Opening A New Shopping Mall in Mumbai, Maharashtra

COURSERA CAPSTONE
IBM APPLIED DATA SCIENCE CAPSTONE
Aman Mittal | May 2020

Introduction

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of Mumbai and many more are being built. Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

BUISNESS PROBLEM

The objective of this capstone project is to analyze and select the best locations in the city of Mumbai, Maharashtra 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: In the city of Mumbai, Maharashtra, if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

TARGET AUDIENCE

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in the capital city of Maharashtra i.e. Mumbai. This project is timely as the city is currently suffering from oversupply of shopping mall. The newspaper *Economic Times* also reported in March last year that the true occupancy rates in malls may be as low as 40 per cent in some areas, quoting a Financial Times (FT) article cataloguing the country's continued obsession with building more shopping space despite chronic oversupply.

Data

To solve the problem, we will need the following data:

- List of neighborhoods in Mumbai. This defines the scope of this project which is confined to the city of Mumbai, the capital city of the state of Maharashtra in South Asia.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

Sources and Methods to extract data:

This Wikipedia page (*https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Mumbai*) contains a list of neighborhoods in Mumbai, with a total of 136 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages.
Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward.

This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Mumbai. Fortunately, the list is available in the Wikipedia page (*https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Mumbai*). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude.

After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Mumbai. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 meters.

We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude.

With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues.

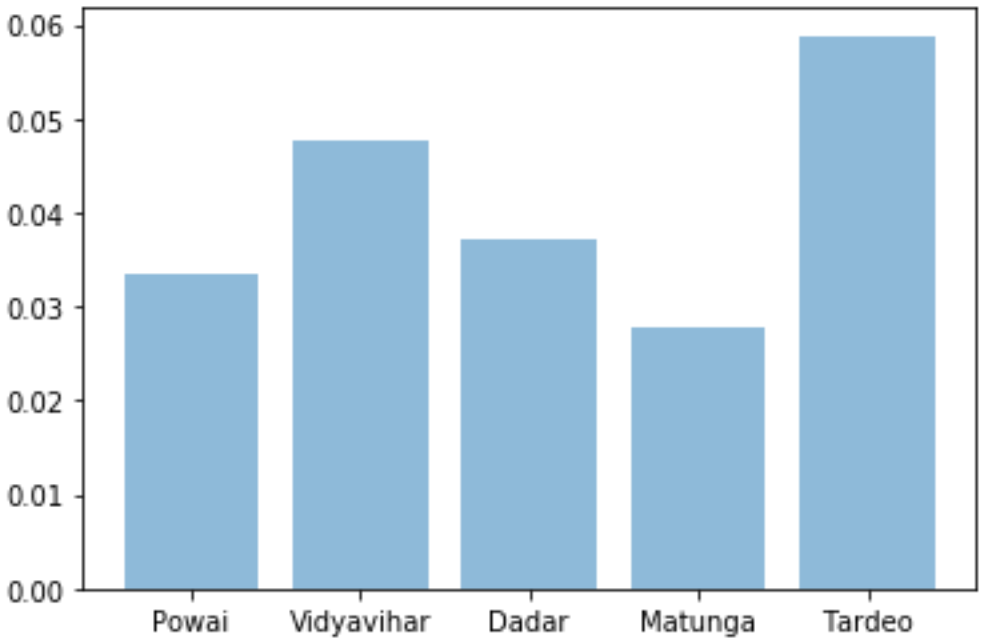
We will also gather information about the population and density per square kilometer in different districts of Mumbai. Fortunately, this information was given in the website (*http://www.indiaonlinepages.com/population/mumbai-population.html*). By using beautifulsoup packages, the data was extracted into a CSV formal and read into the notebook. Different tools for analyzing the population aspect with Shopping Malls in different neighborhoods like Bar graphs, Pie charts, Normalization were used. This data was used to calculate how density of population affected availability of space, reachability to the Shopping Mall.

Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue

category for the neighborhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 4 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

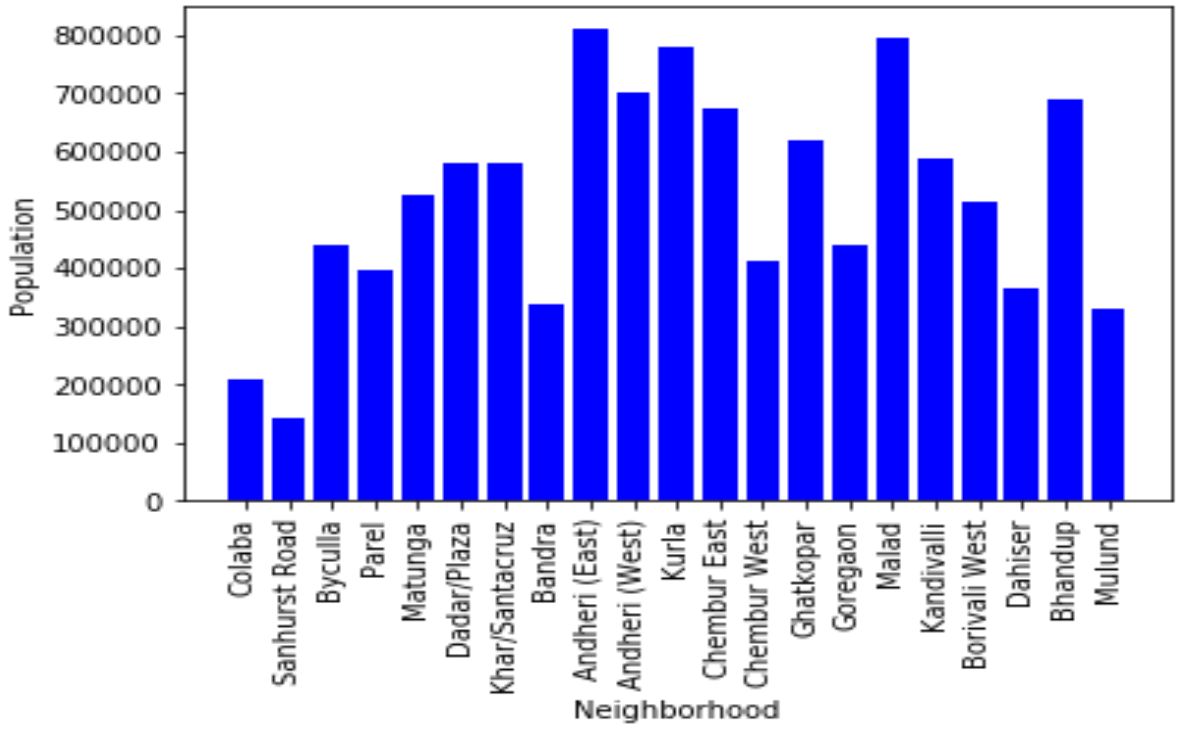
Results

The results from the Dataset regarding the Shopping Mall venues gave us a visualization:

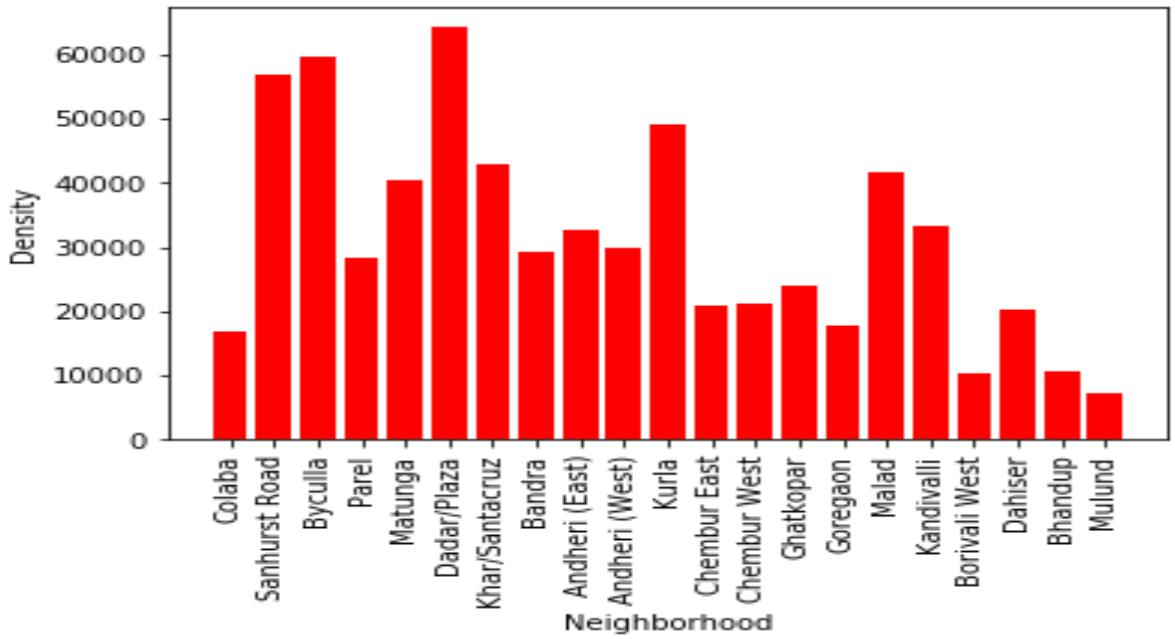


These neighborhoods had Shopping Malls already and the frequency of the venue in these localities is shown, Tardeo giving the highest.

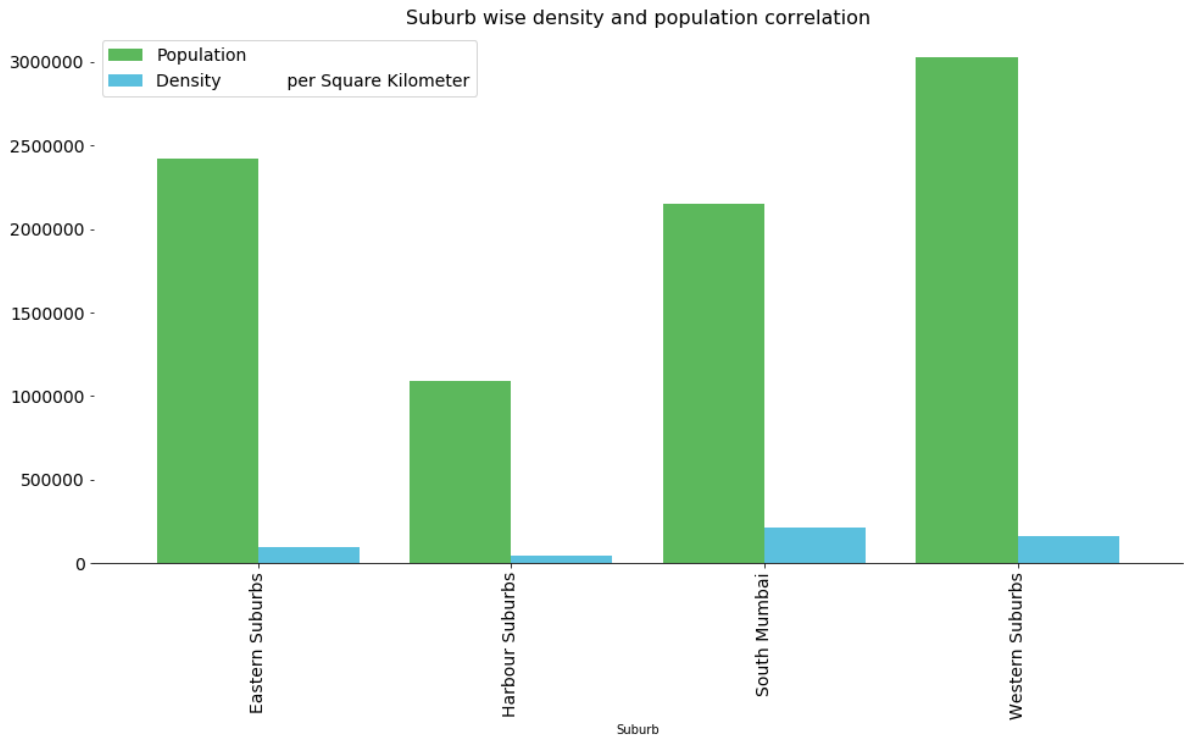
We have data regarding the population in each neighborhood:



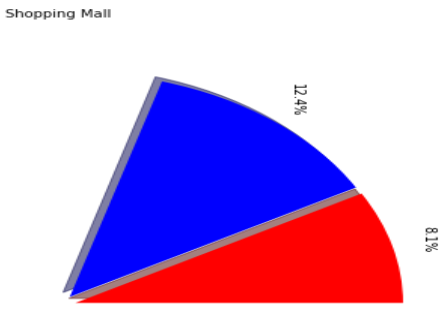
We have data regarding 'density per square kilometer' from all the neighborhoods:



We also have data to segregate population and density data with respect to different suburbs in Mumbai. To visualize the data:

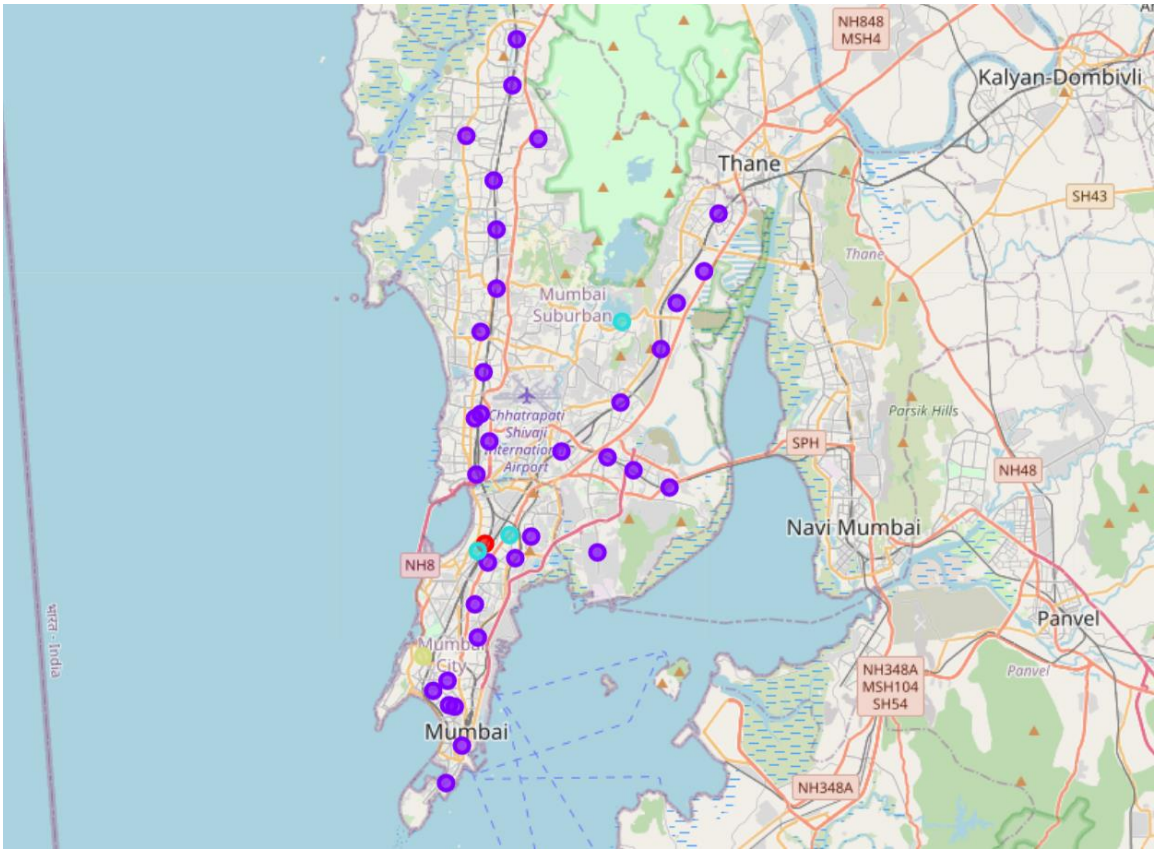


We also have a visualization regarding the suburbs' abundance to Shopping Malls:



The results from the k-means clustering show that we can categorize different neighborhoods in 4 clusters based on the frequency of Shopping Mall.

- Cluster 0: The cluster with minimum or zero number of shopping malls.
- Cluster 1: Cluster with high number of Shopping Malls.
- Cluster 2: The cluster with high to moderate number of shopping malls.
- Cluster 3: Cluster with moderate to low number of Shopping malls.



Discussion

We understood the Suburb and neighborhood data for shopping malls, and also included population data into our analysis. We can say that Eastern Suburb and South Mumbai are the two apt location for setting up a mall. South Mumbai does have a greater number of Shopping malls but it is not so densely populated and

there is space to create a plot for a shopping mall. On the other hand, Eastern Suburb has less space as compared to South Mumbai but have lesser number of Shopping Malls which means low competition level. South Mumbai will face a greater competition level but it is so vastly spread that if a mall is setup in an apt central location, people from all over South Mumbai can reach it.

Limitations and Suggestions for future research

In this project, we only consider two factors i.e. frequency of occurrence of shopping malls and population on different neighborhoods. There are other factors such as income of residents that could influence the location decision of a new shopping mall. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 4 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The optimum location for opening a shopping mall are:

1. **Powai – Eastern Suburb**
2. **Matunga – South Mumbai**

REFERENCES

- The list regarding the neighborhoods in Mumbai, Maharashtra was recovered from the Wikipedia page: *https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Mumbai*
- The population data for different neighborhoods in Mumbai was obtained through: *http://www.indiaonlinepages.com/population/mumbai-population.html*

Thank You.