IBM Applied Data Science Capstone

City Of Opportunity: Mumbai

Introduction:

Mumbai is the commercial capital of India. It is also known as the city that never sleeps. Mumbai is the perfect blend of culture, customs, and lifestyles. Mumbai is India's most cosmopolitan city, its financial powerhouse, and the nerve center of India's fashion industry. Mumbai is also dotted with plenty of architectural landmarks from the Victorian era and the days of Raj. Mumbai is also the birthplace of Indian Cinema.

Located on Maharashtra's coast, Mumbai is India's most-populous city, and it is one of the largest and most densely populated urban areas in the world. Mumbai developed a highly diversified infrastructure.

It suffers, however, from some of the perennial problems of many large expanding industrial cities: air and water pollution, widespread areas of substandard housing, and overcrowding. With its diverse society, comes diverse infrastructure which decides the quality of living. There are many infrastructures in Mumbai, each belonging to different categories like Drinking Water Plant, Waste Water/ Sewage, Hospitals, Schools, Colleges, Railway Network, Electricity Power Plants, Telecommunication Support, Bank, Shopping malls, Supermarket, Gas Station, Hotels, Police Station, Café, medical shops, grocery shops, theatre, etc. One of the main problems, when one moves to a new city, is where to find a good area to build and grow prosperously.

Business Problem:

The questions I aim to answer in this project are the following:

---

1. List and visualize all major parts of Mumbai City with top existing infrastructure.

2. What are the best locations in Mumbai as per infrastructure?

3. Which areas have the potential for the development of infrastructure of different kinds?

4. Which all areas lack the infrastructure facilities?

5. What is the best place to stay within a city for all vital infrastructure facilities?

Target Audience:

The purpose of this project is to help people in exploring better facilities around their neighborhood. It will help people making a smart and efficient decision on selecting great neighborhoods out of numbers of other postal areas in Mumbai, India.

Lots of people are migrating from various states of India and needed lots of research for good housing prices, new business, and reputed professional places for their children. This project is for those people who are looking for better neighborhoods and businesses.

It will help people to get the awareness of the area and neighborhood before moving to a new city, state, country, or place for their work or to start a new fresh life.

Data Description:

Mumbai City's demographics show that it is a large and ethnically diverse metropolis. With its diverse society, comes diverse infrastructure. There are many different kinds of infrastructure in Mumbai City, each belonging to different categories like Hospitals, Schools, Colleges, Hotels, etc.

For this project we need the following data:

- Mumbai Pincode (Scraped from web source)
    - Data source: https://mumbai7.com/postal-codes-in-mumbai/
    - Description: Contain a list of pin codes, postal office names, city which can be used to discover all postal office of Mumbai.
- Mumbai City data (GeoSpace data on Mumbai Pincode)
    - Data source : https://github.com/geospace-code/pymap3d
    - Description: By using this geospace data, we will get the latitude and longitude coordinates of the postal office of Mumbai. We will use this data set to explore various neighborhoods of Mumbai City.
- Different kinds of infrastructures in each neighborhood of Mumbai City.
    - Data source: Foursquare API
    - Description: By using this API we will get all the venues in each postal office. We can filter these venues to get different infrastructures and venues.

Using this data will allow exploration and examination to answer the questions. This is a project that will make use of many data science skills, from web scraping (mumbai7.com), working with API (Foursquare), data cleaning, data wrangling and map visualization (Folium) and to machine learning (K-means clustering).

Methodology:

DATA EXPLORATION -

Firstly, we need to get the list of neighborhoods in Mumbai. Fortunately, the list is available on the web page (https://mumbai7.com/postal-codes-in-mumbai/). We have to do web scraping using Python requests to extract the list of neighborhood data. However, this is just a list of pin codes, postal office names, and cities.

DATA GEOCODING-

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 the address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame.

DATA VISUALIZATION -

Visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the city of Mumbai.

FINDING TOP 225 INFRASTRUCTURE EXPLORATION -

Next, we make use of Foursquare API to get the top 225 venues that are within a radius of 625 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.

DATA WRANGLING -

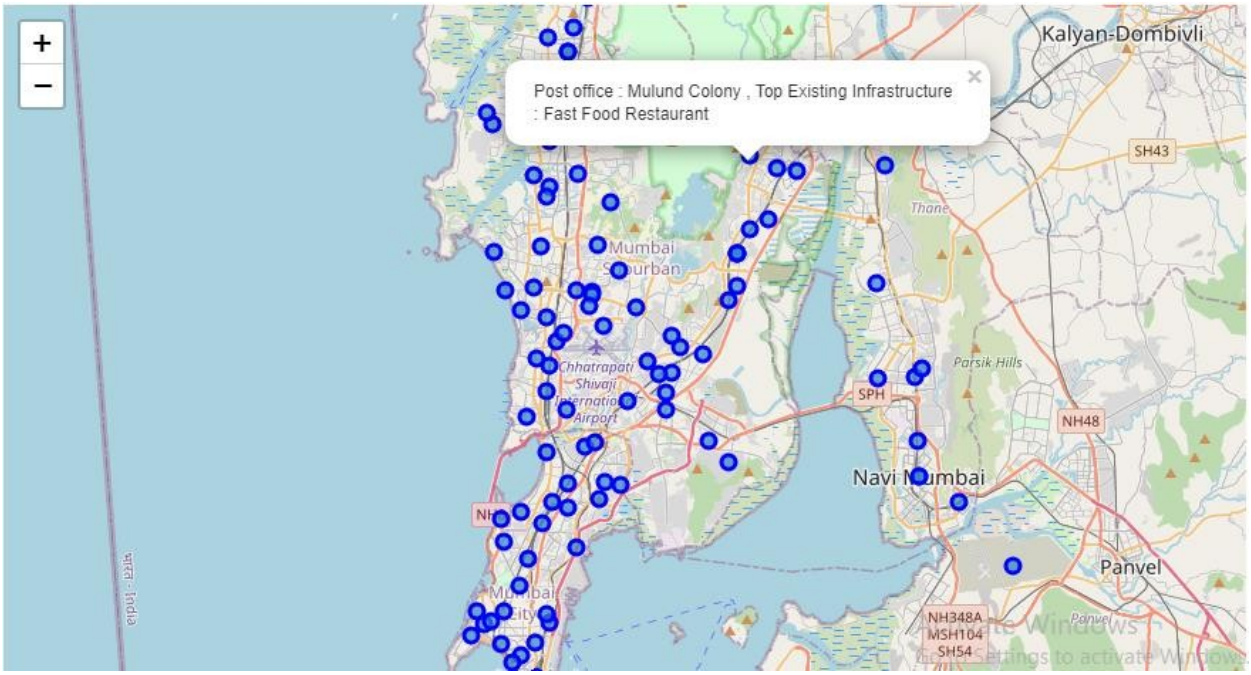
We are also preparing the data for use in selection. Based on the occurrence of infrastructures in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new infrastructures and which neighborhoods are most suitable to visitors to stay.

DATA CLUSTERING -

Finally, 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 "no. of existing infrastructures". The results will allow us to identify which neighborhoods have higher, medium and lower concentration of infrastructures. Based on the occurrence of infrastructures in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new infrastructures.

Result :

1. *Display the top existing infrastructure for each Postal Office in Mumbai.*



2. *What are the best locations in Mumbai as per infrastructure?*

	9
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

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

3. *Which all areas lack the infrastructure facilities?*

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

4. *Which of your choice areas have the potential for the development of infrastructure of different kinds?*

```
These are infrastructures with highest potential in Mantralaya area :
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
```

5. *What is the best place to stay within a city for all vital infrastructure facilities?*

Best place to stay within a city for vital infrastructure facilities :

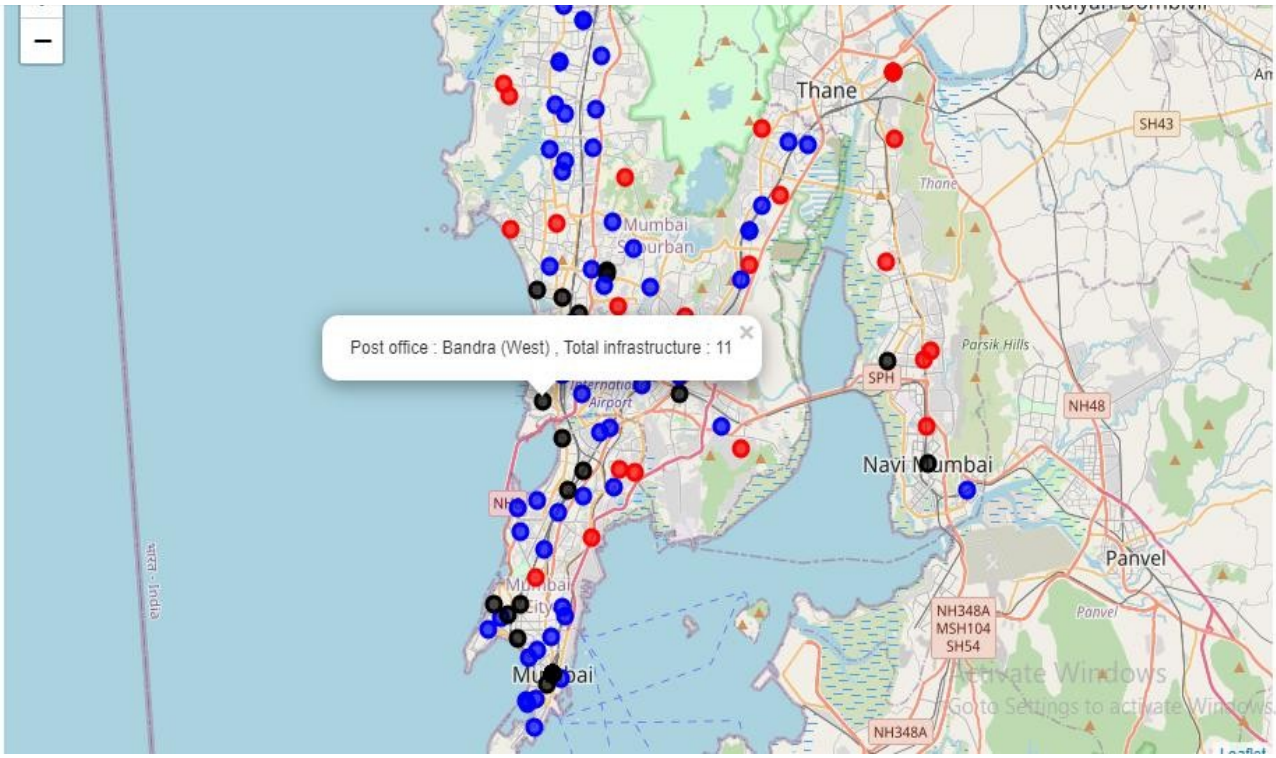
	Post Office	Total infrastructure
18	Bhavani Shankar Road	13
28	Council Hall	13
29	Cumballa Hill	13
30	Dadar	13
34	F C I Mumbai	13
35	Ganeshpuri	13
38	Girgaon	13
43	I I T Mumbai	13
44	J B Nagar	13
45	JNPT Town Ship	13
67	Manor	13
79	Mumbai G P O	13
80	N I T I E	13
88	Papdi	13
92	Rajbhavan	13
97	Santacruz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

Clustering Based On Total Infrastructure:

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “no. of existing infrastructures”:

- **Cluster 1**: Neighborhoods with a low number to no existence of infrastructures.
- **Cluster 0**: Neighborhoods with a high number of infrastructures
- **Cluster 2**: Neighborhoods with a moderate number of infrastructures.

(The results of the clustering are visualized in the map below with cluster 0 in black color, cluster 1 in red color, and cluster 2 in blue color.)



Discussion :

(Based to constraint on API calls and search radius, the true result might vary.)

- Bandra (West) is the best location in Mumbai as per infrastructure, with 10 Café boosting the bar.
- 19 postal offices are the best postal office area for essential varietyinfrastructure. (shown in the result section)

Most of the infrastructures are concentrated in the Southern areas of Mumbai city, with the highest number in cluster 0 and moderate number in cluster 2. On the other hand, cluster 1 has a very low number of infrastructures in the neighborhoods. This represents a great opportunity and high potential areas to open new infrastructures as it is very little to no competition from existing varied infrastructures. Meanwhile, one can specifically check the infrastructure of choice against the postal office choice area.

Infrastructures in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of already established Infrastructures. From another perspective, this also shows that the oversupply of infrastructures mostly happened in the developed parts like Bandra(West) in Mumbai city; with the suburb areas like South Mumbai still have a very high

frequency of established infrastructure. Therefore, this project recommends a person who is planning to build infrastructure to capitalize on these findings to open new Infrastructures in neighborhoods in cluster 1 with little to no competition. A person who is planning to build infrastructure with unique selling propositions and lives prosperously to stand out from the competition can also open new infrastructures in neighborhoods in cluster 2 with moderate competition and supporting adequate no. of infrastructures. Lastly, people with planning to settle in the city are advised to start in cluster 0 which already has a high concentration of infrastructures.

Limitations and Suggestions for Future Research:

In this project, I tried to cover most of the factors affecting the people. But there are other factors which could be included to improve for realistic business development like Quality of Infrastructures, Population and Income of the residents that could easily influence the desired location decision of new infrastructures. However, to the best knowledge of mine, such data are not available to the neighborhood 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 new infrastructures.

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 a paid account to bypass these limitations and obtain more results.

Conclusion:

In this project, I have gone through the process of identifying the business problems, specifying the data required, extracting and preparing the data, visualizing the results, performing machine learning by clustering the data into 3 clusters based on their frequency similarities, tackling and reaching to a definitive solution to business problems (mentioned in results). Lastly, the project is providing recommendations to the relevant stakeholders i.e. **business developers** regarding the best locations to open a new infrastructure. The project also provides **visitors and immigrants** to the city regarding postal office areas for growth and living prosperously.