Introduction

For this capstone project, I have taken a scenario where a businessman wants to start a Indian restaurant in Toronto City. The business would like to find the best area to start this new restaurant such that it has the least competition from other Indian Restaurants in the neighbourhood. I design a project to help achieve this by finding places that are viable and places with have the most competition in the City of Toronto for Indian Restaurants.

Business Problem

The goal is to find options of the best places for a businessman to start a Indian Restaurant. This project should be able to answer the question which places are good options for them to start their business.

Data

The data that is going to be used are as follows

1.List of neighbourhoods using postcodes taken
from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
This contains all the postcodes used in Toronto Region and by categorizing we can get the neighbourhoods for each postal code provided to form a data frame.

2.Geographic data for the respective neighbourhoods which is taken
from https://cocl.us/Geospatial_data
This is a table which contains the geographic data(latitudes and longitudes) for each of the postal codes which can be used to map the neighbourhoods.

3.Foursquare API venue search for Indian Restaurants in Toronto Region. This Data comprises of results of the query for Indian Restaurants in the City of Toronto which can be used to find areas with least competition for the businessman.

Example of one result obtained from foursquare API is as follows


```
{'id': '4b2a634af964a52020a824e3',
   'name': 'Indian Flavour',
   'location': {'address': '123 Dundas St W',
    'crossStreet': 'btw Elizabeth & Bay',
    'lat': 43.65564910619165,
    'lng': -79.38411937886697,
    'labeledLatLngs': [{'label': 'display',
      'lat': 43.65564910619165,
      'lng': -79.38411937886697}],
    'distance': 241,
    'cc': 'CA',
    'city': 'Toronto',
    'state': 'ON',
    'country': 'Canada',
    'formattedAddress': ['123 Dundas St W (btw Elizabeth & Bay)',
     'Toronto ON',
     'Canada']},
   'categories': [{'id': '4bf58dd8d48988d10f941735',
     'name': 'Indian Restaurant',
     'pluralName': 'Indian Restaurants',
     'shortName': 'Indian',
     'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/indian_',
```

'suffix': '.png'},
'primary': True}],
'referralId': 'v-1591074627',
'hasPerk': False}

Target

The main target for this project is the entrepreneurs and businessman who would like to start a Indian restaurants in City of Toronto.

Methodology

First the data from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M which is a Wikipedia page needs to extracted. By using the Wikipedia API for python, I was able to get the html source for the page. Then, by using the pandas read html function, I extracted the data and converted them into a dataframe stored as a csv file.

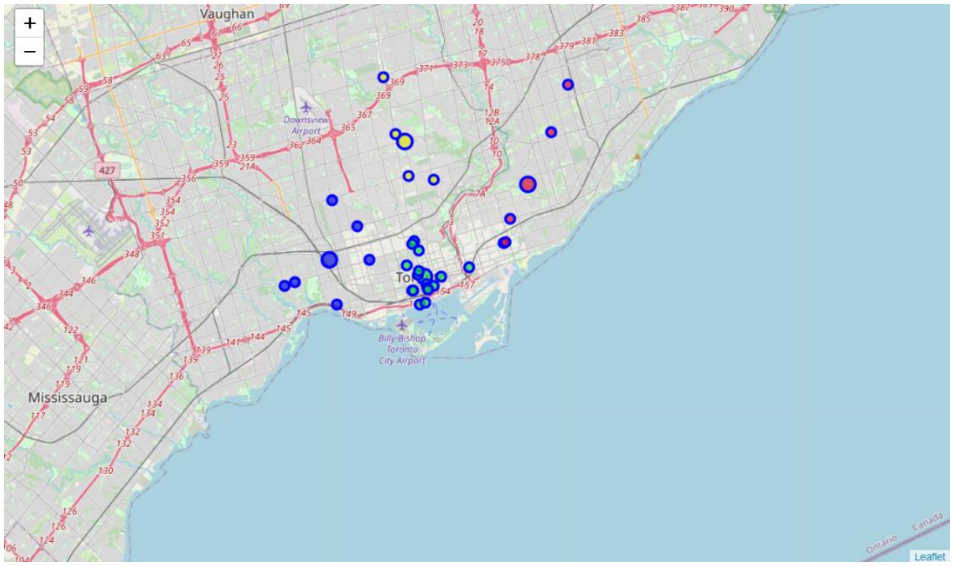
After the data is extracted, the data is cleaned by removing the postal codes which are unassigned to any neighbourhoods. Since the tables do not have the latitudes and longitudes for the neighbourhoods, the data from https://cocl.us/Geospatial_data is used. The latitudes and longitudes from the data is used to get the geographic co-ordinates for all the assigned postal codes of Toronto and a new dataframe is formed.

Using folium the map is generated to visualize the postal codes on map and verify the data. Now, I use the Foursquare API to request all the Indian Restaurants in the Toronto City for a radius of 25 KM from the centre. The requests are obtained as a json file and this json file is converted into a dataframe so that it is easy to perform operations.

This new dataframe with all the data obtained from Foursquare API is cleaned by removing all the unnecessary data. Other data such as restaurants outside the Toronto city and also venues that do not have proper postal codes to keep the accuracy intact.

The dataframe is then passed through a k-means clustering algorithm to put them into 4 clusters so we can find how many restaurants are present in each clusters(where the restaurants are clustered). Finally the centres of the clusters are plotted in a map to visualize the clusters location. Based on the 4 clusters we will be able to determine the best place to open a Indian restaurant with least competition.

Results



We can see 4 cluster centres along with the locations of the Indian restaurants in their respective clusters. The 4 clusters are colour coded with green, red, yellow and blue.

Recommendations

Based on the clustering data, we can say that the green cluster is the most competitive with most number of restaurants within it.
The yellow cluster has the least number of restaurants within, so it would be the ideal neighbourhood for opening the Indian Restaurant.
The other clusters blue and red are in between the blue clusters, (i.e.) while they are not heavily competed like the green clusters, they are also not as least competitive as the yellow clusters.

Limitations

The main limitation for this solution is that it takes into consideration only one factor of number of restaurants present in the neighbourhoods which is then formed into clusters. Other factors such as population density, income of people, etc. could change the solution differently.

Conclusion

In this project, we are able to determine the ideal neighbourhoods to start an Indian restaurant taking only one factor in consideration by performing data extraction, data cleaning , machine learning and data visualization and provide an adequate recommendation to the stakeholders.