Data Science Capstone Project

By Aayush Kamath

Introduction

As a resident of Mumbai, I have personally witnessed a lot of cafes opening up especially in and around my neighborhood. But this phenomenon is not just limited to my neighborhood. It seems to be the case for the entirety of Mumbai as well as other metropolitan cities in India. Cafes possess a certain charm and an inviting quality especially with regards to the youth. People flock to cafes because it provides a certain level of comfort through its cozy ambience. Not just for light meals, but cafes are also a hotspot for a lot of working professionals who find the environment comfortable enough to carry out their work while maybe enjoying a cup of tea or coffee. Running a café is a business and for any business belonging to the restaurant/café industry location plays an important factor in whether or not the business succeeds in the long run.

Business Problem

Having identified location as a key component in running a café it is necessary that through careful analysis we arrive at the optimum location. This project intends to analyze suburbs/neighborhoods in Mumbai, and come up with location of suburbs/neighborhoods which might be ideal for opening a café. The solution to this business problem is arrived at by using the key concepts that we have accustomed ourselves with throughout this capstone project course.

The aforementioned business problem is tackled in a step by step manner that we learnt about in our week 3 assignment of the Battle of Neighborhoods.

It includes scraping data, accessing location data using Foursquare API and clustering method. Part 2 of this report explains in detail about data used for this business problem

Stakeholders

Finally, the stakeholders for this business problem could be the people who are looking to open a café but lack an idea about what the optimum location for the café would be. With the help of this data there could be better clarity on the business owners part, as far as decision making is concerned. Another set of stakeholders would be brokers/property agents who intend to use this data to guide people who are keen on opening a café. Property brokers could use this data to gain better insights and add to their repertoire of knowledge about locations. This could help them in brokering a deal for the person looking to open a café.

Data Required to solve the business problem:

1. The list of suburbs/neighborhoods in Mumbai, India.

2. The geographical co-ordinates, that is, the latitude and longitude of the neighborhoods in Mumbai.

3. Data of venues near all neighborhoods, especially data of cafes.

Data Sources:

- The list of suburbs/neighborhoods in Mumbai is scraped from the wikipedia page of suburbs in Mumbai.

  (Link: https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai)

  The scraping of data is done using the Beautiful Soup library in Python.

- Once we have the list of all neighborhoods, we find the latitude and longitude for each of them using the Geocoder package in Python. The location data is necessary as it is used for creating a map of the neighborhoods of Mumbai. While the data remains stored in the dataframe, a quick glance at the map gives us an idea of the distribution of neighborhoods in Mumbai.

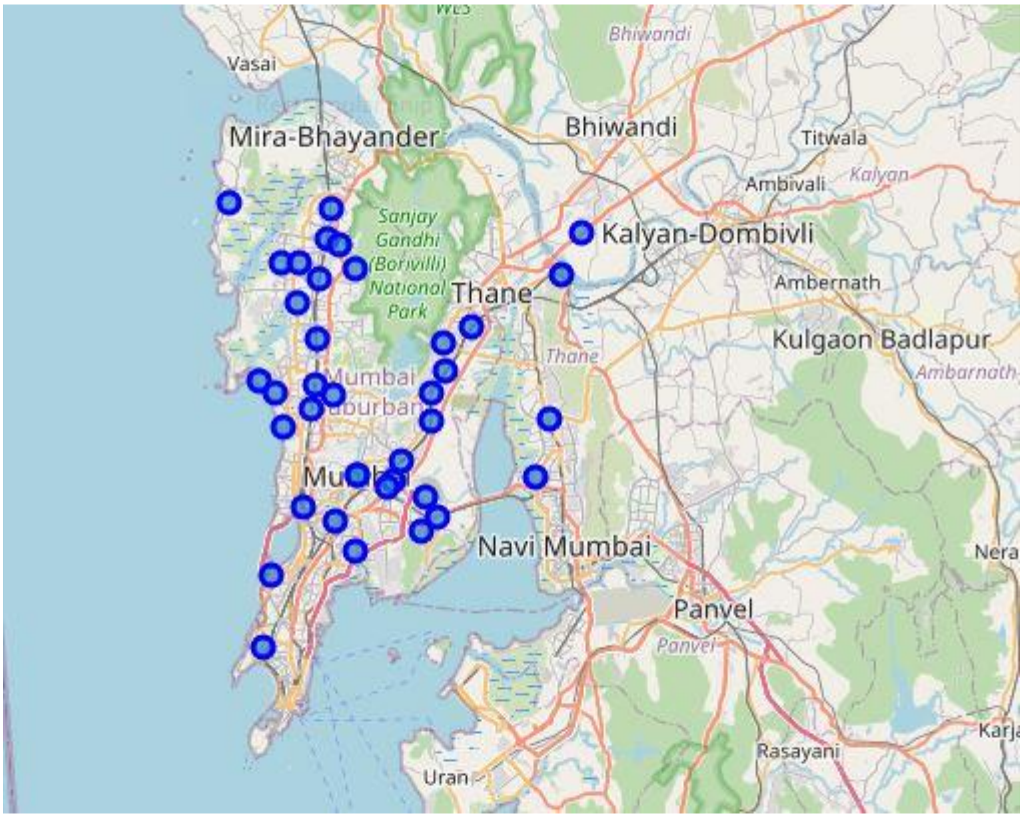
- The venue data for all neighborhoods are obtained by using the Foursquare API. By writing appropriate queries, we can access the necessary venue data through this API. Once the data is obtained, we sift through it to find the data that is necessary for arriving at a solution for our business problem. In this case, the necessary data being talked about is the data about cafes.

Data being used:

- The scraped data from the wikipedia page of Suburbs of mumbai for creation of an intial dataframe that displays the names of all the suburbs/neighborhoods in Mumbai.
- The latitude and longitude of every neighborhood, which is found using the Geocoder package
- Data of the top venues (limit 100) in a 1km radius of every neighborhood accessed using foursquare API.
- Finally, working on the top venues data obtained through the Foursquare API leads us to necessary data regarding cafes.

Methodology

The Suburbs of Mumbai wikipedia page is scraped using the Beautiful Soup library available in Python. This gives us a list of names of the suburbs/neighborhoods in Mumbai. To further build upon this, we need the geographical co-ordinates of every neighborhood. Using the Geocoder package we find the latitude and longitude values of all the neighborhoods in our list. We append these values to our existing dataframe. So far, we have gathered the basic location data which is the name, latitude and longitude of every neighborhood. Using this data we create a map with markers to show the neighborhoods located in Mumbai.

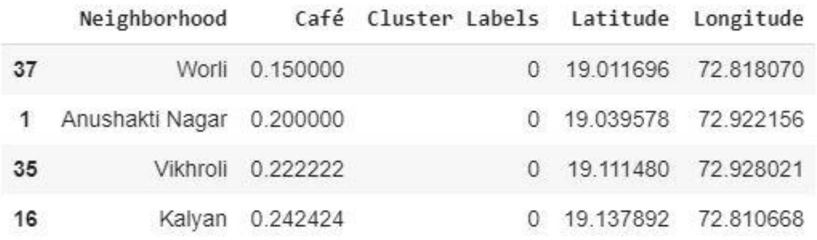


Now that we know we have the right location markers for neighborhoods in Mumbai, we make use of the Foursquare API to get top venues (limit 100) in a 1 kilometer radius for each neighborhood. This data is accessed by writing a query and making an API call. A foursquare developer account is necessary as keys such as Foursquare ID and Foursquare Secret (Both account specific) are to be passed with the query along with the version(date) in order to access data. A json file is returned by foursquare from which we extract the venue name, its longitude and latitude and the venue category(Ex: Restaurant, gym,etc.). Then, we analyze the frequency of every type of venue category for a neighborhood. For Example, let us say that the neighborhood of Andheri has 50 venues in a 1km radius out of which 3 are Indian Restaurants. Then, the frequency of Indian Restaurants for Andheri would be 0.06. Now, that we have all frequencies, we extract the frequency of cafes for each neighborhood as our business problem involves finding suitable locations for a café. Based on these frequencies, we divide all the neighborhoods into 3 different clusters using k-means clustering method. The result obtained helps us analyze about the suitability of each neighborhood as a location for a new café.

Results

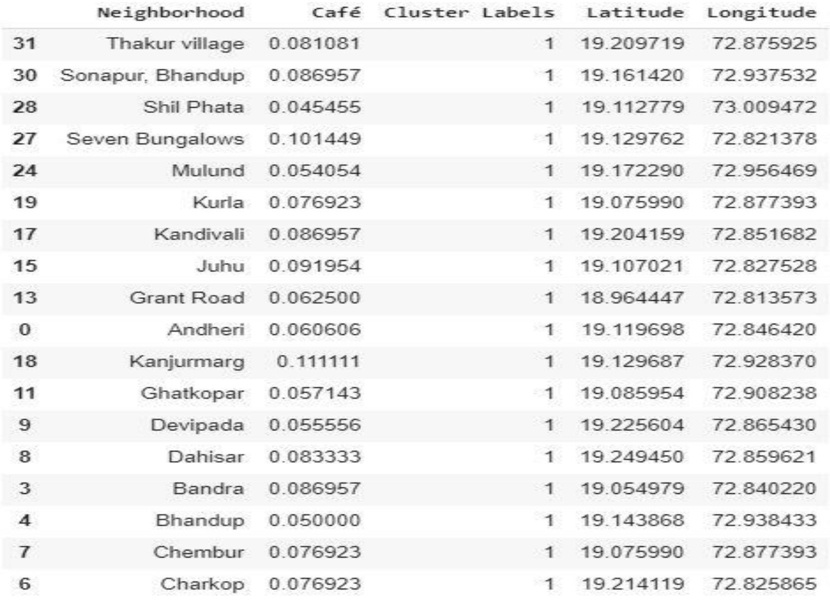
We obtained 3 clusters:

Cluster 0 had neighborhoods with high frequency of cafes.



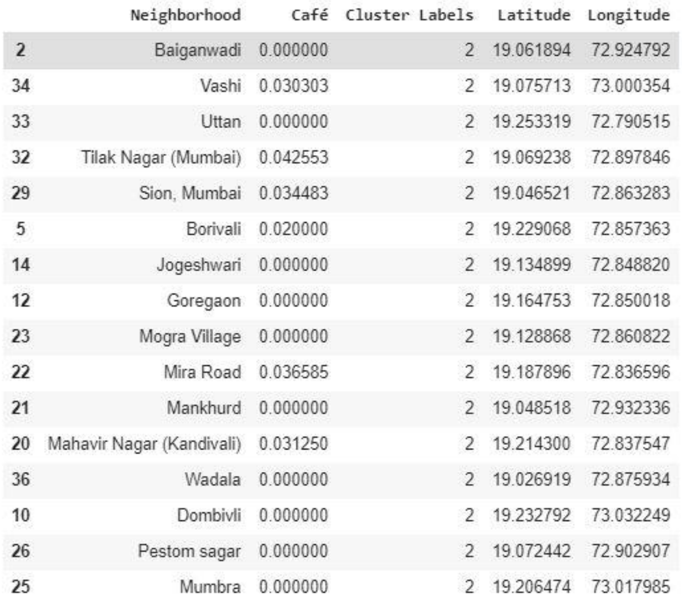
	Neighborhood	Café	Cluster Labels	Latitude	Longitude
37	Worli	0.150000	0	19.011696	72.818070
1	Anushakti Nagar	0.200000	0	19.039578	72.922156
35	Vikhroli	0.222222	0	19.111480	72.928021
16	Kalyan	0.242424	0	19.137892	72.810668

Cluster 1 had neighborhoods with moderate frequency of cafes.



	Neighborhood	Café	Cluster Labels	Latitude	Longitude
31	Thakur village	0.081081	1	19.209719	72.875925
30	Sonapur, Bhandup	0.086957	1	19.161420	72.937532
28	Shil Phata	0.045455	1	19.112779	73.009472
27	Seven Bungalows	0.101449	1	19.129762	72.821378
24	Mulund	0.054054	1	19.172290	72.956469
19	Kurla	0.076923	1	19.075990	72.877393
17	Kandivali	0.086957	1	19.204159	72.851682
15	Juhu	0.091954	1	19.107021	72.827528
13	Grant Road	0.062500	1	18.964447	72.813573
0	Andheri	0.060606	1	19.119698	72.846420
18	Kanjurmarg	0.111111	1	19.129687	72.928370
11	Ghatkopar	0.057143	1	19.085954	72.908238
9	Devipada	0.055556	1	19.225604	72.865430
8	Dahisar	0.083333	1	19.249450	72.859621
3	Bandra	0.086957	1	19.054979	72.840220
4	Bhandup	0.050000	1	19.143868	72.938433
7	Chembur	0.076923	1	19.075990	72.877393
6	Charkop	0.076923	1	19.214119	72.825865

Cluster 2 had neighborhoods with low frequency of cafes



	Neighborhood	Café	Cluster Labels	Latitude	Longitude
2	Baiganwadi	0.000000	2	19.061894	72.924792
34	Vashi	0.030303	2	19.075713	73.000354
33	Uttan	0.000000	2	19.253319	72.790515
32	Tilak Nagar (Mumbai)	0.042553	2	19.069238	72.897846
29	Sion, Mumbai	0.034483	2	19.046521	72.863283
5	Borivali	0.020000	2	19.229068	72.857363
14	Jogeshwari	0.000000	2	19.134899	72.848820
12	Goregaon	0.000000	2	19.164753	72.850018
23	Mogra Village	0.000000	2	19.128868	72.860822
22	Mira Road	0.036585	2	19.187896	72.836596
21	Mankhurd	0.000000	2	19.048518	72.932336
20	Mahavir Nagar (Kandivali)	0.031250	2	19.214300	72.837547
36	Wadala	0.000000	2	19.026919	72.875934
10	Dombivli	0.000000	2	19.232792	73.032249
26	Pestom sagar	0.000000	2	19.072442	72.902907
25	Mumbra	0.000000	2	19.206474	73.017985

Discussion

The neighborhoods were divided into 3 clusters based on their frequency value for cafes. Neighborhoods in Cluster 0 show the highest frequency of cafes. For a person looking to open a new cafe, neighborhoods belonging to cluster 0 might not be the best choice for business as there are already a lot of cafes in those neighborhoods. Cluster 1 shows neighborhoods with moderate frequency of cafes. Opening a cafe in these neighborhoods can be considered as the market isn't saturated and at the same time there seems to be a decent amount of demand for cafes. Cluster 2 has neighborhoods with either low frequency of cafes or zero cafes. Opening a cafe in any one of these neighborhood presents a high risk high reward scenario. On one hand, a cafe might be considered as a new experience for the residents which could lead to good revenue. Whereas, on the other hand, people might simply not react to the opening of a cafe in their neighborhood as it might not be considered as a necessity by the residents of the neighborhood in question.

Limitations

The factor taken into account here is frequency of cafes in a neighborhood. But various factors such as population, availability of real estate for opening a café, and average income of the neighborhood would also play a key role in deciding about the suitable locations for a café. Further restrictions exist in the form of data that can be accessed from Foursquare. For a free developer account, a limited number of API calls can be made. Extraction of data for acquiring in depth knowledge would require the researcher to opt for a paid Foursquare account.

Conclusion

In this Data Science Capstone Project we have identified a business problem, provided a brief introduction, identified the data requirements and the data sources, extracted data from those sources and prepared it and clustered it in order to gain the necessary insights and accordingly, provide recommendations to the concerned stakeholders. Finally, the inference made from this project could be that cluster 2 is saturated due to high frequency of cafes, cluster 0 is unpredictable due to low frequency of cafes and cluster 1 is probably the most suitable group of neighborhoods where a café could be opened as the frequency of cafes in those neighborhoods are moderate.