Coursera Capstone Project IBM Data Science

Battle of the Neighborhoods Week-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.

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

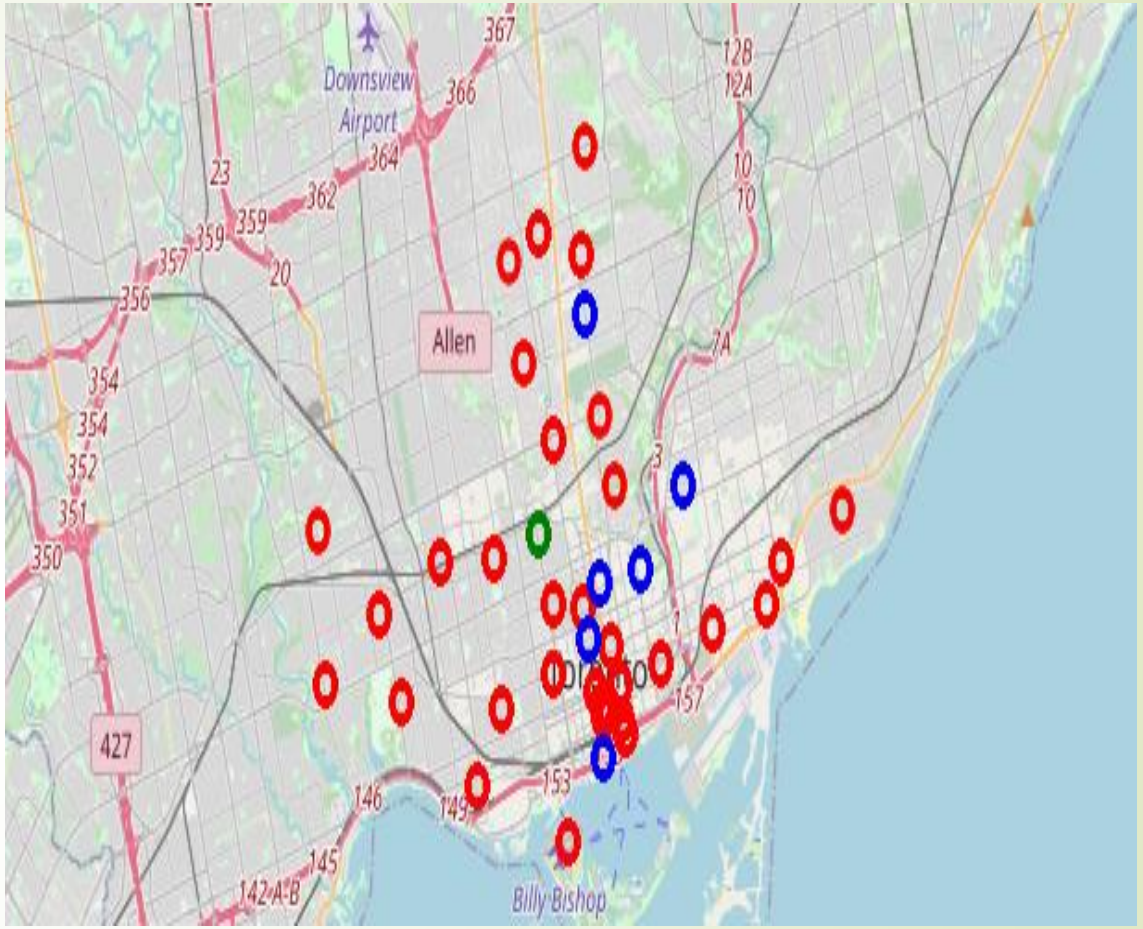
- Installing the required packages and then scraping the data to get the dataset that contains postalcode and neighborhood names.
- Obtaining the coordinates of the neighborhoods with the help of the csv file of the geospatial data and then merging both the data to obtain a data set that has neighborhood names along with their coordinates.
- Using foursquare credentials accessing the API to generate the venue data for all the neighborhoods that limit to 100 venues across a 500m radius.
- Clustering the data on basis of grouping by neighborhoods and setting the key to “Indian Restaurants” with the mean frequency by K-means clustering.
- Generating 3 clusters that provide data about Indian Restaurants in each one of them showcasing in which neighborhood its best to open the restaurant.

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.

Thank You