**COURSERA CAPSTONE PROJECT**

**THE BATTLE OF THE NEIGHBOURHOODS-REPORT**

**Opening a new Indian Sweet Shop**

**in Delhi, India**

**Prepared by:** Akshat Kumar

**Place:** India

**INTRODUCTION**

When it comes to Indian Cuisine and food one thing cannot be overlooked is Indian’s love for Sweets. One can get a beautiful scene of colourful sweets in any Indian sweet shop. It’s not uncommon to see huge crowds at Sweet stores across the Country. And with the variety and sheer number of sweets available it's no wonder that it's such an important part of an Indian's lives.

Sweets are part of any Indian celebration or festivity of any kind. Sweets are a symbol of good gesture in India.

In the city like Delhi having a population of 1.9 crores (till 2012) it provides a great opportunity to open a new Indian sweet shop. But the question arises “where to open the new sweet shop”, so that we can get the maximum profits or benefits out of it.

**Business Problem**

The objective of this capstone project is to analyse and select the best locations that are available in Delhi, India to open a new Indian Sweet Shop .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 a Sweet Shop in Delhi, India, what would be some best location available?

**Audience Interested in Project**

Sweet Shop chain owners, Businessman and 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 a Sweet Shop and flourish them in near future. Also, all the people who are in food industry might also be interested in knowing the location.

**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.

• 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 sweet shop. 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, India from the Wikipedia. There were about 185 neighbourhoods available on the website. We used the requests and beauifulsoap 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 data to dataframe and get the dataset ready.

We further used the folium library to plot the map of Delhi, India and superimpose the neighbourhood’s data on them in order 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 occurrence.

We finally made a dataframe by taking Indian Sweet Shop as a feature and then we used Kmeans 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**

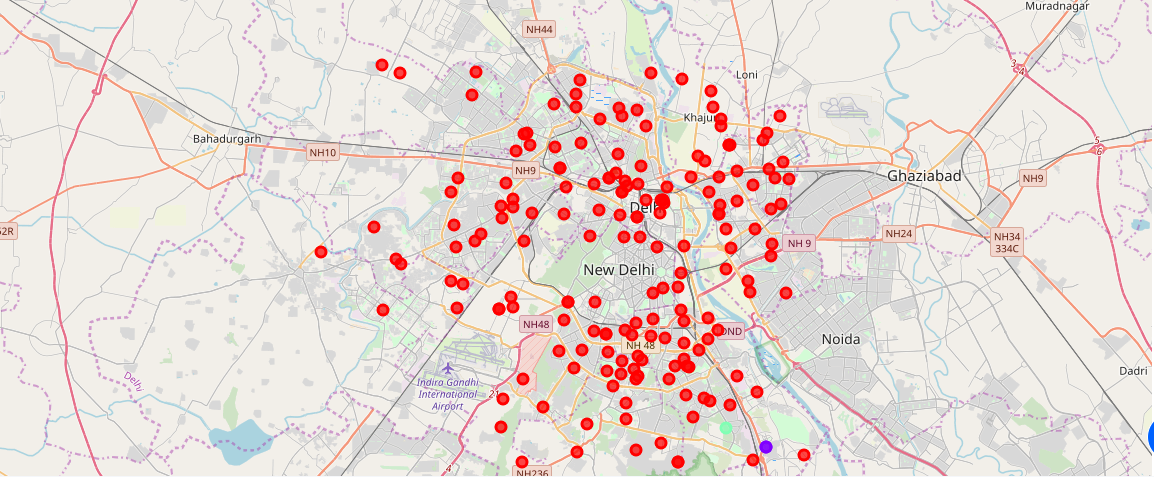
The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Indian Sweet Shop”:

• Cluster 0: Neighbourhoods with maximum number of sweet shops

• Cluster 1: Neighbourhoods with low number of sweet shops

• Cluster 2: Neighbourhoods with low to no concentration of sweet shops

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 sweet shops are located at very faraway places. Also it can be seen that the locality in cluster 0 have large number of sweet shops available so opening a sweet shop in such place would lead you to survive a very tough competition resulting in less profit.

Cluster 1 and 2 have low to negligible sweet shops so opening sweet shop 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 sweet shop in locality was taken in consideration for clustering. Many other factors 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 factors 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 sweet shop in the locality. Now answering the question, it is best to open the sweet shop in locations which fall under cluster 1 and 2 as these locations have minimal or no sweet shops in their locality.