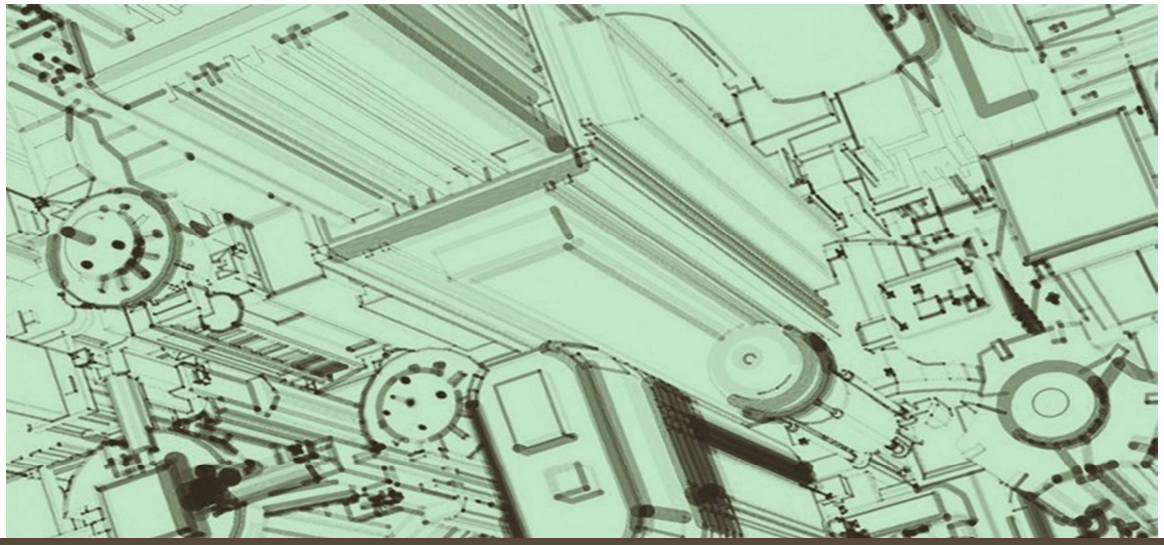


Battle of Neighborhoods

IBM Applied Data Science Capstone

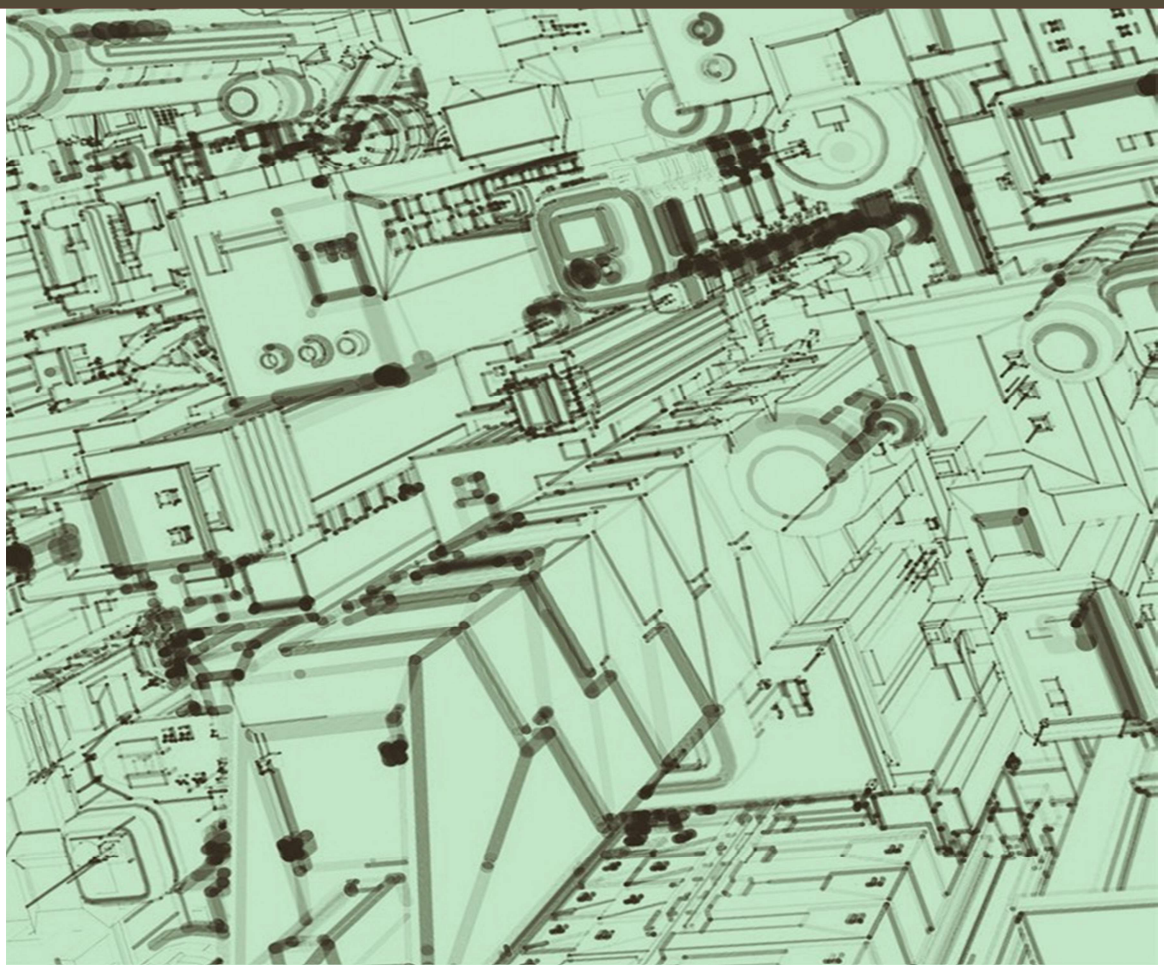


Table of Contents

Table of Figures

List of Tables

Introduction

1.1 Business Problem

In this dataset I will explore two cities of India - Mumbai and Banglore. Suppose a person gets a job in Mumbai and/or in Banglore and they need to find a new place to live along with all the places to explore like trending restaurants, food joints, parks, movie theatre etc. Based on this assignment they can find a suitable place to live in the city with most favourable options.

The datasets used for this case study is taken from Kaggle (Link provided below for both datasets). Both the dataset includes neighbourhoods, type of properties, price of that properties, total area of the properties with their latitude and longitude. This will help us extract venues from the Foursquare API and divide them into different clusters. Here for the clusters, we will use K-means clustering.

- Mumbai data set - Kaggle

- Banglore data set- Kaggle

1.2 Data

Mumbai Dataset

Structure of the dataset:

```
-----------------------------------------------------------------------

The dataset has 34348 rows and 6 columns.

Total elements in this dataset are 206088
```

Missing Values Check

```
-----------------------------------------------------------------------
There are 984 missing values in the dataset! Need further checks!
```

Info of the dataset

```
-----------------------------------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34348 entries, 0 to 34347
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Neighbourhood  34140 non-null  object
 1   Price          34348 non-null  int64
```

```
 2   Area (sq.ft)    33572 non-null  float64
 3   Type            34348 non-null  object
 4   Latitude        34348 non-null  float64
 5   Longitude       34348 non-null  float64
dtypes: float64(3), int64(1), object(2)
memory usage: 1.6+ MB
None
```

Duplicate Values Check

```
----------------------------------------------------------------------
Number of duplicate rows = 3583
```

	Neighbourhood	Price	Area (sq.ft)	Type	Latitude	Longitude
51	NaN	8000	650.0	Apartment	0.000000	0.000000
88	Bhayandar East	9000	NaN	Apartment	19.304779	72.860413
115	Virar West	8000	630.0	Apartment	19.470536	72.808309
124	Virar West	8000	630.0	Apartment	19.470536	72.808309
126	Virar West	8000	630.0	Apartment	19.471822	72.805372

This dataset contains 2 object data type `Neighbourhood` and `Type` . `Price` , `Area (sq.ft)` , `Latitude` and `Longitude` are numeric type. There are 34348 rows and 6 columns with 206088 total elements. There are missing and duplicate values that needs further analysis. The memory usage by this data set is 1.6+ MB

Table 1 Top 5 rows of Mumbai Dataset

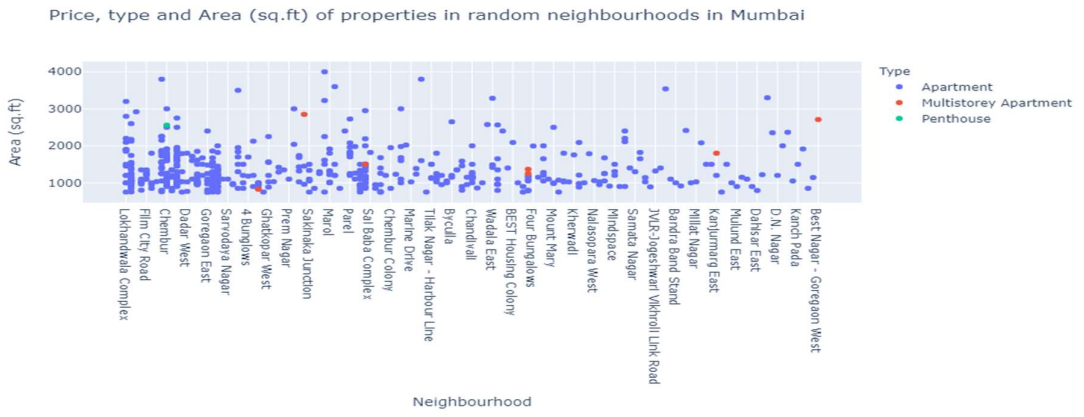


Figure 0-1 Area (sq.ft) in Mumbai

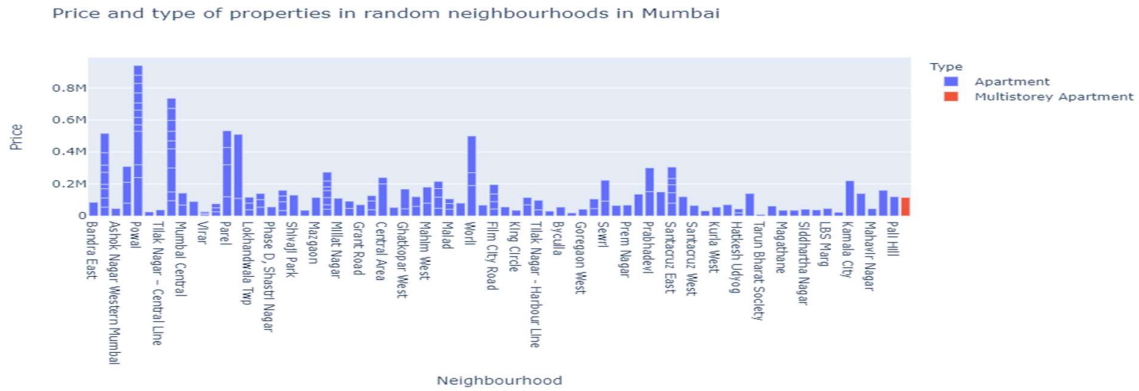


Figure 0-2 Price range for Mumbai properties

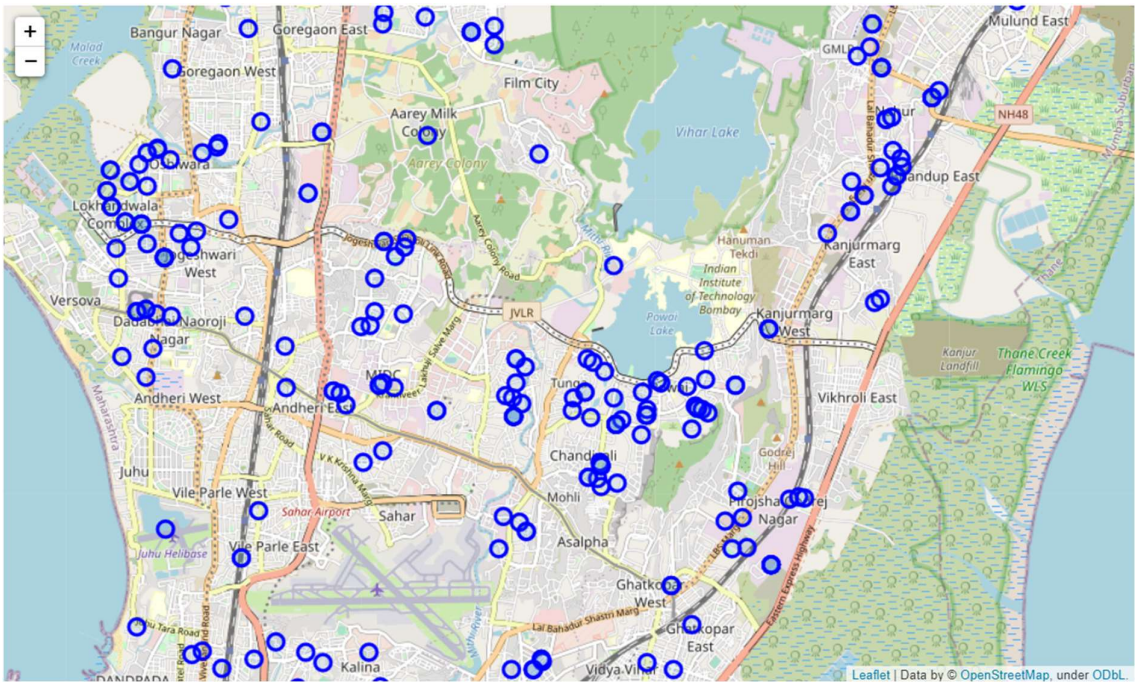


Figure 0-3 Property location in map of Mumbai

With the property rates and the area (sq.ft) we can conclude that Mumbai city mostly has apartments and few multistorey buildings. Most of the clusters are near suburban areas and by Powai Lake. There aren't many residential apartments near the airport. There is a small cluster near Bhandup East. Property rates are higher in the locations near Powai Lake than in the suburbs of Mumbai.

Banglore Dataset

Structure of the dataset:

```
--------------------------------------------------------------------------

The dataset has 2881 rows and 6 columns.

Total elements in this dataset are 17286
```

Missing Values Check

```
--------------------------------------------------------------------------
There are 895 missing values in the dataset! Need further checks!
```

Info of the dataset

```
--------------------------------------------------------------------------
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2881 entries, 0 to 2880
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Neighbourhood  2881 non-null   object
```

```
 1   Price           2758 non-null   object
 2   Area (sq.ft)    2711 non-null   object
 3   Type            2881 non-null   object
 4   Latitude        2580 non-null   float64
 5   Longitude       2580 non-null   float64
dtypes: float64(2), object(4)
memory usage: 157.6+ KB
None
```

Duplicate Values Check

```
-------------------------------------------------------------------------
Number of duplicate rows = 229
```

	Neighbourhood	Price	Area (sq.ft)	Type	Latitude	Longitude
184	The Central Regency Address	85 L -1.24 Cr	1454-1609 sq.ft	2 BHK Apartment	12.921034	77.670640
186	The Central Regency Address	1.05 Cr -1.64 Cr	1920-2105 sq.ft	3 BHK Apartment	12.921034	77.670640
188	The Central Regency Address	1.50 Cr	1920 sq.ft	4 BHK Apartment	12.921034	77.670640
203	Premier Inspira Maplewood	60 L -83.97 L	1168-1369 sq.ft	2 BHK Apartment	12.905424	77.699319
205	Premier Inspira Maplewood	79 L -1.04 Cr	1402-1633 sq.ft	3 BHK Apartment	12.905424	77.699319

The dataset has 2881 rows and 6 columns. Total elements in this dataset are 17286. There are same columns in this dataset as in Mumbai with price and Area(sq.ft) columns as category. There are missing and duplicate values in the data that needs further analysis. Total memory usage by the data set is 157.6+ KB.

Table 2 Top 5 rows of Banglore dataset

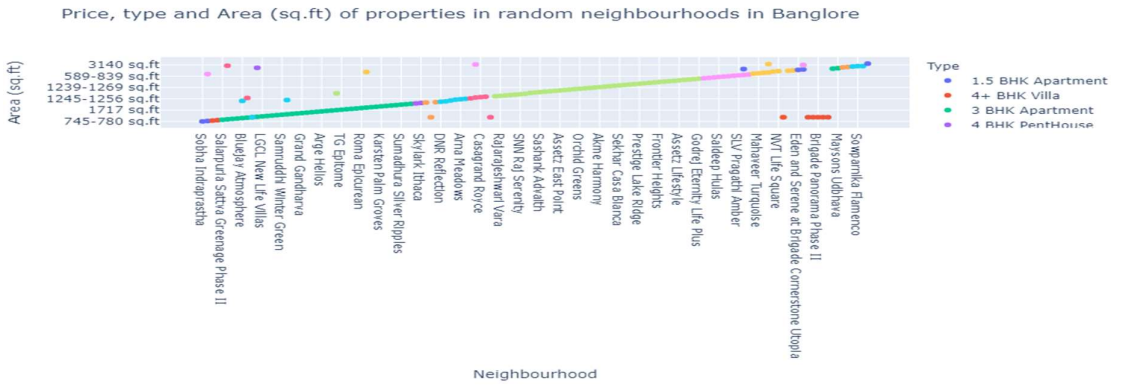


Figure 0-4 Area (sq.ft) in Banglore

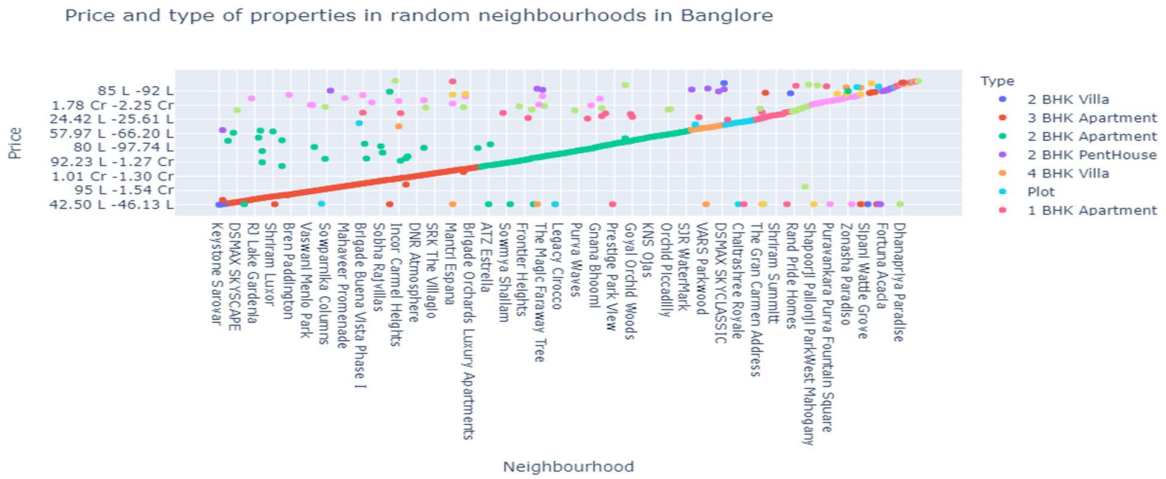


Figure 0-5 Price range of Banglore properties

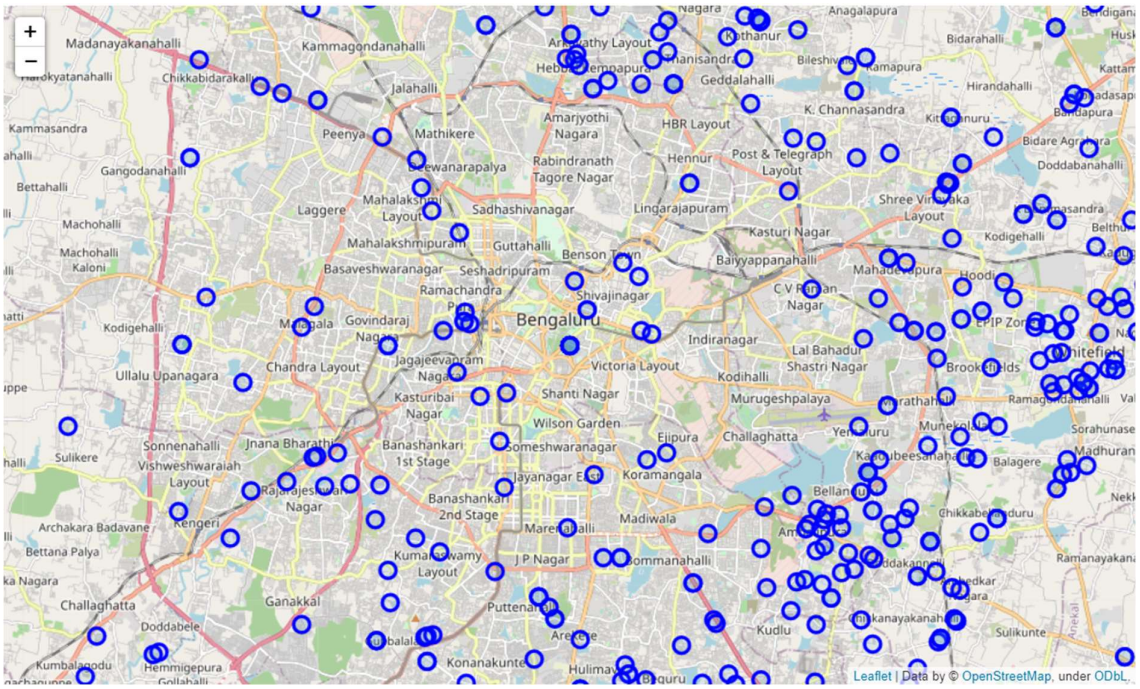


Figure 0-6 Property location in Banglore map

Like Mumbai, we can see that most of the properties listed are apartments and the area (sq.ft) and prices are more or less similar to the apartments listed in Mumbai. This makes you think that both the cities are somewhat similar to each other. Though there are more villas in Banglore compared to Mumbai which has high rates and more area (sq.ft).

EDA (Exploratory Data Analysis)

Performing EDA for Mumbai and Banglore dataset before proceeding to clustering.

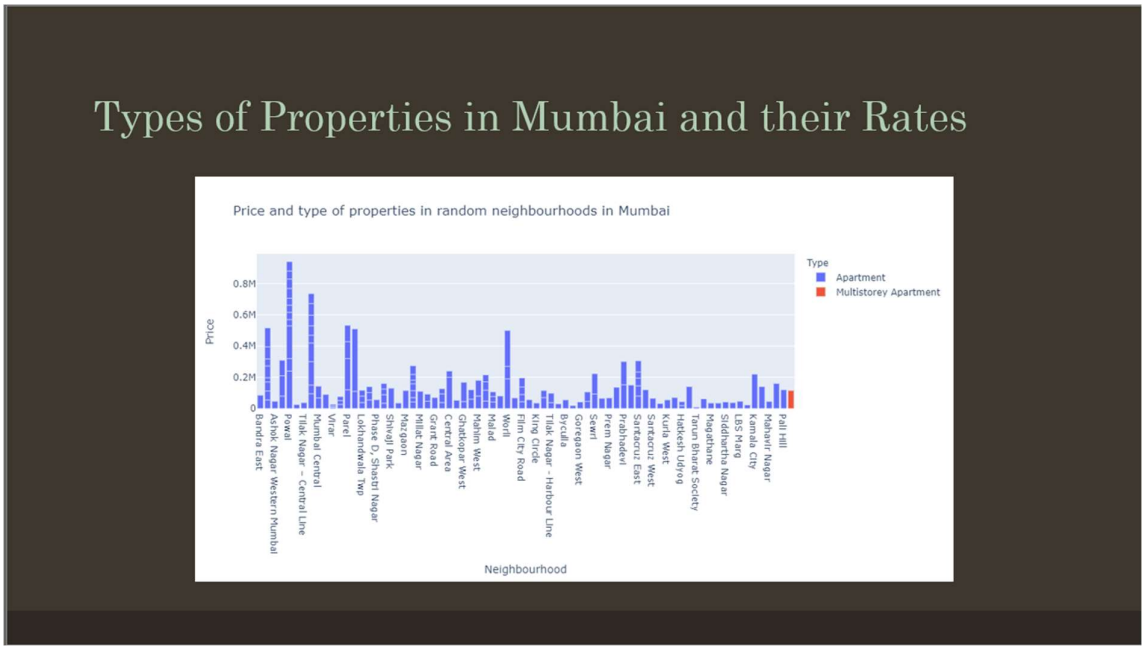


Figure 0-1 Types of properties in Mumbai

Indian Restaurants are the most common venues in the neighbourhoods of Mumbai. Most of the venues are either restaurants, fast food joints or pizza place. There are tons of coffee shops and café where the clusters are formed. The neighbourhoods with high property rates like Lokhandwala complex, Andheri West, Worli and Powai have the most common venues as the places listed above.

1

[1] Vada Pav – food delicacy in Mumbai

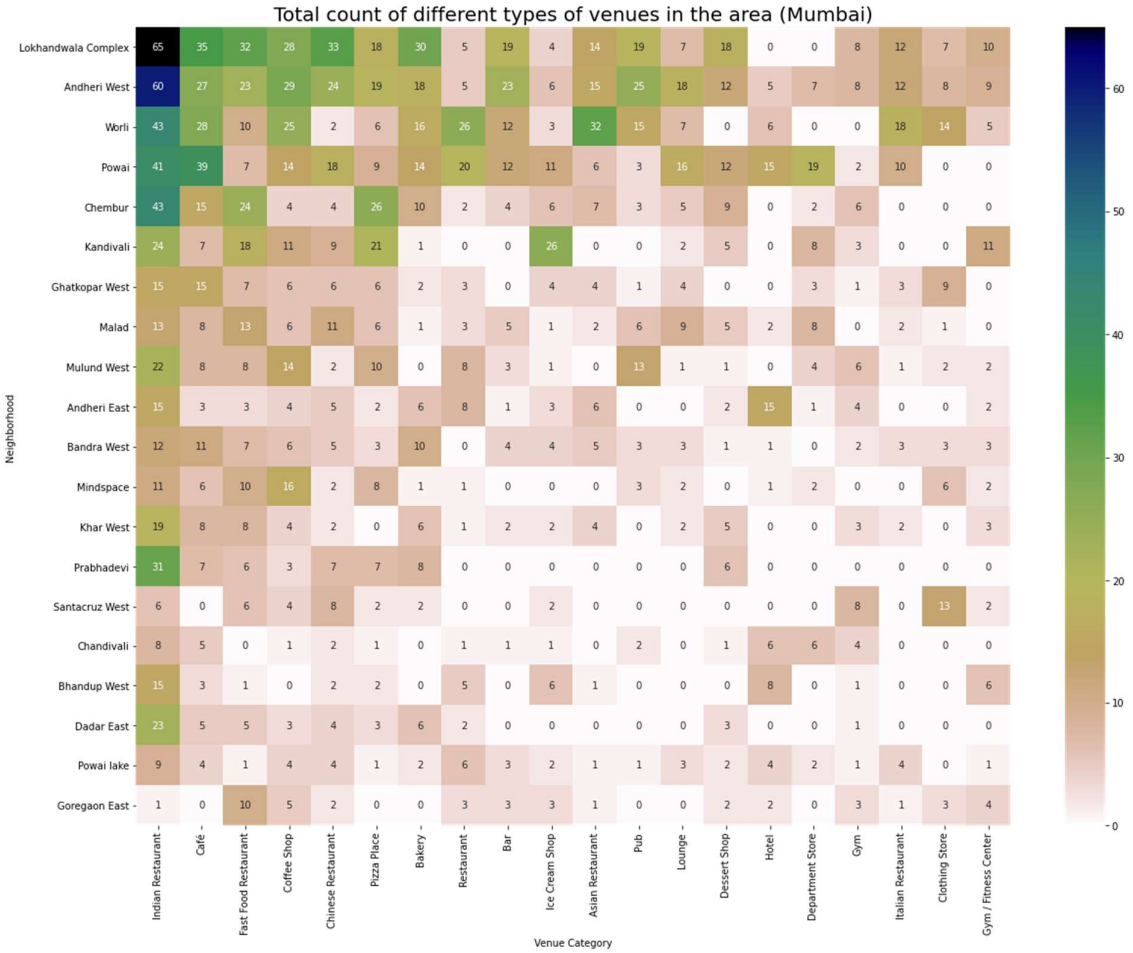


Figure 0-2 Count of venues in Mumbai

In Banglore, we can see that Dhanpriya Paradise, Citrus Florence, SKS Garden and Incor Carmel Heights are the top neighbourhoods with highest number of venues. Those venues include Indian Restaurants (which is the most common in Mumbai as well), Café, Pizza Place and Hotel. Most of the Indian Restaurants are located in Sobha Tulip, Asrithas Grand Living and Century Ethos. Overall we can say that each venues are moderately distributed in each neighbourhood.

Figure 0-3 Types of properties in Banglore

[2] Food delicacy in Banglore

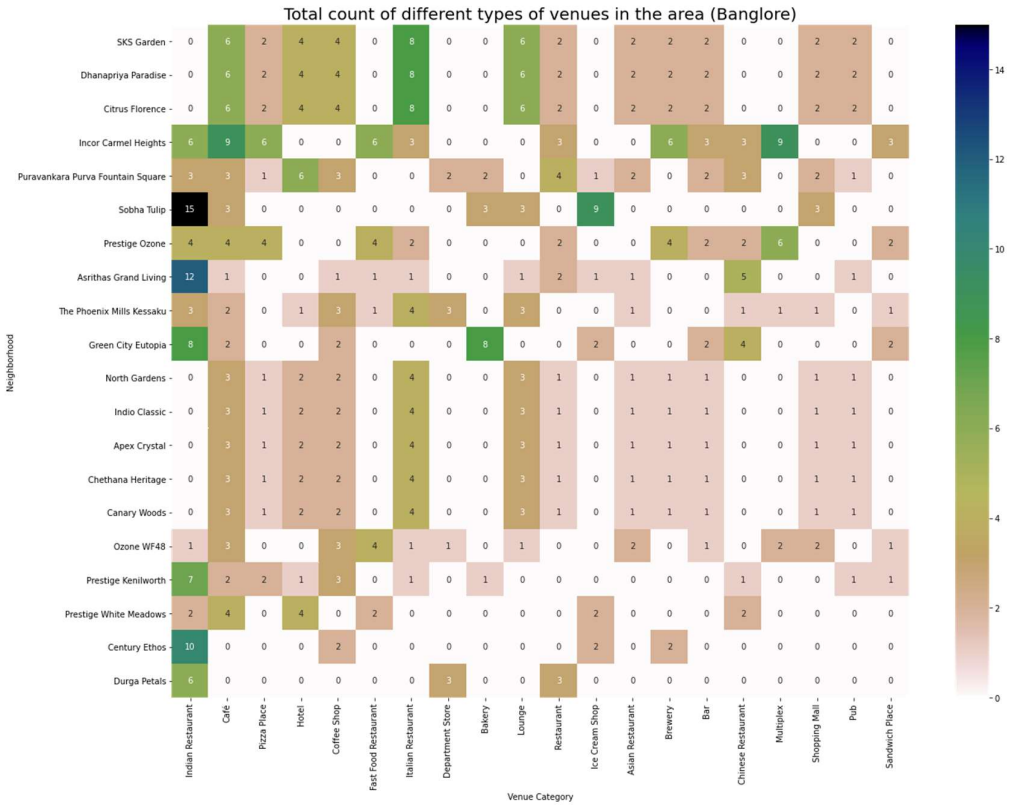


Figure 0-4 Count of properties in Banglore

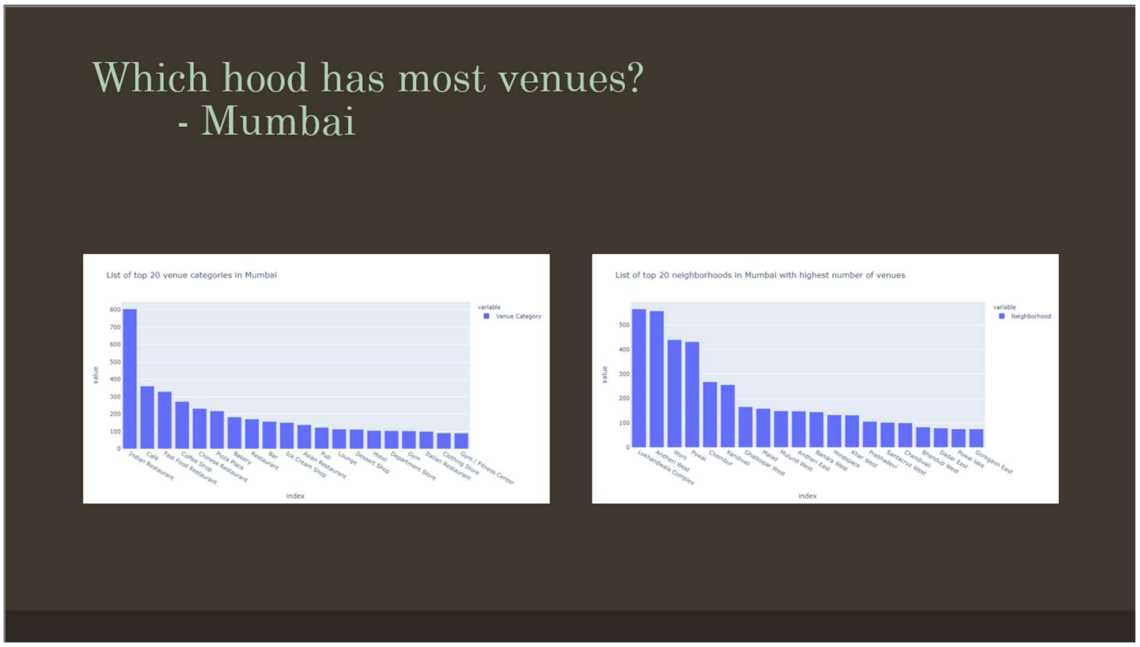
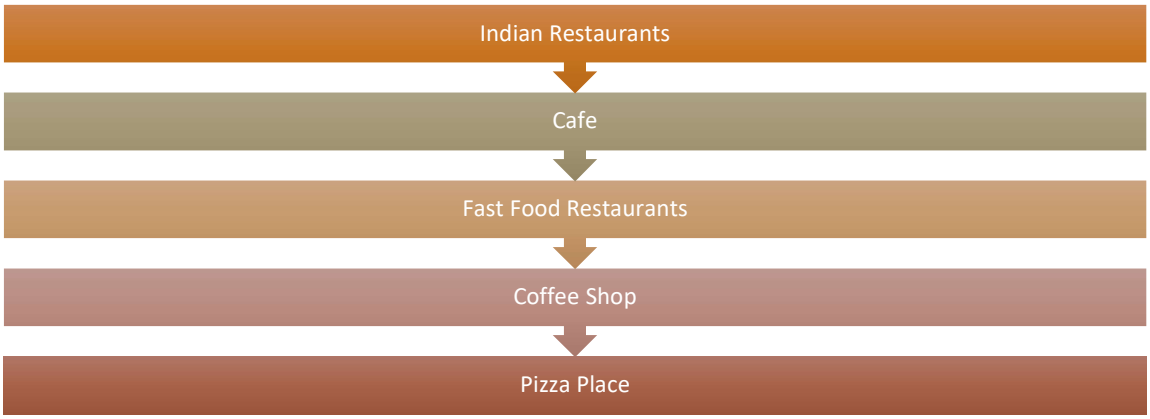


Figure 0-5 Most venues - Mumbai

Figure 0-6 Most venues - Banglore



Clustering using K-means

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Within sum of squares (WSS)

The within-cluster sum of squares is a measure of the variability of the observations within each cluster. In general, a cluster that has a small sum of squares is more compact than a cluster that has a large sum of squares. Clusters that have higher values exhibit greater variability of the observations within the cluster.

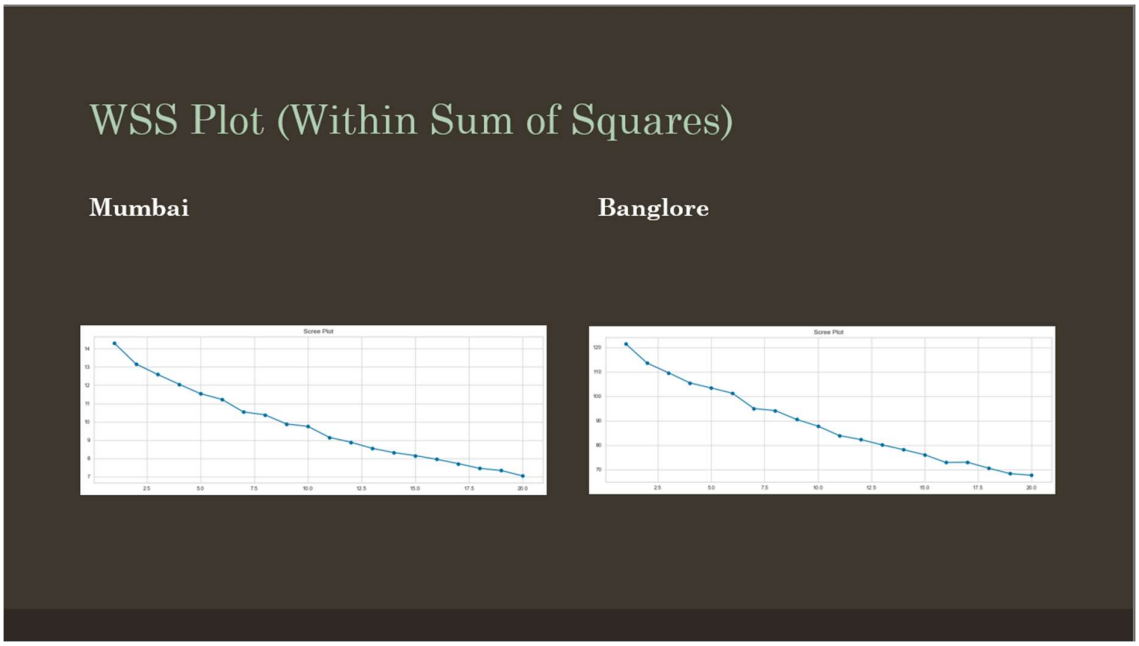


Figure 0-1 WSS Plot

For better analysis let's take the k as 5 i.e., total number of clusters as 5

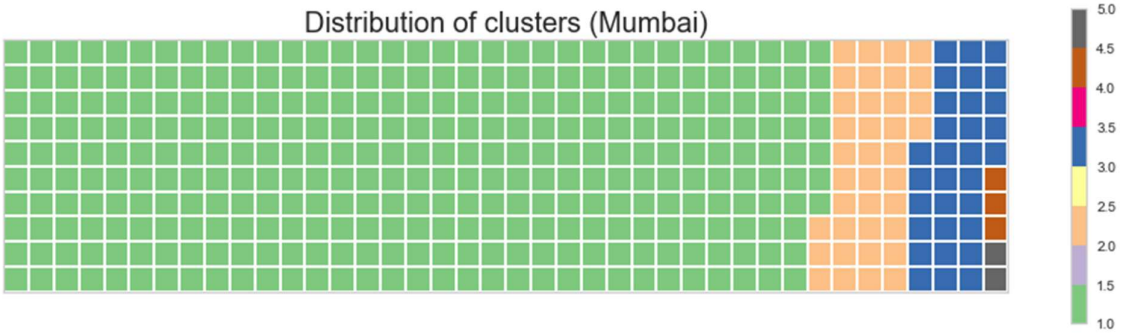


Figure 0-2 Distribution of clusters - Mumbai

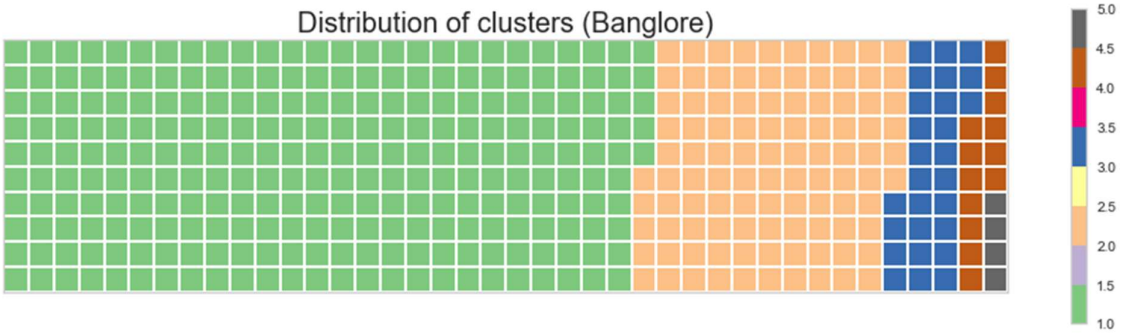
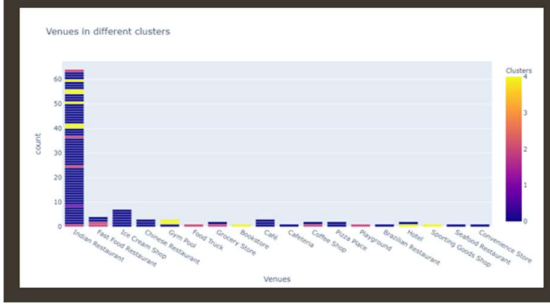


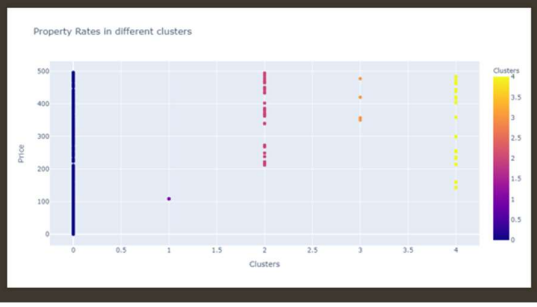
Figure 0-3 Distribution of clusters – Banglore

Mumbai Clusters

Common Venues



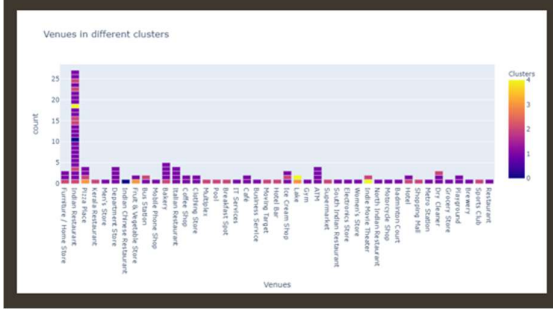
Property Rates



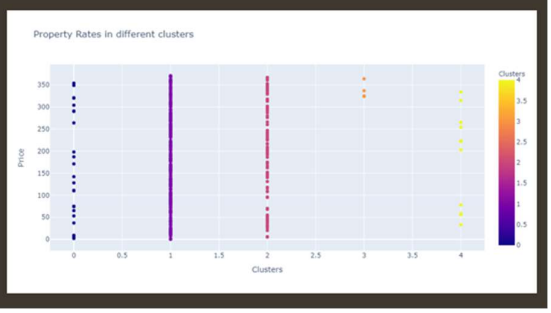
- *In Mumbai, we can see that the places with Indian Restaurants have diversified property rates.*
- *Places with shopping malls, Hotels and Department Stores have high property rates.*
- *Places with fast food restaurants, pizza place, café, theme park or historical sites have moderately high property rates.*
- *In Banglore, we can see that the places with Indian Restaurants and Karnataka Restaurant have high property rates.*
- *Places with Supermarket, Hotel, Café have diversified property rates since these venues are common and likely to be in all neighborhoods.*
- *Places with Women's Store, Food Truck or Food Court have high property rates.*
- *Places with pool, motorcycle shop, clothing store or departmental store have moderately high property rates.*

Banglore Clusters

Common Venues



Property Rates



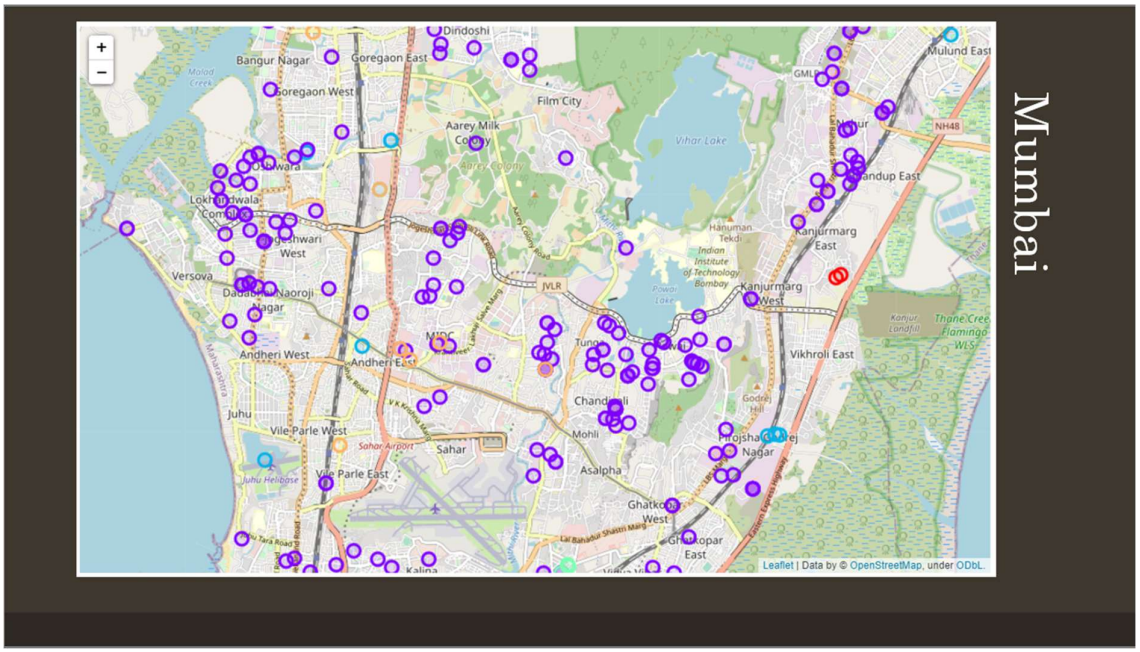


Figure 0-4 Location cluster map - Mumbai

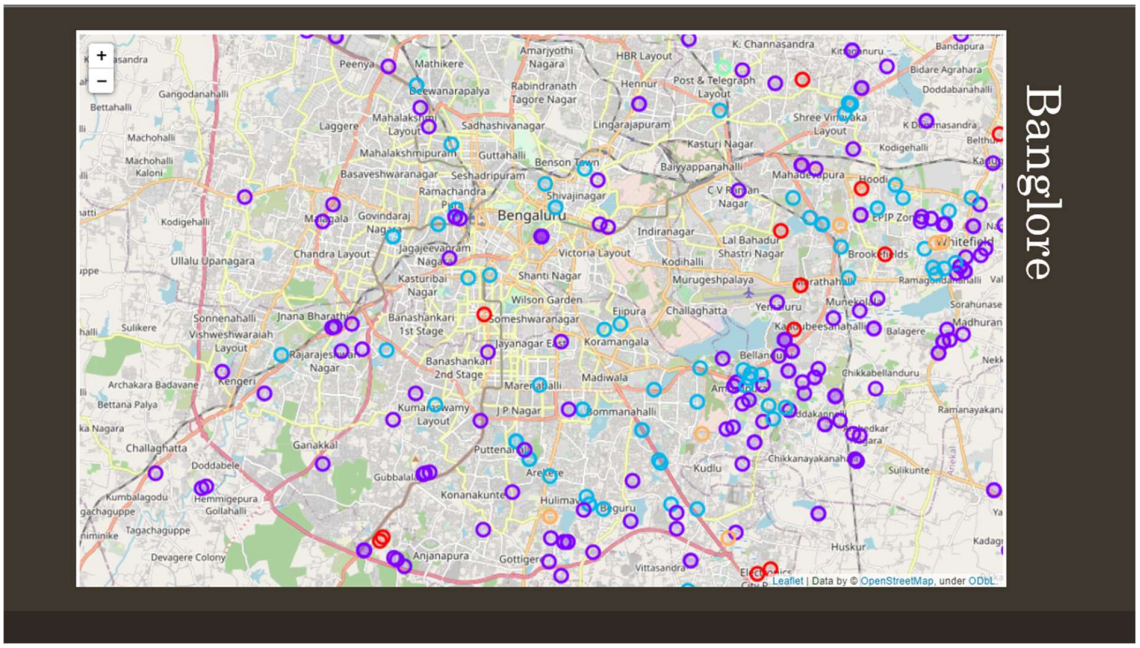


Figure 0-5 Location cluster map - Banglore

Conclusion

On analysing both the cities Mumbai and Banglore, I come to the conclusion that both the cities are similar when it comes to the properties rate though some properties in Banglore have more area (sq.ft) like a 4BHK villa which is not available in Mumbai. Similarly, there are multistorey buildings in Mumbai that are not available in Banglore. This concludes that though similar in venues and rates there are few unique things which makes these cities unique on its own.

Figure 0-1 Most common places