IBM - Coursera
Data Science Specialization

Capstone - Final Report

Identifying the suitable location to open an Indian Restaurant in Toronto, Canada

Vinil M. Thombrey

Introduction:

A successful Businessman with more than 50 restaurants in India is planning to expand his business in Toronto, Canada considering there are lot of Asians (Indians) staying there. There are already lot of Indian restaurants in Canada, especially in Toronto and he is not sure where exactly to open. He needs an analysis of the neighborhoods in Toronto and want to understand the type of population in the neighborhood. Thus, as a Data Scientist, I am helping him in analyzing all the neighborhoods in Canada and identify the best location to open an Indian restaurant.

Business Problem:

The objective of this project is to identify the best location to open an Indian restaurant in Toronto, Canada. Thus, with the help of data science methods and machine learning techniques, the project aims in finding the most suitable location to open an Indian restaurant by considering the type of already existing restaurants, type of locality and demographics of that area.

Data:

In this project, I will be scraping the location data from Wikipedia which was explained during the course and then combining this data with the venues related data to Indian restaurants using FourSquare API.

- Get the neighborhood data from the Wikipedia page for Toronto for Latitude and Longitudes of each neighborhood in Json format and use Python to clean and convert it into the pandas dataframe.
- Combine this data with the venues related data for Indian restaurants using the Foursquare API which was introduced in this course.

Methodology:

1. Data Extraction and Cleaning

The first step in the project was to extract the data and get the location data for all the Neighborhoods in Toronto, Canada. This neighborhood Data was obtained by using the web scraping methodologies covered in the module. I have used Beautiful Soup library to extract and read the HTML page **https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M**

After reading the data from the HTML file using the beautiful soup library, I checked the dataset for the null values and there were few Boroughs and Neighborhoods which were null and thus, they were removed from the analysis. This data was then converted to a pandas data frame

After cleaning the data and converting it into the pandas data frame, the location data was extracted for all the neighbourhoods. This lat/long data was joined with the pandas data frame and a new data frame was created

2. Extracting Venues Data using Four Square

Created a developer account on Four square to get the client ID and secrets ID to interact with the Four Square API.

Created a function to get all the nearby venues and the total venues per neighborhood

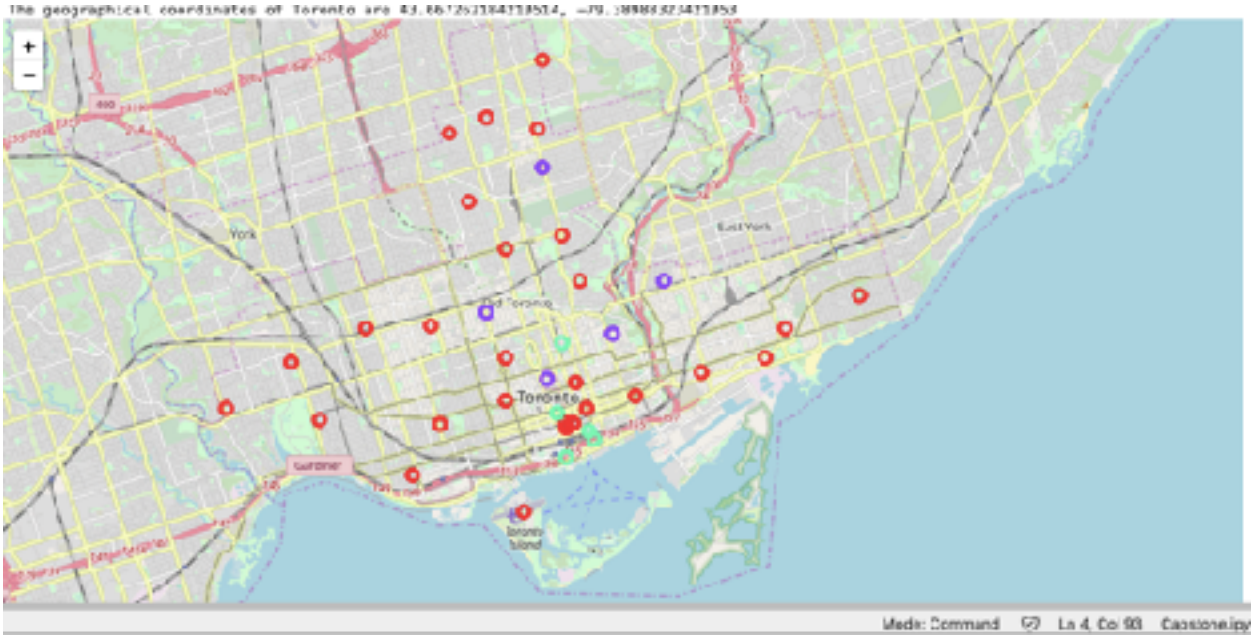
Created a separate data Frame with only Indian restaurants and its locations

3. Analysis

I have used k-means clustering in order to understand similarities and dissimilarities between the clusters.

Also used elbow cure method to find the optimum number of clusters that the data can handle.

A visualization was built using the Folium library and the locations were plotted on the Map in order to better understand the clusters.



Observation and Conclusion:

There are lot of Indian restaurants in Toronto but after the cluster analysis, we see that there are no Indian restaurants in cluster 0 while there are many of them in cluster 1 and cluster 2. Cluster 0 consists of areas like Dominion Centre, St. James street, North Toronto, Davisville and thus these areas lack Indian restaurants.
There are many Asian restaurants in the area which serves India/Chinese/Thai and Malaysian food as well, but lacks authentic Indian restaurant. Thus, there is a scope to have a restaurant in these areas.

Future Work

We can have further analysis about the demographics of that region which will help us in making this decision quicker. Understanding the average age in the area, income and the ethnicity can add value to our analysis but the data for demographics is not publicly available.