IBM Applied Data Science Capstone Project (Final Report)

Topic - ***Setting up a coffee shop in Mumbai, Maharashtra***

INTRODUCTION

This is a part of the capstone project offered by IBM Applied Data Science Course on Coursera. In this project I have selected Mumbai as my target city. Mumbai is one of the most populous cities and also the financial, commercial and entertainment capital of India. The tremendous opportunities offered by this city attracts a lot of youth here.

Coffee is a beverage that never goes out of style and is consumed daily by millions of people specially the youth. Moreover, coffee shops are a great place for people to sit and relax or get their work done. When set up in the right location it can be a great business. One can always add some creativity to make the shop stand out amongst others. Therefore, I decided to pick a coffee shop business in Mumbai for my project.

BUSINESS PROBLEM

The aim of this project is to find a suitable location in the city of Mumbai to set up a coffee shop business. Now, there are various factors to be considered when trying to set up a coffee shop like the cost of living of that place, the competition in the market, the population there, the kind of neighborhood of that location etc. In this project I have solved the problem using two parameters-

1. The competition in the market
2. The neighborhood i.e. the kind of places present there like bookstores, malls etc.

A location having places like bookstores, banks, shops and movie theaters is more likely to offer a larger number of customers. This is because people like to have a cup of coffee while reading a book, during a movie or while waiting at the shops or bank. This is why analyzing the neighborhood becomes an important factor while deciding on the location. So keeping the competition and the kind of neighborhood in mind, I have tried to formulate a solution using datasets from Wikipedia and Foursquare API and machine learning algorithms like K-means clustering.

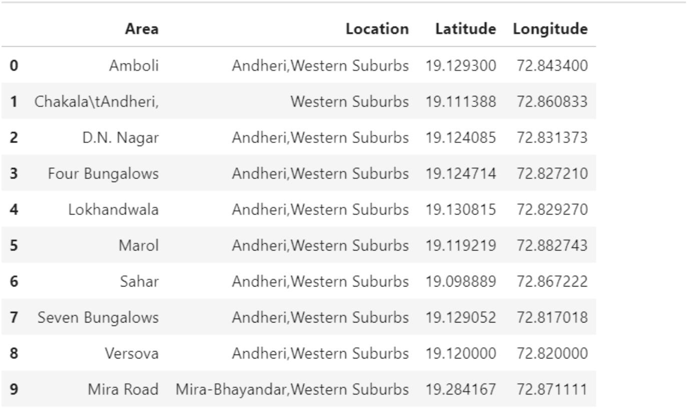
TARGET AUDIENCE

The target audience of this project is anyone who is looking to set up a coffee shop in the city of Mumbai. It can act as the main business or even as an extra source of income. So it can also be useful to people looking for some extra money. This project will be helpful in narrowing down the choices of locations which will benefit the entrepreneur immensely as location is one of the major factors to be considered while setting up the shop.

DATA SECTION

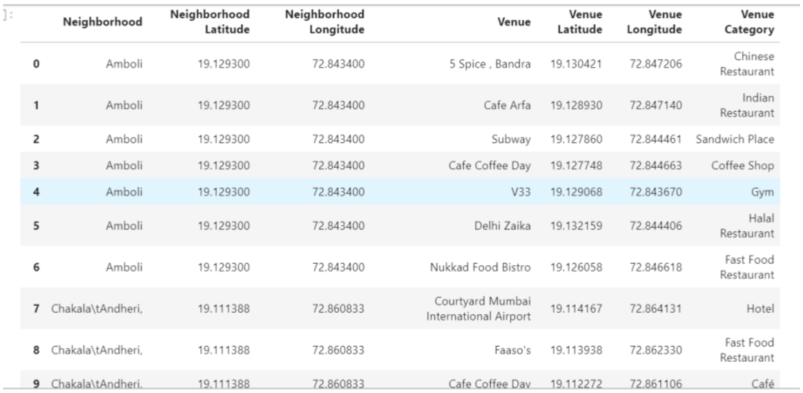
1. List of neighborhoods in Mumbai
   Data source: The dataset was obtained from Wikipedia:
   https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai
   Data was scraped from the web using Beautiful Soup and extracted into a dataframe.

2. The location co-ordinates (Latitudes and Longitudes)
   Data Source: The data was already present in the dataset extracted from Wikipedia.



	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. The venues present in Mumbai
   Data Source: Foursquare API



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Amboli	19.129300	72.843400	5 Spice , Bandra	19.130421	72.847206	Chinese Restaurant
1	Amboli	19.129300	72.843400	Cafe Arfa	19.128930	72.847140	Indian Restaurant
2	Amboli	19.129300	72.843400	Subway	19.127860	72.844461	Sandwich Place
3	Amboli	19.129300	72.843400	Cafe Coffee Day	19.127748	72.844663	Coffee Shop
4	Amboli	19.129300	72.843400	V33	19.129068	72.843670	Gym
5	Amboli	19.129300	72.843400	Delhi Zaika	19.132159	72.844406	Halal Restaurant
6	Amboli	19.129300	72.843400	Nukkad Food Bistro	19.126058	72.846618	Fast Food Restaurant
7	Chakala\tAndheri,	19.111388	72.860833	Courtyard Mumbai International Airport	19.114167	72.864131	Hotel
8	Chakala\tAndheri,	19.111388	72.860833	Faaso's	19.113938	72.862330	Fast Food Restaurant
9	Chakala\tAndheri,	19.111388	72.860833	Cafe Coffee Day	19.112272	72.861106	Café

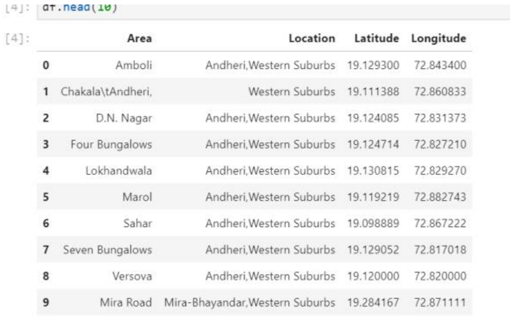
METHODOLOGY

Exploratory data analysis used- *Groupby(), describe()*

Inferential Statistics used- *Bar Chart*

Machine Learning Algorithm used- *K-Means Clustering*

1. First, we import all the libraries required for this project.
2. Next, we use Beautiful Soup package and requests to scrape the data from the web. The neighborhoods of Mumbai along with their latitudes and longitudes were found on Wikipedia.
3. Next, we extract the required information into a dataframe.

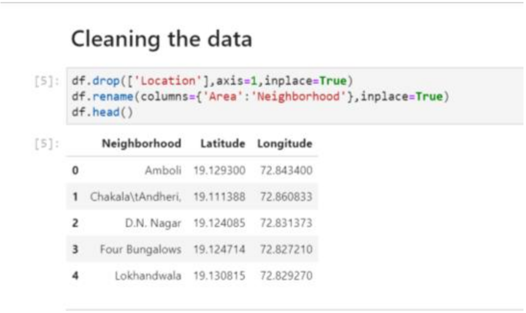


```
[4]: df.head(10)
```

[4]:

	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

4. After the dataframe is created, we clean it for accurate results.



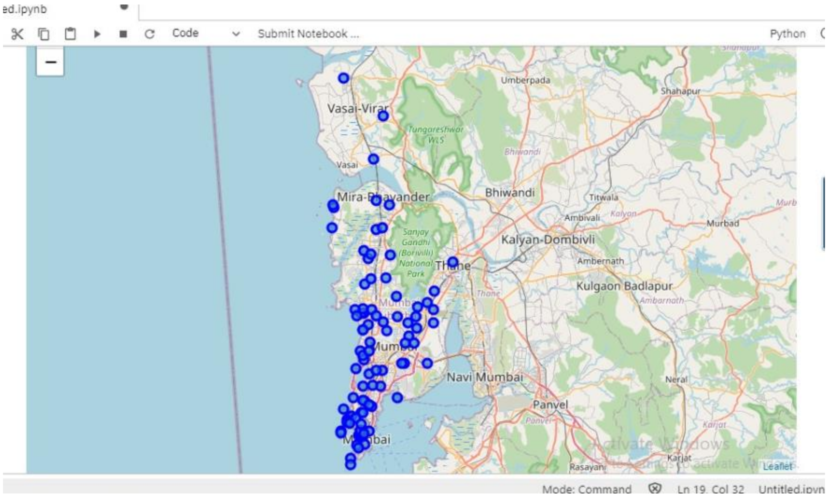
Cleaning the data

```
[5]: df.drop(['Location'],axis=1,inplace=True)
     df.rename(columns={'Area':'Neighborhood'},inplace=True)
     df.head()
```

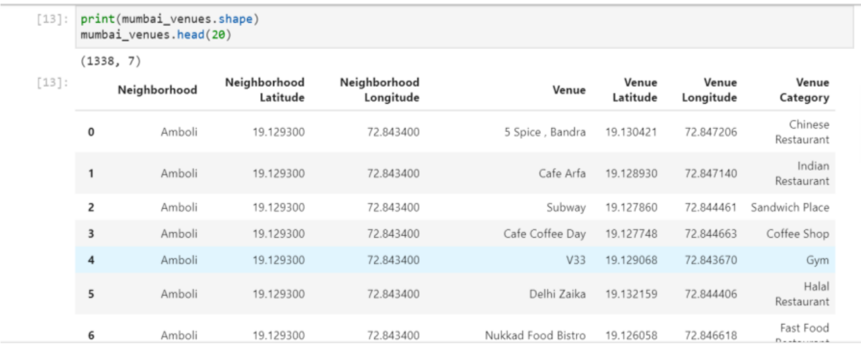
[5]:

	Neighborhood	Latitude	Longitude
0	Amboli	19.129300	72.843400
1	Chakala\tAndheri,	19.111388	72.860833
2	D.N. Nagar	19.124085	72.831373
3	Four Bungalows	19.124714	72.827210
4	Lokhandwala	19.130815	72.829270

5. We use Folium to plot the map of Mumbai and display the neighborhoods on the map.



6. Then we use the Foursquare API to fetch data of venues.



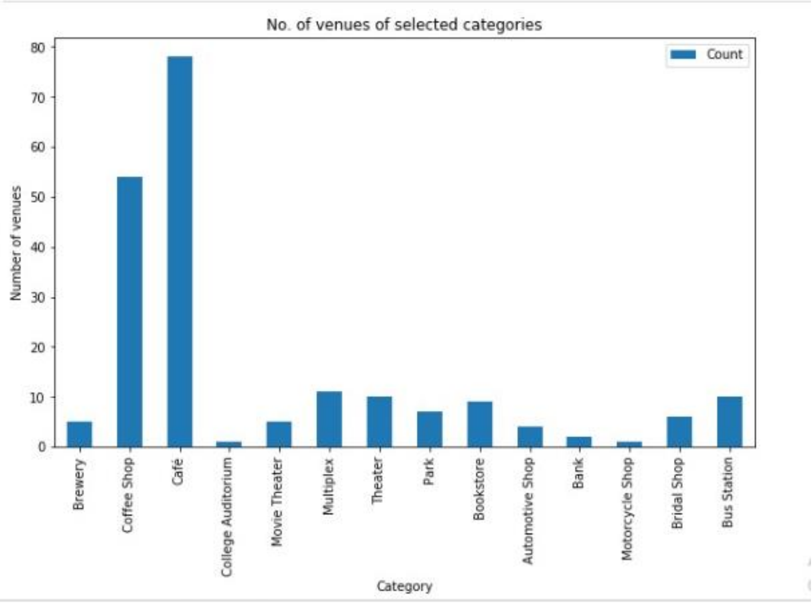
```
[13]: print(mumbai_venues.shape)
      mumbai_venues.head(20)

(1338, 7)
```

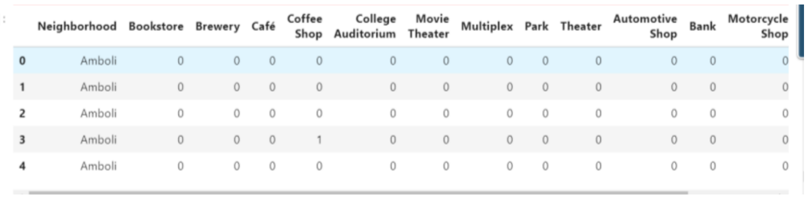
[13]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Amboli	19.129300	72.843400	5 Spice , Bandra	19.130421	72.847206	Chinese Restaurant
1	Amboli	19.129300	72.843400	Cafe Arfa	19.128930	72.847140	Indian Restaurant
2	Amboli	19.129300	72.843400	Subway	19.127860	72.844461	Sandwich Place
3	Amboli	19.129300	72.843400	Cafe Coffee Day	19.127748	72.844663	Coffee Shop
4	Amboli	19.129300	72.843400	V33	19.129068	72.843670	Gym
5	Amboli	19.129300	72.843400	Delhi Zaika	19.132159	72.844406	Halal Restaurant
6	Amboli	19.129300	72.843400	Nukkad Food Bistro	19.126058	72.846618	Fast Food

7. Analyze different venue categories:
    - First, we create a dataframe categorized by the venues using the groupby() function by the count method to get number of venues in each category.
    - Then, we extract the required categories.
    - Finally, we plot the bar graph to get a better idea of the different venue categories.

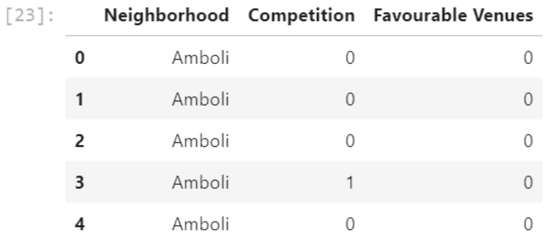


8. The bar graph depicts the different categories we are going to use and the number of venues in each category. We then proceed to create our dataframe for clustering.
9. We first perform one hot encoding on the dataframe of venues created using Foursquare API by creating dummies and extract the columns required for analysis.



	Neighborhood	Bookstore	Brewery	Café	Coffee Shop	College Auditorium	Movie Theater	Multiplex	Park	Theater	Automotive Shop	Bank	Motorcycle Shop
0	Amboli	0	0	0	0	0	0	0	0	0	0	0	0
1	Amboli	0	0	0	0	0	0	0	0	0	0	0	0
2	Amboli	0	0	0	0	0	0	0	0	0	0	0	0
3	Amboli	0	0	0	1	0	0	0	0	0	0	0	0
4	Amboli	0	0	0	0	0	0	0	0	0	0	0	0

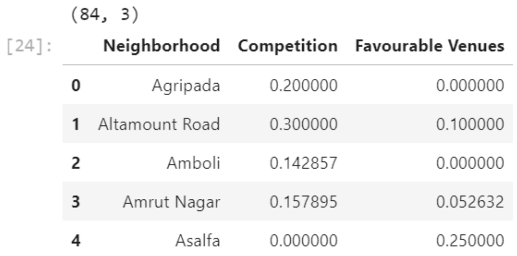
10. After that, we create two columns- one which will include all categories of coffee shops (this will act as our competition) and the other which will consist of places that are likely to attract customers to a coffee shop (the favorable venues which will be beneficial for our shop)



[23]:

	Neighborhood	Competition	Favourable Venues
0	Amboli	0	0
1	Amboli	0	0
2	Amboli	0	0
3	Amboli	1	0
4	Amboli	0	0

11. Then we group the neighborhoods by averaging the rows belonging to the same group.

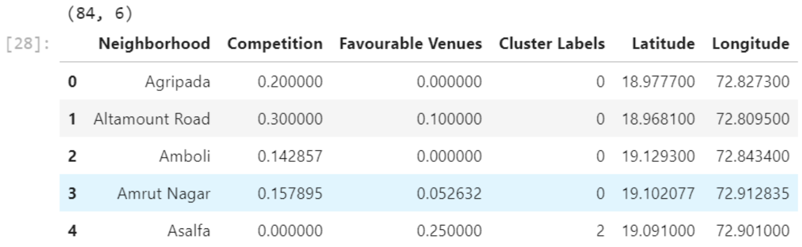


```
(84, 3)
```

[24]:

	Neighborhood	Competition	Favourable Venues
0	Agripada	0.200000	0.000000
1	Altamount Road	0.300000	0.100000
2	Amboli	0.142857	0.000000
3	Amrut Nagar	0.157895	0.052632
4	Asalfa	0.000000	0.250000

12. Finally, we perform k-clustering on the dataframe and assign the cluster labels. We merge this dataframe with the one we created in the beginning.

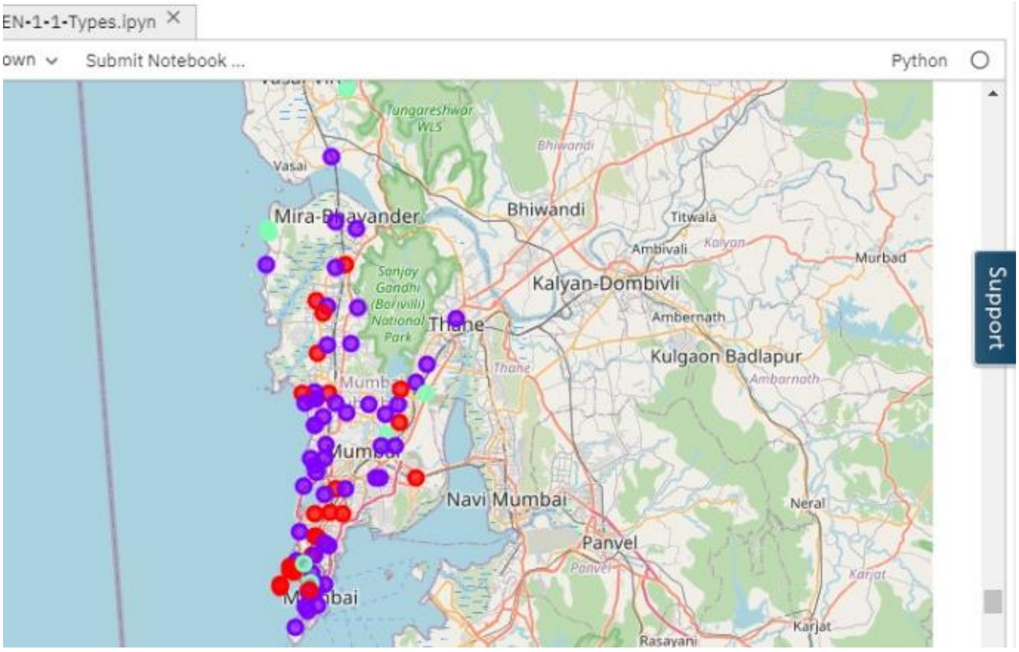


```
(84, 6)
```

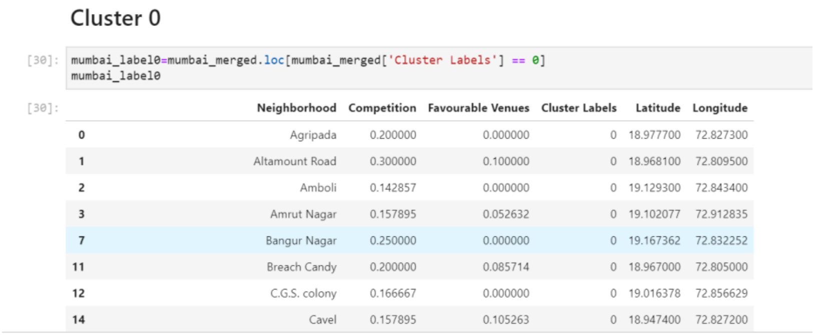
[28]:

	Neighborhood	Competition	Favourable Venues	Cluster Labels	Latitude	Longitude
0	Agripada	0.200000	0.000000	0	18.977700	72.827300
1	Altamount Road	0.300000	0.100000	0	18.968100	72.809500
2	Amboli	0.142857	0.000000	0	19.129300	72.843400
3	Amrut Nagar	0.157895	0.052632	0	19.102077	72.912835
4	Asalfa	0.000000	0.250000	2	19.091000	72.901000

13. We use Folium to plot these clusters on the map of Mumbai.



14. Display the clusters:

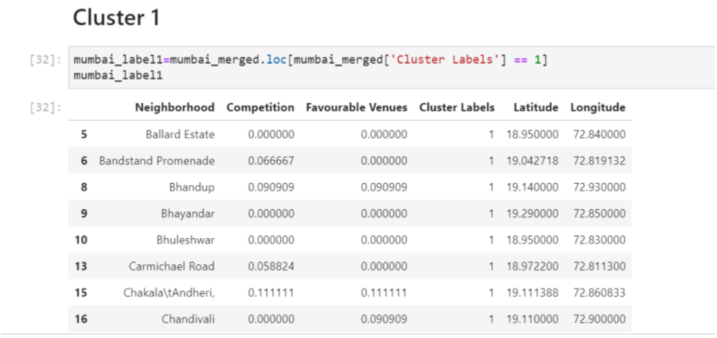


Cluster 0

```
[30]: mumbai_label0=mumbai_merged.loc[mumbai_merged['Cluster Labels'] == 0]
      mumbai_label0
```

[30]:

	Neighborhood	Competition	Favourable Venues	Cluster Labels	Latitude	Longitude
0	Agripada	0.200000	0.000000	0	18.977700	72.827300
1	Altamount Road	0.300000	0.100000	0	18.968100	72.809500
2	Amboli	0.142857	0.000000	0	19.129300	72.843400
3	Amrut Nagar	0.157895	0.052632	0	19.102077	72.912835
7	Bangur Nagar	0.250000	0.000000	0	19.167362	72.832252
11	Breach Candy	0.200000	0.085714	0	18.967000	72.805000
12	C.G.S. colony	0.166667	0.000000	0	19.016378	72.856629
14	Cavel	0.157895	0.105263	0	18.947400	72.827200

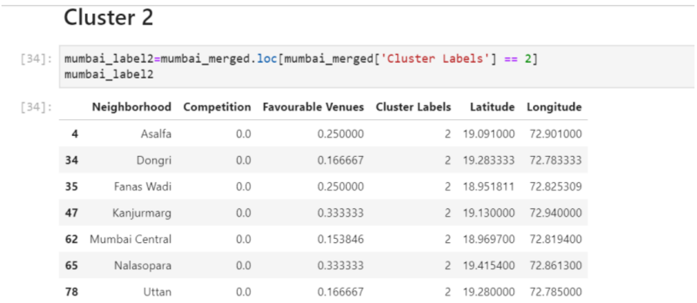


Cluster 1

```
[32]: mumbai_label1=mumbai_merged.loc[mumbai_merged['Cluster Labels'] == 1]
      mumbai_label1
```

[32]:

	Neighborhood	Competition	Favourable Venues	Cluster Labels	Latitude	Longitude
5	Ballard Estate	0.000000	0.000000	1	18.950000	72.840000
6	Bandstand Promenade	0.066667	0.000000	1	19.042718	72.819132
8	Bhandup	0.090909	0.090909	1	19.140000	72.930000
9	Bhayandar	0.000000	0.000000	1	19.290000	72.850000
10	Bhuleshwar	0.000000	0.000000	1	18.950000	72.830000
13	Carmichael Road	0.058824	0.000000	1	18.972200	72.811300
15	Chakala\tAndheri,	0.111111	0.111111	1	19.111388	72.860833
16	Chandivali	0.000000	0.090909	1	19.110000	72.900000



Cluster 2

```
[34]: mumbai_label2=mumbai_merged.loc[mumbai_merged['Cluster Labels'] == 2]
      mumbai_label2
```

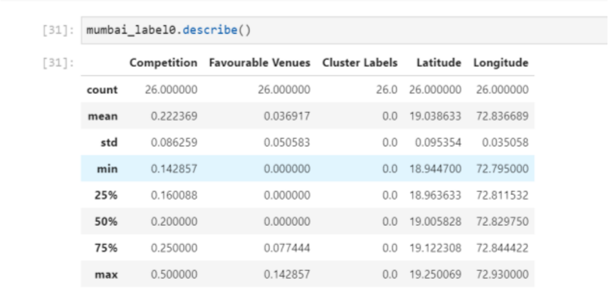
[34]:

	Neighborhood	Competition	Favourable Venues	Cluster Labels	Latitude	Longitude
4	Asalfa	0.0	0.250000	2	19.091000	72.901000
34	Dongri	0.0	0.166667	2	19.283333	72.783333
35	Fanas Wadi	0.0	0.250000	2	18.951811	72.825309
47	Kanjurmarg	0.0	0.333333	2	19.130000	72.940000
62	Mumbai Central	0.0	0.153846	2	18.969700	72.819400
65	Nalasopara	0.0	0.333333	2	19.415400	72.861300
78	Uttan	0.0	0.166667	2	19.280000	72.785000

RESULTS

To examine the results I used the describe() method which displayed the statistics of the clusters. Here is what they look like-

Cluster 0

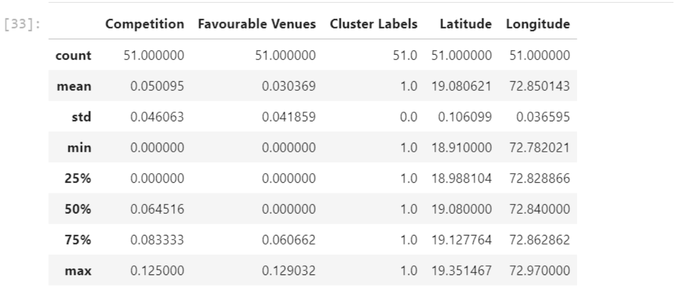


```
[31]: mumbai_label0.describe()
```

[31]:

	Competition	Favourable Venues	Cluster Labels	Latitude	Longitude
count	26.000000	26.000000	26.0	26.000000	26.000000
mean	0.222369	0.036917	0.0	19.038633	72.836689
std	0.086259	0.050583	0.0	0.095354	0.035058
min	0.142857	0.000000	0.0	18.944700	72.795000
25%	0.160088	0.000000	0.0	18.963633	72.811532
50%	0.200000	0.000000	0.0	19.005828	72.829750
75%	0.250000	0.077444	0.0	19.122308	72.844422
max	0.500000	0.142857	0.0	19.250069	72.930000

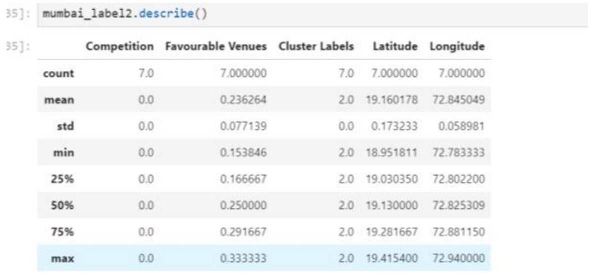
Cluster 1



[33]:

	Competition	Favourable Venues	Cluster Labels	Latitude	Longitude
count	51.000000	51.000000	51.0	51.000000	51.000000
mean	0.050095	0.030369	1.0	19.080621	72.850143
std	0.046063	0.041859	0.0	0.106099	0.036595
min	0.000000	0.000000	1.0	18.910000	72.782021
25%	0.000000	0.000000	1.0	18.988104	72.828866
50%	0.064516	0.000000	1.0	19.080000	72.840000
75%	0.083333	0.060662	1.0	19.127764	72.862862
max	0.125000	0.129032	1.0	19.351467	72.970000

Cluster 2



```
35]: mumbai_label2.describe()
```

35]:

	Competition	Favourable Venues	Cluster Labels	Latitude	Longitude
count	7.0	7.000000	7.0	7.000000	7.000000
mean	0.0	0.236264	2.0	19.160178	72.845049
std	0.0	0.077139	0.0	0.173233	0.058981
min	0.0	0.153846	2.0	18.951811	72.783333
25%	0.0	0.166667	2.0	19.030350	72.802200
50%	0.0	0.250000	2.0	19.130000	72.825309
75%	0.0	0.291667	2.0	19.281667	72.881150
max	0.0	0.333333	2.0	19.415400	72.940000

Now, from the mean of the competition and favorable venues of the three clusters we can observe the following-:

1. Cluster 0 has average competition 0.22 and favorable venues 0.03

2. Cluster 1 has average competition 0.05 and favorable venues 0.03

3. Cluster 2 has average competition 0.00 and favorable venues 0.23

DISCUSSIONS

Mumbai is the financial, commercial and entertainment capital of India which makes it the perfect choice for setting up a coffee shop business. It's immense population is a plus point. Through this project I have tried to examine some factors that can affect the coffee shop business. A location having less competition and more venues to attract people to the shop would be the ideal spot.

The bar graph clearly showed that Mumbai is a happening place as it is full of different kind of places that attract foot traffic with coffee shop topping the chart. This shows great possibility of a successful coffee business. The statistics returned by the cluster showed promising results. A cluster was identified to be a possible set of ideal locations.

CONCLUSION

- Cluster 0 has the highest competition.
- Cluster 1 has less competition but the number of favorable venues is also less.
- Cluster 2 has negligible competition and moreover has a fair amount of favorable venues to bring customers into the shop.

So a preferred neighborhood to set up a coffee shop would be one present in Cluster 2. Selecting a location is one of the major factors when setting up the business. However, there are other factors too which need to be taken into consideration for example the cost of living in that area, the population of that area, finding the space that suits the business needs etc. Though this project doesn’t cover every factor but does provide an useful insight which can help the entrepreneurs.