Capstone Project

Determining Ideal Location to Open a Coffee Shop in Mumbai

By- Adit Bangera

Table of Contents

Introduction

Mumbai is the most populated city in India and one of the most densely populated cities in the world. Mumbai is known for its diversity and people with different tastes and lifestyle live here. There is large assortment of restaurants and fast food joints for people to choose from and apart from that there are many street vendors as well. Due to this there is intense competition amongst these various entities and it gets difficult for a business to stay profitable if the location is packed with competitors as well as other alternatives. Even though there are many restaurants there is still a lack of speciality coffee shops where people can have a relaxing cup of coffee and some snacks while working or talking to their friends. If by using data science we would be able to determine areas within Mumbai with less competition it would benefit anyone opening a new shop to know such areas within Mumbai.

Business Problem

The business problem is to determine suburbs within Mumbai with coffee shops and cluster them within groups to be able to determine groups of suburbs which have lower frequency of shops and hence it become a better option to build a coffee shop at that particular. It would be better to even visualize these spots in a map to have a clearer picture. Since even restaurants can have impact on sales it would be better to consider even that data in decision making process.

Target Audience

Restaurants and other small outlets are a major source of income for many residents in Mumbai and as said before in order to be profitable lower number of competitors in an important factor. Also, large commercial chains like Starbucks already use such data science to plan their strategy of expansion and moving into new locations. This type of evaluation will hence help middle level businessmen to plan their ventures in order to provide a good service and also be profitable at the same time. This can also help people who plan on opening new restaurants.

Data Requirements

1) List of suburbs in Mumbai
2) Corresponding latitude and longitudinal data of the suburbs.
3) Data on the coffee shops and restaurants in each suburb.

Methodology

Following are the lists of tasks that's needed to be completed in order to solve this particular problem

1) **Getting the list of Suburbs-**

   This task can be achieved by web-scrapping which can be done using numerous methods. The method used by me was to BeautifulSoup to get the information from the following Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai)

A
- Andheri
- Anushakti Nagar

B
- Baiganwadi
- Bandra
- Bhandup
- Borivali

C
- Charkop
- Chembur

D
- Dahisar
- Devipada
- Dombivli

E
- Eastern Suburbs (Mumbai)

G
- Ghatkopar
- Goregaon
- Grant Road

J
- Jogeshwari
- Juhu

K
- Kalyan
- Kandivali
- Kanjurmarg
- Kausa
- Kurla

M
- Mahavir Nagar (Kandivali)
- Mankhurd
- Matharpacady, Mumbai
- Mira Road
- Mogra Village
- Mulund
- Mumbra

P
- Pestom sagar

S
- Seven Bungalows
- Shil Phata
- Sion, Mumbai

T
- Thakur village
- Tilak Nagar (Mumbai)

V
- Vashi
- Vikhroli

W
- Wadala
- Western Suburbs (Mumbai)
- Worli

Fig 1: List of Suburbs from the Wikipedia page

This list of suburbs is to be obtained and converted into a dataframe. The list of suburbs is scrapped using BeautifulSoup and the dataframe is created suing pandas. The following figure shows the dataframe.

	Neighborhood
0	Andheri
1	Anushakti Nagar
2	Baiganwadi
3	Bandra
4	Bhandup
5	Borivali
6	Charkop
7	Chembur
8	Dahisar
9	Devipada

Fig 2: Suburbs Dataframe

2) Getting the longitude and latitude for each suburb-

The latitude and the longitude values are needed in order to be able to plot the locations on a map and further get the venues data from the Foursquare API. The geocoder and the geopy library is being used to get those values. We get those values and append it to the dataframe previously created. The dataframe looks the following after appending the longitude and latitude data.

	Neighborhood	Latitude	Longitude
0	Andheri	19.118483	72.841774
1	Anushakti Nagar	19.042830	72.927340
2	Baiganwadi	19.062930	72.926660
3	Bandra	19.054220	72.840190
4	Bhandup	19.145560	72.948560

Fig 3: Dataframe with longitude and latitude

This dataframe will be now used to get the venues data from Foursquare API and also to create the map visualization using folium.

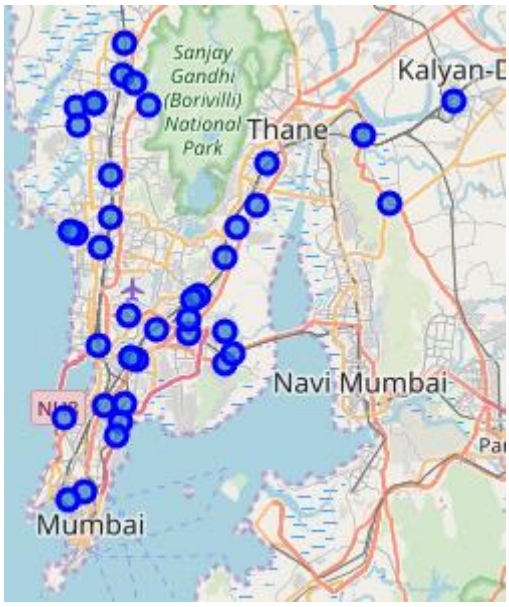

Fig: 4 Visualization of Mumbai with suburbs

3) Using Foursquare API for venues data-

Since the objective of this project is to suggest locations to open a coffee shop we would venues data for each location. We are interested in finding location data on the coffee shop for each suburb and do analysis on that data to suggest potential locations to open a coffee shop. Since the presence of a restaurant may also affect the profitability of a coffee shop we are also interested in finding the location data of restaurants for each suburb in order to do further analysis. The method used is given in detail in the notebook. The data is looked at individually for both the coffee shop venues and the restaurant venues.

Using the groupby count for venues we can see the number of coffee shops for each venue.

Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
Andheri	2	2	2	2	2	2
Anushakti Nagar	1	1	1	1	1	1
Baiganwadi	1	1	1	1	1	1
Bandra	2	2	2	2	2	2
Borivali	2	2	2	2	2	2
Charkop	3	3	3	3	3	3

Fig 5: Coffee Shops count for each suburb

Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
Bandra	2	2	2	2	2	2
Bhandup	2	2	2	2	2	2
Borivali	6	6	6	6	6	6
Charkop	1	1	1	1	1	1
Chembur	3	3	3	3	3	3
Dahisar	3	3	3	3	3	3

Fig 6: Restaurants count for each suburb

4) Using k-means clustering on the location data-

Since we are looking to find pattern in the data which was previously unknown to allow is to be able to make the decision on where to open a coffee shop we will use k-means clustering to divide the suburbs into clusters where each cluster will be looked to see what the underlying pattern is to help us to decide the ideal locations. The same method will also be used for the restaurants. We hope to find clusters with high and low frequency of coffee shops and restaurants which will help us decide on an ideal location to open a coffee shop.

Results

1) Visualization of coffee shop clusters

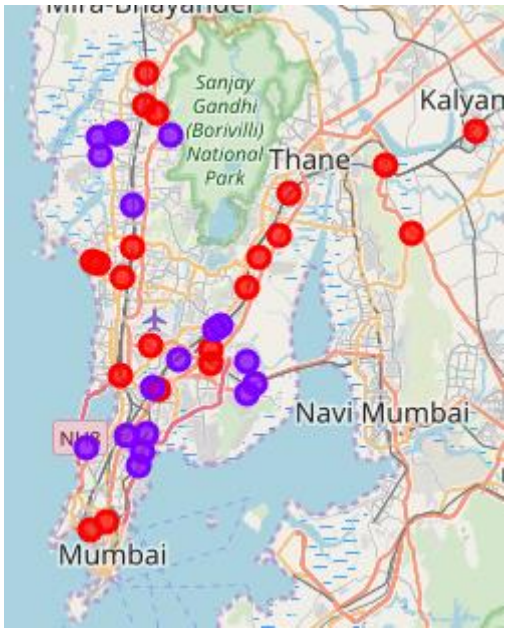

Fig 7: Coffee Shop Clusters

The red markers in the map are of cluster 0 and the purple marker is of cluster 1. On closer examination of individual clusters it can be seen that the clusters are made according to the frequency of the coffee shops in the suburb. The cluster 0 has either zero or lowers number of coffee shops while cluster 1 has a higher frequency of coffee shops.

2) Visualization for restaurant clusters

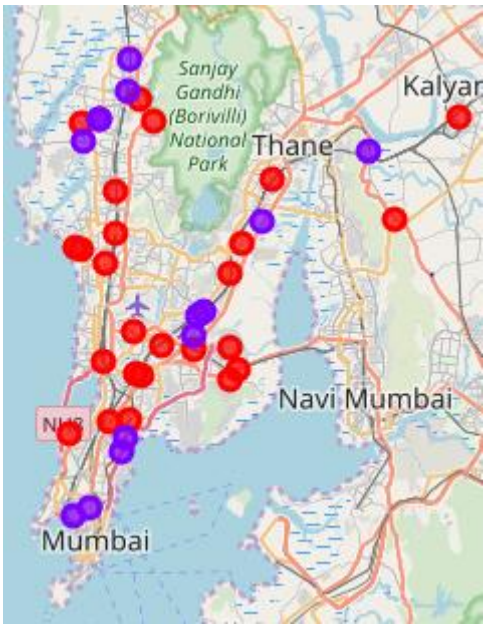

Fig 7: Restaurant Clusters

The red markers in the map are of cluster 0 and the purple marker is of cluster 1. On closer examination of individual clusters it can be seen that the clusters are made according to the frequency of the restaurants in the suburb. The cluster 0 has either zero or lowers number of restaurants while cluster 1 has a higher frequency of restaurants.

3) Comparing the clusters

From the above observations we can see that cluster 0 of both the coffee shop data and the restaurant data contains suburbs with either zero or lower frequencies of coffee shops and restaurant. Since we are interested in finding an ideal location for coffee shops we also need to consider the effect of having a more number of restaurants around since they can take away the customers of the coffee shops. Hence an intersection of both the clusters is taken to find locations with lower number of both coffee shops and restaurants. We get the following results.

:

	Suburb
0	Mulund
1	Kanjurmarg
2	Matharpacady, Mumbai
3	Andheri
4	Seven Bungalows
5	Kausa
6	Chembur
7	Mira Road
8	Jogeshwari
9	Devipada
10	Dombivli
11	Vikhroli
12	Shil Phata
13	Mogra Village
14	Bandra

Fig 8: Ideal Locations

Conclusion

The following analysis includes using many different libraries and method in order to get the final result. By using the venues data and the clustering algorithm we were able to find ideal locations for opening a new coffee shop. Such analysis is subject to the data available and the results are also dependent on it. If the data available is not correct or is insufficient the model will yield inaccurate results. This type of method can be used to ideal locations for other types of venues as well like a gym, garden, hospitals etc. Using the data that is available to us it is clear that the locations in Fig 8 are ideal to open a coffee shop due to lack of competitors. If more data is available then the results can change according using the same method. This project showcases the use of various python libraries to solve a particular problem.

Discussions

In this particular project the clusters where created based on the frequency of shops or restaurants in a particular area. Additional features can also be used to further increase the validity of the results. Demographic information can be included in the model to increase validity but a more complex model would be needed. The population of individual suburbs can also be used for creating clusters as an area with more population means more customers for the coffee shop. Adding average income as a feature is another possibility. Adding more features can increase the validity of the model but can also increase the complexity. Also, getting more data can benefit the model as the quality of the analysis is mainly dependent on the data. So, more data can be added in order to improve the model.