Real Estate Price Prediction Using Data Analysis

Abstract:

This project involves analyzing and predicting real estate prices based on a dataset containing information on various housing attributes. The primary goal is to clean and preprocess the data, engineer relevant features, detect and handle outliers, and finally, perform exploratory data analysis (EDA) to understand the key factors influencing house prices. The data is processed using Python, leveraging libraries such as Pandas, NumPy, and Matplotlib. The output includes a processed dataset that can be used for further modeling and predictive analysis.

Architecture:

The architecture of this project is divided into several stages:

1. **Data Loading and Initial Exploration**:
   - I have loaded the dataset and performed an initial examination of the data structure.
   - Later I understood the distribution of various features, especially categorical ones.
2. **Data Cleaning**:
   - Performed data cleaning including handling missing values by removing rows with null values.
   - Dropped irrelevant columns that do not contribute to the analysis.
3. **Feature Engineering**:
   - Created new features such as 'BHK' (number of bedrooms) from existing ones.
   - Converted non-numeric data (e.g., ranges of square footage) into numeric values for easier analysis.
4. **Outlier Detection and Removal**:
   - Identified and removed outliers based on domain-specific knowledge (e.g., BHK-to-square-feet ratio, price per square foot).
   - Applied statistical methods to further clean the data by removing anomalies.
5. **Data Visualization**:
   - Generate scatter plots and histograms to visualize the relationship between different features and the target variable (price per square foot).
   - Use these visualizations to understand the impact of various factors on house prices.

Tools Used

- Python: The primary programming language used for data processing and analysis.
- Pandas: For data manipulation, cleaning, and preprocessing.
- NumPy: For numerical operations and handling large datasets.

- Matplotlib: For creating visualizations like scatter plots and histograms to understand data distributions and relationships.

Detailed Explanation

1. **Data Loading and Initial Exploration**

- The dataset is loaded using Pandas, and an initial exploration is done to understand the shape and structure of the data.

```
data = pd.read_csv("house_prices.csv") print(data.shape) data.head()
```

2. Data Cleaning

- Unnecessary columns such as 'area_type', 'society', 'balcony', and 'availability' are dropped.
- Missing values are handled by dropping rows with NA values.

```
nd = data.drop(['area_type', 'society', 'balcony', 'availability'], axis='columns')

nd1 = nd.dropna()
```

3. **Feature Engineering**

- New columns are created based on existing ones. For example, the 'BHK' column is derived from the 'size' column.
- Non-numeric data, such as square footage given in ranges, is converted into numeric values for analysis.

```
nd1['bhk'] = nd1['size'].apply(lambda x: int(x.split(' ')[0]))

nd2['total_sqft'] = nd2['total_sqft'].apply(convert_sqft_to_num)
```

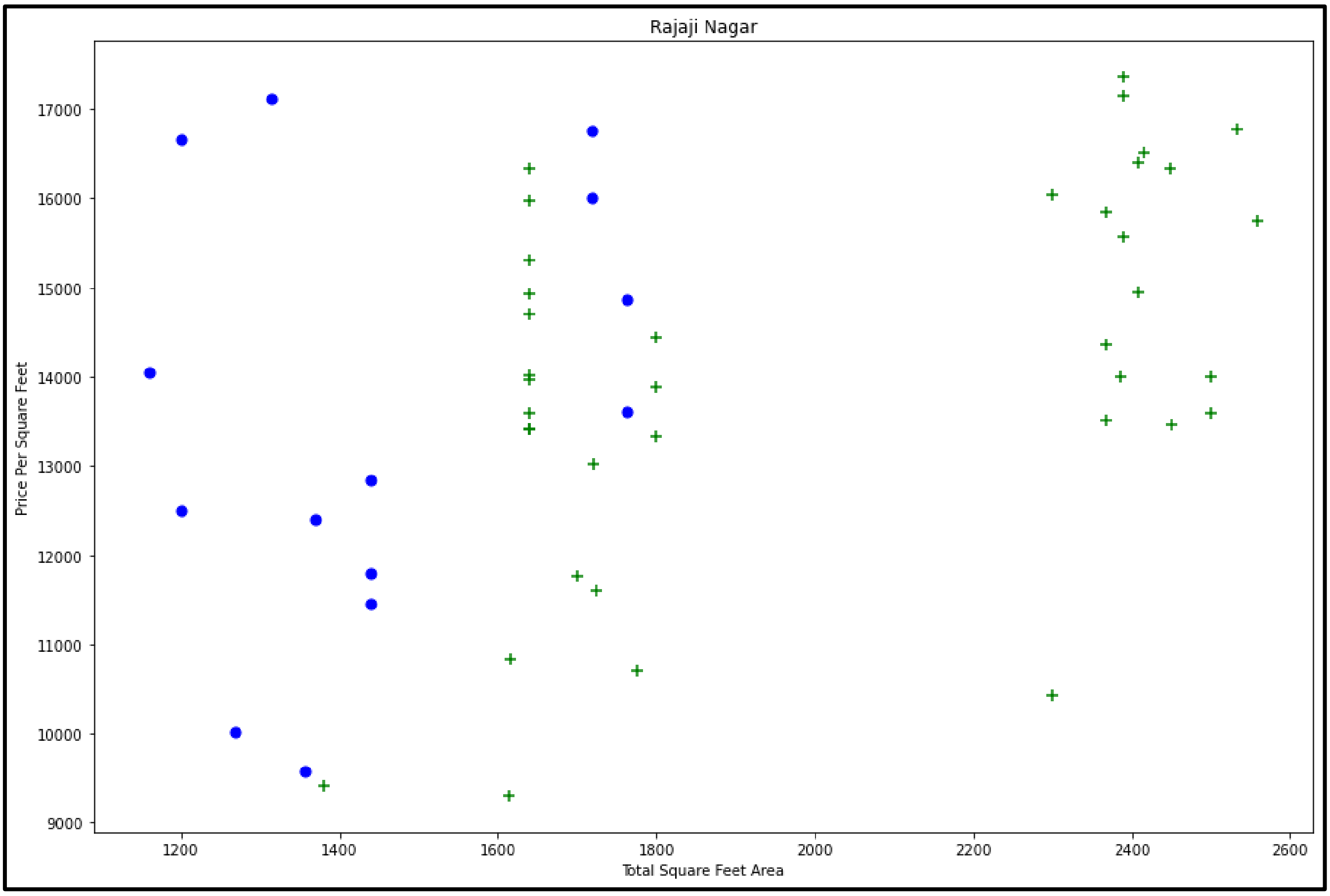
4. **Outlier Detection and Removal**

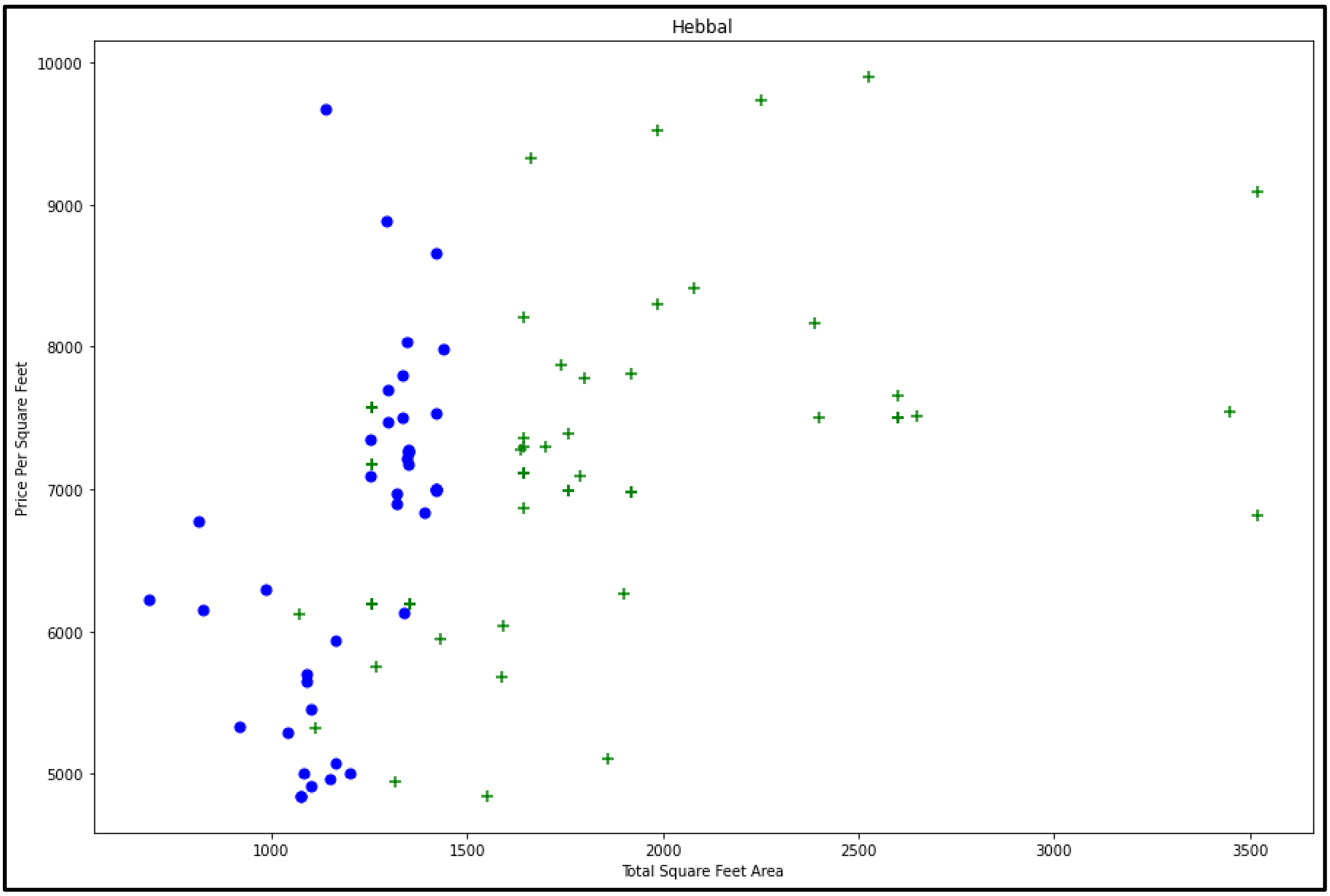
- Outliers are identified and removed based on logical constraints (e.g., minimum square footage per BHK) and statistical methods (e.g., standard deviation).

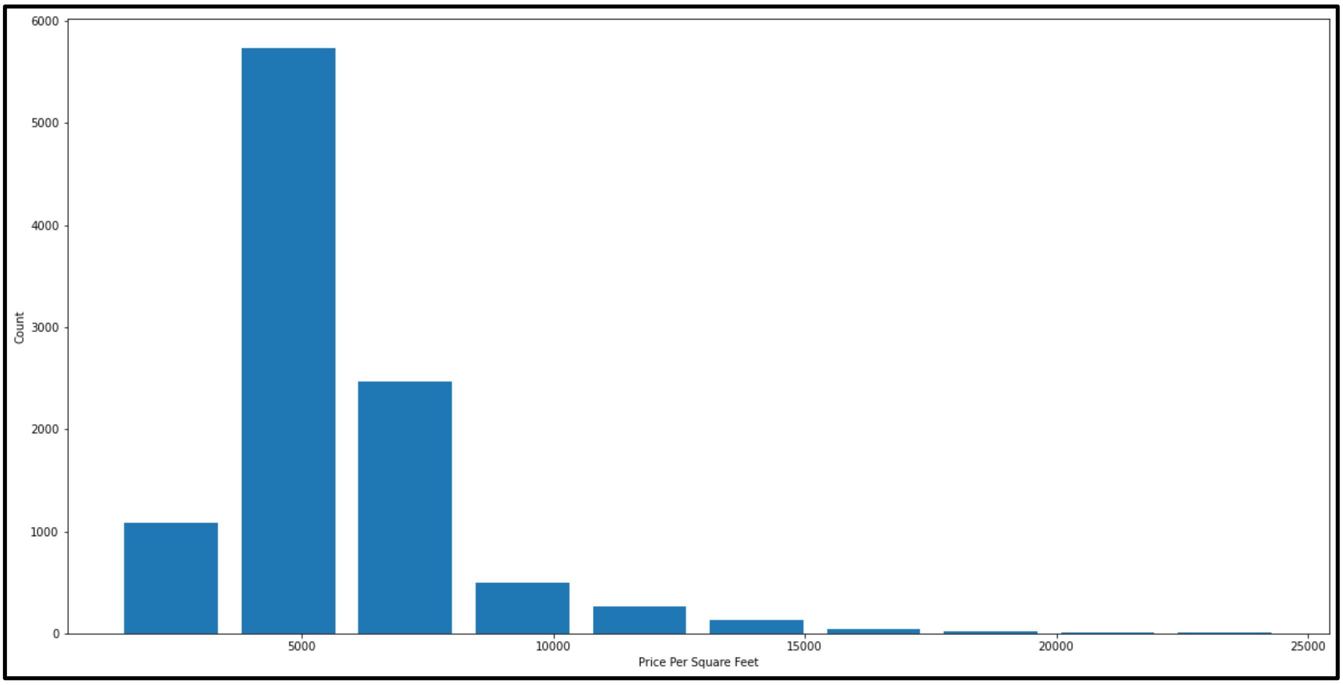
```
df1 = df[~(df.total_sqft/df.bhk < 300)]

df3 = remove_outliers(df1)
```

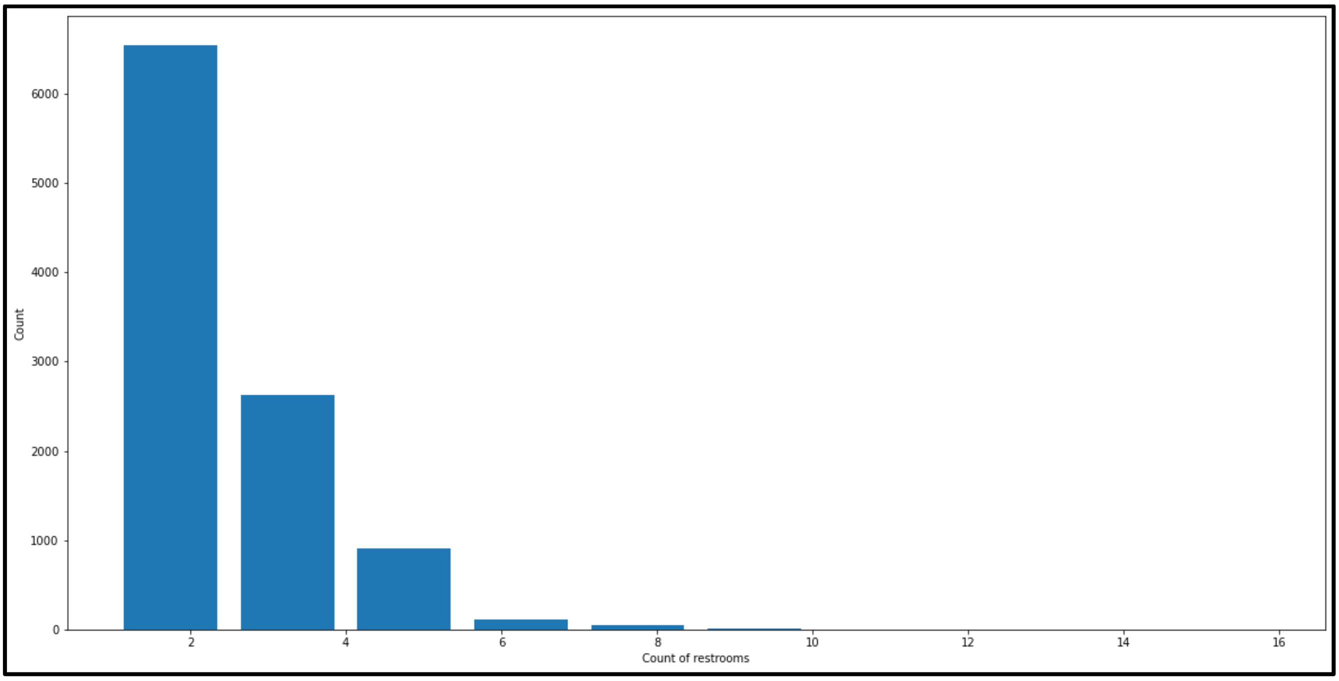
5. **Data Visualization**

- Visualizations are created to understand the distribution of the 'price per square foot' and other key metrics.


Rajaji Nagar
Price Per Square Feet
17000
16000
15000
14000
13000
12000
11000
10000
9000
1200
1400
1600
1800
2000
2200
2400
2600
Total Square Feet Area


Hebbal
Price Per Square Feet
10000
9000
8000
7000
6000
5000
1000
1500
2000
2500
3000
3500
Total Square Feet Area





Conclusion This project provides a comprehensive approach to preprocessing and analyzing real estate data. The final cleaned and processed dataset is ready for further modeling, such as predictive analysis using machine learning algorithms. The techniques used in this project are essential for ensuring that the data is of high quality, which is critical for building accurate predictive models.