2020

CAPSTONE PROJECT

The Battle of the Neighborhoods

FARAH PANKHANIA

Table of Contents

Opening a new Cafe in Mumbai, India

1. Introduction

1.1. Background

Coffee drinking is a big part of Indian food culture. For centuries past, coffee shops have played a big part in shaping communities by bringing people together and affording the space for a community to ‘be a community’.

When you invest in a coffee shop franchise, you are essentially creating a location for your neighbours to meet and socialize, thus playing a crucial part in building a strong and close-knit community. They prove to be a form of a relaxation, entertainment and a hangout place for people of all ages. Thus, local budding entrepreneurs look forward to setting up and investing in a cafe shop.

In order for a cafe to profitable, there should be enough customers and to get enough customers, it is not beneficial to set up a cafe in the vicinity of the existing ones.

Also, the location of a cafe has a significant impact on the expected returns.

1.2. Business Problem

The main objective of this Capstone Project is to analyse and select the best location in Mumbai city to open a cafe. We will be using data science methodology and machine learning techniques to answer the following business problem- If a local entrepreneur/ investor wishes to open a coffee shop in the city of Mumbai, India; where would you recommend that they open it?

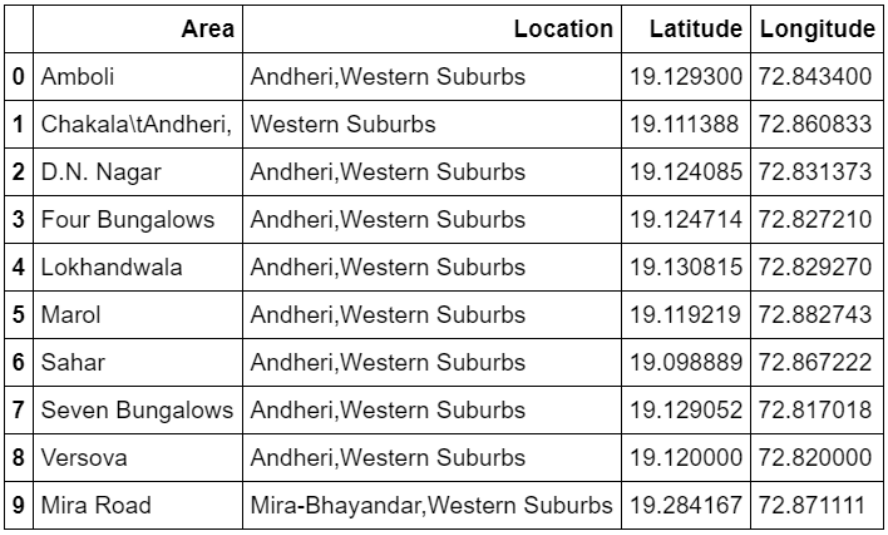
1.3. Target Audience

Mumbai is the financial, commercial and entertainment capital of India. It is one of the biggest cities in India with a large population. Opening a cafe here would be beneficial to anyone who is a local entrepreneur/investors or who is looking forward to expand their businesses.

2. Data and Sources

To solve the given Business Problem- we would need the following data:

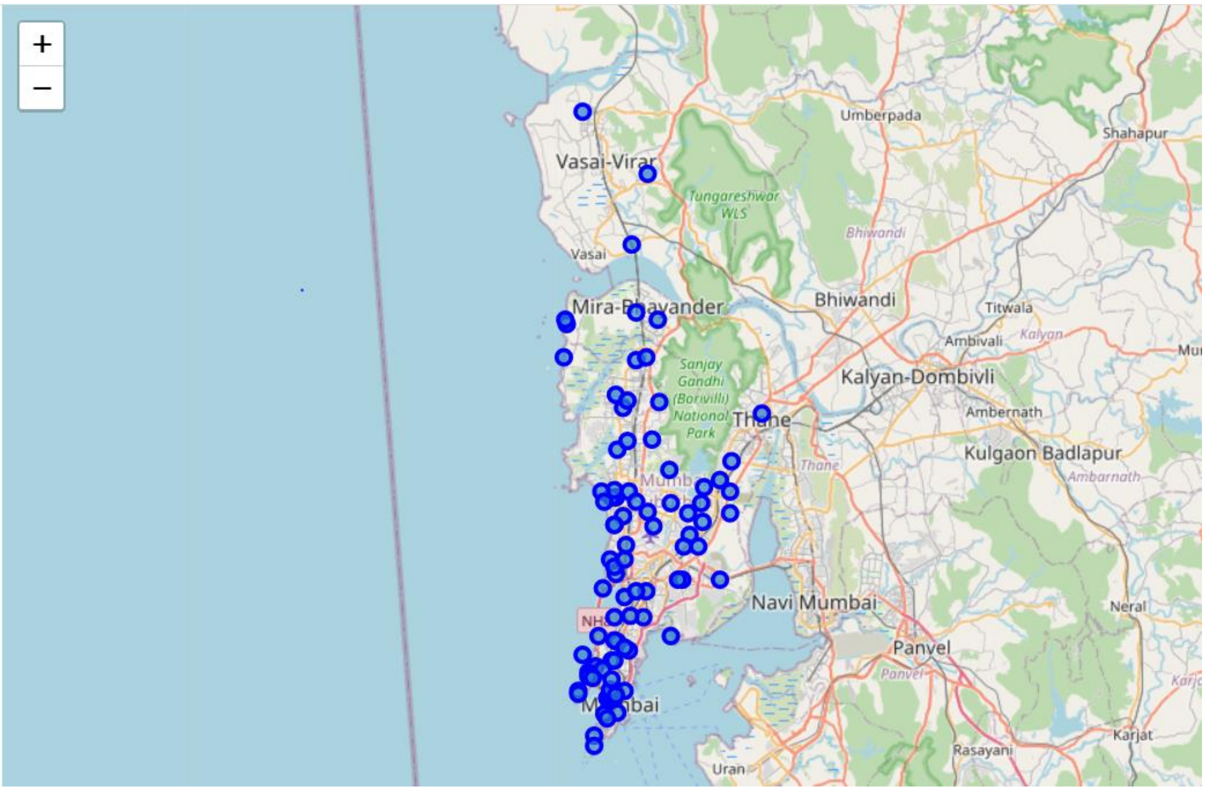
- List of neighbourhoods in Mumbai, India. We can get this from Wikipedia- https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai
- List of Latitudes and Longitudes of the corresponding neighbourhoods. These are required in order to plot them on the map for visualizations and also get the venue data.
- Venue data, particularly for coffee shops. We will use this data to perform clustering on neighbourhoods. We can obtain this data using the **Foursquare API**



	Area	Location	Latitude	Longitude
0	Amboli	Andheri,Western Suburbs	19.129300	72.843400
1	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833
2	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373
3	Four Bungalows	Andheri,Western Suburbs	19.124714	72.827210
4	Lokhandwala	Andheri,Western Suburbs	19.130815	72.829270
5	Marol	Andheri,Western Suburbs	19.119219	72.882743
6	Sahar	Andheri,Western Suburbs	19.098889	72.867222
7	Seven Bungalows	Andheri,Western Suburbs	19.129052	72.817018
8	Versova	Andheri,Western Suburbs	19.120000	72.820000
9	Mira Road	Mira-Bhayandar,Western Suburbs	19.284167	72.871111

3. Methodology

Our first step is to get the list of neighbourhoods in the city of Mumbai. It is available on Wikipedia page link mentioned. We perform web scraping technique using pandas read_html method. This list also contains the geographical coordinates of the neighbourhoods in Mumbai thereby making our task easier. After gathering the data, we use the Folium library package to visualize all the neighbourhoods on a map.



Next, we use the Foursqaure API to obtain the top 100 venues that are within a radius of 500 metres. We need to register and create a Foursqaure Developer account to obtain our

Foursquare ID and secret key. We can then make API calls to Foursquare by passing in the location coordinates. Foursquare then returns venue data in JSON format and from that we can extract venue name, venue category, and venue latitude and venue longitude. With this data, we can check how many venues are returned for each neighbourhood and find out how many unique categories can be curated from this data. Then, we analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. We can filter our data only for 'Coffee Shop' category in the neighbourhoods for our problem statement.

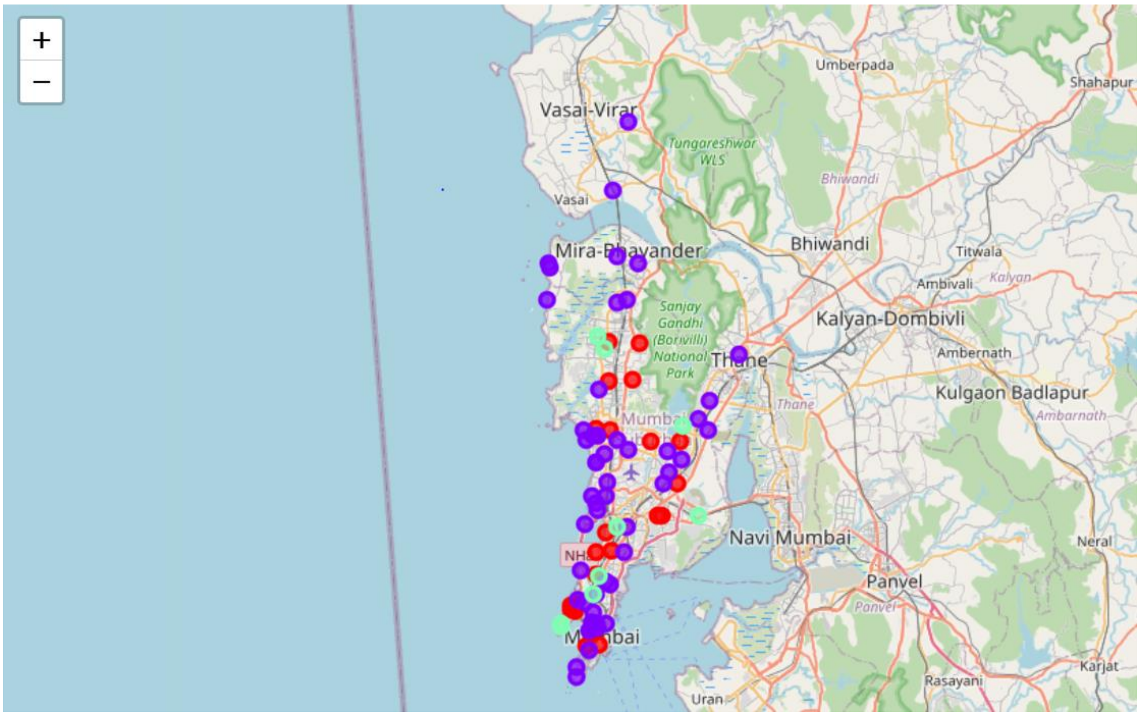
Lastly, we perform clustering on the data obtained in the previous step using k-means clustering algorithm. Identify k number of centroids or clusters you want to create and then allocate every data point to the nearest cluster, while keeping number of centroids to minimum. It is an unsupervised machine learning algorithm and best suited for solving this kind of a problem. The results help us to identify which neighbourhoods have high number of cafes and which of them have lower number of cafes. Based on this concentration of neighbourhoods, it helps us to answer the question that is which location would be best suited to open a new café restaurant in the city.

4. Results

So, we were able to categorize neighbourhoods into 3 categories:

1. Cluster 0: Neighbourhoods with moderate number of Coffee shops.
2. Cluster 1: Neighbourhoods with high concentration of coffee shops.
3. Cluster 2: Neighbourhoods with low number to no existence of coffee shops.

Also, we can observe that there is a high concentration of cafes in the southern part of Mumbai. The results of clustering are visualized in the map below where Cluster 0 is in red color, Cluster 1 is in purple color and Cluster 2 is in green color.



5. Discussion

Most of the cafes are concentrated in the southern area of Mumbai city, with the highest number in cluster 1 and moderate number in cluster 0. On the other hand, cluster 2 has very low number to totally no coffee shops in the neighbourhoods. This represents a great opportunity and high potential areas to open new cafes as there is very little to no competition from existing ones.

Meanwhile, cafes in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of cafes. The suburb areas still have very few coffee shops.

Therefore, this project recommends local cafe entrepreneurs to utilize these findings to open new shops in neighbourhoods in cluster 2 with little to no competition. Entrepreneurs who have a unique selling proposition can also open cafes in Cluster 1 which has moderate competition. Lastly, they are advised to avoid neighbourhoods in cluster 2 which already have high concentration of cafes and are already suffering from high competition.

6. Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting the data and performing a machine learning algorithm K-means clustering on the data to produce clusters based on similarity index and finally making recommendation to target audience for this project.

Thus, we can use our findings to answer the question which was asked in the Business problem, concluding that neighbourhoods in cluster 2 would be best locations to open a new café in Mumbai city. Of course, there can be other parameters such as population and income of people in that area which can further enhance the results obtained. The observations and results obtained from this project can help entrepreneurs and investors to make better decisions to open a new café.