# Coursera Capstone

## IBM Applied Data Science

### Capstone Opening a New Shopping Mall in Mumbai, India

**By: Abhay Kotal**

**November 2019**

## Introduction:

Shopping mall is a great place where people can find all sort of things at one place from clothing ,grocery , daily needs, movies etc. many people spend huge amount of money in shopping mall. But opening a shopping mall at a right location is very important. There are already many shopping malls in the city opening the best one at the right place and attracting more crowd. Attracting different crowd group based on their age financial status, etc into the mall and making profit out of those crowd is also an important factor .So opening a shopping mall with all features at a right place is very important and challenging. Opening a shopping mall in Mumbai is very challenging because finding suitable property at a given place is very challenging and also costly .Most important factor is the shopping mall location that will determine the success o the shopping mall or failure of a shopping mall

## 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: 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 for the investor and for the developers those who want to build a shopping mall at a right location and invest for this project to make this project successful. This project will show the no of clustered shopping mall at a place due to which the city is suffering from over supply of shopping mall and a large no of shopping mall are built at close by place. Due to which there is a clustered of shopping mall at one place. In the city shopping malls always target for the people with financial status so most of the shopping mall end up at the costly area of the city.

## Data:

**To solve the problem we require the following data:**

- List of neighbourhoods in Mumbai. This defines the scope of this project which is confined to the city of Mumbai, the capital city of the country of India .
- Latitude and longitude coordinates of those neighbourhoods. 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 neighbourhoods

## Sources of data and method to extract them:

Sources of data and methods to extract them This Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai) contains a list of neighbourhoods in Mumbai. 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 neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods. After that, we will use Foursquare API to get the venue data for those neighbourhoods. 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.

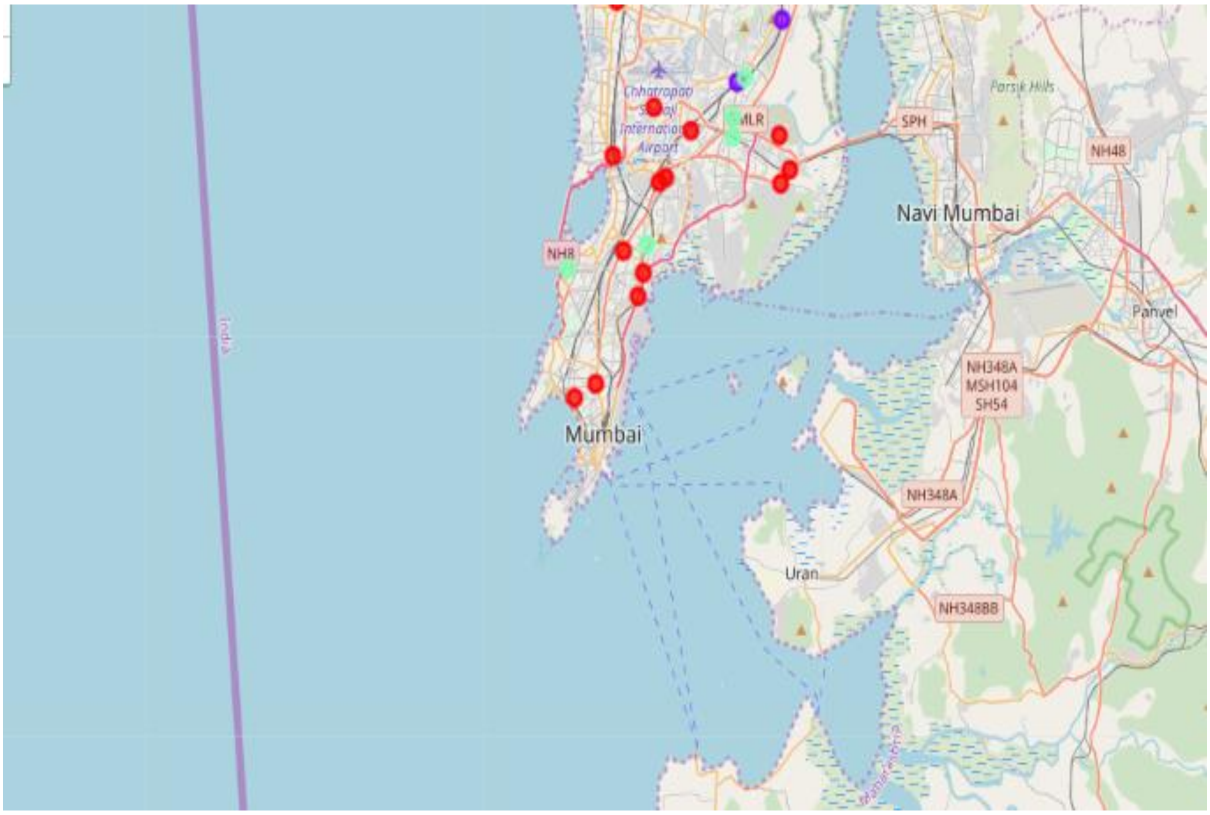
## Methodology:

Firstly, we need to get the list of neighbourhoods in the city of Mumbai . Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods 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 neighbourhoods 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 neighbourhoods 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 neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 analysing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighbourhoods. 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 neighbourhoods into 5 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

**Results :**

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”:

- Cluster 0: Neighbourhoods with high number of shopping malls
- Cluster 1: Neighbourhoods with low number of shopping malls
- Cluster 2: Neighbourhoods with medium number of shopping malls



## Discussion:

Most of the shopping malls are concentrated in the western area of Mumbai city, with the highest number in cluster 0 and moderate number in cluster 2. On the other hand, cluster 1 has low number of shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 0 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 western 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 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 0 which already have high concentration of shopping malls and suffering from intense competition.

## Limitation and Suggestion for Future Research:

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls, there are other factors such as population and 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 3 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 neighbourhoods in cluster 1 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

**References:**

Category:Sururbs in Mumbai, India. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai

Foursquare Developers Documentation. Retrieved from https://developer.foursquare.com/docs

Project Reference: Chia Hooi Lim https://github.com/limchiahooi/Coursera_Capstone.git