Coursera Capstone Project
IBM Data Science Specialization

Battle of the Neighborhoods Part-2

By Ashmitha S S

Introduction:

This is a capstone project for my IBM Data Science Professional Certificate. In this situation I am creating a hypothetical situation where a client wants to open an Indian Restaurant in Toronto, Canada and the data scientist has to provide the best neighborhood to open one in Toronto.

Indian food is one of the most popular cuisines in the world and the client is an up and coming chef who puts a spin on Indian food and wants to make it popular growing the business in that neighborhood. With the purpose of a good location where people will frequent Indian food with some edge , its upto the data scientist to perform analysis and predict a good location to open the restaurant.

Business Problem:

The main objective of this capstone project is for the data scientist to predict the most suitable location for the client to open an Indian Restaurant in Toronto , Canada where people will want Indian food and is a good market to run the business in. With the help of data analysis ,data visualisation and machine learning we should provide the best neighborhood for the client to open an Indian Restaurant.

Target Audience:

The target audience here is the Client, an up and coming chef who wants to open an edgy Indian Restaurant that caters to the palettes of people who are familiar with Indian food and also to those who would be introduced to the food and to grow the business .

Data Requirement:

The data that is required for this project are,

- The wikipedia data of the list of postal codes of all the neighborhoods in Toronto, Canada. This data provides the information about the names of the neighborhoods along with their postal codes. [https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=1011037969]
- The geospatial data that lists the latitudes and the longitudes of the neighborhoods.[https://cocl.us/Geospatial_data]
- The data of the venues in the neighborhood that helps link the Indian Restaurants in all the neighborhoods which is found using Foursquare API.

Data Extraction:

The data is extracted by the means of,

- Web Scraping of wikipedia data with the help of Beautiful Soup.
- Extracting geospatial data using the file and the merging the datasets using data analysis and getting required data without duplicate values obtaining a dataset that has the latitudes and the longitudes of the neighborhoods.
- With the help of the Foursquare API and credentials the venue data is generated to know more about the Indian Restaurants in each neighborhood so it would aid in finding the best neighborhood to open one and the predict it with the help of K-means clustering.

Methodology:

- First all the required packages that are required for implementing the code are said to be imported. Like the pandas package for creating and manipulation of the dataset, numpy so it could handle the data structuring part, Beautiful soup package so that it could perform web scraping and extract all the data from the Wikipedia to get the dataset, a package to also integrate with json so the web application is accessed for the foursquare API, geopy from the geocoders package so we can get the coordinates of the neighborhoods in Toronto , matplotlib to perform the visualisation part of the project with the help of folium so that we can render the data into maps and finally we perform machine

learning with the help of K-means clustering from the sklearn package which is going to cluster the data forming various clusters of neighborhoods that have Indian restaurants in them so that we could look at the resultant data to infer where we can open a good Indian Restaurant.

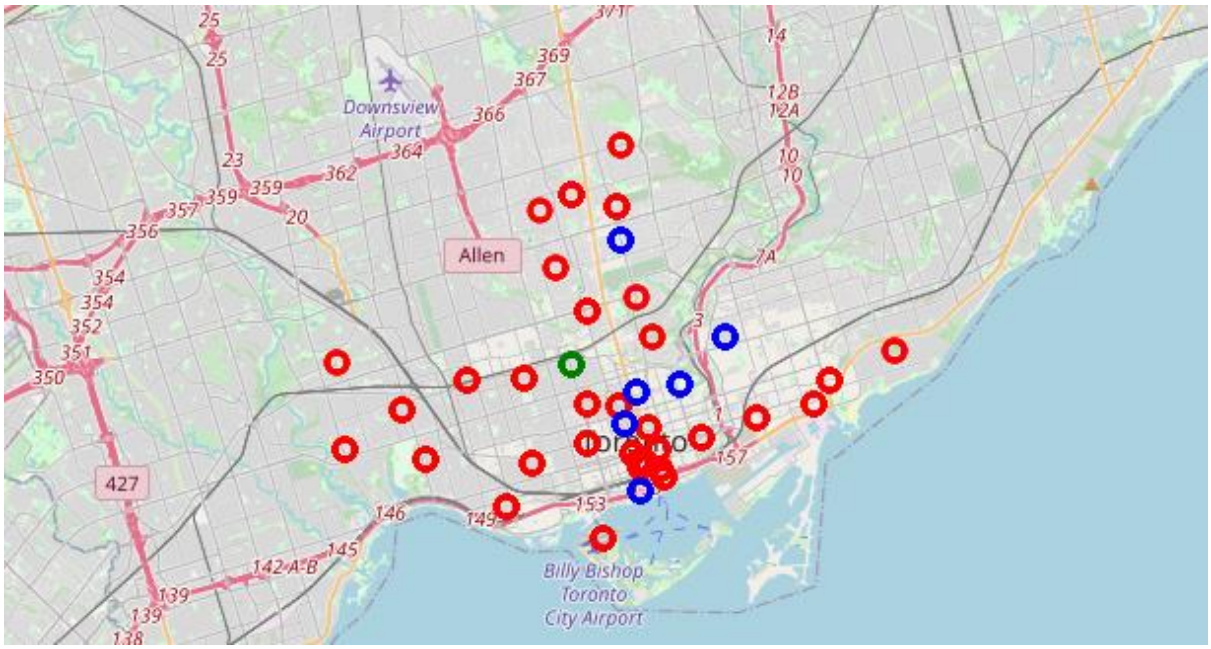
- After importing the packages we extract the data set with the help of beautiful soup and get the Wikipedia data of all the neighborhoods and their postal codes in Toronto[https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=1011037969], with the data we use webscraping and generate the data set and we generate the final dataset after Exploratory data analysis where we eliminate duplicate and null data to obtain the clean data set.
- Once we get that we generate another dataset that generates the coordinates of all the postal codes of the neighborhoods using the geospatial data csv file [https://cocl.us/Geospatial_data] with which once cleaned we merge both the data sets to generate the final dataset which is a table that consists of the postal codes,names and coordinates of the neighborhoods in Toronto.
- With the help of the credentials that consist of an access ID and an API key the Foursquare API is accessed with the help of which venue data is obtained with the coordinates of neighborhoods in the dataset which generates the list of all the venues in the neighborhoods

of Toronto where the top 100 venues within a 500 meter radius of the location.

- After generating a venue dataframe for the neighborhoods we sort out and group them that have Indian Restaurants in them before which we check if there are any Indian Restaurants in the list of venues. This is done when the venues are grouped with each neighborhood. This is to generate **mean of the frequency of occurrence** of Indian Restaurants as the key for each venue category in order to perform clustering .
- Lastly the clustering is performed by the method of **K-means clustering** which is said to identify k number of centroids which them identifies to allocating the data points to the nearest cluster while keeping the centroid as small as possible. This is one of most effective unsupervised machine learning algorithms and is very effective for this project.
- There are as a total of 3 clusters of the neighborhoods in Toronto which are on the basis of the frequency for “Indian Restaurants” on the concentration of the cluster markers are to be mapped. Viewing these clusters we can infer an ideal location for us to open an Indian Restaurant in there.

Result:

The result is a folium map that consists of a cluster data of various Indian restaurants in each cluster that is represented by different color markers on the map.



Here it represents three different clusters of neighborhoods in Toronto where,

- Cluster 0(green markers): Are the neighborhoods that have no Indian Restaurants in them.
- Cluster 1(red markers): Are the cluster of neighborhoods that have the most number of Indian Restaurants
- Cluster 2(blue markers): Are the cluster of neighborhoods that have some Indian Restaurants in them.

Discussion:

From the above results of clusters we can infer that setting up an Indian Restaurant in a neighborhood in cluster 0 would

be the best choice as there are no Indian restaurants there and is a best location for the chef for introduce people to the food there. Also setting up a restaurant in a neighborhood in cluster 2 isn’t a bad option as there are not many restaurants there and could be the next best option. Setting up one in a neighborhood in cluster 1 will not be considered as an ideal choice as there are many Indian Restaurants located there already.

Conclusion:

From here, we can infer that neighborhoods in cluster 1 have the most number of Indian Restaurants followed by neighborhoods in cluster 2 which are the Annex,North Midtown and Yorkville having fewer Indian restaurants. So, it would be the best choice to start an Indian Restaurant in cluster 0 as there aren't any in those neighborhoods which would be a good location to open a new restaurant as it would be new in the area to open an edgy Indian Restaurant and generate better revenue. Thus we come to the conclusion that by K-means clustering we helped predict the most ideal location to open a new Indian Restaurant in Toronto.