Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Shopping Mall in Mumbai, India

By Monu
July 2021

Introduction

Mumbai is the financial capital of India and is one of the most densely populated cities in the world. It lies on the west coast of India and attracts heavy tourism from all over the globe every year. Personally, I have been brought up in Mumbai and have loved the city from the bottom of my heart. It is one of the major hubs of the world and is extremely diverse with people from various ethnicities residing here. The multi-cultural nature of the city of Mumbai has brought along with it numerous cuisines from all over the world. The people of India generally love food and I personally love to try different cuisines and experience different flavors. Thus, the aim of this project is to study the neighborhoods in Mumbai to determine possible locations for opening a restaurant. This project can be useful for business owners and entrepreneurs who are looking to invest in a restaurant in Mumbai. The main objective of this project is to carefully analyze appropriate data and find recommendations for the stakeholders.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Mumbai, India to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question.

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in the city of Mumbai, India. This project is timely as the city is currently suffering from oversupply of shopping malls.

Data

The data required for this project has been collected from multiple sources. A summary of the data required for this project is given below.

- The data of the neighborhoods in Mumbai was scraped from https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai. The data is read into a pandas data frame using the read_html() method. The main reason for doing so is that the Wikipedia page provides a comprehensive and detailed table of the data which can easily be scraped using the read_html() method of pandas.

- The geographical coordinates for Mumbai data 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 geocoder library in python has been used to obtain latitude and longitude data for various neighborhoods in Mumbai. The coordinates of all neighborhoods in Mumbai are used to check the accuracy of coordinates given on Wikipedia and replace them in our data frame if the absolute difference is more than 0.001. These coordinates are then further used for plotting using the Folium library in python.
- The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all neighborhoods in Mumbai and is used to study the popular venues of different neighborhoods.

Data Wrangling

Lets look at the different values for Location present in the Location column.

```
South Mumbai                    30
Andheri,Western Suburbs          8
Western Suburbs                  6
Eastern Suburbs                  4
Bandra,Western Suburbs           3
Kandivali West,Western Suburbs   3
Ghatkopar,Eastern Suburbs        3
Mira-Bhayandar,Western Suburbs   3
Powai,Eastern Suburbs            3
Borivali (West),Western Suburbs  2
Kalbadevi,South Mumbai           2
Mumbai                           2
Harbour Suburbs                  2
Goregaon,Western Suburbs         2
Vasai,Western Suburbs            2
Khar,Western Suburbs             2
Malad,Western Suburbs            2
Sanctacruz,Western Suburbs       1
Dadar,South Mumbai               1
Antop Hill,South Mumbai          1
Govandi,Harbour Suburbs          1
Fort,South Mumbai                1
```

We can see that there are many locations that appear only once or twice. This is because the main locations like "Western Suburbs" or "South Mumbai" are being further divided by the area within these locations. Lets clean the Location column to make it easier to understand. Although the data we gathered contained latitude and longitude information, we can reconfirm these coordinates using Geocoder. We can create new columns to see the difference between coordinate values obtained from wikipedia and those obtained from geocoder. We will take the absolute difference between these values and store them in our dataframe.

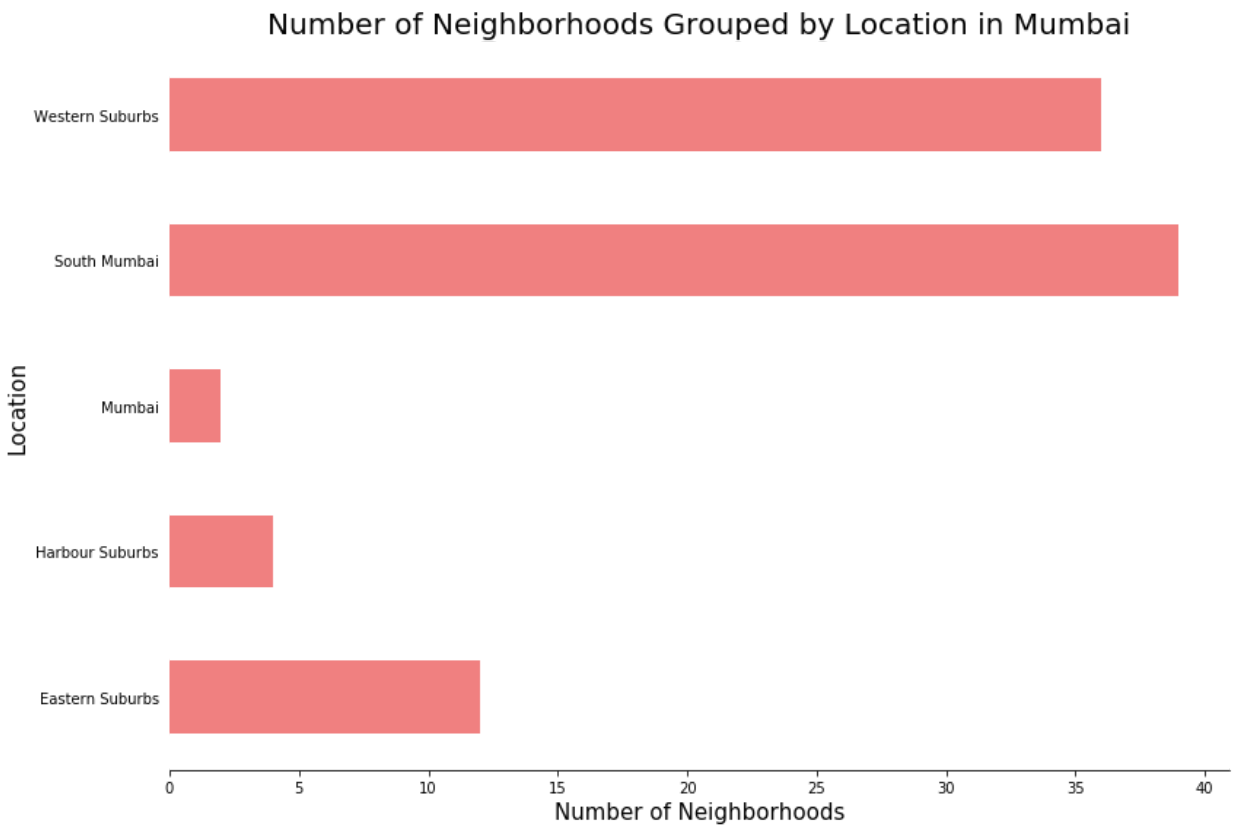
	Neighborhood	Location	Latitude	Longitude	Latitude1	Longitude1	Latdiff	Longdiff
0	Amboli	Western Suburbs	19.129300	72.843400	19.1291	72.8464	0.00024	0.00304
1	Chakala, Andheri	Western Suburbs	19.111388	72.860833	19.1084	72.8623	0.003028	0.001497
2	D.N. Nagar	Western Suburbs	19.124085	72.831373	19.1251	72.8325	0.000965	0.001107
3	Four Bungalows	Western Suburbs	19.124714	72.827210	19.1263	72.8243	0.001606	0.00288
4	Lokhandwala	Western Suburbs	19.130815	72.829270	19.1432	72.8249	0.012345	0.0044

5	Marol	Western Suburbs	19.119219	72.882743	19.1191	72.8828	0.000169	6.7e-05
6	Sahar	Western Suburbs	19.098889	72.867222	19.1027	72.8626	0.00376476	0.00464166
7	Seven Bungalows	Western Suburbs	19.129052	72.817018	19.1315	72.8165	0.00240802	0.000558001
8	Versova	Western Suburbs	19.120000	72.820000	19.1377	72.8135	0.01769	0.00652
9	Mira Road	Western Suburbs	19.284167	72.871111	19.2657	72.8707	0.0184624	0.000418149

We can see that the latitude and longitudes from wikipedia and geocoder are very similar, yet there are some differences. We will replace the values with the coordinates obtained from geocoder if the absolute difference is more than 0.001.

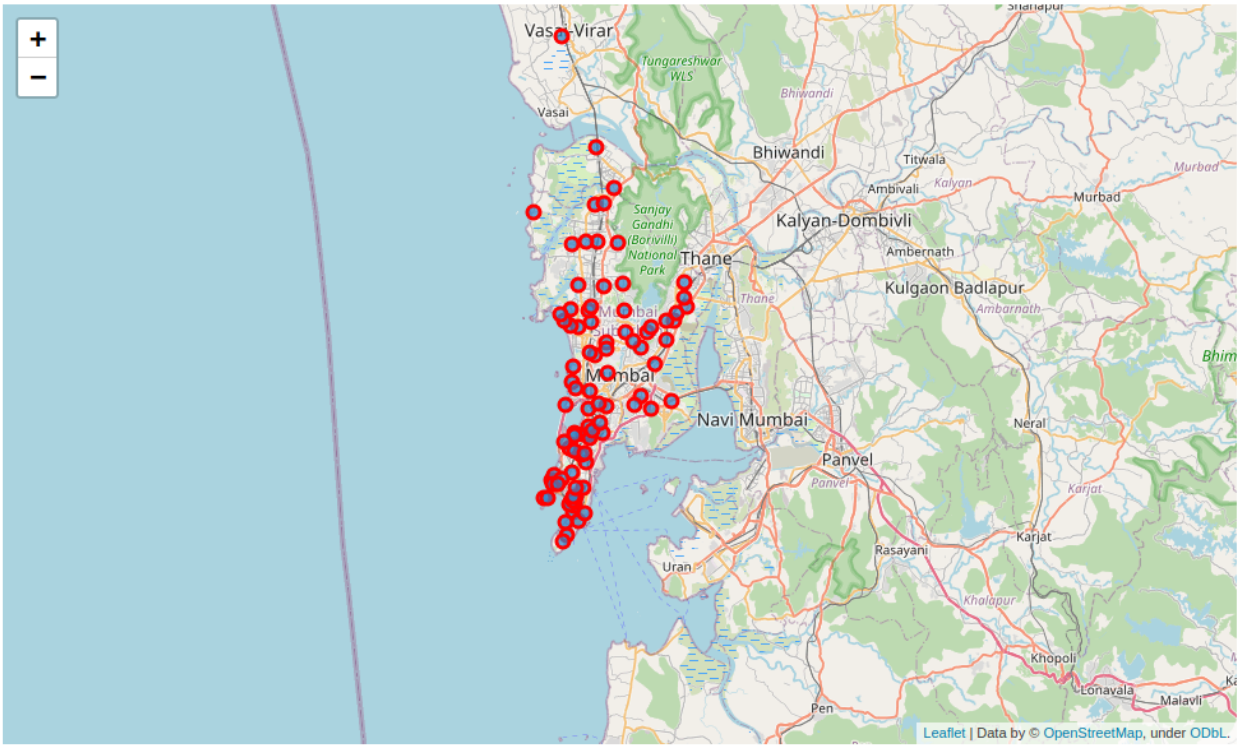
Data Visualization

To understand our data better, we can see how many neighborhoods are in each location.



Clearly we can see that South Mumbai and Western Suburbs have the most number of neighborhoods. Notice how we see one of the locations as Mumbai itself? This is because the neighborhoods contained in this location are located at the outskirts of Mumbai and thus have been grouped as just Mumbai.

Now lets visualize the neighborhoods on a map using Folium. First we will obtain the geographical coordinates of Mumbai using GeoPy.



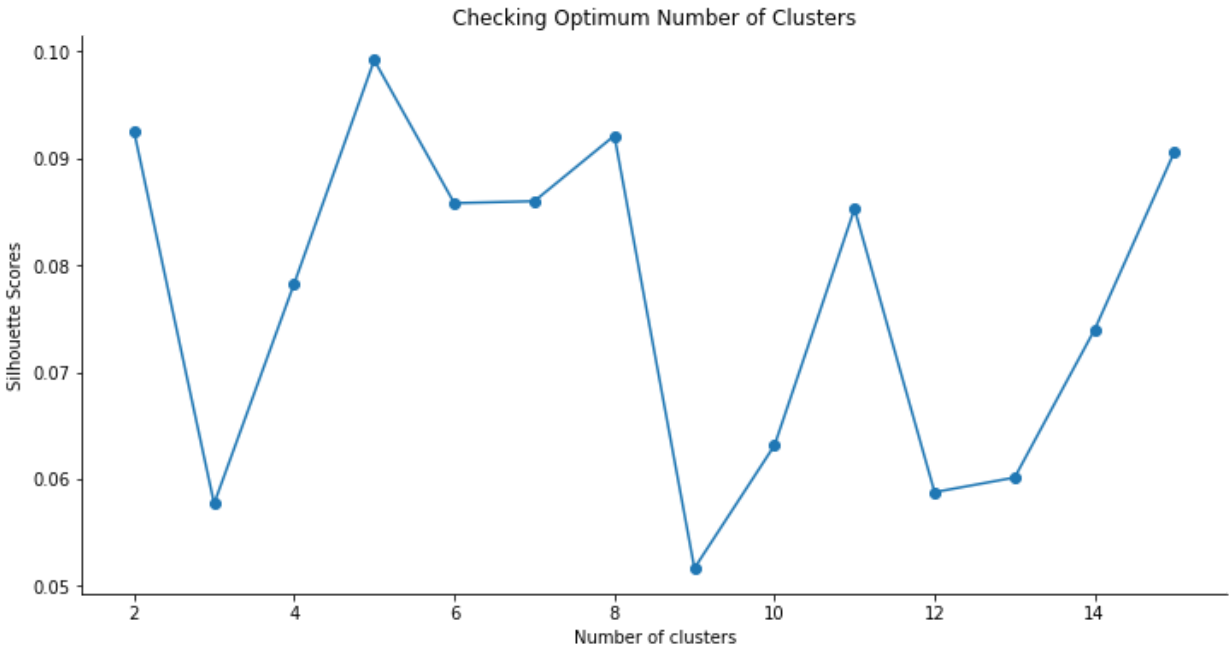
Analyzing each neighborhood

Now we can start working with the Foursquare API to obtain venue recommendations. We can start analyzing each neighborhood by One-hot Encoding to see which categories belong in which neighborhoods. We can groupby neighborhood and take the mean for all categories.

	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 Ba
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
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
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
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
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
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
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
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
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
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

Clustering neighborhoods

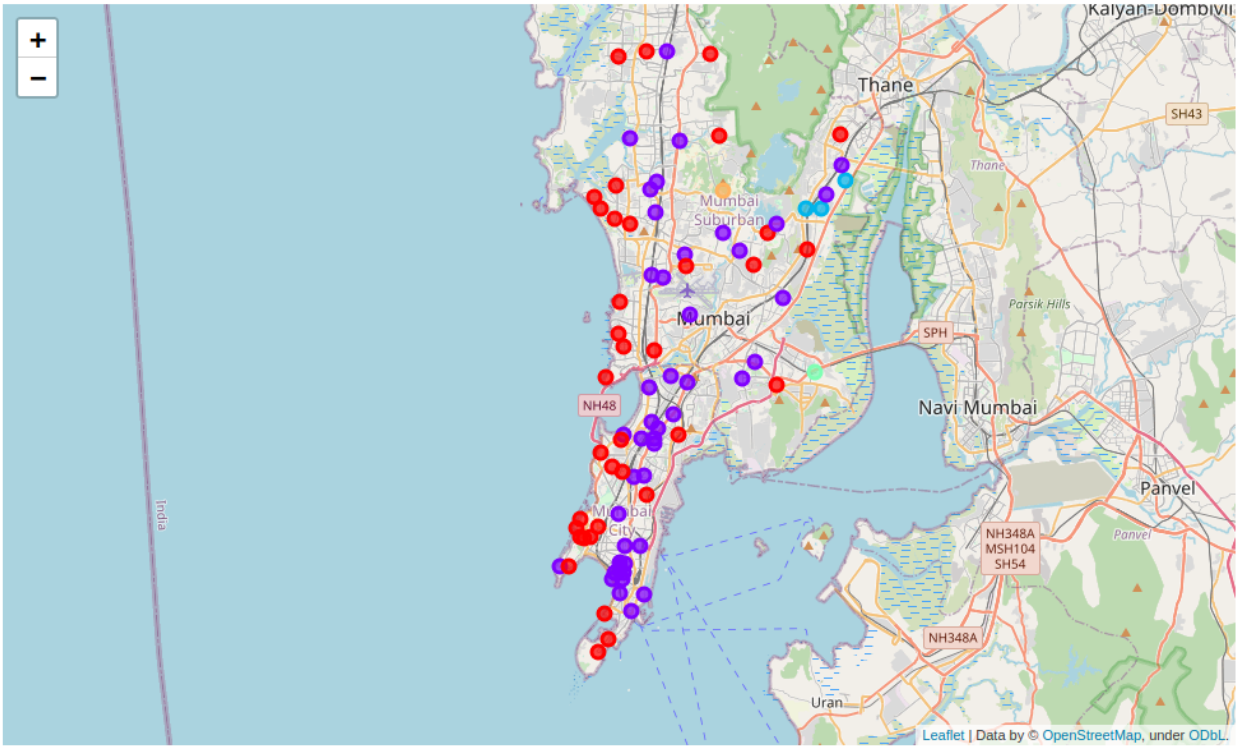
Now we can use KMeans clustering method to cluster the neighborhoods. First we need to determine how many clusters to use. This will be done using the Silhouette Score. We will define a function to plot the Silhouette Score that will be calculated using a different number of clusters. We can now display the scores for different number of clusters and plot the data as well.



We can see that the silhouette scores are not very high even as we increase the number of clusters. This means that the inter-cluster distance between different clusters is not very high over the range of k-values. However, we will try to cluster our data as best as we can. For this, we will use 5 clusters for our clustering model since it provides the highest silhouette score as seen above.

Now we can create a new dataframe that includes cluster labels and the top 10 venues.

We can visualize the clustering by creating a map.

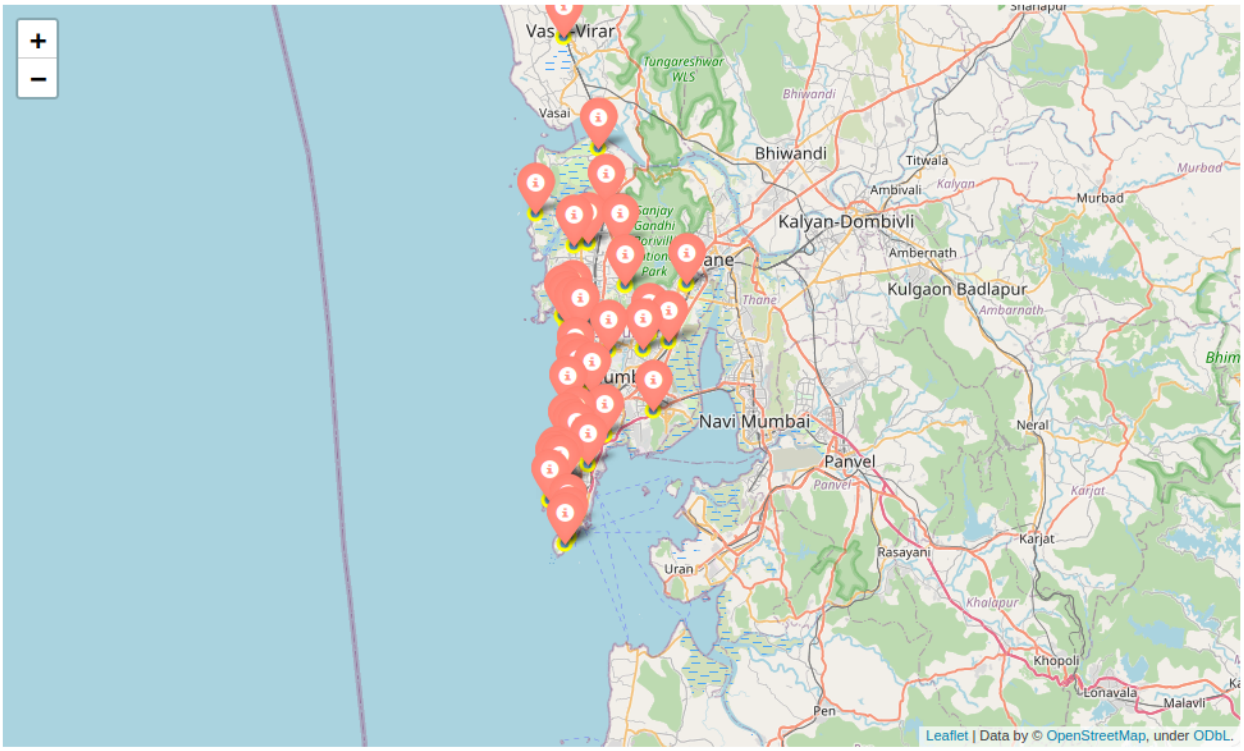


Results and Discussion

By analyzing the five clusters obtained we can see that some of the clusters are more suited for restaurants and hotels, whereas, other clusters are less suited. Neighborhoods in clusters 3, 4, and 5 contain a small percentage of restaurants, 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 restaurant. On the other hand, neighborhoods in clusters 1 and 2 contain a much higher degree of restaurants, hotels, multiplex, cafes, bars and other food joints. Thus, the neighborhoods in these clusters would be well suited for opening a new restaurant.

Comparing clusters 1 and 2, neighborhoods in cluster 1 seem to be more suited for starting a restaurant since they contains a larger percentage of food joints in the top 10 most common venues than cluster 2. The neighborhoods in cluster 1 contain a variety of food joints like restaurants, tea rooms, bakery, cafe, steakhouse and pubs and also contain very diverse cuisines like Japanese, Indian, Chinese, Italian and seafood

restaurants. Most neighborhoods in cluster 2 seem to have Indian Restaurant as their top most common venue; however, on careful analysis we can see that neighborhoods in cluster 2 also contain other venues like soccer field, flea market, smoke shop, gym, train station, dance studio, music store, cosmetics shop and so on. Thus, it is recommended that the new restaurant can be opened in the neighborhoods belonging to cluster 1. This neighborhood can be further plotted on a map as shown below



Conclusion

We have successfully analyzed the neighborhoods in Mumbai, India for determining which would be the best neighborhoods for opening a new restaurant. Based on our analysis, neighborhoods in cluster 1 are recommended as locations for the new restaurant. This has also been plotted in the map above. The stakeholders and investors can further tune this by considering various other factors like transport, legal requirements, and costs associated. These were out of the scope for this project and thus were not considered.

THANK YOU