# IBM Applied Data Science Capstone Project Week 5

# Mumbai, The City of Dreams

## Introduction

Mumbai (formerly known as Bombay) is the commercial capital of India. It is also known as the city that never sleeps and the city of dreams. It is the most populous city in all of India. Mumbai is located on an island off the west coast of India. The city, which has a deep natural harbor, is also the largest port in western India, handling over half of India's passenger traffic. It also has the highest number of millionaires and billionaires among all cities in India. Mumbai is home to three UNESCO World Heritage Sites: the Elephanta Caves, Chhatrapati Shivaji Maharaj Terminus, and the city's distinctive ensemble of Victorian and Art Deco buildings.

The city is the birthplace of Indian Cinema. Hindi cinema, often known as Bollywood and formerly as Bombay cinema, is the Indian Hindi-language film industry based in Mumbai. The term is a portmanteau of "Bombay" and "Hollywood". The industry is related to Cinema of South India and other Indian film industries, making up Indian Cinema- the world's largest by number of feature films produced.

However, all places are not perfect, and Mumbai has its fair share of issues and short comings. It suffers from the expansion of large cities leading to air and water pollution and overcrowding. Since it has a diverse society, the infrastructure pertaining to this diversity is vastly different. There are many infrastructures in Mumbai, each belonging to different categories like drinking water plant, hospitals, schools and hotels.

**Business Problem**

The following are the questions that are going to be answered in the project.

1) What are the major infrastructure sites in Mumbai?
2) The top locations by way of design and infrastructure
3) What are the areas that have the hope for improvement in infrastructure in future?
4) Areas with poor infrastructure
5) Best places to stay with the required infrastructure needs

## Description of the Data to be Used in the Project

There is a various amount of infrastructure to be considered in Mumbai.

For this project, the following data is required.

- Mumbai pin code (scraped from web source)- Source is Mumbai7.com, contains the list pin codes and postal office names.
- For the kinds of infrastructure in each neighborhood, Foursquare API will be used. The venues can be easily filtered therefore, and it will be easy to access the various places.
- Geospace data is needed lastly, to get the latitudes and longitudes based on the postal offices in Mumbai.

This project will make use of skills in data science such as web scraping from Mumbai7.com, working with location data from the Foursquare API, data cleaning, wrangling, and using Folium maps to visualize the data. K-means clustering will also be used to cluster up the good infrastructure locations.

## Methodology

**Data exploration**- The list of neighborhoods in Mumbai must be retrieved first. The list is available in the web page- https://mumbai7.com/postal-codes-in-mumbai/

Using Python and web scraping, the list of data on the neighborhoods can be extracted. This list contains the postal codes, postal office names and cities.

**Geocoding**- The geographical coordinates must be received in the form of latitude and longitude so that we can use the Foursquare API. Using the Geocoder package, we can convert the address into geographical coordinates with the latitude and longitude. After extracting the required data, we can convert it into a Pandas DataFrame.

**Visualizing the data**- To visualize the results, we will use Folium. With Folium we can generate maps that pertain to the coordinates and receive locations. Markers with different colors can be used to differentiate various locations as well.

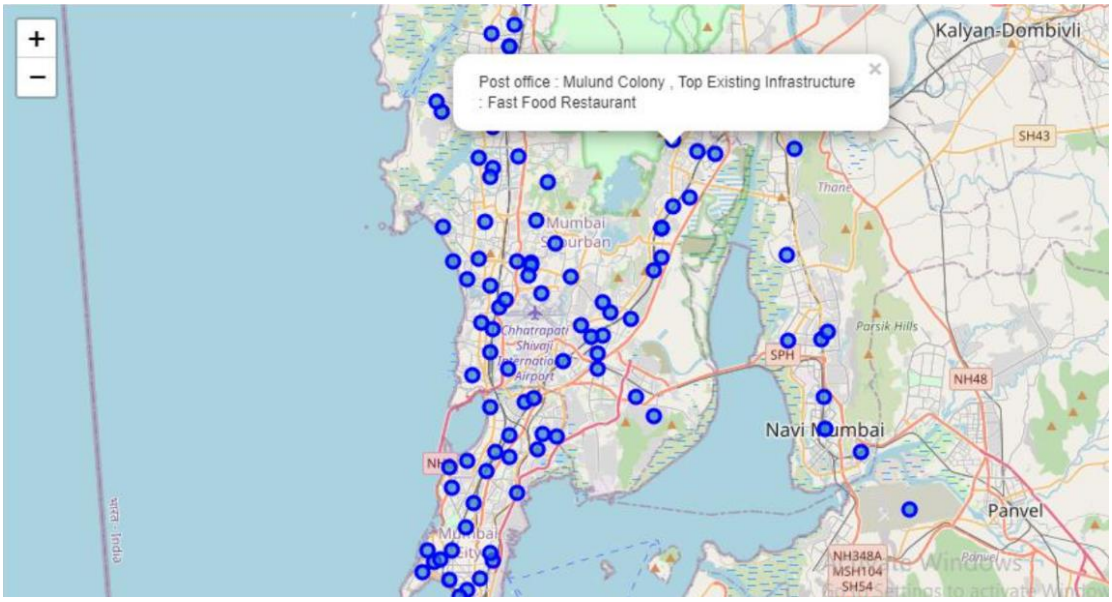
**Finding the top infrastructures**- We will use Foursquare API to get the top venues in Mumbai. Once API callas are made, Foursquare will return the venue data in JSON format and then the venue name, latitude and longitude can be received. Once all the data is received, data will be filtered into rows and columns and will be analyzed.

**Data wrangling**- The data is prepared to be used in selecting an appropriate area with top infrastructures.

**Clustering**- Using k-means clustering, we will cluster the data and show which areas are high in infrastructures and areas that are low in infrastructures. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “no. of existing infrastructures”. So, with this classification we will be able to identify whether neighborhoods have a high, medium, or lower concentration of infrastructures. So, according to the occurrence of infrastructures in different neighborhoods, we can answer the question as to which neighborhoods are best for new infrastructures.

## Result Extractions

Below is the best existing infrastructure for each postal office in Mumbai.



Best places in Mumbai with good infrastructure

| | |
|---|---|
| **Post Office** | Bandra (West) |
| **Pin Code** | 400050 |
| **City** | Mumbai |
| **Airport Terminal** | 0 |
| **Bank** | 0 |
| **Bus Station** | 0 |
| **Business Service** | 0 |
| **Café** | 10 |
| **College Auditorium** | 1 |
| **Electronics Store** | 1 |
| **Farmers Market** | 1 |
| **Garden** | 0 |
| **Government Building** | 0 |
| **Gym / Fitness Center** | 3 |
| **Hotel** | 1 |
| **Indie Movie Theater** | 1 |
| **Light Rail Station** | 0 |
| **Market** | 0 |
| **Monument / Landmark** | 0 |
| **Park** | 1 |
| **Pharmacy** | 0 |
| **Playground** | 0 |
| **Resort** | 0 |
| **Restaurant** | 1 |
| **Shopping Mall** | 1 |
| **Theater** | 0 |
| **Train Station** | 0 |
| **Total infrastructure** | 21 |

Areas with the worst or unsatisfactory infrastructure

| Post Office | Pin Code | City |
|---|---|---|
| Agashi | 401301 | Thane |
| Anu Shakti Nagar | 400094 | Mumbai |
| Bassien | 401201 | Thane |
| Bhandup (East) | 400042 | Mumbai |
| Bhayander (East) | 401105 | Thane |
| Boisar | 401501 | Thane |
| Ghansoli | 400701 | Navi Mumbai |
| Jacob Circle | 400011 | Mumbai |
| Jakegram | 400606 | Thane |
| Jawhar | 401603 | Thane |
| Jawhar | 401603 | Thane |
| Kopri Colony | 400603 | Thane |
| Krishi Utpanna Bazar | 400705 | Navi Mumbai |
| Mahim | 400016 | Mumbai |
| Nerul Mode | 400706 | Navi Mumbai |
| Santacruz P&T Colony | 400029 | Mumbai |
| Sopara | 401203 | Thane |
| Tagore Nagar | 400083 | Mumbai |
| Talasari | 401606 | Thane |
| Umbarpada | 401102 | Thane |
| Uran | 400702 | Navi Mumbai |
| Vasai East I/E | 401208 | Thane |
| Wadala | 400031 | Mumbai |

Areas that can develop in future and have great prospects to have various types of infrastructure be built upon

```
Airport Terminal
Bank
Bus Station
Business Service
College Auditorium
Farmers Market
Garden
Government Building
Indie Movie Theater
Light Rail Station
Market
Monument / Landmark
Park
Pharmacy
Playground
Resort
Train Station
```

Best areas to stay for important and necessary facilities

| | Post Office | Total infrastructure |
|---|---|---|
| **18** | Bhavani Shankar Road | 13 |
| **28** | Council Hall | 13 |
| **29** | Cumballa Hill | 13 |
| **30** | Dadar | 13 |
| **34** | F C I Mumbai | 13 |
| **35** | Ganeshpuri | 13 |
| **38** | Girgaon | 13 |
| **43** | I I T Mumbai | 13 |
| **44** | J B Nagar | 13 |
| **45** | JNPT Town Ship | 13 |
| **67** | Manor | 13 |
| **79** | Mumbai G P O | 13 |
| **80** | N I T I E | 13 |
| **88** | Papdi | 13 |
| **92** | Rajbhavan | 13 |
| **97** | Santacruz (East) | 13 |

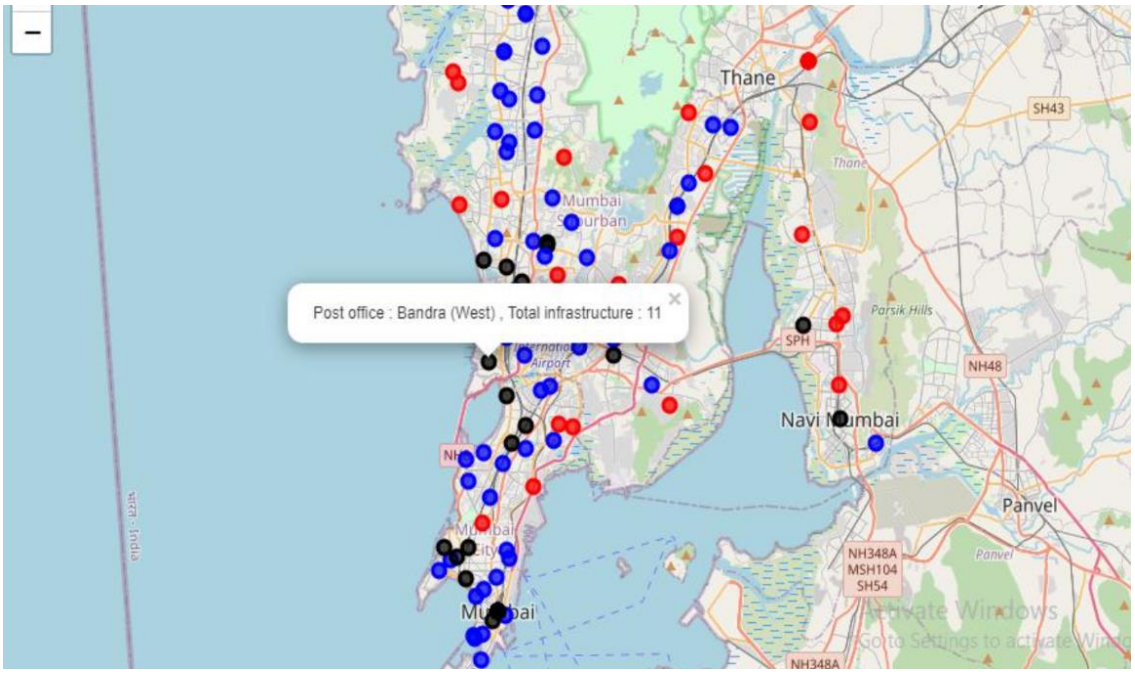
# Clustering Total Infrastructure

Results received from k-means clustering shows that 3 clusters can be made based on the occurrence of the number of existing infrastructures.

Cluster 1- Neighborhoods with a low number of infrastructures (shown in red)

Cluster 0- Neighborhoods with a high number (shown in black)

Cluster 2- Neighborhood with a medium number (shown in blue)




According to the results, West Bandra is the best location in Mumbai with the best infrastructure. 10 Café is the best location.

Most of the infrastructures and constructions are in the South of Mumbai. The highest number is in cluster 0 and there is a moderate number in cluster 2. Cluster 1 has a very low number of construction and infrastructures in the neighborhoods. This could mean that there is a great chance of the areas being good for future developments with little competition from existing areas.

Construction and infrastructures in cluster 0 are most probably facing competition due to a high concentration of places that have already established themselves. West Bandra also has a lot of infrastructures and is well developed.

I would recommend for a person who is trying or planning to build infrastructures to capitalize on these findings to open new places in cluster 1 where there is almost no competition. In cluster 2, there is moderate competition but has a little of supporting infrastructure which could be advantageous. Finally, in cluster 0, this area is well developed and for someone

wanting to build here, there will be high competition because there are high concentrations of infrastructures.

## Limitations of the Research

Though most factors are taken into consideration in this project, there are factors that cannot all be covered. Quality of the infrastructure, incomes of people living in the same communities as the infrastructures etc.

This project was done using the free Sandbox Tier Account of the Foursquare API, which restricts the number of calls that can be made and if the service was the paid one, more calls could have been made and therefore more results could have been filtered etc.

## Conclusion

In this project I have used various data science and machine learning tools, extracted data, scraped off the web and used location data to show points on geographical maps. This project will come in handy for those who want to set up businesses and want to know where exactly to set up in Mumbai based on the support infrastructures located near by and could also help tourists understand the environment of the land and help foreigners understand the setting of the city of Mumbai better.