Coursera Capstone

IBM Applied Data Science Capstone Project

***Recommend places to open a new Indian Restaurant in New Delhi, India.***

From: Piyush Jain

Date: January,2020

1. **Business Problem Description:**

To open a new restaurant, the one of the most important factors which contribute to the success of any restaurant is its location. To determine, if a person wants to open a new restaurant, this project will recommend the places where they need to open it. According to the CNBC report, 45% of restaurant failure is because the owner chose bad location for the restaurant. Poor visibility, no parking, and no foot traffic is a combination that makes it nearly impossible to turn a profit. Picking the right location can make up for some of the other deficits in this list which include food quality, staff friendliness, etc.

1. **Business Requirement Description:**

To analyze and select the best locations in New Delhi, India to open a new Indian restaurant. Picking up the best location includes three important factors:

* Frequency of the Indian Restaurants in the location.
* Population Base
* Accessibility
* Parking Facility

This project is timely as the city is currently suffering from the oversupplying of Indian restaurants.

1. **Data Resources:**

Following are the data resources which define the scope of the project.

* + Excel file from the website <https://simplemaps.com/data/in-cities> which include the Indian Geocodes for the places in Delhi.
  + Geocoder package for latitude and longitude coordinates.
  + We will be using the Foursquare API for this project. The reason for using the Foursquare API is because it has one of the largest databases of 105+ million places and is used widely by many developers too. Foursquare will provide different categories of the venues data and Indian Restaurants is among one of the categories, we are interested in for this project. Foursquare API to get the venue details related to the Indian restaurants.
  + This project will use make use of Data Science skills which we learnt under this professional certification program like Reporting, Python, Data Analyzing, Data Cleaning, Data Wrangling, Project Methodology, Machine Learning Skills.

1. **Project Methodology**

The first thing we need the list of places in New Delhi. Luckily, I got an excel from the mentioned web site <https://simplemaps.com/data/in-cities>. This includes all the places from India. We now get the name of the locations in Delhi. Now, we need to apply the data cleansing and pre-processing steps. The first step is to extract all the places belong to Delhi and store in a new data frame. Once, we store the information, we need to have the longitude and latitude of the places. We will be using the **Geocoder package** which converts the address to longitude and latitude. Once, we get the coordinates,

we need to populate it on the map using the **Folium Maps.**

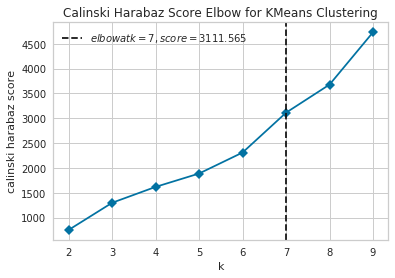
By doing this exercise, we will make sure that the geographical coordinates belong to Delhi only. The next step is to use the Foursquare API to get the top 100 venue data within the 1000 mts radius. To use the Foursquare API, we need to first register under the Foursquare developer account. Once you registered, you will be issued a client ID and password. You need to use these Foursquare credentials to call the Foursquare API. Foursquare will return the venue data in JSON file and then we need to extract the venue details, venue categories, venue latitude and longitude. We will then analyze all the locations by grouping neighborhood and taking the mean of the frequency of occurrence of each venue category.

Filter the venue category by the Indian Restaurants which includes

* South Indian Restaurant
* Indian Restaurant
* North Indian Restaurant
* Indian Chinese Restaurant
* Multicuisine Indian Restaurant
* Northeast Indian Restaurant

After the completion of the data analysis, we need to apply the Machine Learning technique. In this project, we are using the K means clustering on the data which identifies the K number of centroids and then allocates every data point to the nearest cluster. The K means is one of the popular and simplest unsupervised machine learning algorithms. Before applying the K means algorithm, we must know what the optimum value of K. There are quite few methods available to know the optimum value of K and we will be using the Elbow method to know the value of K suited for the K means algorithm.

We will be using the Yellow brick function to visualize the elbow graph. After applying the Elbow method, the optimum value of K come as 7.



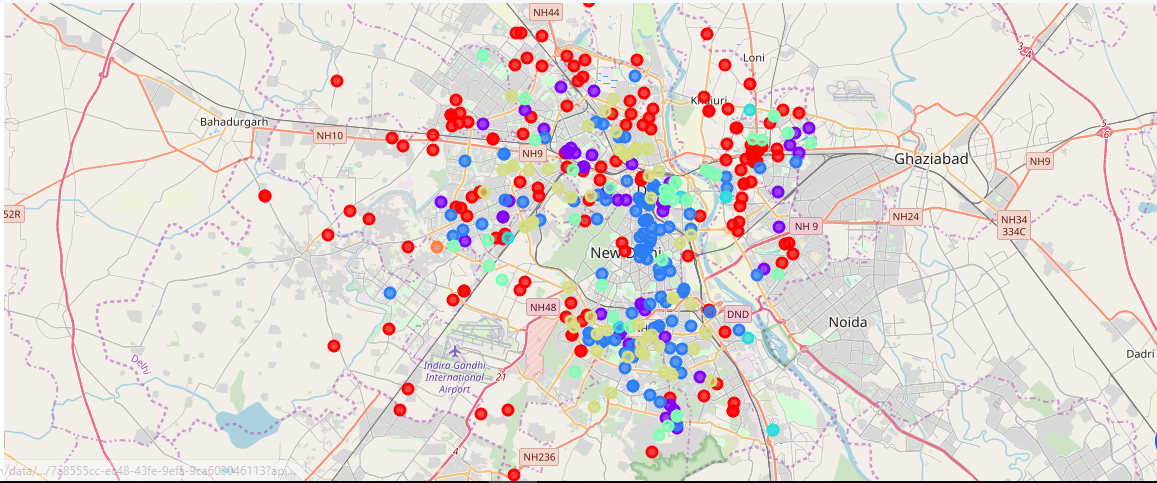
Apply the Machine Learning Technique using K means algorithm. We will cluster the neighborhoods in 7 clusters based on the ‘Indian Restaurants’. The result will allow us to identify which cluster have the higher concentration of the Indian restaurants and which one have moderate to less concentration of restaurants. Based on the occurrence of the Indian restaurants in different neighborhoods, it will help us to identify the places to open a new Indian Restaurant. Visualize the clusters in the map using the Folium Maps.

1. **Results of Project:**

The results from the K-means clustering shows that Cluster 0 have the least concentration of Indian restaurants

1. Clusters with least number of Indian Restaurants – Cluster 0
2. Clusters with moderate number of Indian Restaurants – Clusters 1, 2 and 5
3. Clusters with high number of Indian Restaurants – Clusters 3, 4 and 6

The result of the clusters is visualized in the map below with cluster 0 is red color, cluster 1 is in purple, cluster 2 is in blue color, cluster 3 is in brown color, cluster 4 is in dark blue color, cluster 5 is in mint color, cluster 6 is in crystal color.



1. **Discussions:**

As observed from the map, most of the Indian restaurants are in the South and Central Delhi with the highest number in clusters 3,4 and 6. On the other hand, cluster 0 has very low numbers for Indian restaurants. This represents a good opportunity to big food chains or the persons thinking to open a new Indian Restaurant in the areas under cluster 0 as there will be a less competition compare to other areas. The people avoid the areas mentioned in the clusters 3,4 and 6 which already have high competition among them.

1. **Recommendations**

In this project, we consider only one factor i.e. frequency of the occurrence of the Indian Restaurants. There are some other important factors as well, like population base, income of residents, parking facility which could influence the location decision to open a new Indian Restaurant. Future research could use of these factors too to make more accurate recommendation of the location to open a restaurant.

1. **Conclusion:**

In this project, we have gone through of identifying the business problem, specifying the data resource, prepare the data and perform the machine learning algorithms to recommend the best places to open a new Indian restaurant. To answer the question raised in the Business problem is to recommend the places in the Cluster 0 which have quite less concentration of the restaurants and with very less competition. Avoid the places in clusters 3, 4 and 6 as there are over supply of the restaurant with high competition.

1. **Appendix:**

Snippet of the places recommend opening a new Indian Restaurant.

