COURSERA CAPSTONE PROJECT REPORT

IBM APPLIED DATA SCIENCE CAPSTONE

Opening a new Café in Delhi, India

Prepared By: Harshit Singhal

Place: India

INTRODUCTION

We all are busy in work and hectic schedule of life. Taking out time to have some fun time enjoy the life we find places to go out have some great time. Café these days have become one of the most common place to hang out with friends and family. A **café** is a type of restaurant which typically serves coffee and tea, in addition to light refreshments such as baked goods or snacks. The term "café" comes from the French word meaning "coffee".

Nowadays café are basically theme based come up with some fantastic concepts. In the city like Delhi having a population size of near about 1.9 crores (as of 2012) it provide a great opportunity to open a new café in such great city. But the question that stuck is “where to open the open”, where the is full with some amazing café.

Business Problem

The objective of this capstone project is to analyse and select the best locations that are available in the Delhi, India to open a new café. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question:- If someone who is looking to open to open a café in Delhi, India, what would be some best location available?

Audience Interested in Project

All the builders, businessman, café chain owners or some great future business tycoon or some entrepreneurs would be interested in knowing the location in the city like Delhi, India which is densely populated and have a great scope to open café and flourish them in near future. Also all the people who are in food industry might also be interested in knowing location.

People would come to this project because many people have land and resources to have business but are not aware which business is best suited for that locality. So such people would be helped a lot.

DATA

Data Description

In order to solve the problem, we will be needing the following data:

- List of neighbourhoods in Delhi. This defines the scope of this project which is confined to the city of Delhi, the capital city of the country India in Asia.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to explore and get the venue data of the neighbourhoods.
- Venue data such as venue category or type and number of all such categories, particularly data related to café. We will use this data to perform clustering on the neighbourhoods.

Data Source

We will be taking the neighbourhoods of Delhi data available on the Wikipedia.

Link: https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi

We will be using the requests library of python in order to scrap the data and beautifulsoap to perform the parsing. For the sake of geographical coordinates of the neighbourhoods we will be using the geopy library.

For the the purpose of exploring the neighbourhoods we will be using the foursquare API. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.

METHODOLOGY

Firstly we have taken the data of the neighbourhoods of Delhi,Ibdia from the Wikipedia. There were about 185 neighbourhoods available on the website. We used the requests and beaauifulsoap libraries of the python for the purpose of data scraping. We gathered the data and put the data into the dataframe.

After that we needed the geographical coordinates of every neighbourhood so we used the geopy library to get the data. We merged the to dataframe and get the dataset ready.

We further used the folium library to plot the map of Delhi,India and superimpose the neighbourhoods data on them inorder to get the visual idea of the locations.

After that we make use of the foursquare API in order to explore each every neighbourhood and find out the trending places in that locality. Before using foursquare API we need to register on its website and setup a developer account in order to get the API credentials. From foursquare we limit it to explore locality with maximum of 200 places in radius of 2 km. After that we got a dataset of about 7101 venues.

We found 214 unique category among 7101 and them we created the table by grouping the based on the neighbour and take the mean of the frequency of occourence.

We finally made a dataframe by taking café as a feature and then we used K-means algorithm (a machine learning algorithm) in order to perform clustering of data in 3 segments.

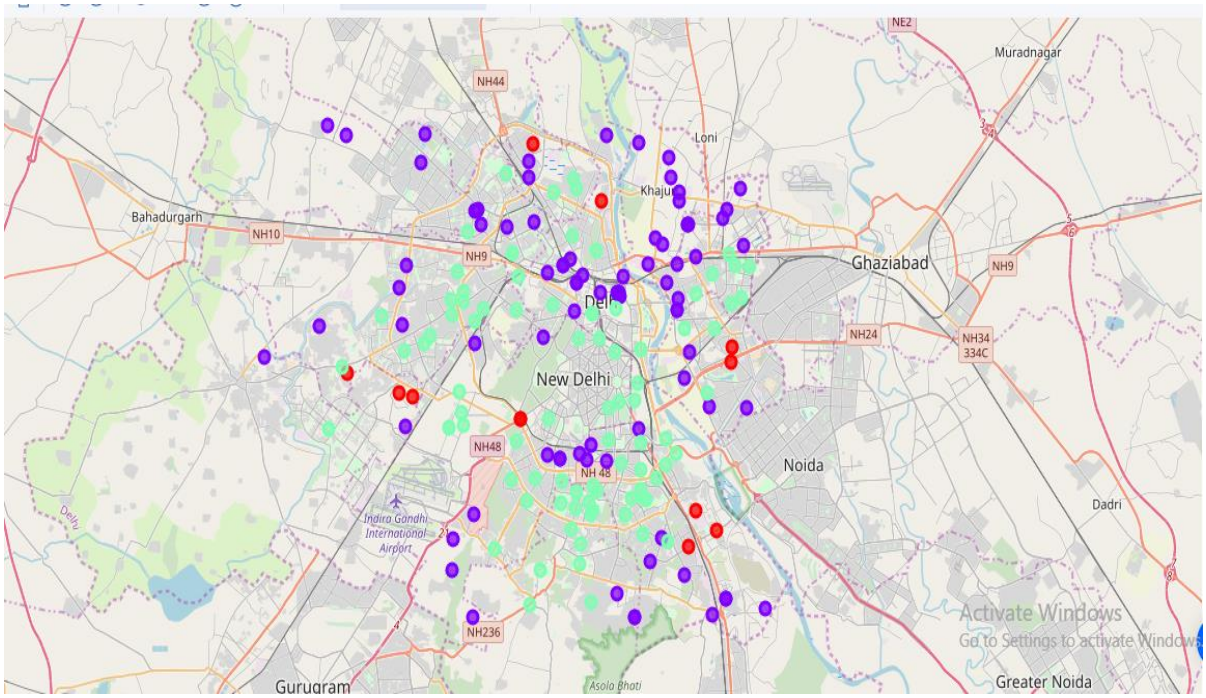
Finally we got results where every neighbourhood was clustered in these segments and presented the result on map with folium library.

RESULTS

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Café":

- Cluster 0: Neighbourhoods with maximum number of café
- Cluster 1: Neighbourhoods with low number to no existence of café
- Cluster 2: Neighbourhoods with moderate concentration of café

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



DISCUSSION

From the above results we can see that the locality having high number of café are located at very faraway places. Also it can be seen that the locality in cluster 0 have large number of café available so opening a café in such place would lead you to survive a very tough competition resulting in less profit.

Cluster 2 have moderate number of café so opening a café in such location would be beneficial if you have a uniqueness in your business model or in the interior/decoration of café.

Cluster 3 have low to negligible café so opening café in such location would lead to fast growth and neither you have to go through high competition of market.

Future Scope

While doing this project only the frequency of café in locality was taken in consideration for clustering. Many other factor such as population of locality, average age factor and income factors could also be taken in consideration to train our model better but for now such data are not available. Also we limited the venue per locality to 100 and radius to 2 km these factor could also be changed to see results.

CONCLUSION

For this project we have gone through the business problem and have solved it by going through the process of data collection, data cleaning, data analysis, data preparation, data visualization, choosing and training the model and finally giving the results by categorising each location in three different cluster based on the frequency of café in the locality. Now answering the question it is best to open the café in locations which fall under cluster 1 as these location have minimal or no café in their locality.