few areas where there is a high density of Indian restaurant. Downtown Toronto is a business district and hub of activities. Low density in those areas present a good opportunity. Also, there are some areas such as Brampton and Mississauga where Indian diaspora is present in big number. Low density in these areas present a good opportunity too. Hence, the results provided convey a good deal of information.

However, the results such provided only aid in making business decisions. Further analysis of the low density areas with regards to costs, business volume and other aspects need to evaluated before a final decision is made. But this map provides a good idea which localities to look for a start.

Conclusion

This report highlights the potential areas where a new Indian restaurant can be set up.

Through the analysis, pandas and numpy provide great tools to extract and manipulate data. Folium is a great way to visualise information on a map. Finally, K-means clustering is a fast and efficient method to cluster the neighborhoods. However, iteration is required to find out the optimum number of clusters. In this case 3 clusters seem to be optimum.

Overall, the analysis presents good insights into distribution of Indian restaurants throughout Toronto. It is not only beneficial for setting up new restaurants, but also for people who want to explore different restaurants in an area and hence may look for high density areas.