EXPLORATORY DATA ANALYSIS ON INDIAN HOUSING DATASET

Project Introduction and About the Dataset

The Project aims to perform Exploratory Data Analysis on the Indian Housing

Dataset. The dataset is originally scrapped from India's largest Real Estate Platform makaan.com, but the scrapped data is downloaded from Kaggle and is used to perform Exploratory Data Analysis.

The dataset comprises of 332096 rows and 32 columns.

Why I choose Indian Housing Dataset

The reasons I decided to work on this dataset are: -

1. The analysis of dataset will solve some real-world problem that belongs to real estate domain.
2. The dataset is quite large, so the is very low probability for the results to be biased.

Detailed description of the Columns of the dataset.

The columns (attributes) present in the dataset are described below.

Column	Description
Property Name	This refers to the name of the property that is listed.
Property Id	This column tells the specific id of the registered property.
Property Type	It tells what type of property it is. Whether it is Villa, Apartment, residential plot etc.
Property Status	It tells whether the property is under construction or ready for sale.
Price_per_unit_area	It gives price per unit area for that particular property.
Posted_On	It tells how many days before the property was listed on the website.
Project_URL	This URL takes you to the property page, where all the details regarding that property can be seen.
builder_id	It gives the unique builder id of the builder who has built that property.
Builder_name	It gives the name of the builder who has built that property.
Property_building_status	It tells about the property status, whether it is active, inactive or unverified.
City_id	It gives the unique id of that city where that property is located.
City_name	It gives the name of that city where that property is located.
No_of_BHK	It gives the size of the property in terms of BHK – Bathroom,Hall,Kitchen.
Locality_ID	It gives the unique id of that locality where the property is present.
Locality_Name	It gives the name of the locality where the property is present.
Longitude	It gives longitude part of that property (geographical corrdinates)
Latitude	It gives latitude part of that property (geographical coordinates).
Price	It gives the actual price of that property.
Size	It gives the size of the property in square feet.

Sub_urban_ID	It gives the unique id corresponding to that specific sub urban region from which the property belongs.
Sub_urban_name	It gives the name of the sub urban region from where the property belongs.
is_furnished	It tells whether the property is furnished or not.
listing_domain_score	It gives the domain score of that property. It basically ranks your property link.
is_plot	It tells whether the property is simple plot or not.
is_RERA_registered	It tells whether the property is registered under Real Estate Regulatory Authority.
is_Apartment	It tells whether the property is Apartment or not.
is_ready_to_move	It tells whether the property is ready for sale or not.
is_commercial_Listing	It tells whether the property is for commercial purpose or not.
is_PentaHouse	It tells whether the property is penthouse or not.
is_studio	It tells whether the property is studio or not.
Listing_Category	It tells whether the property is to be sold or to give for rental etc.

What are the insights I want to retrieve from the dataset?

The insights that I want to draw from this dataset are –

1. Total number of properties in each city.
2. Number of properties as per BHK and RK in each city.
3. Properties in different regions and states across India.
4. Price of Properties in different regions.
5. Ready for sale and under construction properties in each city.
6. Average price per square feet for each locality.

7. Average price of property as per locality.
8. Different types of property (by quantity) in each locality.

Approach and Steps that will be followed

1. Importing dataset and the required libraries.
2. Interpretation of different columns.
3. Performing various techniques for missing value imputation
4. Standardising the data.
5. Finding outliers and removing duplicates and irrelevant data from the dataset
6. performing univariate analysis
7. performing bivariate analysis and multivariate analysis

Univariate Analysis

Types –

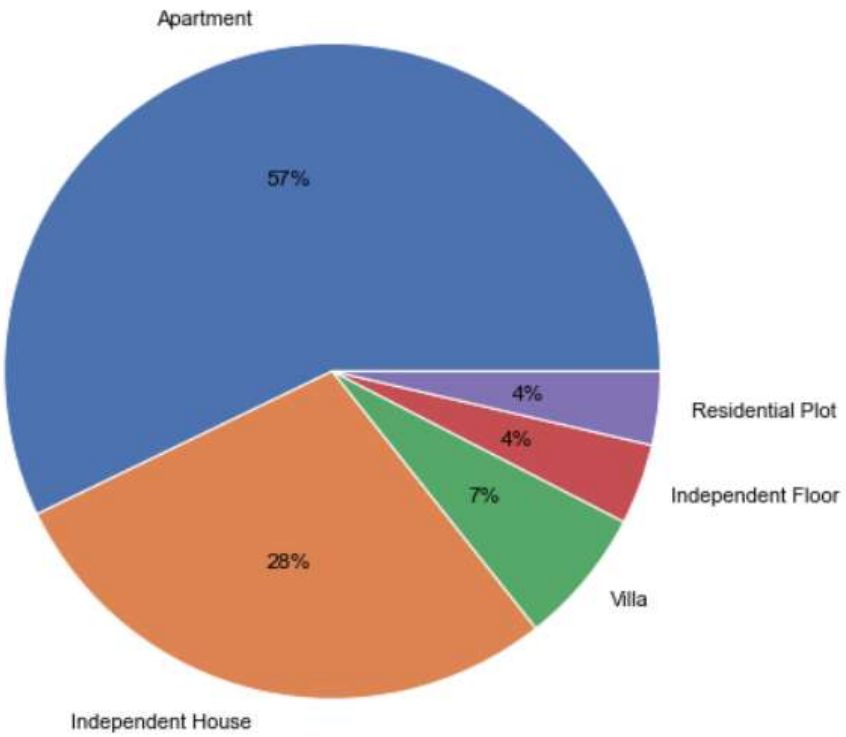
Categorical Unordered Univariate Analysis

Categorical Ordered Univariate Analysis

Statistics on Numerical features

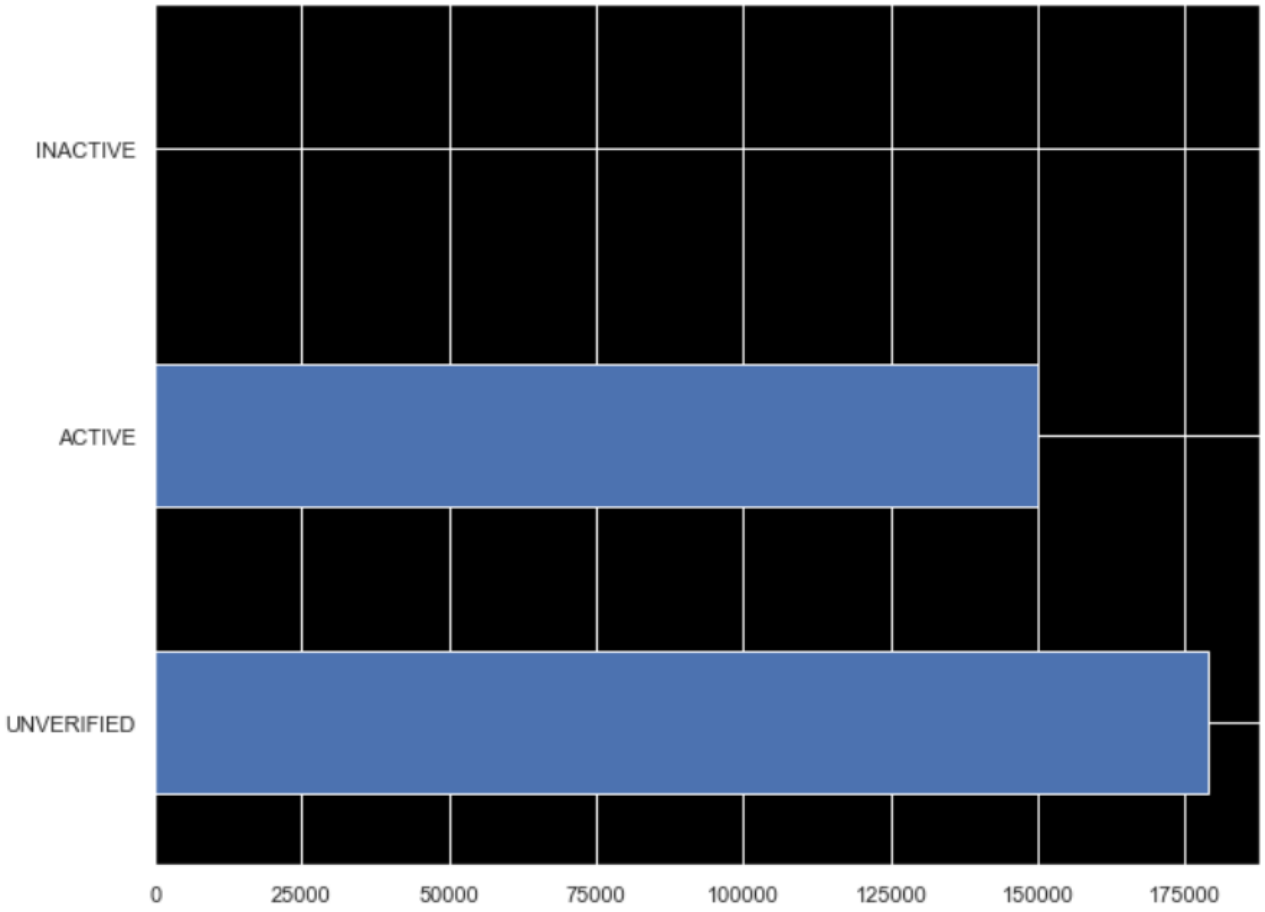
A. Categorical Unordered Univariate Analysis

1. Percentage share of different kinds of properties



The above chart shows what is the percentage share of each type of property among all the properties.
This shows that Apartment comprises of most of the properties, whereas residential plots and independent floors are least.

2. Current Status of Properties.



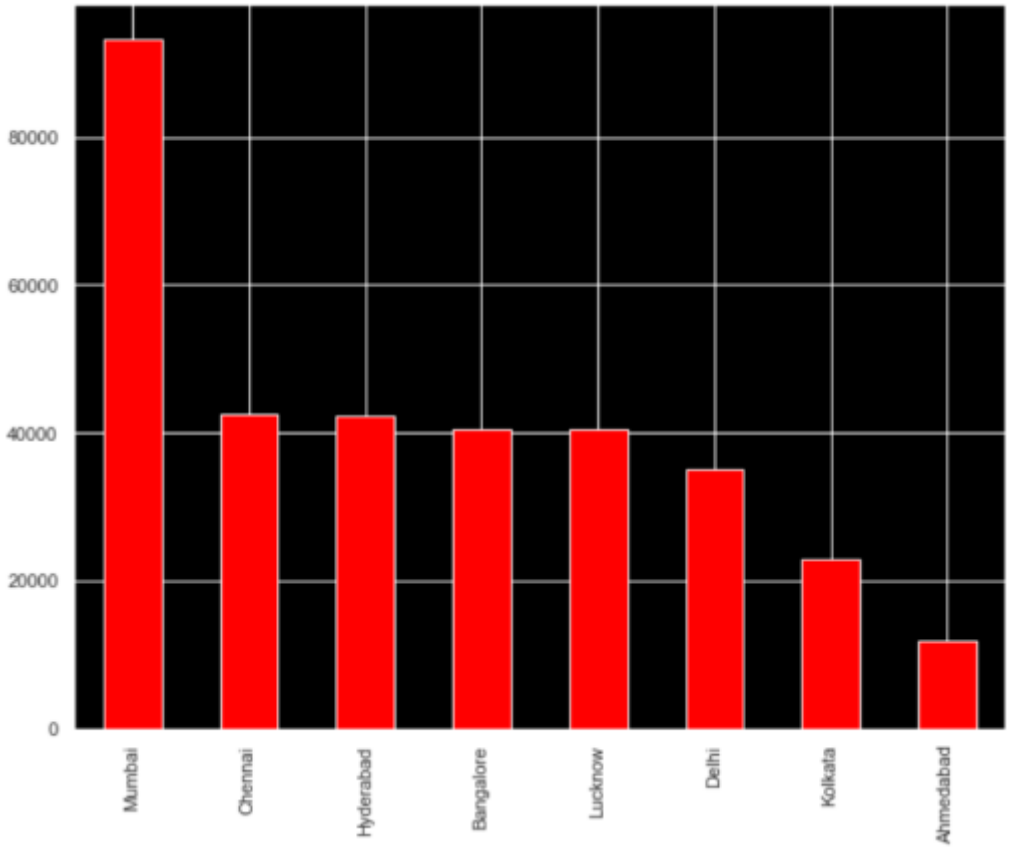
This shows almost negligible properties are inactive, whereas most of the properties are unverified.

```
print(df.Property_building_status.value_counts())
```

```
UNVERIFIED    178702
ACTIVE        149857
INACTIVE         126
Name: Property_building_status, dtype: int64
```

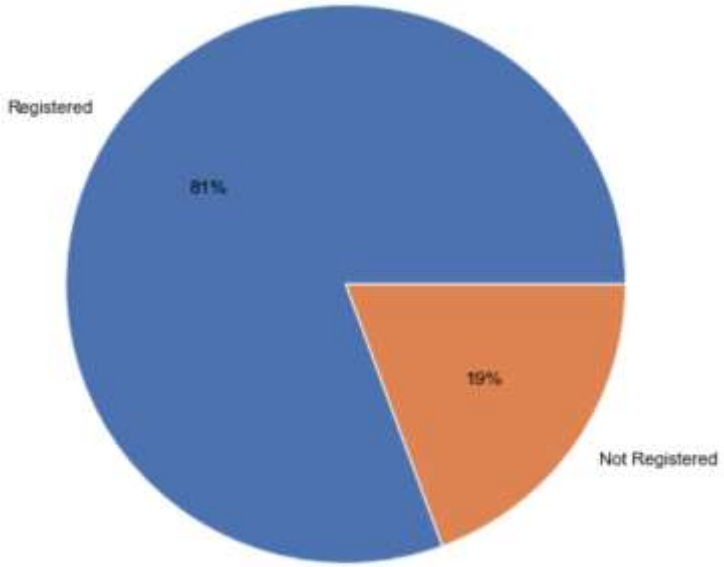
We can see 178702 are unverified, 149857 are verified and only 126 are inactive.

3. **Number of Properties in each city**



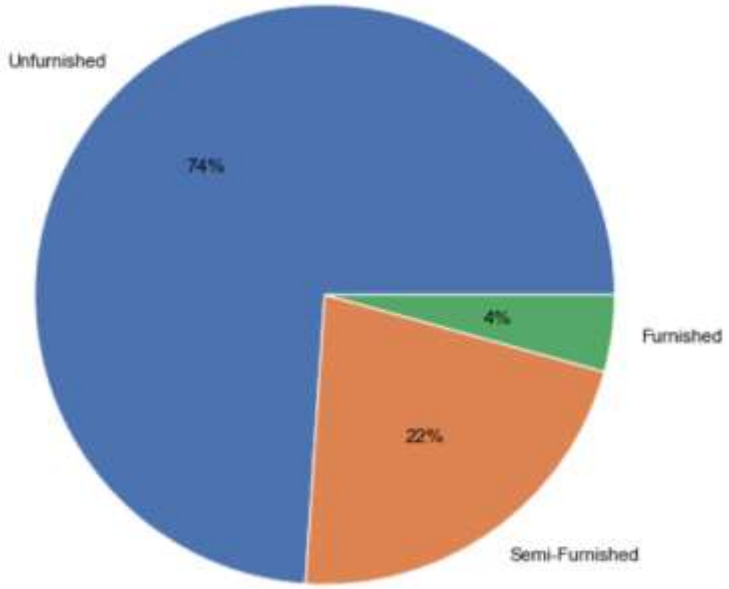
This shows that Mumbai has highest number of properties to be sold.

4. **Percentage of Properties Registered under RERA (Real Estate Regulatory Authority)**



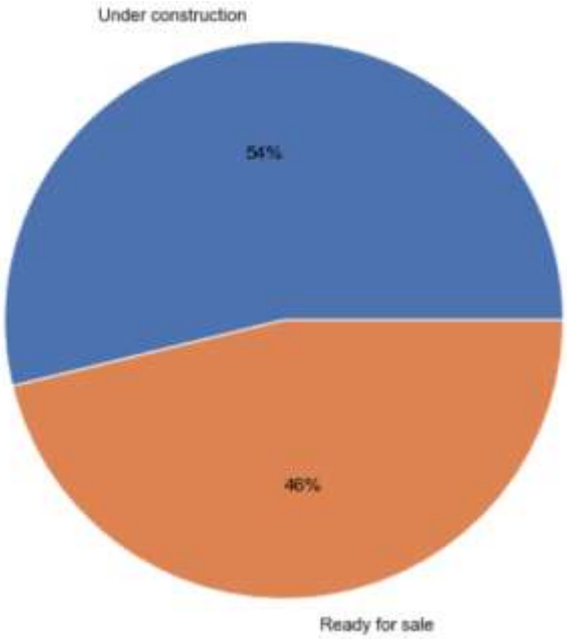
This shows only 81% of all the properties are officially registered under RERA.

5. **Percentage share of furnished and unfurnished properties**



Out of all the properties, almost 74% properties are unfurnished and 22% semi-furnished, whereas only 4% properties are fully furnished.

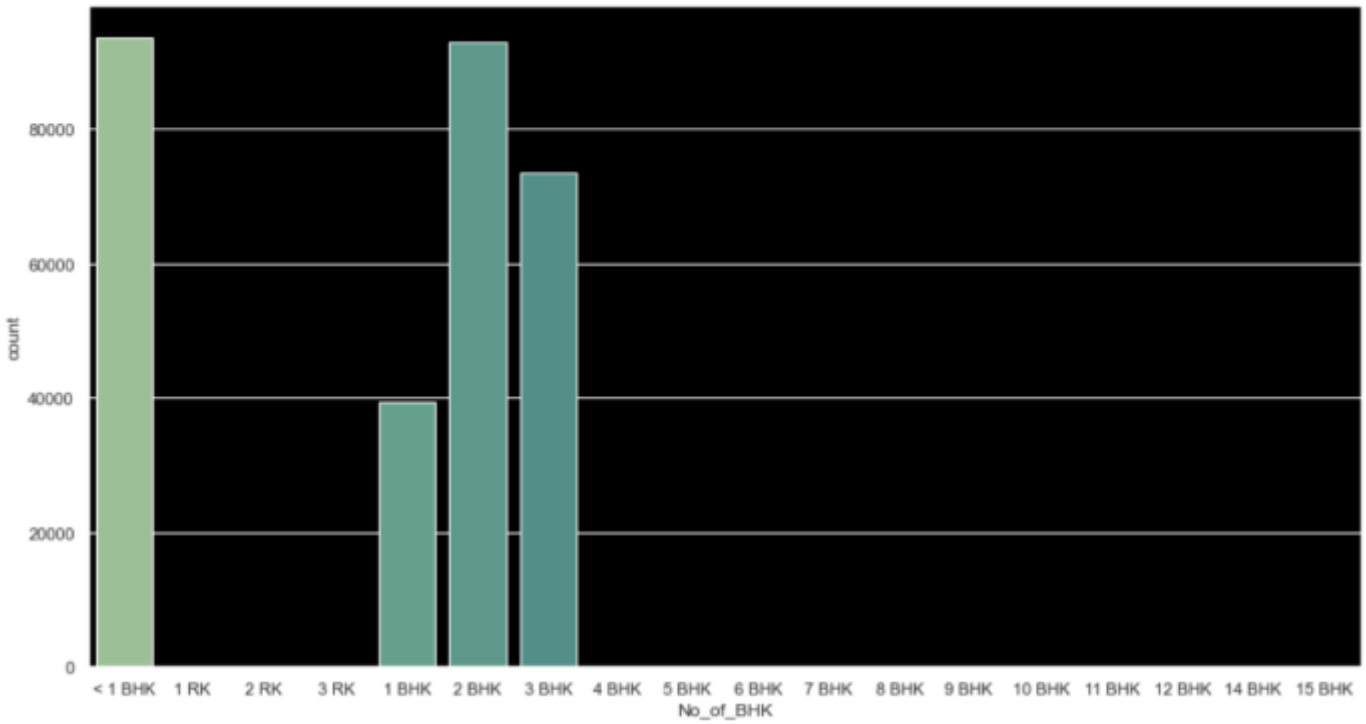
6. **Percentage of properties that are under – construction and those which are ready for sale**



Out of all the listed properties, around 54% properties are under construction and not yet ready for sale.

B. Categorical Ordered Univariate Analysis

1. Number of properties available in different BHK (bathroom, hall, kitchen) and RK (room, kitchen) category




```
df['No_of_BHK'].value_counts()
```

```
< 1 BHK     93586
2 BHK       92959
3 BHK       73492
1 BHK       39486
4 BHK       20978
5 BHK        3191
1 RK         3054
6 BHK         904
10 BHK        303
7 BHK         298
8 BHK         232
9 BHK         155
12 BHK         14
11 BHK         14
15 BHK         13
3 RK            2
14 BHK          2
2 RK            2
```

This shows that most of the properties are 1 BHK, 2 BHK, 3 BHK, 4 BHK or less than 1 BHK.

C. Statistics on Numerical features.

1. Price per unit area

```
df.Price_per_unit_area.describe()
```

```
count    328685.000000
mean       7935.322190
std        7821.205728
min           0.000000
25%        2950.000000
50%        5517.000000
75%        9990.000000
max       52272.000000
```

Average price per unit area among all the cities is 7935 Rs, whereas almost 50% (median) properties are less than 5517 Rs.

2. Size of Property

```
df.Size_in_sq_feet.describe()
```

```
count    328685.000000
mean       1417.297613
std        1220.240405
min          10.000000
25%         830.000000
50%        1100.000000
75%        1600.000000
max       20000.000000
```

Average size of properties is 1417 square feet, whereas maximum size of any property is 20000 square feet.

3. Domain score of property

```
df.listing_domain_score.describe()
```

```
count    328685.000000
mean          4.005617
std           0.124663
min           4.000000
25%           4.000000
50%           4.000000
75%           4.000000
max           9.107140
```

The minimum domain score of any property is 4, whereas 25th, 50th, and 75th percentile is also 4, but the maximum value of any domain is 9.1

Bivariate Analysis

Types –

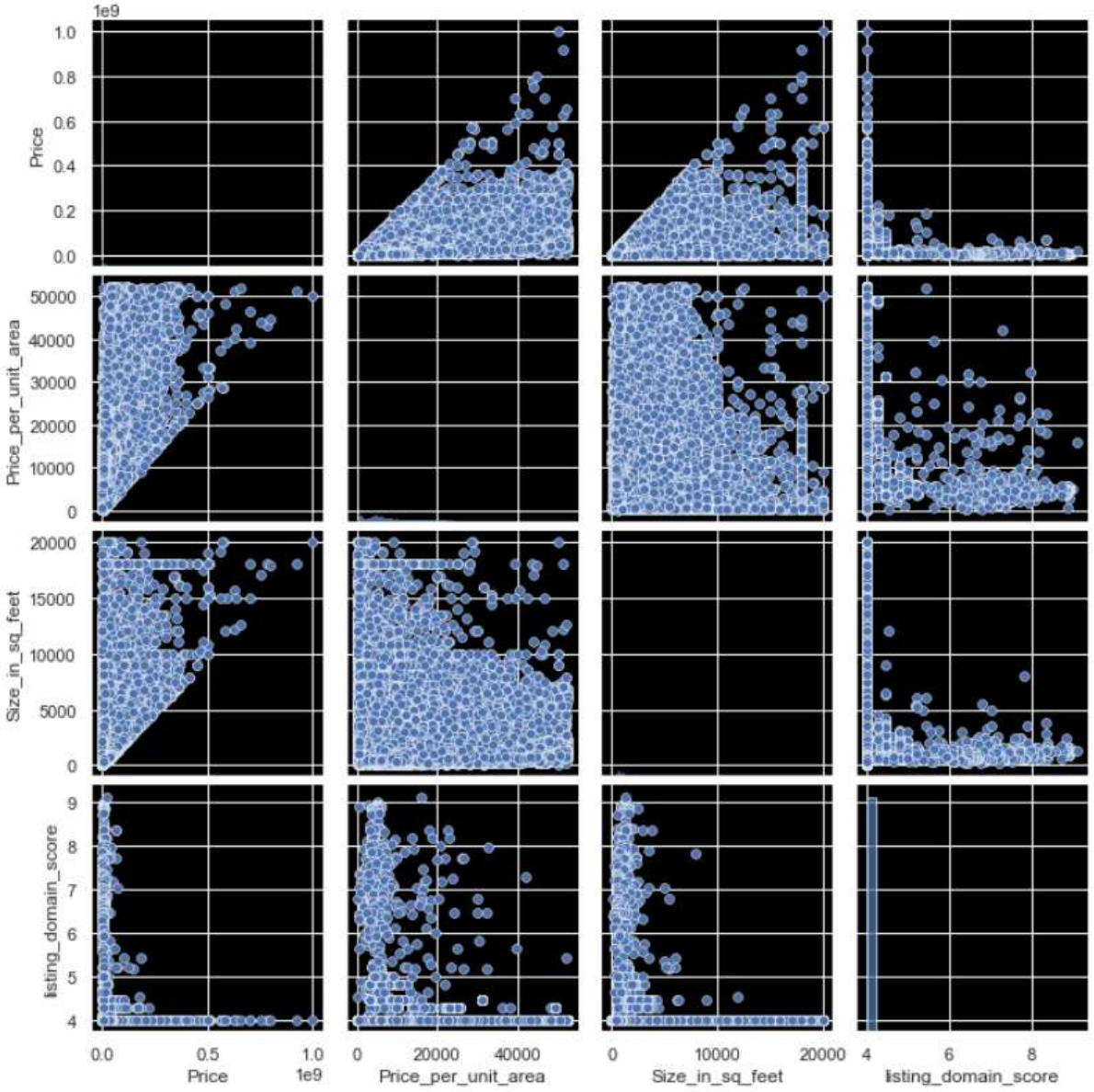
Numerical – Numerical Analysis

Numerical – Categorical Analysis

Categorical – Categorical Analysis

A. Numerical – Numerical Analysis

1. Correlation matrix of different attributes



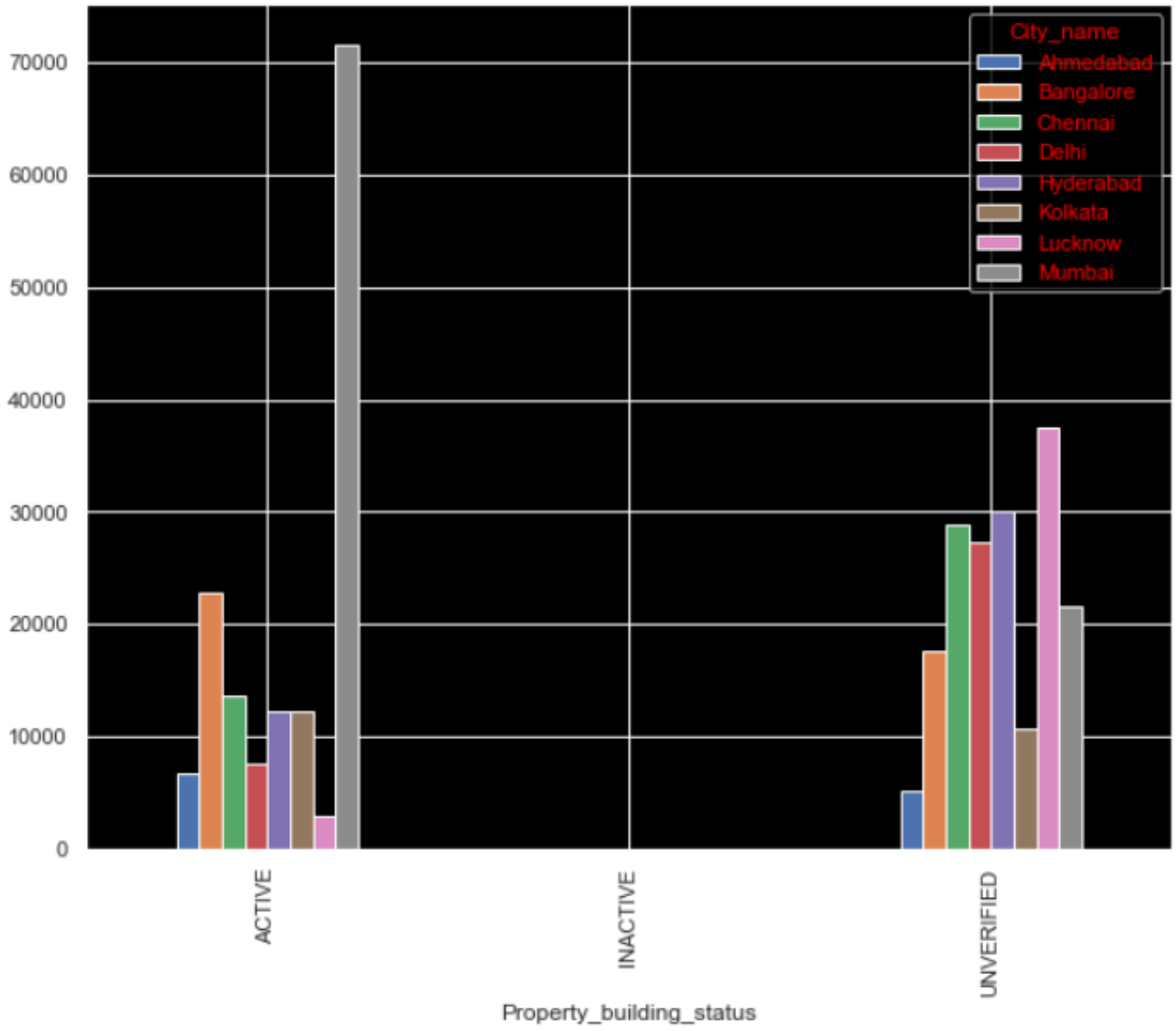
There does not seem any significant correlation between any of the numerical features.
Also, it is important to note that the high correlation does not always imply causation.

B. Numerical – Categorical Analysis

1. Property building status in each city

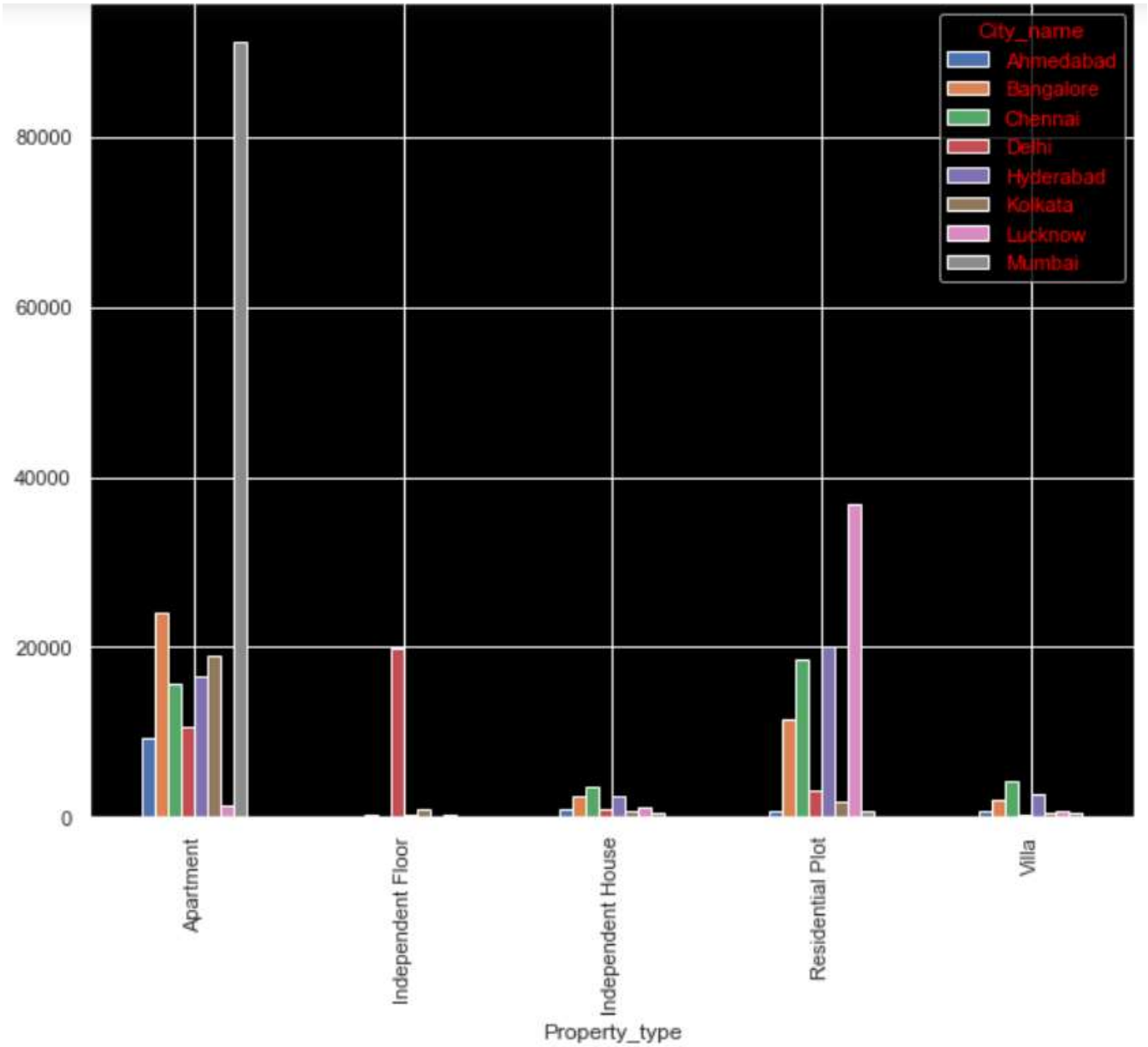
```
df.groupby('City_name').Property_building_status.value_counts().unstack(0).plot.bar()
```

```
<AxesSubplot:xlabel='Property_building_status'>
```



Mumbai has the highest number of active properties, whereas Lucknow has highest number of unverified properties.

2. Property type in each city

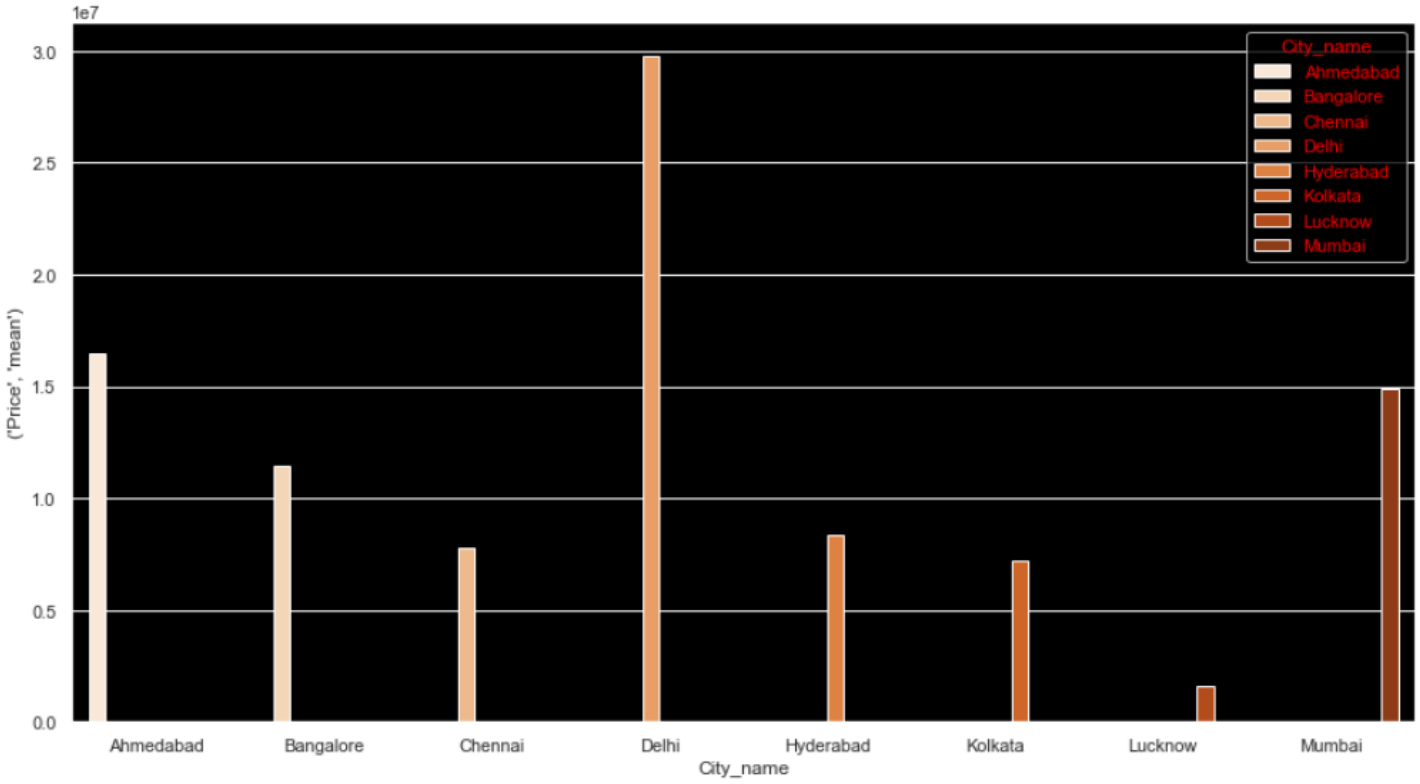


Cities with highest number in property type

- Apartment = Mumbai
- Independent floor = Delhi
- Independent House = Chennai
- Residential plot = Lucknow
- Villa = Chennai

3. Average Price of property in any City.

```
df.groupby('City_name').agg([np.mean])[['Price']]
```

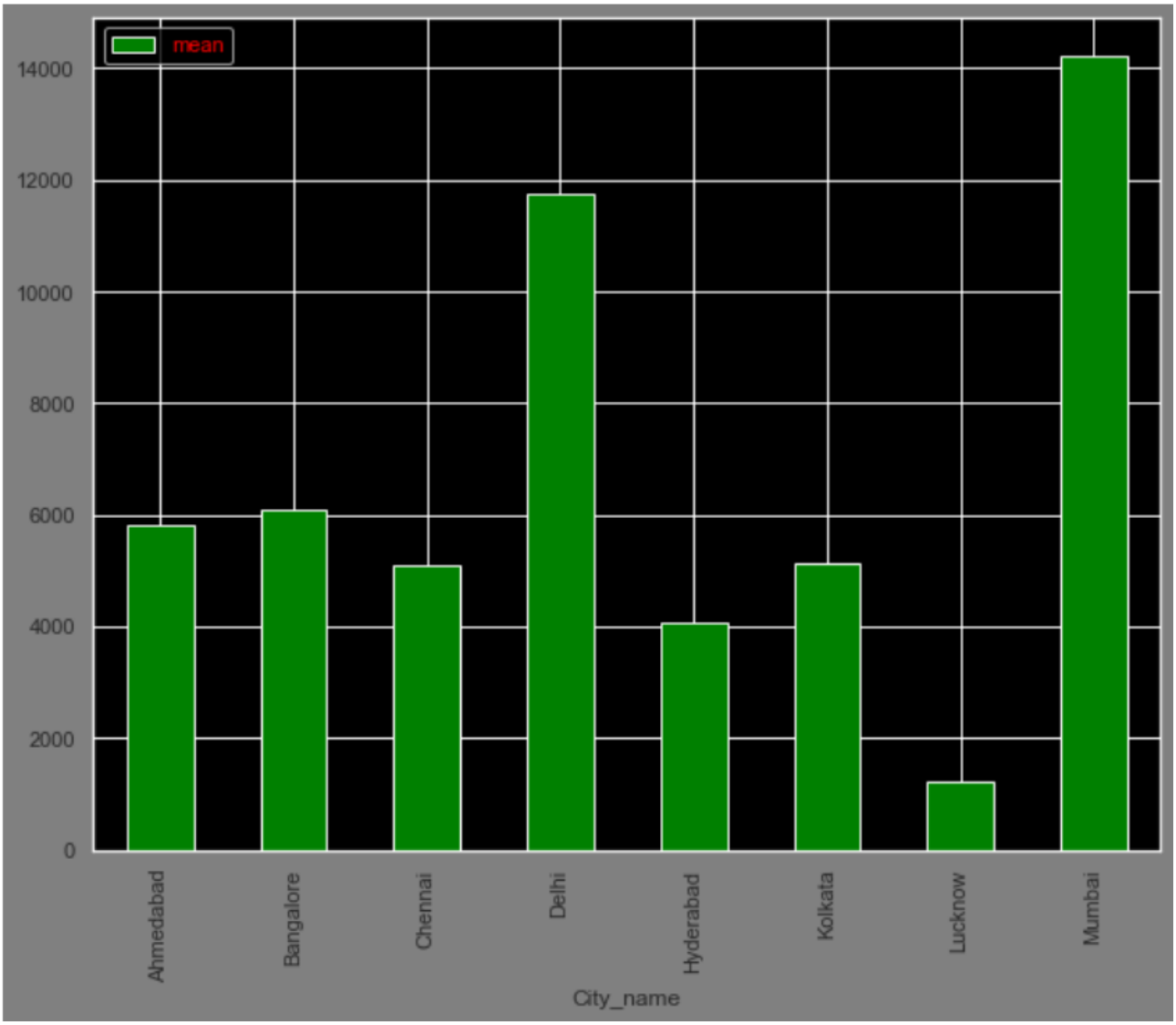


Delhi has the highest average price among all the cities. But this does not necessarily mean that Delhi is the most expensive city.

Let’s check the next feature to find this out.

4. Average Price per unit area in each city

```
df.groupby('City_name').agg([np.mean])['Price_per_unit_area'].plot.bar(color = 'Green')
```



As price per unit area is a better metric than average price, we can say that Mumbai the most expensive city in India in terms of real estate properties.

C. Categorical – Categorical Analysis

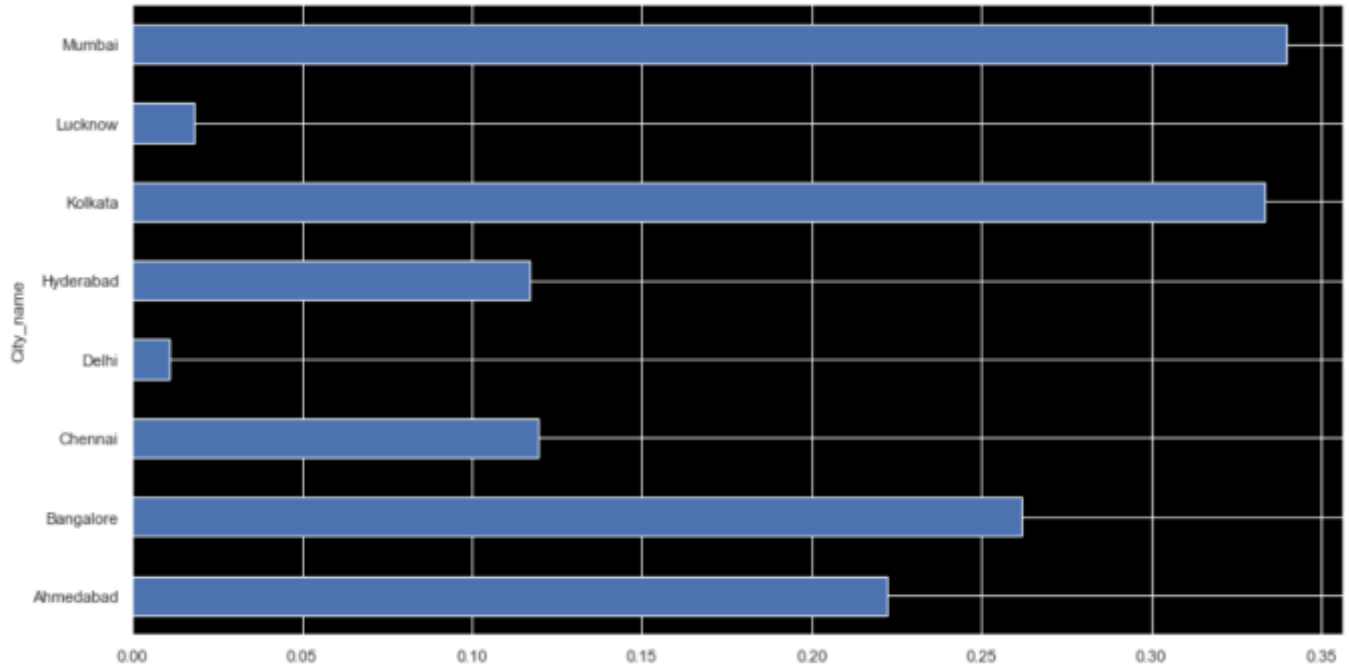
1. Properties registered under RERA v/s City name

```
df["registered_property"] = np.where(df.is_RERA_registered==True,1,0)
df["registered_property"].value_counts()
```

```
0    264979
1     63706
```

```
df.groupby(["City_name"])["registered_property"].mean().plot.barh()
```

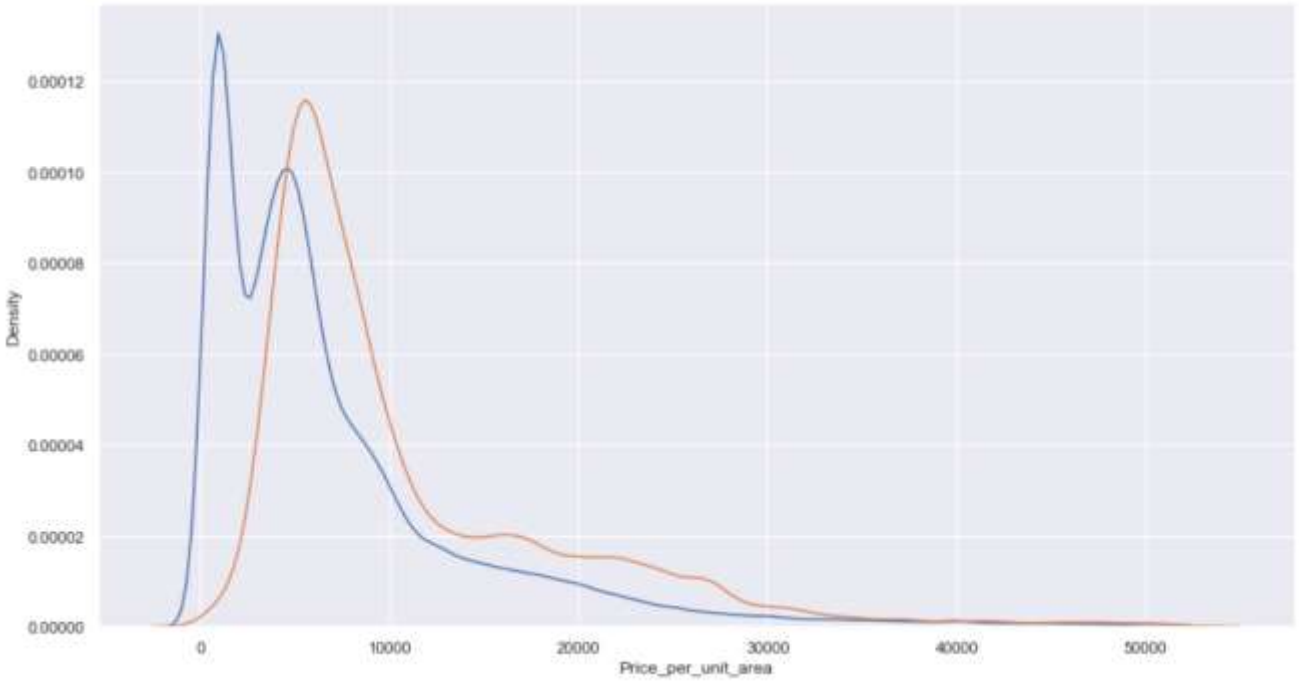
```
<AxesSubplot:ylabel='City_name'>
```



Mumbai has the highest percentage of registered properties, whereas in Delhi most of the properties are not registered officially.

2. **How Registration of property under Real Estate Authority varies with its Price per unit area**.

```
sns.distplot(df[df["registered_property"]==0]["Price_per_unit_area"],hist=False)
sns.distplot(df[df["registered_property"]==1]["Price_per_unit_area"],hist=False)
```

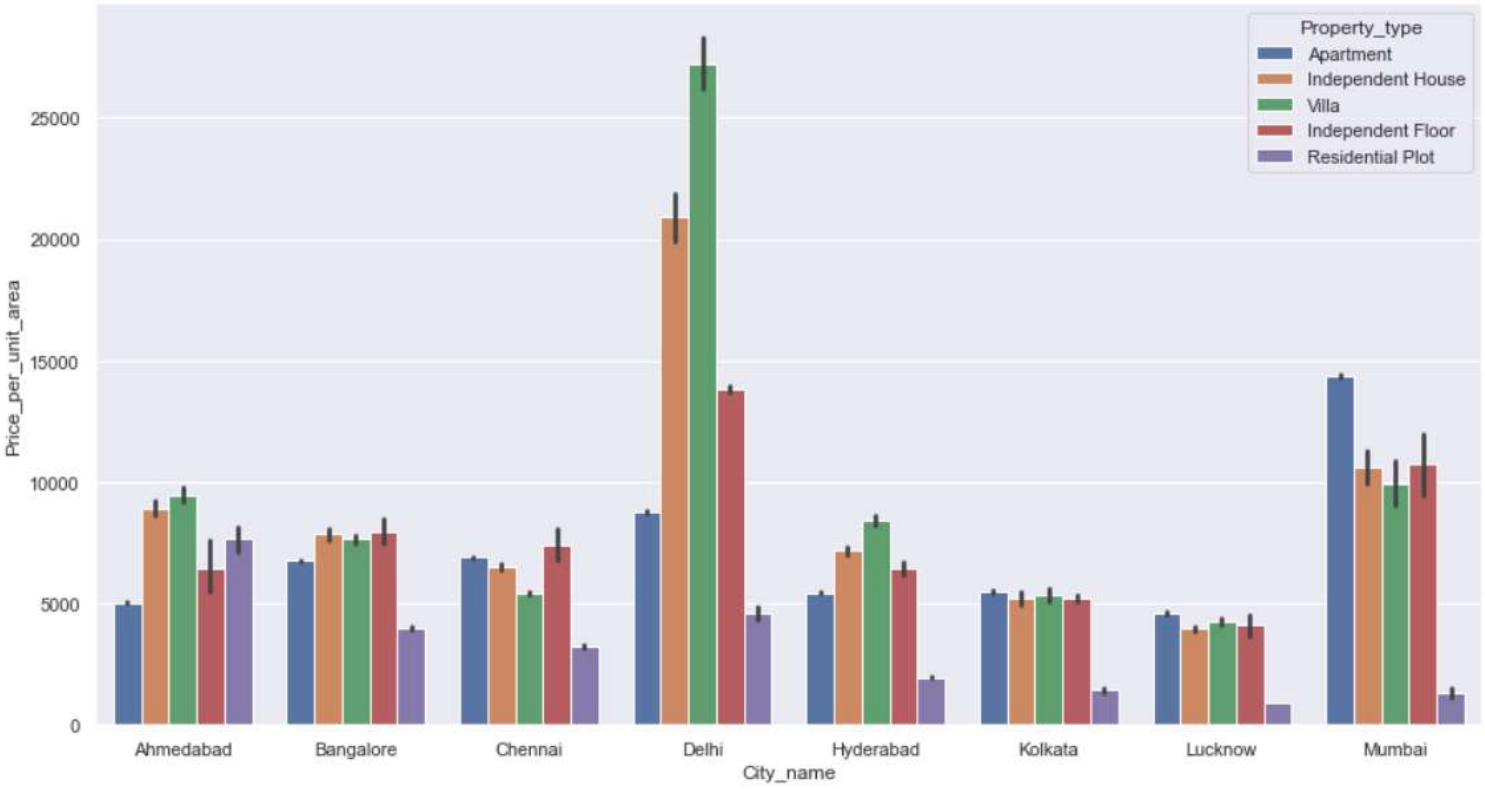


- It can be seen from the plot below that, for buildings with lower Price per unit area, we have a greater number of unregistered buildings, but as the Price per unit area increases, the number of registered buildings are more than unregistered buildings.
- This shows that most of the builders who have expensive properties register those properties with RERA (Real Estate Regulatory Authority).

Multivariate Analysis

1. **Different types of property with their price per unit area in different cities.**

```
sns.barplot(df["City_name"],df["Price_per_unit_area"],hue=df["Property_type"])
```

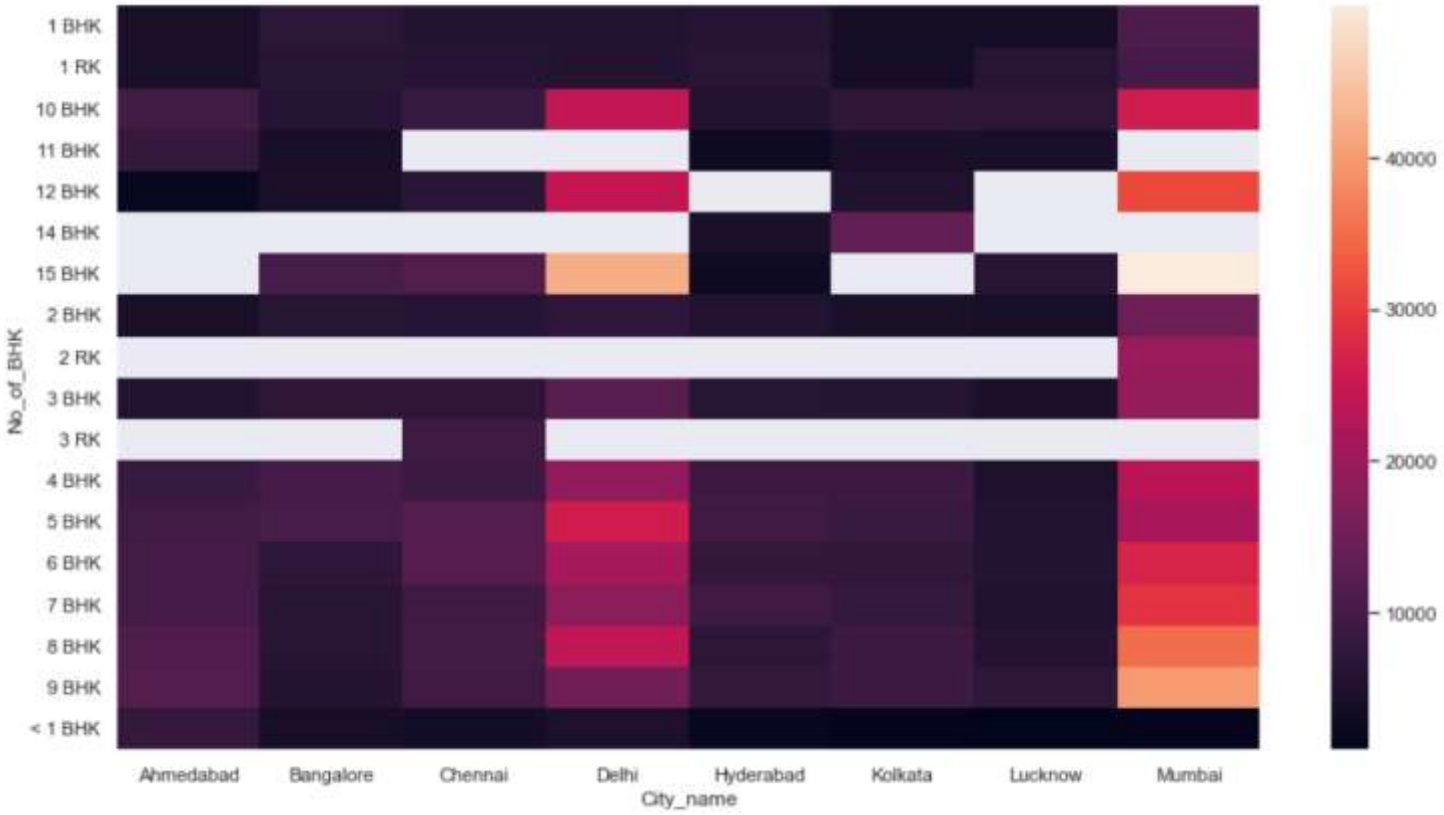


On Average,

- Villa is most expensive Property type in Delhi.
- Apartments are most expensive in Mumbai.
- Independent houses are most expensive in Delhi.
- Independent floors are most expensive in Delhi.
- Residential plots are most expensive in Ahmedabad.

2. Price per unit area as per BHK and City

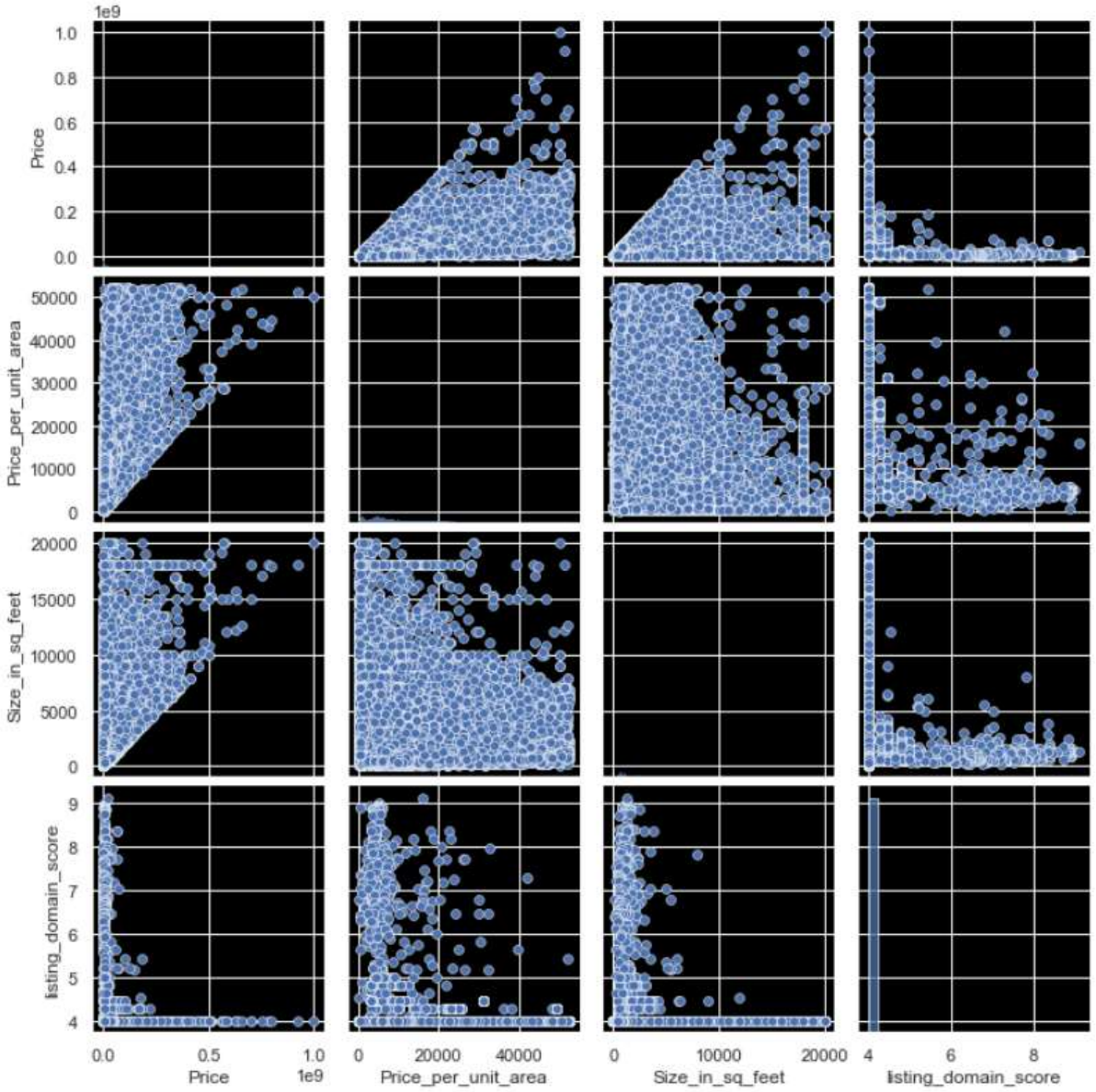
```
sns.heatmap(df.pivot_table(index="No_of_BHK",columns="City_name",values="Price_per_unit_area"))
```



- This shows that for 3 RK house, Price per unit area is very high except in Chennai.
- Similarly, for 2 RK house, Price per unit area is very high except in Mumbai.
- For 1 BHK, 2 BHK, 3BHK, 4BHK, Mumbai has the highest Price per unit area.
- For 5BHK, Delhi seems to have highest Price per unit area.
- For 6,7,8,9,10 BHK, Mumbai seems to have highest Price per unit area.

3. Pair Plots of all Numerical columns to find correlation between them.

```
sns.pairplot(data=df,vars=["Price", "Price_per_unit_area","Size_in_sq_feet","listing_domain_score"])
plt.show()
```



From this we can observe that there is no correlation between any numerical variables.

Drawing Important Inferences from the complete process of Exploratory Data Analysis.

1. **How Registration of property under Real Estate Authority varies with its Price per unit area.**
   - For buildings with lower Price per unit area, we have a greater number of unregistered buildings, but as the Price per unit area increases, the number of registered buildings is more than unregistered buildings.
   - This shows that most of the builders who have expensive properties register those properties with RERA (Real Estate Regulatory Authority).

2. **Property building status in each city**
   - Mumbai has the highest number of active properties, whereas Lucknow has highest number of unverified properties.

3. **Properties registered under RERA v/s City name**
   - Mumbai has the highest percentage of registered properties, whereas in Delhi most of the properties are not registered officially.

4. **Percentage of properties that are under – construction and those which are ready for sale**
   - Out of all the listed properties, around 54% properties are under construction and not yet ready for sale.

5. **Percentage share of furnished and unfurnished properties**
   - Out of all the properties, almost 74% properties are unfurnished and 22% semi-furnished, whereas only 4% properties are fully furnished.

6. **Price per unit area as per BHK and City**
   - For 3 RK house, Price per unit area is very high except in Chennai.
   - Similarly, for 2 RK house, Price per unit area is very high except in Mumbai.
   - For 1 BHK, 2 BHK, 3BHK, 4BHK, Mumbai has the highest Price per unit area.
   - For 5BHK, Delhi seems to have highest Price per unit area.
   - For 6,7,8,9,10 BHK, Mumbai seems to have highest Price per unit area.

7. **Percentage of Properties Registered under RERA (Real Estate Regulatory Authority)**
   - This shows only 81% of all the properties are officially registered under RERA, whereas 19% are not.

8. **Different types of property with their price per unit area in different cities.**
   On Average,
   - Villa is most expensive Property type in Delhi.
   - Apartments are most expensive in Mumbai.
   - Independent houses are most expensive in Delhi.
   - Independent floors are most expensive in Delhi.
   - Residential plots are most expensive in Ahmedabad.

9. **Mumbai has highest number of properties to be sold.**

The End