Coursera Capstone

Opening a Snack shop in Mumbai, India

By

Vatsal Sheth

2021

Problem Statement:

Mumbai being one of the densest cities in the world, opening a snack place can be a potentially huge investment due to the people's love for food in the country. The main aim of this project is finding a neighborhood which suits best for opening the snack shop. As any business decision, opening a snack shop involves a lot of variables. And the most important variable, the location can make or break the shop as a thriving business.

Target Audience:

This project is particularly useful for any business investor or people planning to open snack shops, as it helps you figure out the optimum position for your snack shop.

Data Solving and Extraction Steps:

The following data would be needed:

- A dataset of all the neighborhoods in the city.
- Co-ordinates for each and every neighborhood.
- Different venues in each neighborhood, particularly snack shops.

Extracting:

The Wikipedia page(https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai) has a list of all the neighborhoods in the city. We will scrape this page to get a dataset of all the neighborhoods. The library Beautiful Soup will be used for parsing the link. Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favourite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

We then will use geocoder library to get all the coordinates for each and every neighborhood which will be used for clustering. Geocoder is a simple and consistent geocoding library written in Python which will help us get all the co-ordinates we need.

The Four Square API will be used to explore the neighborhood the compile the dataset required and will help us determine the best neighborhood suited for a snack place. Foursquare is the most trusted, independent location data platform for understanding how people move through the real world.
Four Square API will give us all the different venues present in different neighborhoods

Methodology:

Firstly, we need to get the list of neighbourhoods in the city of Mumbai , India. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai) . We use Python requests and beautiful soup packages for web scraping to extract the list of neighbourhoods data. However, we need more than a list of names, we need to get the geographical coordinates in the form of latitude and longitude to use Foursquare API. To do so, Geocoder package will be used which 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 map shows us that we have indeed received the location coordinates of Mumbai.

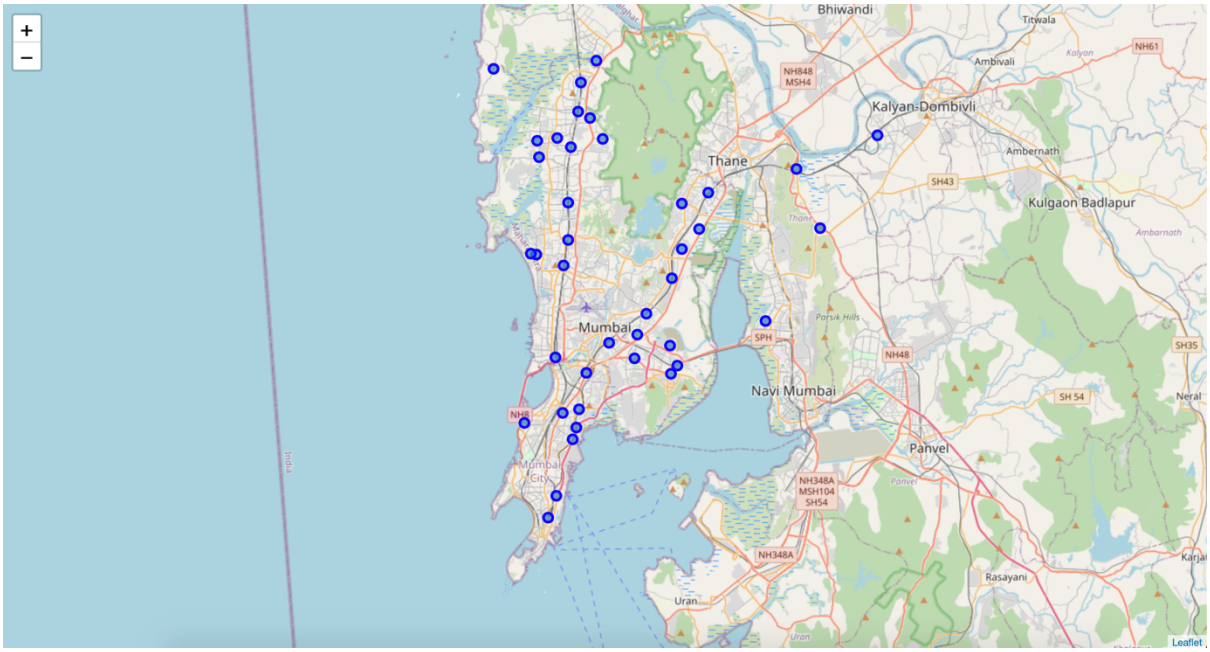

Fig 1: All neighbourhoods in Mumbai, India

Next, we will use Foursquare API to get the top 150 venues that are within a radius of 2000 meters. We need a Foursquare client ID and secret to use the API. We then make a Python loop in which API calls to Foursquare passing in the geographical coordinates of the neighbourhoods and it will return the venue data in JSON format from which we extract the venue name, venue category, venue latitude and longitude into the dataset. We check how much data is returned and how many unique venues are present. Then the data is normalised using the mean metric. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Snack Place” data, we will filter the “Snack Place” as venue category for the neighbourhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K It a very popular unsupervised machine learning algorithms and is particularly suited to solve the problem for our needs. We will cluster the neighbourhoods into 5 clusters based on their frequency of occurrence for “Snack Place”. The results will allow us to identify which neighbourhoods have higher concentration of snack places while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of snack places in different neighbourhoods, we can perform our analysis and conclude which neighbourhood will be most suited for opening a snack shop.

Results:

The data is divided in to 5 clusters based on the density of snack places. All the clusters can be visualised in the attached screenshot of the map. Each cluster has different densities. Increasing the number of clusters helped us figure out the neighbourhood which has the most density of snack places, which is defined by cluster 2. Sonapur, Bhandup has the most concentration for the number of snack places. While cluster 3 has the most number of neighbourhoods. As the project aimed at finding the best place for opening a snack shop, I believe cluster 1 has the lowest infiltration of snack shops and as seen that the snack shop formula works in Mumbai, any neighbourhood in cluster 1 will be an optimum destination to open a snack shop as it will not face any competition.

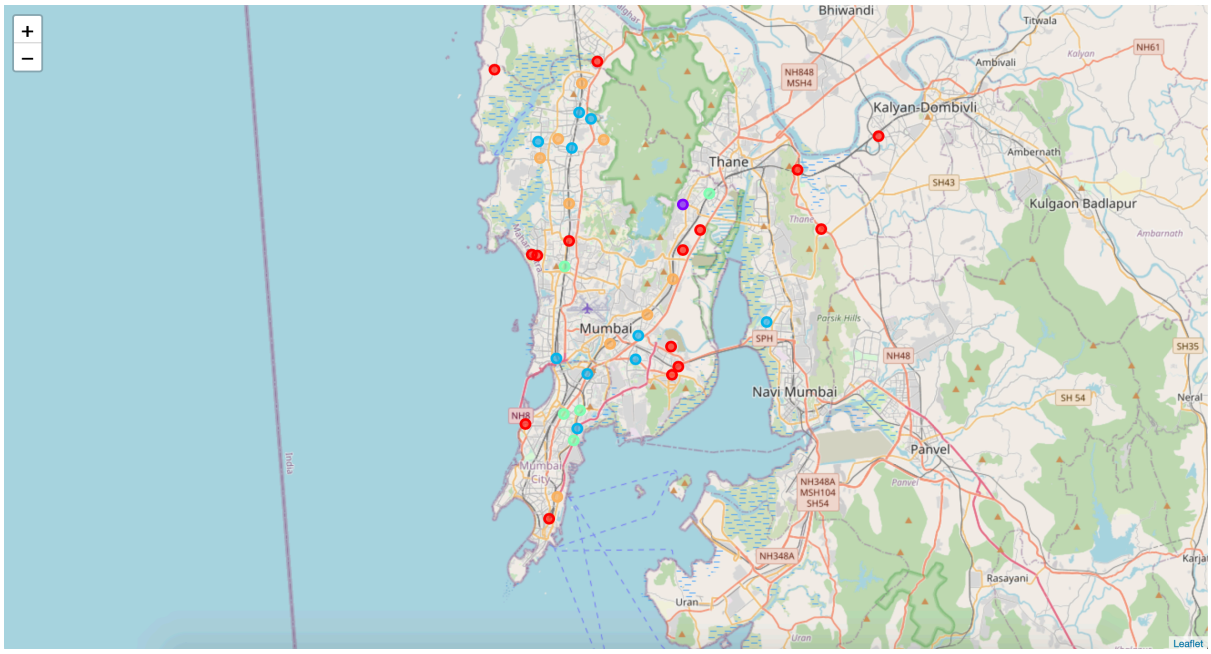

Fig 2: 5 Density based clusters

Conclusion:

Hence, we have performed analysis on the city of Mumbai, India with the aim to find the optimum place to open a snack shop. We have been successful in doing the same and the results are presented in this report. 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 snacking place. Neighbourhoods in cluster 1 are presented in the appendix.

Appendix

Cluster 1:

	Neighborhood	Snack Place	Cluster Labels	Latitude	Longitude
19	Kausa	0.0	0	19.127580	72.825390
34	Uttan	0.0	0	19.260040	72.793412
29	Shil Phata	0.0	0	19.146580	73.040050
28	Seven Bungalows	0.0	0	19.128560	72.821180
26	Mumbra	0.0	0	19.188413	73.022011
24	Mira Road	0.0	0	19.265705	72.870693
23	Matharpacady, Mumbai	0.0	0	18.940170	72.834830
22	Mankhurd	0.0	0	19.048530	72.932220
18	Kanjurmarg	0.0	0	19.131380	72.935680
14	Jogeshwari	0.0	0	19.137920	72.849410
39	Worli	0.0	0	19.007440	72.816880
10	Dombivli	0.0	0	19.212750	73.083240
1	Anushakti Nagar	0.0	0	19.042830	72.927340
4	Bhandup	0.0	0	19.145560	72.948560
2	Baiganwadi	0.0	0	19.062940	72.926630

Cluster 2:

	Neighborhood	Snack Place	Cluster Labels	Latitude	Longitude
31	Sonapur, Bhandup	0.095238	1	19.16394	72.93544

Cluster 3:

	Neighborhood	Snack Place	Cluster Labels	Latitude	Longitude
11	Eastern Suburbs (Mumbai)	0.027778	2	19.004272	72.855790
5	Borivali	0.030000	2	19.229360	72.857510
6	Charkop	0.020833	2	19.208660	72.826120
7	Chembur	0.024691	2	19.053995	72.899675
3	Bandra	0.020000	2	19.054370	72.840170
30	Sion, Mumbai	0.025974	2	19.043410	72.863320
9	Devipada	0.032609	2	19.224690	72.866050
35	Vashi	0.022727	2	19.080060	72.998610
17	Kandivali	0.027397	2	19.204160	72.851690
27	Pestom sagar	0.025000	2	19.070640	72.902170

Cluster 4:

	Neighborhood	Snack Place	Cluster Labels	Latitude	Longitude
33	Tilak Nagar (Mumbai)	0.058824	3	18.99616	72.85279
37	Wadala	0.040816	3	19.01720	72.85816
25	Mulund	0.039474	3	19.17183	72.95565
15	Juhu	0.040000	3	19.01492	72.84522
0	Andheri	0.042553	3	19.11977	72.84629

Cluster 5:

	Neighborhood	Snack Place	Cluster Labels	Latitude	Longitude
20	Kurla	0.011364	4	19.064980	72.880690
38	Western Suburbs (Mumbai)	0.013158	4	19.197010	72.827680
8	Dahisar	0.016667	4	19.250030	72.859070
32	Thakur village	0.017241	4	19.210200	72.875410
16	Kalyan	0.015873	4	18.955437	72.840447
13	Goregaon	0.010000	4	19.164550	72.849460
36	Vikhroli	0.010000	4	19.111090	72.927810
12	Ghatkopar	0.010417	4	19.085357	72.908623
21	Mahavir Nagar (Kandivali)	0.013514	4	19.210940	72.841370