Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Bakery in Delhi, India

By: Anurag Kandalkar

May 2020

Introduction

People of all ages are affectionate of different bakery products, because of their taste, color and easy to digest nature. They eat and serve different bakery products in their parties and festivals. Celebrating any moment of happiness is incomplete with bakery products.Bakery products are becoming prominent day by day. They are very popular because of its taste and simple to digest. Bakery items are usually loved by all. Nowadays individuals have virtually no time to invest much on making breakfast it is the bread and bun or biscuits which had occurred instead of other sorts of stuff.. As a result, there are many Bakery in the city of Delhi and many more are being built. Opening Bakery allows Bakers to earn consistent income. Particularly, the location of the Bakery is one of the most important decisions that will determine whether the mall 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: In the city of Delhi, India, if a Baker is looking to open a new Bakery, where would you recommend that they open it?

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 in Asia.
- 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 Bakery. 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/Neighbourhoods_of_Delhi) contains a list of neighbourhoods in Delhi, with a total of 174 neighbourhoods. We will use Json file mentioning Borough and neighbourhood. 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.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Bakery category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Delhi. Fortunately, the list is available in theJSON file. 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 wonderful 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.

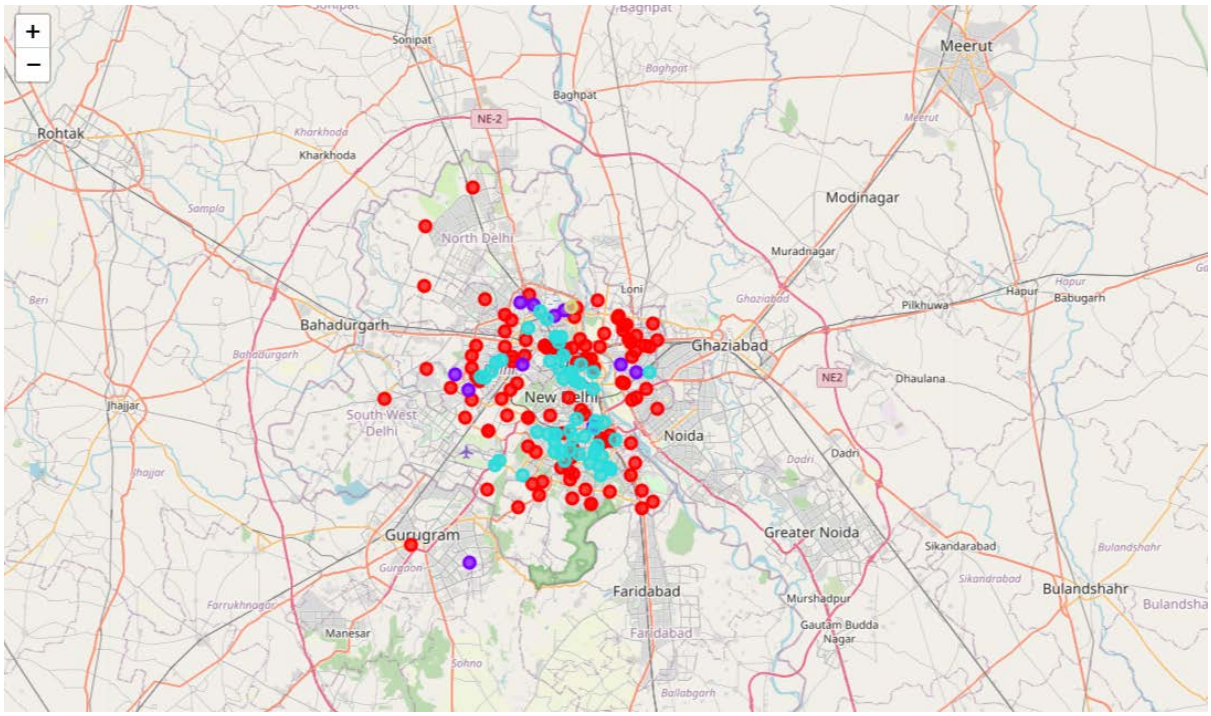
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 4 clusters based on their frequency of occurrence for “Bakery”. The results will allow us to identify which neighbourhoods have higher concentration of Bakery while which neighbourhoods have fewer number of Bakery. Based on the occurrence of Bakery in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Bakery.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 4 clusters based on the frequency of occurrence for “Bakery”:

- Cluster 0: Neighbourhoods with low number to no existence of Bakery
- Cluster 1: Neighbourhoods with moderate number of Bakery
- Cluster 2: Neighbourhoods with high concentration of Bakery
- Cluster 3: Neighbourhoods with very high concentration of Bakery
-

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, cluster 2 in mint blue colour and cluster 3 in light yellow colour



Discussion

As observations noted from the map in the Results section, most of the Bakery are concentrated in the central area of Delhi city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no Bakery in the neighbourhoods. This represents a great opportunity and high potential areas to open new Bakery as there is very little to no competition from existing malls. Meanwhile, Bakery in cluster 3 are likely suffering from intense competition due to oversupply and high concentration of Bakery. From another perspective, the results also show that the oversupply of Bakery mostly happened in the central area of the city, with the suburb area still have very few Bakery. Therefore, this project recommends Baker to capitalize on these findings to open new Bakery in neighbourhoods in cluster 0 with little to no competition. Baker with unique selling propositions to stand out from the competition can also open new Bakery in neighbourhoods in cluster 1 with moderate competition. Lastly, Baker are advised to avoid neighbourhoods in cluster 2 andcluster 3 which already have high concentration of Bakery 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 Bakery, there are other factors such as population 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.

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 4 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e.Bakers regarding the best locations 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 0 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.

References

Category:Neighbourhoods of Delhi. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi

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

Appendix

Cluster 0

- Bangsar South
- Bukit Bintang
- Bukit Nanas
- Bukit Tunku
- Chow Kit
- Damansara Heights
- Damansara Town Centre
- Damansara, Delhi
- Dang Wangi
- Jalan Cochrane, Delhi
- Jalan Duta
- Kampung Baru, Delhi
- Medan Tuanku
- Mont Kiara
- Segambut
- Setiawangsa
- Shamelin
- Taman Desa
- Taman Tun Dr Ismail

Cluster 1

- Alam Damai
- Ampang, DelhiBandar Menjalara
- Bandar Sri Permaisuri
- Bandar Tasik Selatan
- Bandar Tun Razak
- Batu 11 Cheras
- Batu, Delhi
- Bukit Jalil
- Bukit Kiara
- Bukit Petaling
- Cheras, Delhi
- Desa Petaling
- Federal Hill, Kuala Lumpur
- Happy Garden
- Jinjang
- Kampung Datuk Keramat
- Kepong
- Kuchai Lama
- Maluri
- Miharja
- Pantai Dalam
- Putrajaya
- Salak South
- Semarak
- Sentul Raya
- Setapak
- Sri Hartamas
- Sri Petaling
- Sungai Besi
- Taman Bukit Maluri
- Taman Cheras Hartamas
- Taman Connaught
- Taman Ibukota
- Taman Len Seng
- Taman Melati
- Taman Midah
- Taman OUG
- Taman P. Ramlee
- Taman Sri Sinar
- Taman Taynton View
- Taman Wahyu
- Titiwangsa
- Wangsa Maju

Cluster 2

- Bangsar
- Bangsar Park
- Brickfields
- KL Eco City
- Lembah Pantai
- Pudu, Delhi
- Taman U-Thant