**COURSERA CAPSTONE PROJECT**

**IBM Applied Data Science Capstone**

# Deciding Best Place to Open New Shopping Centre in Mumbai , India

By : Mayur Rindhe

August , 2019

# Introduction

Shopping centre are buildings form a complex of shops with interconnecting walkways, usually indoors. In the 21th century, with the rise of the suburb and automobile culture in the India, a new style of shopping centre was created away from downtown.

For many shoppers, visiting shopping Centres 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 Centres are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping Centres provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping Centres to cater to the demand. As a result, there are many shopping Centres in the city of Mumbai and many more are being built. Being the economic capital of India , Mumbai got its own uniqueness .

Opening shopping Centre allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping Centre requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping centre is one of the most important decisions that will determine whether the Centre will be a success or a failure.As India is a place of opportunity and if somebody pick the right thing at right time at right location then there is nothing will be there to stop him.

# Business Problem :

The objective of this capstone project is to analyse and select the best locations in the city of Mumbai , Maharashtra (India) to open a new shopping Centre. Using data science methodology and machine learning techniques like clustering also using web scrapping , this project aims to provide solutions to answer the business question: In the city of Mumbai , India  if a property developer , builders , foreign investors are looking to open a new shopping Centre, 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 Property like Shopping Cenre in the economic capital city of India i.e. Mumbai. This project is timely as the city is currently suffering from oversupply of shopping Centres.

This project may also help the government agencies to decide and take control over development in city .

# Data :

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

• List of neighbourhoods in Mumbai city. This defines the scope of this project which is confined to the city of Mumbai, the Economic 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 Centres. We will use this data to perform clustering on the neighbourhoods.

**Sources of data and methods to extract them :**

This Wikipedia page mentioned in the following bracket (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in Mumbai) contains a list of neighbourhoods in Mumbai, with a total of 50 neighbourhoods. 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 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 Centre 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:Neighbourhoods_in_Mumbai). We have done 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.

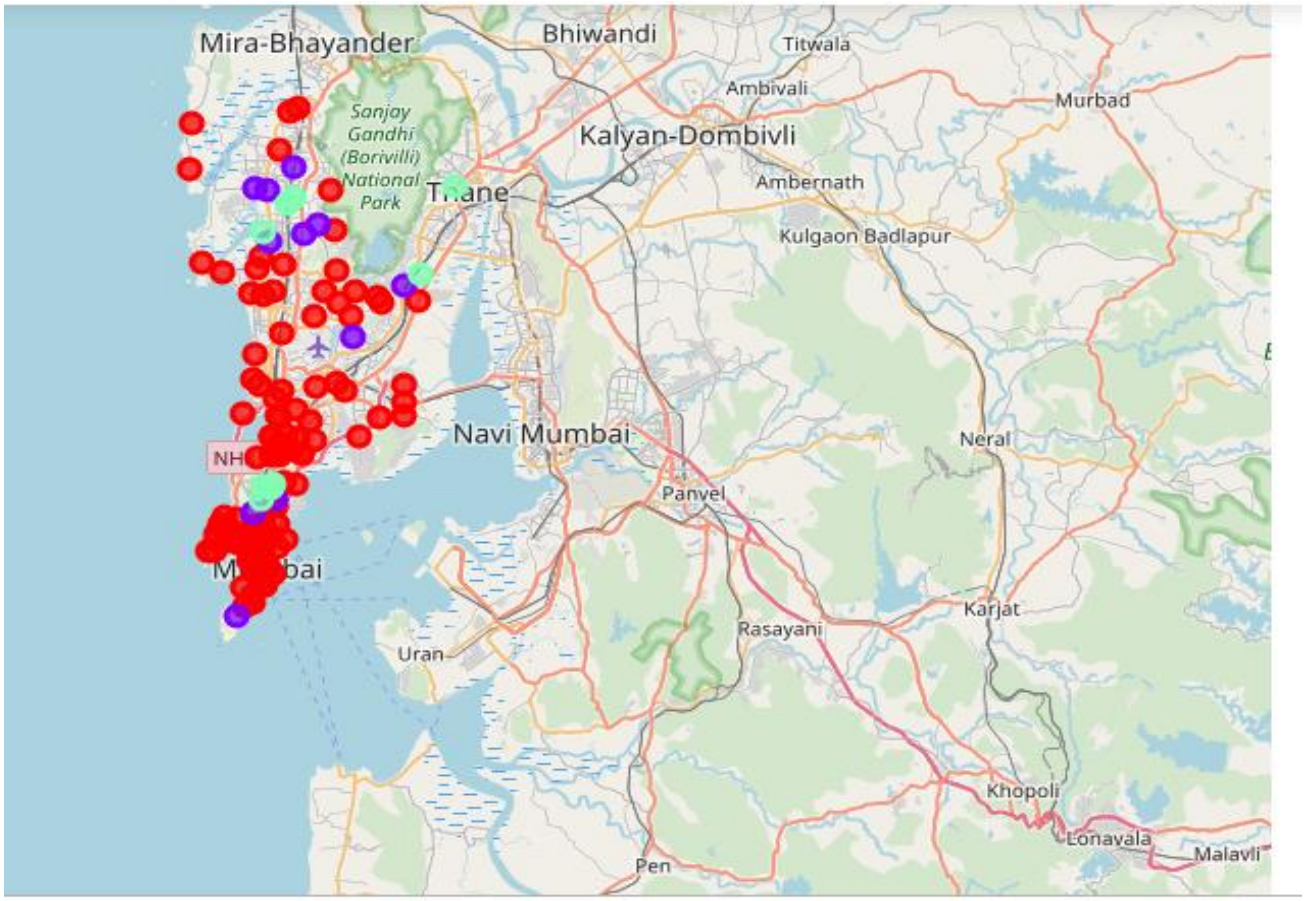
Next, we will use Foursquare API to get the top 50 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods 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 Centre” data, we will filter the “Shopping Centre” 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 Centre 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 3 clusters based on their frequency of occurrence for “Shopping Centre”. The results will allow us to identify which neighbourhoods have higher concentration of shopping Centres while which neighbourhoods have fewer number of shopping Centres. Based on the occurrence of shopping Centres in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping Centres.

## 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 Centres":

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




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

## Discussion :

Most of the shopping Centers are concentrated in the central area of Mumbai city, with the highest number in cluster 0 and moderate number in cluster 1. On the other hand, cluster 2 has very low number to totally no shopping Center in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping Center as there is very little to no competition from existing Shopping centers. Meanwhile, shopping Centers in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of shops. From another perspective, this also shows that the oversupply of shopping centers mostly happened in the central area of the city, with the suburb area still have very few shopping centers. Therefore, this project recommends property developers to capitalize on these findings to open new shopping centers in neighborhoods in cluster 2 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping centers in neighborhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 0 which already have high concentration of shopping centres and suffering from intense competition.

## 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 Centre. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to open a new shopping centre. 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 centre.