



**Opening a New Shopping Mall in Mumbai,  
India**

**Coursera Capstone**

**IBM Applied Data Science  
Capstone**

**Sravya Kodavati**

## **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.

## **Business Problem**

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

## **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 capital city of India i.e. Mumbai. This project is timely as the city is currently suffering from oversupply of shopping malls. Data from the National Property Information Centre (NAPIC) released last year showed that an additional 15 per cent will be added to existing mall space, and the agency predicted that total occupancy may dip below 86 per cent. The local newspaper The Malay Mail 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 financial capital city of the country of India.
- 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 of data and methods to extract them**

This Wikipedia page ([https://en.wikipedia.org/wiki/Category:Suburbs\\_of\\_Mumbai](https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai)) contains a list of neighborhoods in Mumbai, with a total of 42 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful soup 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 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).

## 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/Category:Suburbs\\_of\\_Mumbai](https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai)).

We will do web scraping using Python requests and beautiful soup 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 2000 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. 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.

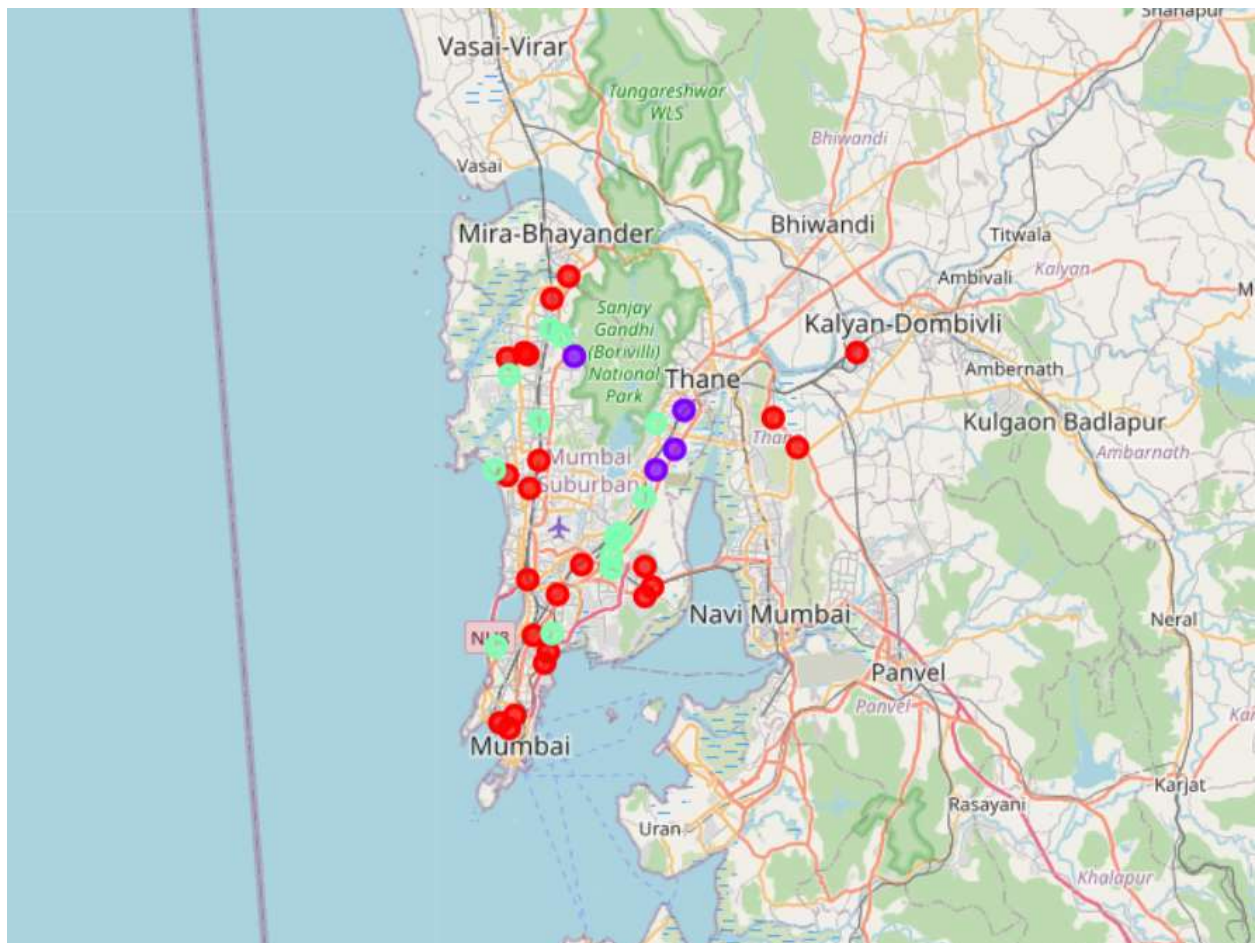
Since we are analyzing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighborhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to 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 3 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 k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”:

- Cluster 0: Neighborhoods with no existence of shopping malls
- Cluster 1: Neighborhoods with high concentration of shopping malls
- Cluster 2: Neighborhoods with moderate number of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.



## **Discussion**

Most of the shopping malls are concentrated in the southern part of Mumbai city, with the highest number in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has no shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is no competition. Meanwhile, shopping malls in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, this also shows that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 0 with no competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.