Capstone Project Week 5

New Spanish café on a great city of India

INDEX

1. INTRODUCTION

- The objective of this project is to determine, through the study of the neighborhood data in one of the greatest towns on India, trying to check good locations for the start-up of a Spanish Café.

- We need a multicultural and tourist character, with a great image in gastronomy, considering all the possible aspects, as population, financial, to extract a geographical candidate for our new coffee.

- Searching on data sources, Mumbai is our city for the study. It´s financial, commercial, and the entertainment capital of India. It is also one of the world's top ten centers of commerce in terms of global financial flow, generating 6.16% of India's GDP, and accounting for 25% of industrial output, and 70% of maritime trade in India. Mumbai Port trust over 70% of capital transactions to India´s economy. Mumbai has the eighth highest number of billionaires of any city in the world, and Mumbai's billionaires had the highest average wealth of any city in the world in 2008. The city houses important financial institutions and the corporate headquarters of numerous Indian companies and multinational corporations. It is also home to some of India's premier scientific and nuclear institutes. The city is also home to Bollywood and Marathi cinema industries. Mumbai's business opportunities attract migrants from all over India, and a great group of tourists, for all of this we considered it a great candidate for the study and the startup of a new Spanish Café.

2. DATA RESOURCES & DATA MANAGEMENT

The essential data that we are going to require for the project will be:

2.1. Mumbai neighborhood data source

2.2. Geographical data and coordinates within Mumbai for those neighborhoods

2.3. Data management with recommendations

2.1. Mumbai neighborhood data source

The data of the Mumbai´s neighborhoods was scraped from https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Mumbai. The data is read into a **panda's data frame** using the **read_html_method**. Doing so, is because Wikipedia page provides a group of detailed city tables with data that can be easily scraped using the read HTML method of pandas. Next picture shows a part of this data frame.

	Neighborhood	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
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

2.2. Geographical data and coordinates

- Coordinates for Mumbai has been obtained from the **GeoPy library in python**. This data is relevant for **plotting the map of Mumbai using the Folium library in python**. The code for getting the geographical coordinates of Mumbai is the next picture.

```
In [29]: for i, neigh in enumerate(df['Neighborhood']):
             lat_lng_coords = None

             while(lat_lng_coords is None):
                 g = geocoder.arcgis('{}, Mumbai, India'.format(neigh))
                 lat_lng_coords = g.latlng

             if lat_lng_coords:
                 latitude = lat_lng_coords[0]
                 longitude = lat_lng_coords[1]

             df.loc[i, 'Latitude1'] = latitude
             df.loc[i, 'Longitude1'] = longitude

         df.head(10)
```

Out[29]:

	Neighborhood	Location	Latitude	Longitude	Latitude1	Longitude1
0	Amboli	Western Suburbs	19.1293	72.8464	19.1291	72.8464
1	Chakala, Andheri	Western Suburbs	19.1084	72.8623	19.1084	72.8623
2	D.N. Nagar	Western Suburbs	19.1241	72.8325	19.1251	72.8325
3	Four Bungalows	Western Suburbs	19.1264	72.8242	19.1264	72.8242
4	Lokhandwala	Western Suburbs	19.1432	72.8249	19.1432	72.8249
5	Marol	Western Suburbs	19.1192	72.8827	19.1191	72.8828
6	Sahar	Western Suburbs	19.1027	72.8626	19.1027	72.8626
7	Seven Bungalows	Western Suburbs	19.1291	72.8212	19.1286	72.8212
8	Versova	Western Suburbs	19.1377	72.8135	19.1377	72.8135
9	Mira Road	Western Suburbs	19.2656	72.8711	19.2656	72.8706

- The **geocoder library in python** has been used to obtain **latitude and longitude** data for various neighborhoods in Mumbai. Cleaned coordinates are then further used for plotting neighborhoods using the **Folium library in python**. Next picture shows the coordinates of neighborhoods in Mumbai obtained from Wikipedia source as ‘Latitude’ and ‘Longitude’ and those obtained from **geocoder as ‘Latitude1’ and ‘Longitude1**’.

	Neighborhood	Location	Latitude	Longitude	Latitude1	Longitude1	Latdiff	Longdiff
0	Amboli	Western Suburbs	19.1293	72.8464	19.1291	72.8464	0.00024	0.00304
1	Chakala, Andheri	Western Suburbs	19.1084	72.8623	19.1084	72.8623	0.003028	0.001497
2	D.N. Nagar	Western Suburbs	19.1241	72.8325	19.1251	72.8325	0.000965	0.001107
3	Four Bungalows	Western Suburbs	19.1263	72.8243	19.1263	72.8243	0.001606	0.00288
4	Lokhandwala	Western Suburbs	19.1432	72.8249	19.1432	72.8249	0.012345	0.0044
5	Marol	Western Suburbs	19.1192	72.8827	19.1191	72.8828	0.000169	6.7e-05
6	Sahar	Western Suburbs	19.1027	72.8626	19.1027	72.8626	0.00376476	0.00464166
7	Seven Bungalows	Western Suburbs	19.1315	72.817	19.1315	72.8165	0.00240802	0.000558001
8	Versova	Western Suburbs	19.1377	72.8135	19.1377	72.8135	0.01769	0.00652
9	Mira Road	Western Suburbs	19.2657	72.8711	19.2657	72.8707	0.0184624	0.000418149

- The new picture shows some lines of the **final data frame** after replacing the latitude and longitude values and cleansing unnecessary columns.

	Neighborhood	Location	Latitude	Longitude
0	Amboli	Western Suburbs	19.1293	72.8464
1	Chakala, Andheri	Western Suburbs	19.1084	72.8623
2	D.N. Nagar	Western Suburbs	19.1241	72.8325
3	Four Bungalows	Western Suburbs	19.1263	72.8243
4	Lokhandwala	Western Suburbs	19.1432	72.8249
5	Marol	Western Suburbs	19.1192	72.8827
6	Sahar	Western Suburbs	19.1027	72.8626
7	Seven Bungalows	Western Suburbs	19.1315	72.817
8	Versova	Western Suburbs	19.1377	72.8135
9	Mira Road	Western Suburbs	19.2657	72.8711

2.3. Data management with recommendations

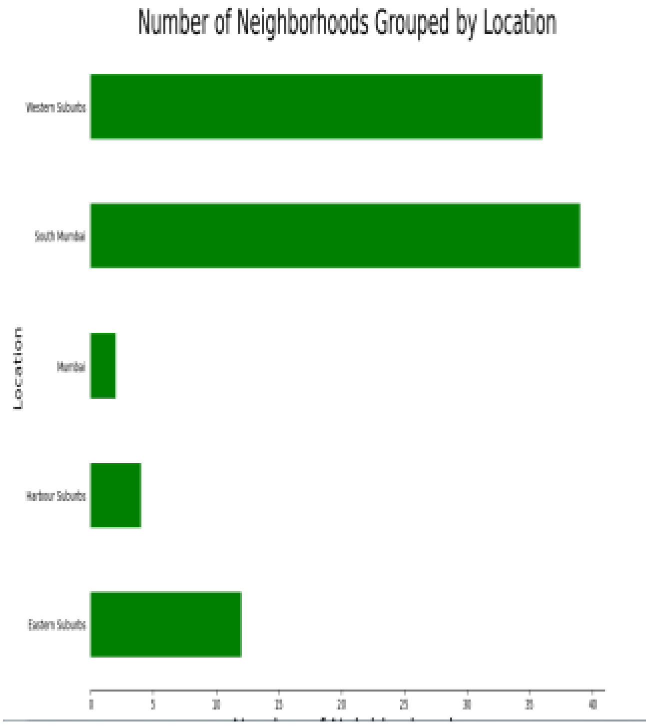
- The recommendations data has been extracted using the **Foursquare API.** This data contains recommendations for all neighborhoods in Mumbai and is used to study the popular venues of different neighborhoods **as well as build the unsupervised learning model to cluster neighborhoods**. Next figure shows some results using **Foursquare API.**

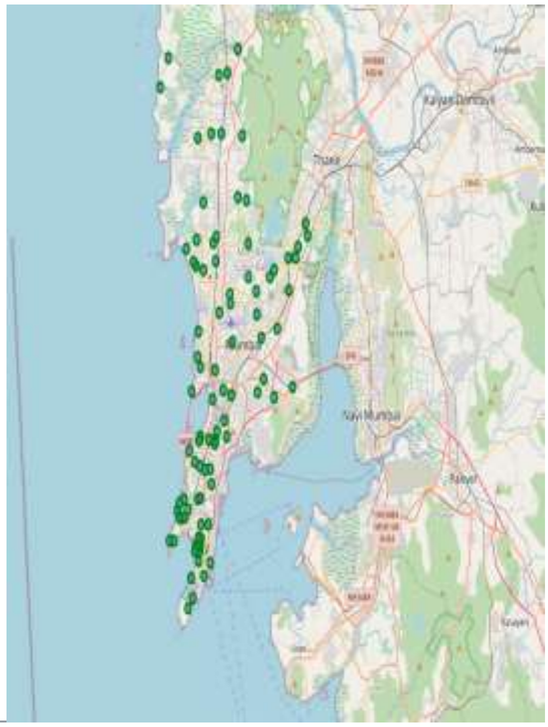
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Amboli	19.1293	72.84644	Cafe Arfa	19.128930	72.847140	Indian Restaurant
1	Amboli	19.1293	72.84644	5 Spice , Bandra	19.130421	72.847206	Chinese Restaurant
2	Amboli	19.1293	72.84644	Shawarma Factory	19.124591	72.840398	Falafel Restaurant
3	Amboli	19.1293	72.84644	Jaffer Bhai's Delhi Darbar	19.137714	72.845909	Mughlai Restaurant
4	Amboli	19.1293	72.84644	Narayan Sandwich	19.121398	72.850270	Sandwich Place
5	Amboli	19.1293	72.84644	Persia Darbar	19.136952	72.846822	Indian Restaurant
6	Amboli	19.1293	72.84644	Domino's Pizza	19.131000	72.848000	Pizza Place
7	Amboli	19.1293	72.84644	Garden Court	19.127188	72.837478	Indian Restaurant
8	Amboli	19.1293	72.84644	Subway	19.127860	72.844461	Sandwich Place
9	Amboli	19.1293	72.84644	Sarvodaya Veg. Restaurant	19.123760	72.850893	Indian Restaurant

3. METODOLOGY

3.1. VISUALIZATION

Visualization was carried out, see left picture shows with a **bar plot** depicting the number of neighborhoods in each location in Mumbai.
In the bar plot we can observe that South Mumbai and Western Suburbs have the greatest number of neighborhoods. Then using **folium**, a map was plotted to show how the different neighborhoods how are spread across Mumbai. This is shown in right picture.





- <u>3.2. FEATURE EXTRACTION</u>

- Feature extraction was carried out to obtain features from the **Foursquare API data** (as shown in Figure 5) which was used for building the unsupervised learning model. To achieve this, the “Venue Category” column had to be converted to some form of numeric value to be used for building the model.

- This was achieved by the **One-hot Encoding method** which takes all the unique categories and creates a column for each category.

- This process was repeated for all venues in all neighborhoods and the result was a sparse matrix containing the neighborhood name and all unique category columns. This data frame was then grouped by the neighborhood name and the average value was taken for all categories. The result is shown in next picture.

	Neighborhood	ATM	Accessories Store	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Arcade	Art Gallery	Arts & Crafts Store	...	Trail	Train	Train Station	Vegetarian / Vegan Restaurant	Whisky Bar	Wine Bar	Wine Shop	Women's Store	Yoga Studio	Zoo
0	Amboli	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000	0.000000	0.0	0.0
1	Chakala, Andheri	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.047619	0.0	0.0	0.000	0.000000	0.0	0.0
2	D.N. Nagar	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.043478	0.0	0.0	0.000	0.021739	0.0	0.0
3	Four Bungalows	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.030303	0.0	0.0	0.000	0.015152	0.0	0.0
4	Lokhandwala	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.010753	0.0	0.0	0.000	0.010753	0.0	0.0
5	Marol	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000	0.000000	0.0	0.0
6	Sahar	0.0	0.0	0.033333	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000	0.000000	0.0	0.0
7	Seven Bungalows	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.014925	...	0.0	0.0	0.0	0.029851	0.0	0.0	0.000	0.000000	0.0	0.0
8	Versova	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.025000	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.025	0.000000	0.0	0.0
9	Mira Road	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.0	0.0	0.0	0.000000	0.0	0.0	0.000	0.066667	0.0	0.0

10 rows × 221 columns

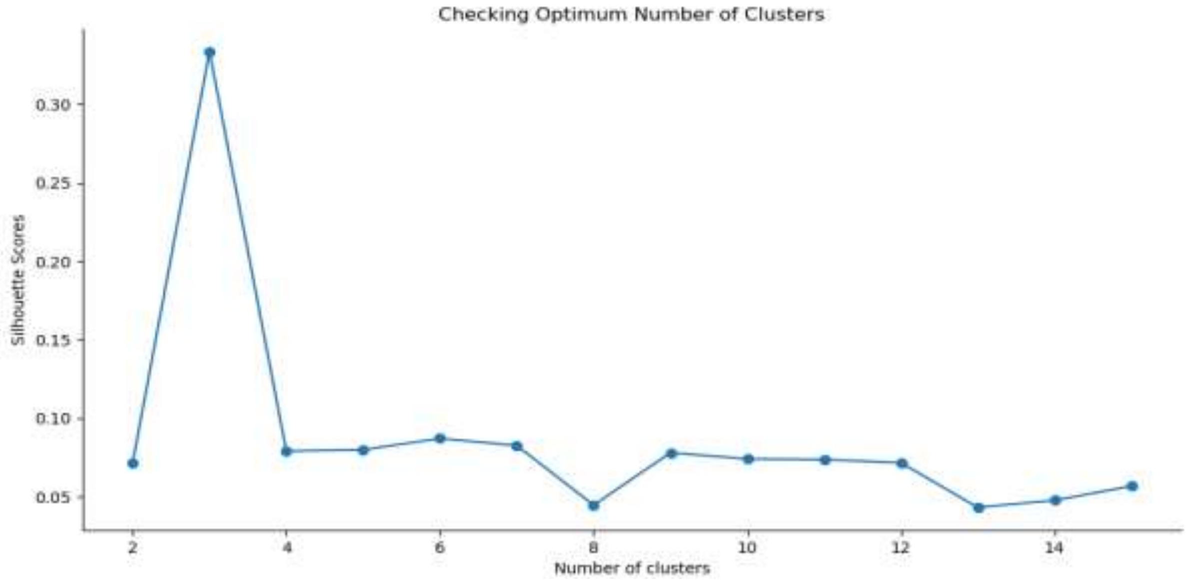
- Notice that most of the values are 0 since there were many unique categories and not all neighborhoods had venues belonging to each category. This data was used for the **unsupervised learning model** with the neighborhood name dropped. The unsupervised learning model is explained in the next section.

- A data frame was also created which contained the top 10 most common venues of all neighborhoods. Though this is not a part of Feature Extraction, it is important to provide a glimpse into what this data frame looks like as it will be used later to combine the results from the unsupervised learning model. The top 10 rows of this data frame are shown in the next picture.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Amboli	Indian Restaurant	Coffee Shop	Bakery	Bar	Asian Restaurant	Pizza Place	Sandwich Place	Bowling Alley	Bus Station	Bike Rental / Bike Share
1	Chakala, Andheri	Hotel	Indian Restaurant	Café	Fast Food Restaurant	Pizza Place	Asian Restaurant	Hotel Bar	Vegetarian / Vegan Restaurant	Restaurant	Gym
2	D.N. Nagar	Bar	Indian Restaurant	Pub	Gym / Fitness Center	Pizza Place	Lounge	Coffee Shop	Vegetarian / Vegan Restaurant	Snack Place	Gym
3	Four Bungalows	Pub	Café	Indian Restaurant	Gym / Fitness Center	Chinese Restaurant	Bar	Seafood Restaurant	Lounge	Vegetarian / Vegan Restaurant	Coffee Shop
4	Lokhandwala	Indian Restaurant	Chinese Restaurant	Café	Pub	Bakery	Bar	Italian Restaurant	Gym / Fitness Center	Coffee Shop	Asian Restaurant
5	Marol	Indian Restaurant	Hotel	Diner	Bakery	Dance Studio	Ice Cream Shop	Chinese Restaurant	Fast Food Restaurant	Restaurant	Lounge
6	Sahar	Hotel	Café	Indian Restaurant	Lounge	Gym	Asian Restaurant	Pizza Place	Seafood Restaurant	Restaurant	Falafel Restaurant
7	Seven Bungalows	Café	Pub	Seafood Restaurant	Chinese Restaurant	Pizza Place	Coffee Shop	Bar	Ice Cream Shop	Asian Restaurant	Bistro
8	Versova	Café	Ice Cream Shop	Beach	Pizza Place	Coffee Shop	Chinese Restaurant	Salon / Barbershop	Frozen Yogurt Shop	Bistro	Sandwich Place
9	Mira Road	Indian Restaurant	Convenience Store	Coffee Shop	Mexican Restaurant	Fast Food Restaurant	Food Truck	Motorcycle Shop	Movie Theater	Basketball Court	Bar

3.3 Machine Learning

- **K-means unsupervised learning technique** was used to cluster the neighborhoods based on the category of venues near the neighborhoods. One important aspect of the k-means model is to determine the number of clusters to use in model development. This was determined by the Silhouette score which was calculated for a range of clusters from 2 to 15. The resulting number of clusters and their respective Silhouette scores are shown in the next picture.



4. **RESULTS**

The Silhouette scores are not extremely high even as the number of clusters increases. This means that the inter-cluster distance is not extremely high over the range of k-values. Despite this, the data will be clustered to the best possible extent. The clustering model then clusters the neighborhoods in Mumbai and provides a label for each neighborhood which is representative of the cluster it belongs to. Coordinates and more date were added to the clustering table for a total representation as shown on the next picture.

	Neighborhood	Location	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Amboli	Western Suburbs	19.1293	72.8464	1	Indian Restaurant	Coffee Shop	Bakery	Bar	Asian Restaurant	Pizza Place	Sandwich Place	Bowling Alley	Bus Station	Bike Rental / Bike Share
1	Chakala, Andheri	Western Suburbs	19.1084	72.8623	1	Hotel	Indian Restaurant	Café	Fast Food Restaurant	Pizza Place	Asian Restaurant	Hotel Bar	Vegetarian / Vegan Restaurant	Restaurant	Gym
2	D.N. Nagar	Western Suburbs	19.1241	72.8325	0	Bar	Indian Restaurant	Pub	Gym / Fitness Center	Pizza Place	Lounge	Coffee Shop	Vegetarian / Vegan Restaurant	Snack Place	Gym
3	Four Bungalows	Western Suburbs	19.1263	72.8243	0	Pub	Café	Indian Restaurant	Gym / Fitness Center	Chinese Restaurant	Bar	Seafood Restaurant	Lounge	Vegetarian / Vegan Restaurant	Coffee Shop
4	Lokhandwala	Western Suburbs	19.1432	72.8249	0	Indian Restaurant	Chinese Restaurant	Café	Pub	Bakery	Bar	Italian Restaurant	Gym / Fitness Center	Coffee Shop	Asian Restaurant
5	Marol	Western Suburbs	19.1192	72.8827	1	Indian Restaurant	Hotel	Diner	Bakery	Dance Studio	Ice Cream Shop	Chinese Restaurant	Fast Food Restaurant	Restaurant	Lounge
6	Sahar	Western Suburbs	19.1027	72.8626	0	Hotel	Café	Indian Restaurant	Lounge	Gym	Asian Restaurant	Pizza Place	Seafood Restaurant	Restaurant	Falafel Restaurant
7	Seven Bungalows	Western Suburbs	19.1315	72.817	0	Café	Pub	Seafood Restaurant	Chinese Restaurant	Pizza Place	Coffee Shop	Bar	Ice Cream Shop	Asian Restaurant	Bistro
8	Versova	Western Suburbs	19.1377	72.8135	0	Café	Ice Cream Shop	Beach	Pizza Place	Coffee Shop	Chinese Restaurant	Salon / Barbershop	Frozen Yogurt Shop	Bistro	Sandwich Place
9	Mira Road	Western Suburbs	19.2657	72.8711	1	Indian Restaurant	Convenience Store	Coffee Shop	Mexican Restaurant	Fast Food Restaurant	Food Truck	Motorcycle Shop	Movie Theater	Basketball Court	Bar

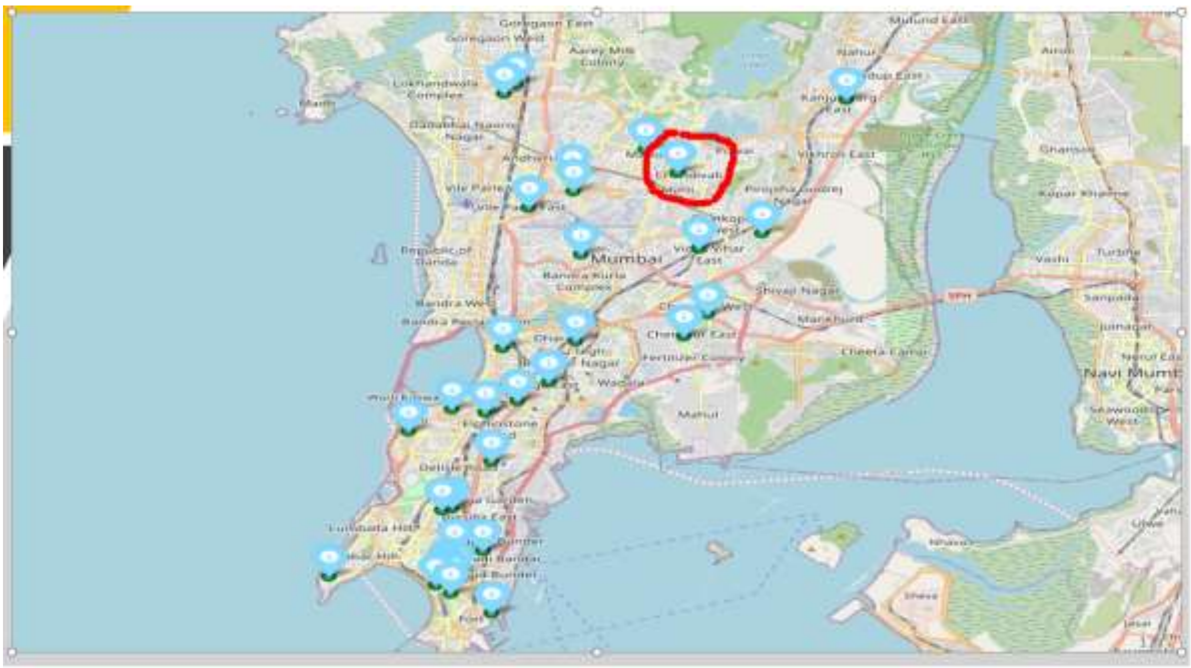
- As result too, furthermore, neighborhoods in each individual cluster can be extracted using cluster labels and thus the details of specific clusters can be seen. This is done below for all clusters with only 10 rows for clusters that contain a high number of neighborhoods.

	Neighborhood	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Chakala, Andheri	Western Suburbs	Hotel	Indian Restaurant	Café	Hotel Bar	Asian Restaurant	Pizza Place	Vegetarian / Vegan Restaurant	Restaurant	Burger Joint	Multiplex
6	Sahar	Western Suburbs	Hotel	Indian Restaurant	Restaurant	Gym	Asian Restaurant	Bar	Coffee Shop	Café	Italian Restaurant	Pub
27	Khar Danda	Western Suburbs	Hotel	Clothing Store	Park	Coffee Shop	Dessert Shop	Bookstore	Bistro	French Restaurant	Boutique	Pool
40	Kanjurmarg	Eastern Suburbs	Train Station	Gym	Hotel	Gift Shop	Chinese Restaurant	French Restaurant	Asian Restaurant	Multiplex	Donut Shop	Electronics Store
70	Malabar Hill	South Mumbai	Gym	Hotel	Park	Convenience Store	Lighthouse	Coffee Shop	Dessert Shop	Indian Restaurant	Cupcake Shop	Cosmetics Shop
77	Walkeshwar	South Mumbai	Gym	Park	Hotel	Convenience Store	Food & Drink Shop	Food Truck	Lighthouse	Restaurant	Dessert Shop	Coffee Shop

5. DISCUSSION

- Based on the clusters shown above, the neighborhoods can once again be plotted on a map of Mumbai, however, this time with different color markers to distinguish between different clusters.

- By analyzing the clusters obtained we can see that some of the clusters are more suited for coffees and hotels, whereas other clusters are less suited. Neighborhoods in clusters 3, 4, and 5 contain a small percentage of coffees, hotels, cafe, and pubs in their top 10 common venues. These clusters contain a higher degree of other venues like train station, bus station, fish market, gym, performing arts venue and smoke shop, to name a few. Thus, they are not well suited for opening a new coffee. **Comparing clusters 1 and 2, neighborhoods in cluster 1 seem to be more suited for starting a coffee** since they contain a larger percentage of food joints in the top 10 most common venues than cluster 2. Recommended to open on cluster 1 as shown in the next picture.



6. CONCLUSION

- In this project, the neighborhoods in Mumbai, India have been successfully analyzed for determining which would be the best neighborhoods for opening a new Spanish coffee. Based on the analysis carried out, neighborhoods in cluster 1 are recommended as locations for the new coffee. This has also been plotted in the map in previous picture. The stakeholders could extract conclusions about the site, between other factors not specifically studied on this project, as economic or administrative factors.