DATA SCIENCE LAB EXPERIMENT 2

NAME: Laukik Padgaonkar DIV: D15C Roll No: 37

AIM: Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn.

THEORY:

Matplotlib is a powerful Python library used for data visualization. It provides a variety of plotting functions to create static, animated, and interactive visualizations. It is widely used for plotting line charts, bar charts, histograms, scatter plots, and more.

- Key module: `pyplot` (`import matplotlib.pyplot as plt`)
- Customization: Supports labels, colors, styles, grids, legends, etc.

Seaborn is a data visualization library built on top of Matplotlib, designed specifically for statistical data visualization. It provides more aesthetically pleasing and informative plots compared to Matplotlib.

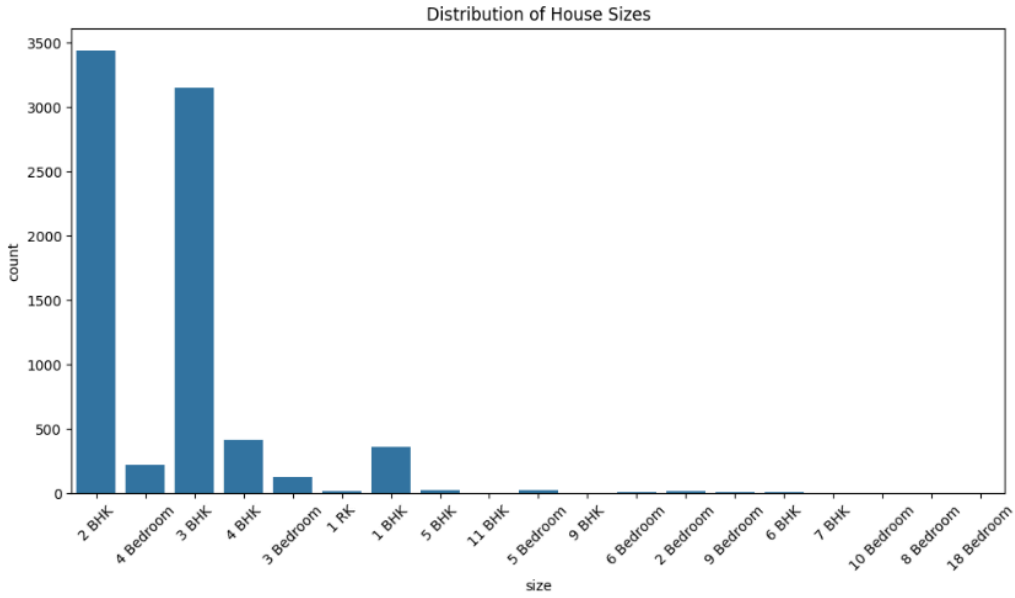
- Key module: `seaborn` (`import seaborn as sns`)
- Built-in themes and color palettes for better visualization.
- Works well with Pandas DataFrames and supports advanced statistical plots like box plots, violin plots, pair plots, heatmaps, etc.

Topic: Bengaluru Housing Prices

DATA SCIENCE LAB EXPERIMENT 2

1. Create a Bar Graph using any 2 features:
   A bar graph (bar chart) is a common visualization technique used to represent categorical data with rectangular bars. When using two features, it helps in comparative analysis between different categories across another variable.

```
# prompt: make a bar graph for this dataset

import matplotlib.pyplot as plt
# Assuming 'data' DataFrame is already loaded and processed as in the provided code.

# Bar plot of 'size' column (assuming it exists)
plt.figure(figsize=(12, 6))
sns.countplot(x='size', data=data)
plt.xticks(rotation=45)
plt.title('Distribution of House Sizes')
plt.show()
```



We can infer from this bar graph that 2 and 3 BHK flats are more popular and more commonly purchased. This indicates that most people can afford 2 and 3 BHK flats and have families consisting of 6-8 members

DATA SCIENCE LAB EXPERIMENT 2

2. Create Contingency Table
   A contingency table (also called a cross-tabulation table) is a type of table that displays the frequency distribution of two or more categorical variables. It helps in understanding relationships between these variables by organizing data in a matrix format.

A contingency table is a structured grid where:

- Rows represent one categorical variable.
- Columns represent another categorical variable.
- Cells contain the frequency/counts of occurrences for the combination of row and column variables.

```
Loading...
# Create a contingency table for 'area_type' vs 'total_sqft'
contingency_table = pd.crosstab(index=data['area_type'], columns='Total_Count')

# Display the contingency table
print(contingency_table)
```

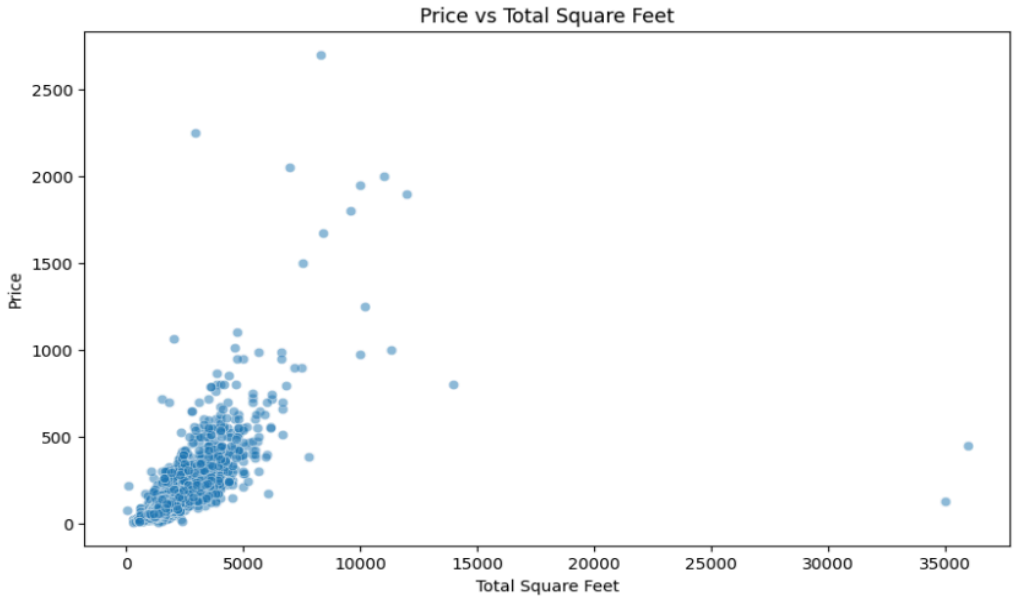
```
col_0                Total_Count
area_type
Built-up  Area              1211
Carpet  Area                  53
Plot  Area                   305
Super built-up  Area        6233
```

3. Scatter Plot

A scatter plot is a graphical representation of the relationship between two numerical variables. Each point on the plot represents a data observation, where:

- The X-axis represents one variable.
- The Y-axis represents another variable.
- The position of each point indicates the values of both variables for a single observation.

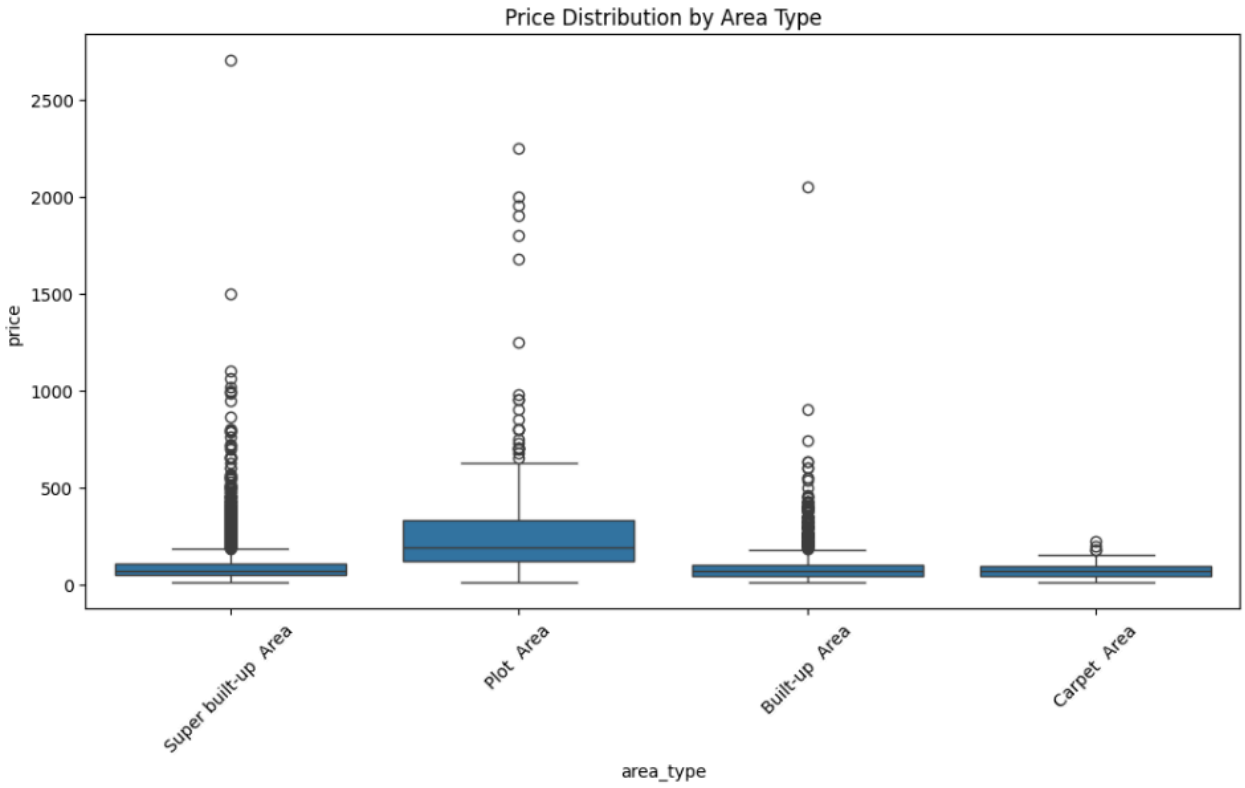
```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.scatterplot(x=data['total_sqft'], y=data['price'], alpha=0.5)
plt.xlabel('Total Square Feet')
plt.ylabel('Price')
plt.title('Price vs Total Square Feet')
plt.show()
```



We can infer that mostly price and area have a direct relation. However if a particular house is located in a popular locality then its price increases dramatically even if its area is less.

4. Box Plot
   A box plot (also known as a box-and-whisker plot) is a statistical visualization used to summarize the distribution, spread, and skewness of a numerical dataset. It is particularly useful for detecting outliers and understanding the data's central tendency.

```
plt.figure(figsize=(12, 6))
sns.boxplot(x='area_type', y='price', data=data)
plt.xticks(rotation=45)  # Rotate labels for better visibility
plt.title('Price Distribution by Area Type')
plt.show()
```



There is a significant variation in prices across different area types. Some area types have a wider spread of prices, indicating price inconsistency. Outliers suggest that a few properties have much higher or lower prices than the majority.

DATA SCIENCE LAB EXPERIMENT 2

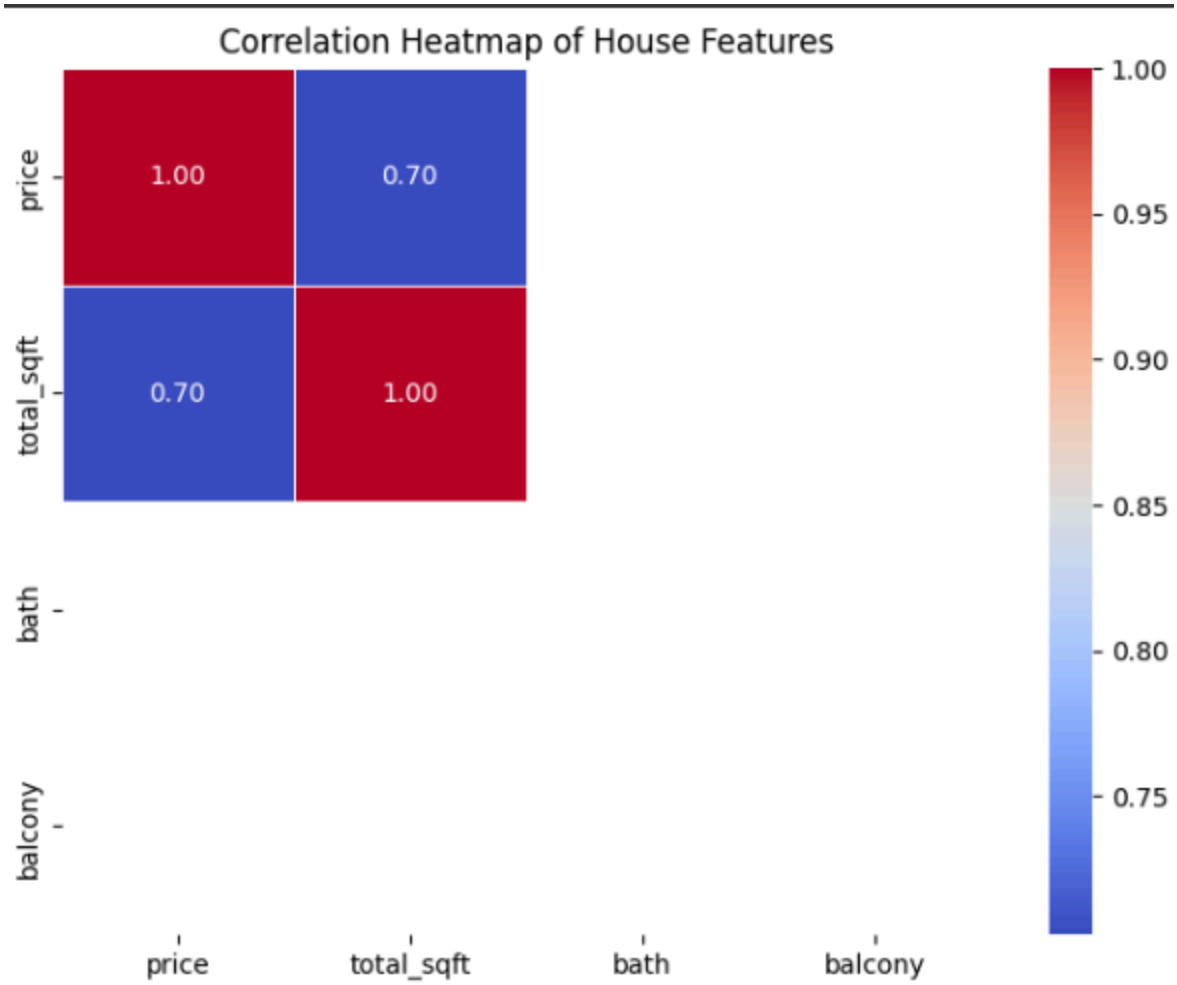
5. Heat Map
   A heatmap is a graphical representation of data where values are depicted using variations in color intensity. It is commonly used to visualize correlations, distributions, and patterns in a dataset, especially for large matrices.

Color Mapping

- The color intensity represents numerical values.
- Darker or more intense colors typically indicate higher values, while lighter colors indicate lower values.
- Color scales can be diverging (e.g., coolwarm, RdBu) or sequential (e.g., viridis, plasma, blues).

```
numeric_cols = ['price', 'total_sqft', 'bath', 'balcony']
data_numeric = data.reindex(columns=numeric_cols, fill_value=0)


# Compute correlation matrix
correlation_matrix = data_numeric.corr()

# Plot heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)

# Add title
plt.title("Correlation Heatmap of House Features")

# Show plot
plt.show()
```

DATA SCIENCE LAB EXPERIMENT 2

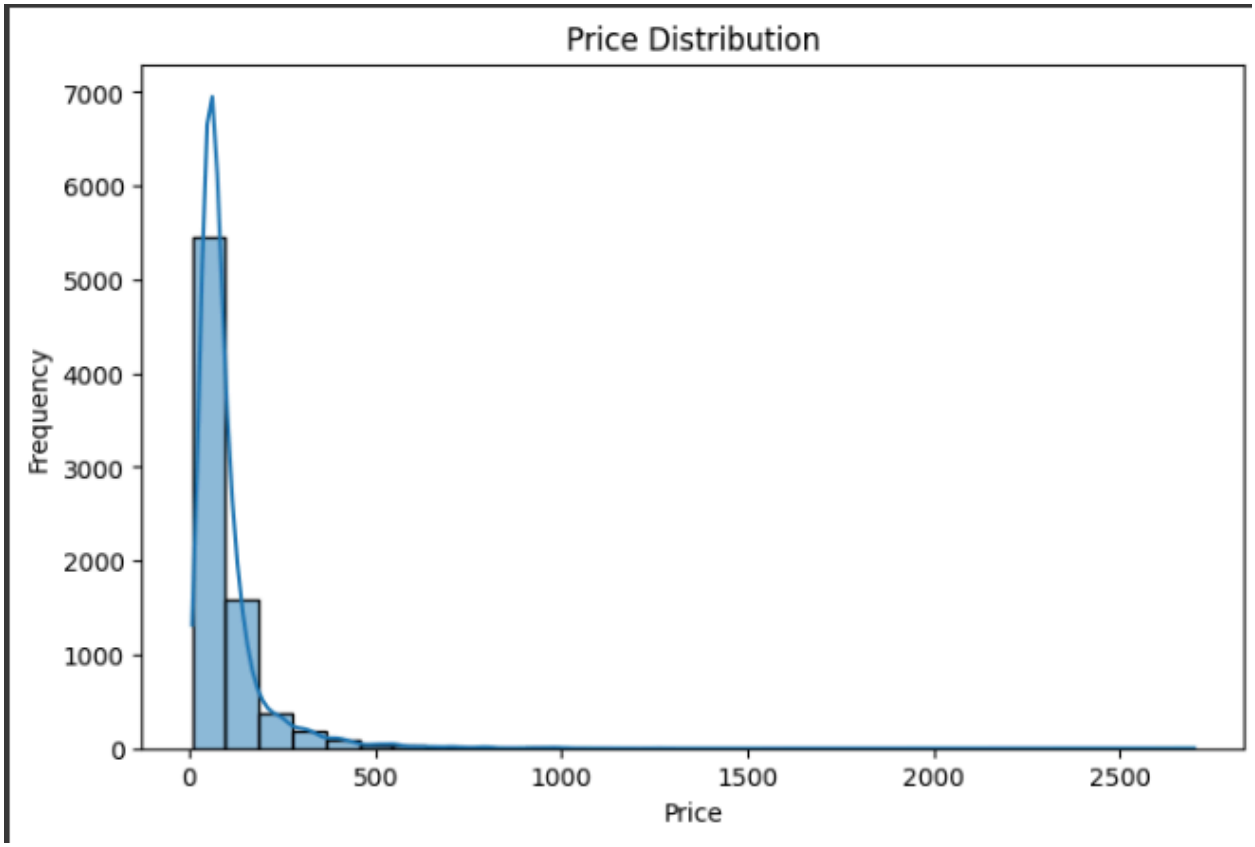


This heatmap shows that price and area have a strong positive correlation. Number of balconies and bathrooms were removed during data cleaning due to large number of null values.

6. Histogram

   A histogram is a type of bar graph that represents the distribution of numerical data. It is used to visualize how frequently different values appear in a dataset.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Plot the histogram
plt.figure(figsize=(8, 5))
sns.histplot(data['price'], bins=30, kde=True)
plt.title('Price Distribution')
plt.xlabel('Price')
plt.ylabel('Frequency')
plt.show()
```



Distribution Shape:

- If the histogram is bell-shaped, the data follows a normal distribution.
- If it is skewed right, it means most prices are on the lower side, with a few high values.
- If it is skewed left, most prices are higher, with fewer lower values.

Price Concentration:

- The tallest bar indicates the most common price range.
- If the bars are evenly spread, prices are uniformly distributed.
- If one side has more bars, the data may have skewness.

Outliers & Anomalies:

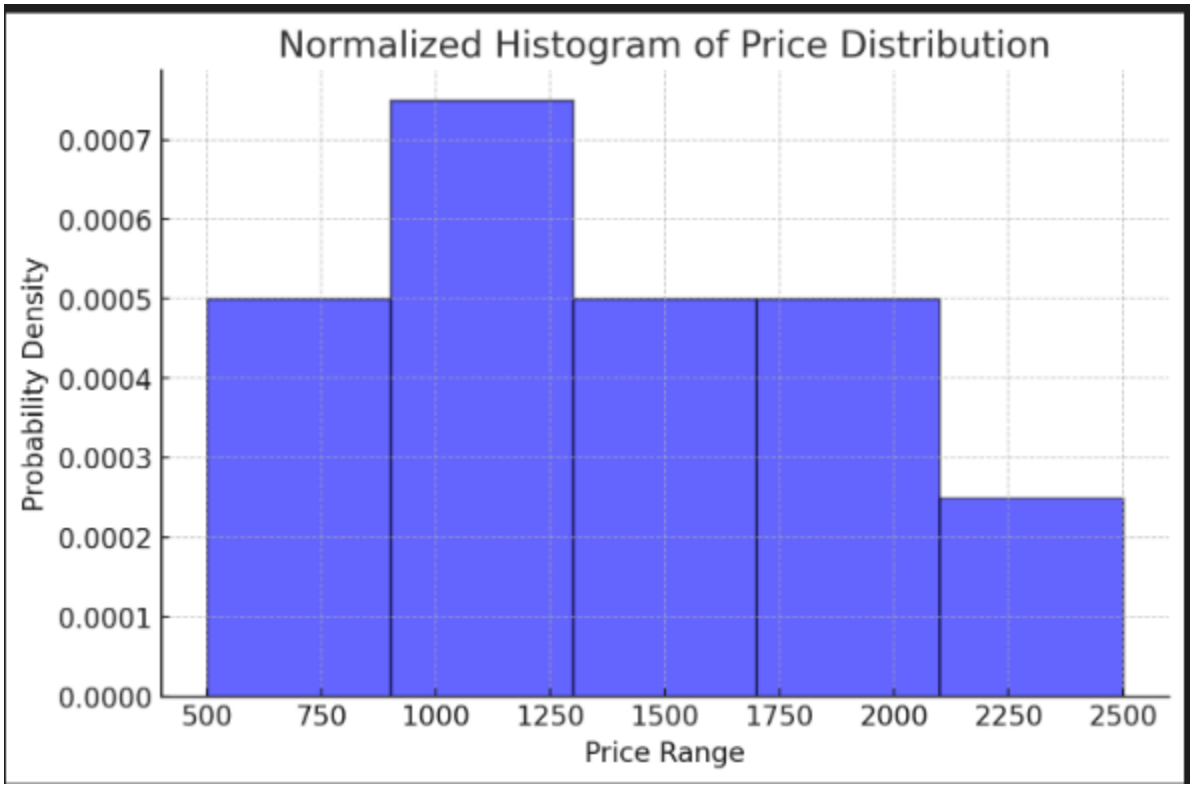
- If there are isolated bars, they might indicate outliers (unusually high or low prices).
- A gap between bars suggests missing values or uncommon price ranges.

7. Normalized Histogram

A normalized histogram is a variation of a standard histogram where the total area under the bars sums to 1. Instead of showing raw counts (frequencies), it represents the relative frequency distribution of data.

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
data['price_normalized'] = scaler.fit_transform(data[['price']])

plt.figure(figsize=(8, 5))
sns.histplot(data['price_normalized'], bins=30, kde=True)
plt.title('Normalized Price Distribution')
plt.show()
```

DATA SCIENCE LAB EXPERIMENT 2



The **most common price ranges** (higher bar heights) indicate where most data points fall. If the histogram is **skewed**, it suggests a bias toward lower or higher prices. The **spread of prices** helps in understanding market variability.

Conclusion: In this study on Data Visualization and Exploratory Data Analysis (EDA) using Matplotlib and Seaborn, we successfully explored various techniques to understand data distribution, trends, and relationships.

Through histograms, box plots, scatter plots, heatmaps, and bar charts, we gained the following insights:

- Histograms helped visualize the distribution of numerical variables and detect skewness or uniformity.
- Box plots provided insights into the spread, quartiles, and outliers in the dataset.

DATA SCIENCE LAB EXPERIMENT 2

- Scatter plots revealed relationships and correlations between different features.
- Heatmaps highlighted correlations between numerical variables using color intensity.
- Bar charts and contingency tables helped compare categorical distributions effectively.