Capstone Project: Coffee Joint in Mumbai

A. Introduction

A.1. Description & Discussion of the Background

Mumbai (formerly called Bombay) is a densely populated city on India's west coast. A **financial centre**, it's **India's largest city**. On the Mumbai Harbour waterfront stands the iconic *Gateway of India* stone arch, built by the British Raj in 1924. Mumbai encloses an area of **603.4 km²** and has a population of nearly **2 crores**.

Mumbai is a city with a high population and population density. Being such a crowded city leads the fast-food chain business and social sharing places in the city where the population is dense. When we think of it by the investor, we expect from them to prefer the areas or suburban places where there is a lower competition and the type of business, they want to install is less intense. If we think of the city residents, they may want to choose the regions where options are more, and prices are economical. However, it is difficult to obtain information that will guide investors in this direction, nowadays.

Me and few of my entrepreneur friends wants to open new coffee hangout joint in Mumbai city. We're interested in a location where there are least No. of coffee shops already present, so that we can maximize the profits.

Mumbai is a costly city and since I'm not much aware about the city, I decided to do a quick study of different areas in Mumbai to see if I can make a decent investment.

In this project, I will try to find an optimal location to set-up a coffee hangout place. Specifically, this report will be targeted to stakeholders interested in opening a Coffee Hangout in Mumbai, India.

Since there are lots of coffee hangouts in Mumbai, I will try to detect locations that are not already crowded with coffee joints.

I'll be using Data Science techniques to generate different visualizations, in forms of graphs and maps, depicting the promising neighbourhoods to open a coffee join in Mumbai.

A.2. Data Description

Following data sources will be needed to extract/generate the required information:

1. **Wikipedia** source for different areas and locations in Mumbai: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai
2. **FourSquare API** to search for existing Coffee Joints in Mumbai.
3. Different **Python Libraries** like *Folium* for visualizing the data on Maps, *Matplotlib* for graphs, *Pandas* and *Numpy* for Data Analysis, etc.

B. Methodology

For start, I identified the below tasks to

1. Scrape Mumbai postal code information from Wikipedia website.
2. Load the data into a dataframe where data is wrangled, cleaned and transformed into a final dataframe.
3. Populate the dataframe with Mumbai Area, Locations and Latitude and Longitude information.

To scrape the Wikipedia website, I imported pandas dataframe and created the dataframe **geodata**:

```
In [1]: # Load url into data objects for analysis:
        import pandas as pd #Library for data analysis
        geodata = pd.read_html("https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai")
        geodata = geodata[0]
        geodata.head()
```

Out[1]:

	Area	Location	Latitude	Longitude
0	Amboli	Andheri,Western Suburbs	19.129300	72.843400
1	Chakala, Andheri	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

This gave me the list of all Areas, Locations and geocoordinates for Mumbai City, as shown above.

Then, I performed a check on the data to ensure that there are no duplicate Area names in the dataset, following which I also, noted the shape of the data frame:

```
In [3]: dup_chk = geodata.apply(lambda col: col.duplicated()).sum()
        print(dup_chk)

Area          0
Location     62
Latitude      8
Longitude    11
dtype: int64
```

The dup_chk lambda function above returned a 0 sum of Area column duplicates, therefore no condition of two or more Locations sharing same Area value exists within dataframe. Let's get the shape of the dataframe to see how many areas are there in Mumbai

```
In [3]: print(geodata.shape)

(93, 4)
```

The shape of **geodata** is **93 x 4** inidicating **93 rows** and **4 columns**.

Once this task was completed and I validated my dataset, I identified the below tasks to perform in order to achieve my goal:

4. Validate the Data further
5. Setup FourSquare Parameters
6. Search for Coffee Places in API Response
7. Explore Neighbourhoods in Mumbai
8. Visualize the Areas with Least No. of Coffee Shops

In order to continue further, imported the below Python Libraries:

```
In [55]: import numpy as np # library to handle data in a vectorized manner

         pd.set_option('display.max_columns', None)
         pd.set_option('display.max_rows', None)

         #import json # library to handle JSON files

         #!conda install -c conda-forge geopy --yes #uncomment this line geopy is not already installed
         from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

         import requests # library to handle requests
         from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

         # Matplotlib and associated plotting modules
         import matplotlib.cm as cm
         import matplotlib.colors as colors
         import matplotlib.pyplot as plt

         # import k-means from clustering stage
         from sklearn.cluster import KMeans

         #!pip install folium #Uncomment this line if Folium is not already installed.
         import folium #map rendering library

         print('Libraries imported.')

         Libraries imported.
```

I validated the data in my dataframe further using the first row of the dataframe and calculating it's geocodes:

```
In [6]: geodata.loc[0,'Area']

Out[6]: 'Amboli'
```

Calculate the No. of Areas and Locations.

```
In [7]: print('The dataframe has {} Areas and {} Locations.'.format(
                 len(geodata['Area'].unique()),
                 geodata.shape[0]
             )
         )

         The dataframe has 93 Areas and 93 Locations.
```

Check Latitude and Longitude values in **geodata** dataframe for first row **(location = 0)**.

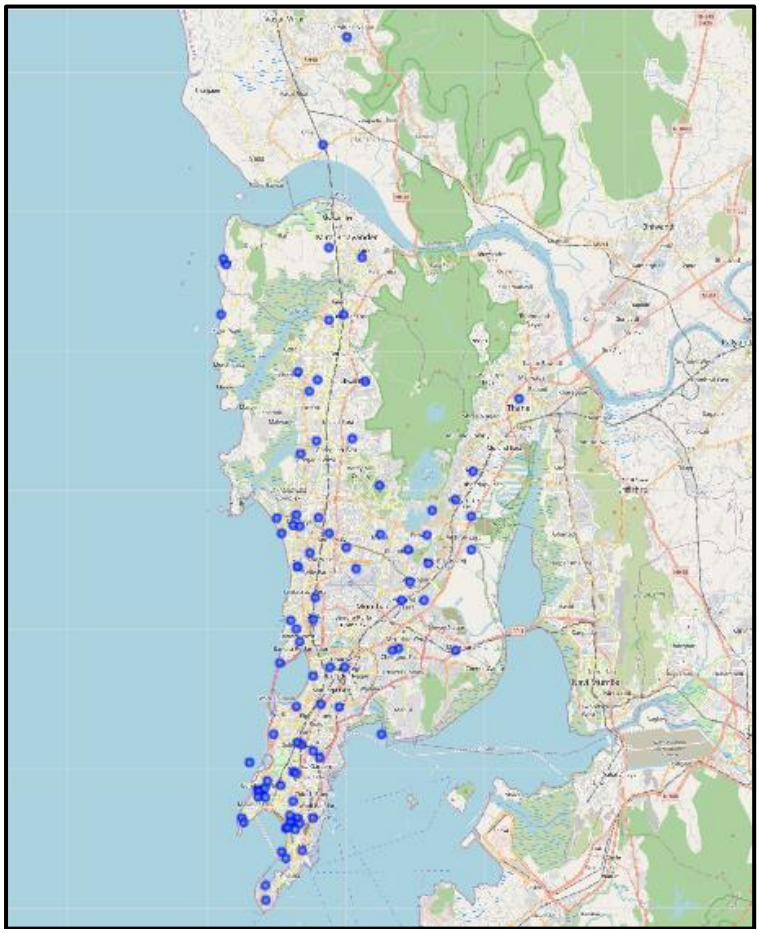
```
In [8]: geodata_latitude = geodata.loc[0, 'Latitude'] # neighborhood latitude value
         geodata_longitude = geodata.loc[0, 'Longitude'] #neighborhood longitude value
         geodata_neighborhood = geodata.loc[0,'Area'] #neighborhood features

         print('Latitude and longitude values of {} are {}, {}.'.format(geodata_latitude,
                                                                        geodata_longitude,
                                                                        geodata_neighborhood))

         Latitude and longitude values of 19.1293 are 72.8434, Amboli.
```

Next, I setup my FourSquare API credentials using my Client ID and Client Secret ID.

I used python **folium** library to visualise geographic details of Mumbai city and its suburban areas and I created a map of Bengaluru with areas superimposed on top. I used latitude and longitude values to get the visual as below:



Next, I used the FourSquare API to explore the areas and locations in Mumbai city. Particularly, I searched a radius of 3 Kms around **Mira Road** in Mumbai to identify different categories of places identified by the API call:

Out[112]:

	name	categories	lat	lng
0	GCC Club	Gym / Fitness Center	19.282988	72.878259
1	Ratnagiri Malwani Food	Indian Restaurant	19.287782	72.864578
2	Vardhman Fantasy	Sculpture Garden	19.289466	72.866234
3	McDonald's	Fast Food Restaurant	19.287089	72.867445
4	N. L. Dalmia Institute of Management Studies a...	General College & University	19.269679	72.870757

Calculate **No. of Venues** returned by Foursquare API.

In [113]:
```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
```

```
48 venues were returned by Foursquare.
```

Next, I used FourSquare API calls to generate a list of all coffee places listed in Mumbai. Below are few of them:

```
In [39]: coffee_places.head(10)
```

Out[39]:

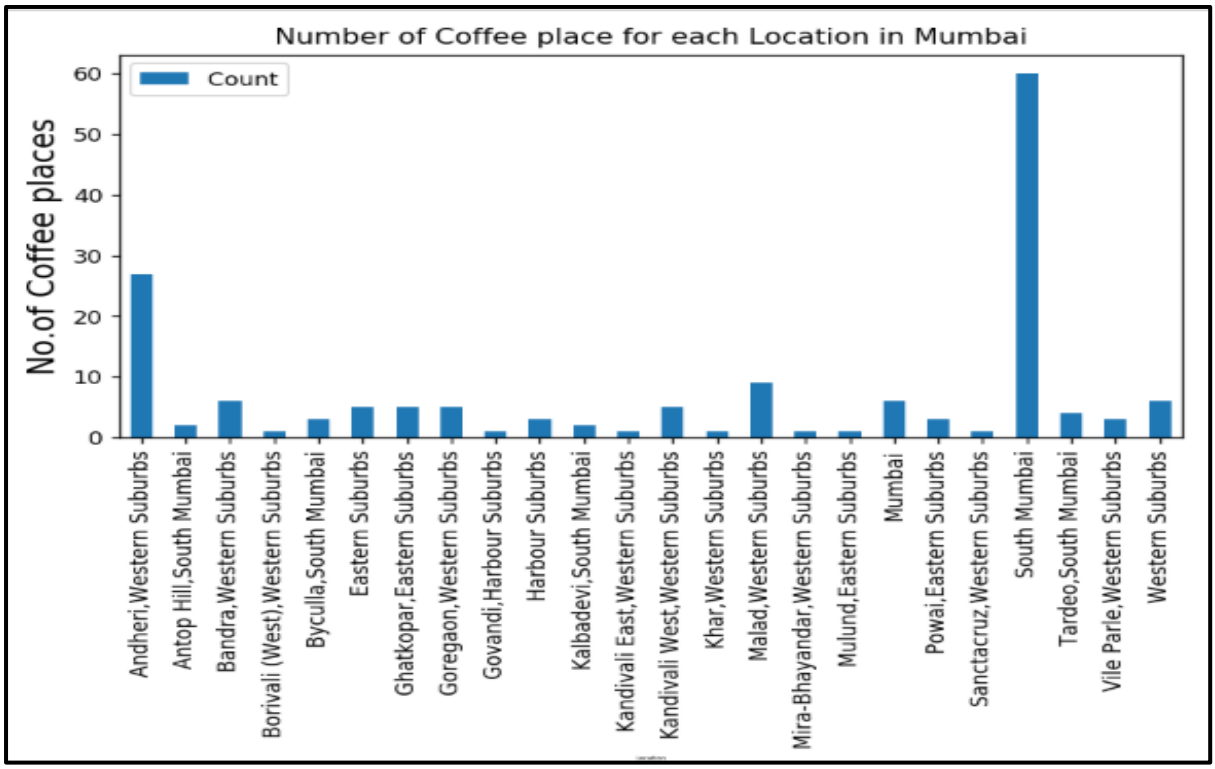
	Area	ID	Location	Name	Count
0	Amboli	4cfa068ac6cca35d13dd8332	Andheri,Western Suburbs	Cafe Coffee Day	2
1	Amboli	4b9fb2aff964a520013637e3	Andheri,Western Suburbs	Cafe Coffee Day	2
2	D.N. Nagar	4cfa068ac6cca35d13dd8332	Andheri,Western Suburbs	Cafe Coffee Day	2
3	D.N. Nagar	4dfb3841b61c6408772e6dd6	Andheri,Western Suburbs	Cafe Coffee Day	2
4	Four Bungalows	4df0d576c65b0270f3a54a0b	Andheri,Western Suburbs	Cafe Coffee Day	3
5	Four Bungalows	4c63efb4e13495218be3c9f0	Andheri,Western Suburbs	cafe coffee day	3
6	Four Bungalows	4dfb3841b61c6408772e6dd6	Andheri,Western Suburbs	Cafe Coffee Day	3
7	Lokhandwala	4dfb3841b61c6408772e6dd6	Andheri,Western Suburbs	Cafe Coffee Day	6
8	Lokhandwala	4df0d576c65b0270f3a54a0b	Andheri,Western Suburbs	Cafe Coffee Day	6
9	Lokhandwala	4c63efb4e13495218be3c9f0	Andheri,Western Suburbs	cafe coffee day	6

A total of 161 Coffee joints were identified by the FourAquare API in Mumbai:

```
In [40]: coffee_places.shape
```

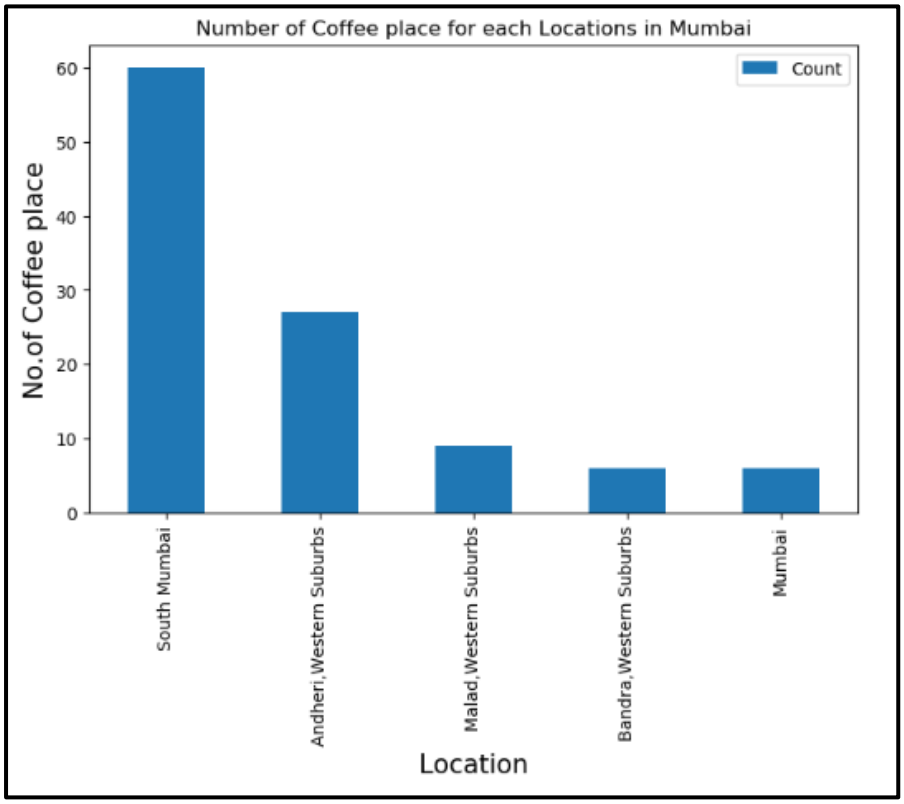
Out[40]: (161, 5)

Based on different locations in Mumbai, I plotted a simple graph to identify which zones has a greater number of coffee places:

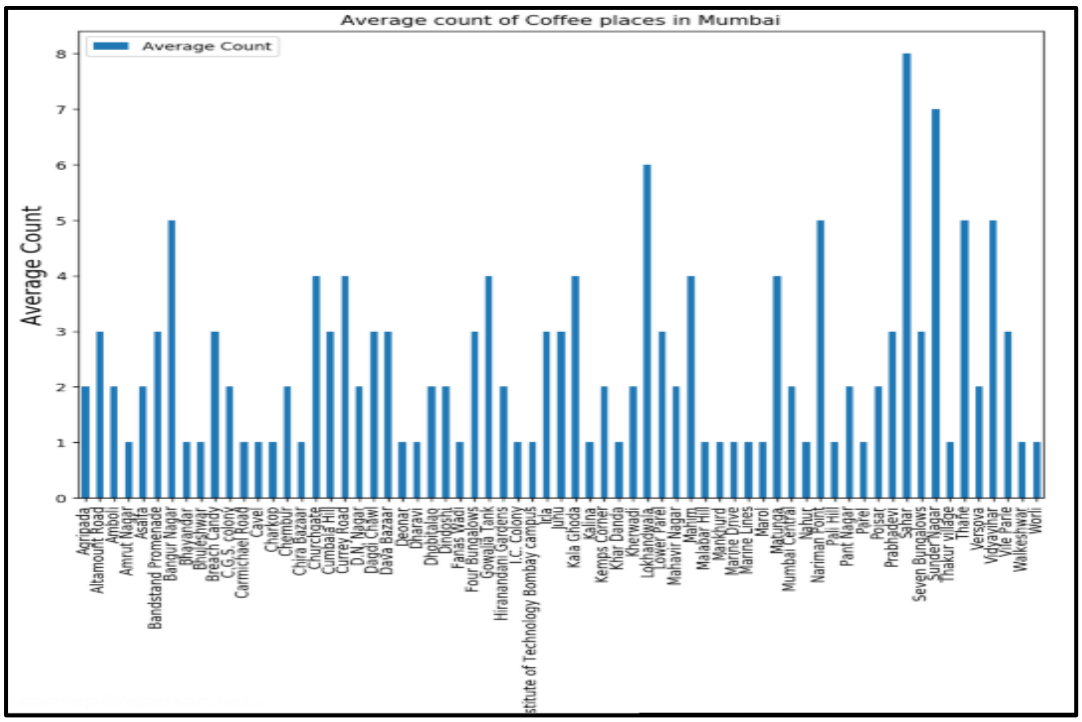


From the above graph, it was evident that **South Mumbai** has a greater number of Coffee shops followed by **Andheri, Western Suburbs**.

Below graph shows the Top 5 Zones having maximum number of Coffee Shops.



Drilling down further, it was found that on an average **Sahar** has the greatest number of coffee shops in Mumbai:

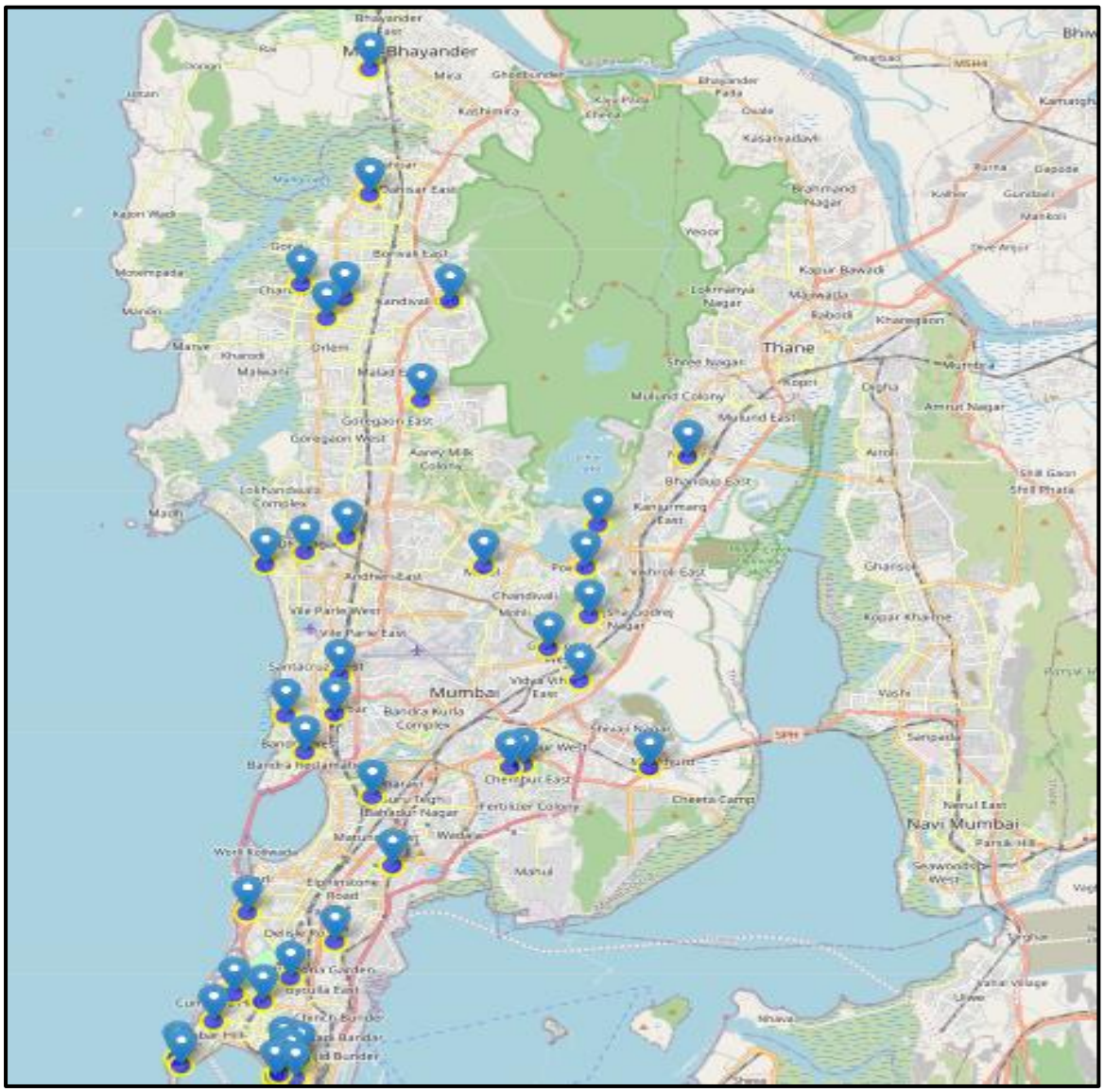


Moving on, I identified the areas having the least average of coffee shops in Mumbai:

Out[168]:

	Area	Latitude	Longitude	Average Count
0	Amrut Nagar	19.102077	72.912835	1.0
1	Bhayandar	19.290000	72.850000	1.0
2	Bhuleshwar	18.950000	72.830000	1.0
3	Carmichael Road	18.972200	72.811300	1.0
4	Cavel	18.947400	72.827200	1.0

To identify an area to invest in coffee shop, I used the above areas which have least average of coffee shops.
Finally, visualizing these areas on map, using Folium, I was able to clearly locate the spots for a good coffee joint where people would come in for a nice coffee:



C. Results

Using Data Science techniques, I was able to identify the below areas in which I can look forward to invest in a coffee shop as these areas have the least number of coffee joints:

Out[73]:

	Area	Latitude	Longitude	Average Count	Label
0	Amrut Nagar	19.102077	72.912835	1.0	Amrut Nagar, (1.0)
13	Khar Danda	19.068598	72.840042	1.0	Khar Danda, (1.0)
23	Walkeshwar	18.947596	72.795957	1.0	Walkeshwar, (1.0)
22	Thakur village	19.210206	72.872980	1.0	Thakur village, (1.0)
21	Parel	18.990000	72.840000	1.0	Parel, (1.0)
20	Pali Hill	19.068000	72.826000	1.0	Pali Hill, (1.0)
19	Nahur	19.157000	72.941000	1.0	Nahur, (1.0)
18	Marol	19.119219	72.882743	1.0	Marol, (1.0)
17	Marine Lines	18.944700	72.824400	1.0	Marine Lines, (1.0)
16	Marine Drive	18.944000	72.823000	1.0	Marine Drive, (1.0)

As we can see below are the list of areas which has less coffee joints and it would be good for business, if we consider these areas to open new coffee hangout.

1. Amrut Nagar
2. Khar Danda
3. Walkeshwar
4. Thakur Village
5. Parel
6. Pali Hill
7. Nahur
8. Marol
9. Marine Lines
10. Marine Drive

Limitations:

The Count is highly dependent on Foursquare API details.
There might be some other coffee hangouts which are not listed in foursquare database.

D. Discussion

Like I mentioned earlier in my problem statement, Mumbai is the largest Indian city which happens to be the costliest with a high population density. Mumbai is known for its variety of food joints for all price ranges.

Probably, it would be best for me to analyse further on the areas where I need to invest to get the best possible profit. For example, which areas cost more to setup a joint compared to others and what best fits my budget.

E. Conclusion

Going through the data science course and having a hands-on this project, I am better familiar with data science techniques and methodology. I'll try to analyse further into my project and see if I can better my analysis and come up with best possible investment areas