**Coursera Capstone**

**IBM Applied Data Science Capstone**

***Opening a new Bakery in Delhi, India***

By: Vaishnavi Gupta

July 2020

1. **INTRODUCTION**

For many people, bakeries are a good place to get the best and exotic delicacies from. They can have various food items that the bakeries provide and satisfies everyone’s sweet tooth. With the improving lifestyle of people, many prefer to opt for good bakeries to satisfy their tummies and with the arrival of new delicacies and sweets from across the world, opening a bakery store has its own brilliant financial returns.

Baking enthusiasts are keener than ever to tap this growing market that has more demand now than ever before. As a result, there are many bakeries in the city of Delhi and many more are being built. Opening bakeries allows property developers to earn rental income as well.

Of course, as with any business decision, opening a new bakery also requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the bakery is one of the most important decisions that will determine whether it will be a success or a failure.

**Business Problem**

The objective of this capstone project is to analyse and select the best locations in the city of Delhi, India to open a new bakery. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question that where we should open a new bakery.

**Target Audience of this project**

This project is particularly useful to help baking enthusiasts who are willing to take further steps in establishing their bakery. As this industry is ever growing and doing well now more than ever, hence this project is timely.

Now that more and more people are looking forward to taking up something that aligns with their passion, this project becomes even more important.

1. **DATA**

**To solve the problem, we will need 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 of India.
* Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
* Venue data, particularly data related to bakeries. We will use this data to perform clustering on the neighbourhoods.

**Sources of data and methods to extract them**

This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods\_in\_Delhi) contains a list of neighbourhoods in Delhi, with a total of 138 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

1. **METHODOLOGY**

Firstly, we need to get the list of neighbourhoods in the city of Delhi. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods\_in\_Delhi). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Delhi.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Bakery” data, we will filter the “Bakery” as venue category for the neighbourhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Bakery”. The results will allow us to identify which neighbourhoods have higher concentration of bakeries while which neighbourhoods have fewer number of bakeries. Based on the occurrence of bakeries in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new bakeries.

1. **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 “Bakery”:

* Cluster 0: Neighbourhoods with moderate number of bakeries
* Cluster 1: Neighbourhoods with low number to no existence of bakeries
* Cluster 2: Neighbourhoods with high concentration of bakeries

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.

A close up of a map

Description automatically generated

1. **DISCUSSION**

Most of the bakeries are concentrated in the central area of Delhi, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number of bakeries in the neighborhoods. This represents a great opportunity and high potential areas to open new bakeries as there is very little to no competition from existing ones.

Meanwhile, bakeries in cluster 2 are likely suffering from intense competition due to high concentration of bakeries.

Therefore, this project recommends bakery owners to capitalize on these findings to open new bakeries in neighborhoods in cluster 1 with little to no competition.

Bakeries can also stand out from the competition if they have new items to offer in neighborhoods in cluster 0 with moderate competition.

Lastly, developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of bakeries and suffering from intense competition.

**Limitations and Suggestions for Future Research**

In this project, we only consider one factor i.e. frequency of occurrence of bakeries, there are other factors such as population, culture and income of residents that could influence the location decision of a new bakery. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project.

Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new bakery. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

1. **CONCLUSION**

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. bakery enthusiasts to open a new bakery.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new bakery. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new bakery.

1. **REFERENCES**

Category: Neighbourhoods in Delhi. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Category:Neighbourhoods\_in\_Delhi

Foursquare Developers Documentation. *Foursquare*. Retrieved from https://developer.foursquare.com/docs