# **House Price Prediction using Machine Learning**

Sanyukta Saha,

Course/batch and Institute name

Period of Internship: 25th August 2025 - 19th September 2025 (Do not change the dates)

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

### **1. Abstract**

This project focuses on predicting house prices in Bengaluru using a dataset of over 13,000 property listings. The dataset includes features like location, size, total square feet, and number of bathrooms. The primary steps in this project included data cleaning, exploratory data analysis, feature engineering, and model building. Several data preprocessing techniques were applied to handle missing values, outliers, and categorical variables. A linear regression model was trained and evaluated to predict house prices based on the cleaned data. The project demonstrates a practical application of machine learning in the real estate sector.

### **2. Introduction**

The real estate market in Bengaluru is a dynamic and growing sector. This project aims to build a model that can predict house prices in Bengaluru based on various property attributes. The project utilizes a dataset of house listings from Bengaluru and employs machine learning techniques to achieve this goal. The relevance of this project lies in its potential to provide valuable insights for buyers, sellers, and real estate agents. The technologies used in this project include Python with libraries like Pandas for data manipulation, and Matplotlib/Seaborn for data visualization, and Scikit-learn for building the machine learning model.

During the first two weeks of the internship, I received training on the following topics:

* Python for data science
* Data manipulation with Pandas
* Data visualization with Matplotlib and Seaborn
* Machine learning fundamentals
* Regression models

### **3. Project Objective**

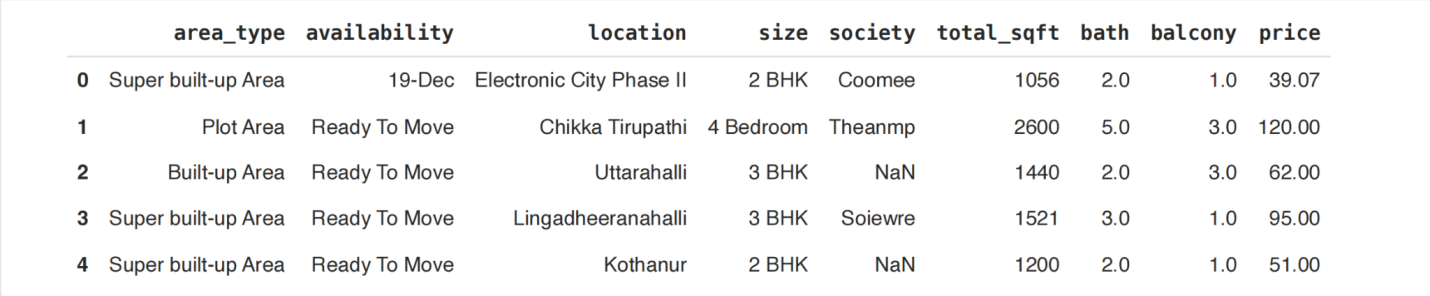
The main objectives of this project are:

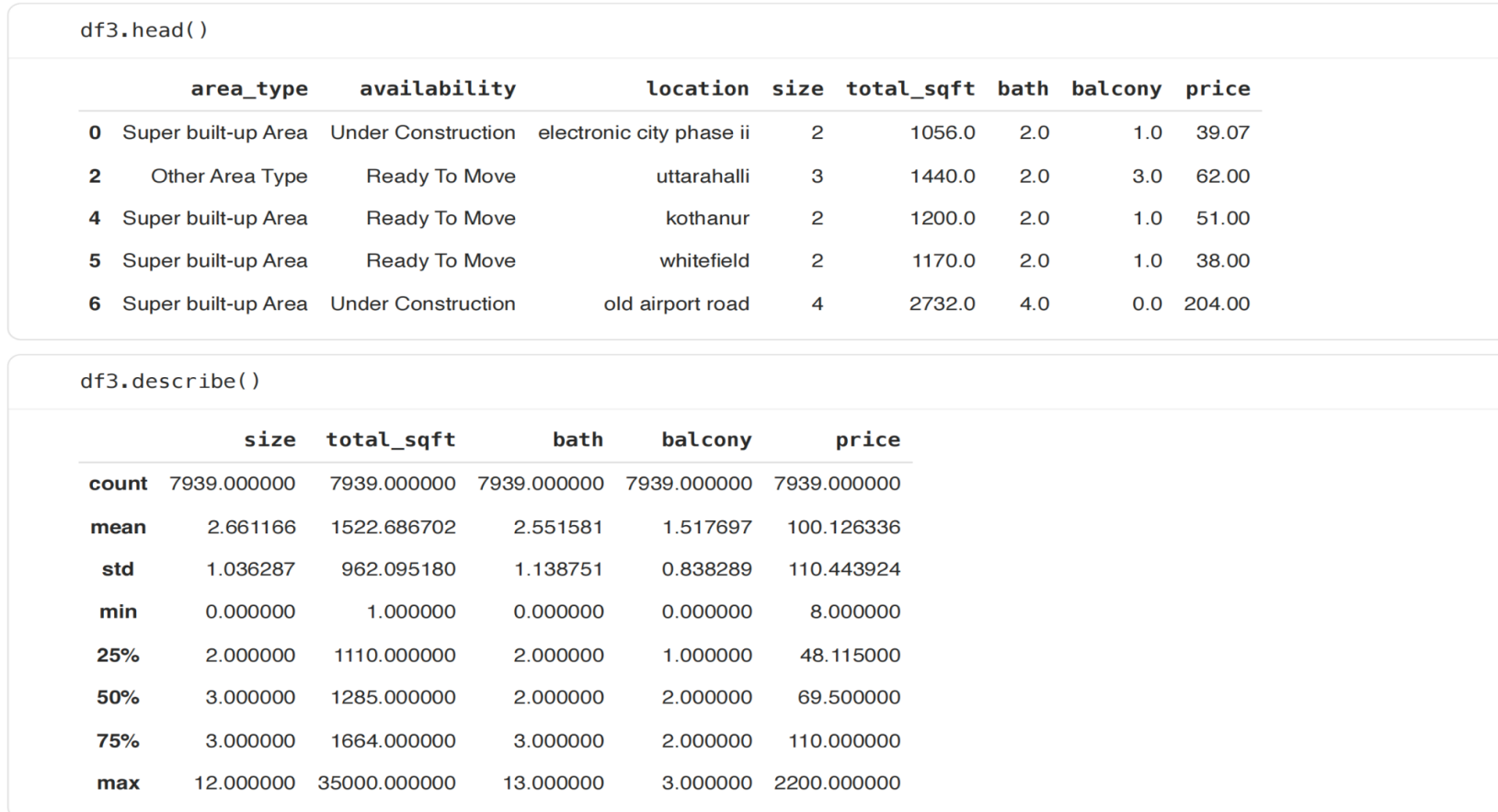
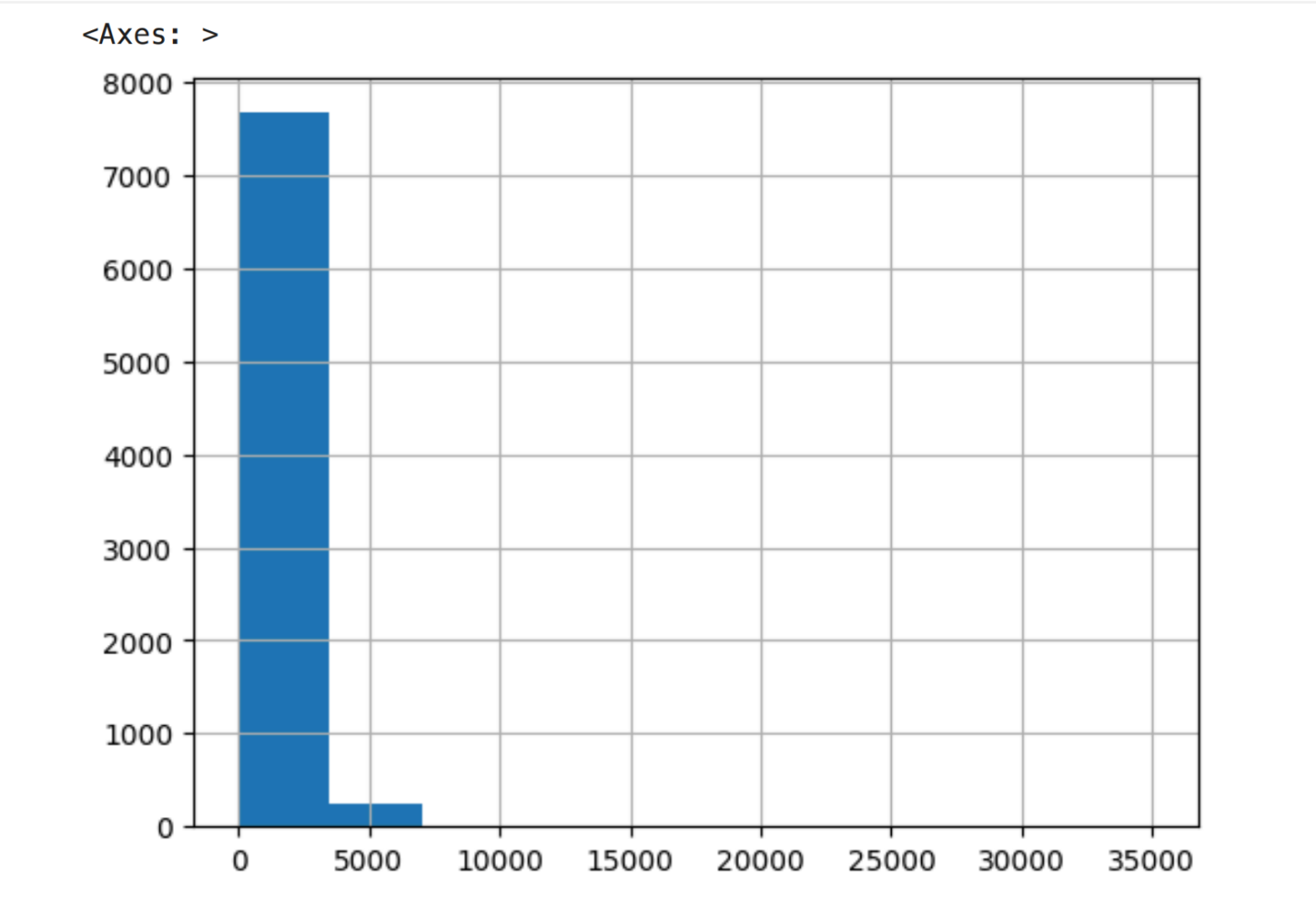
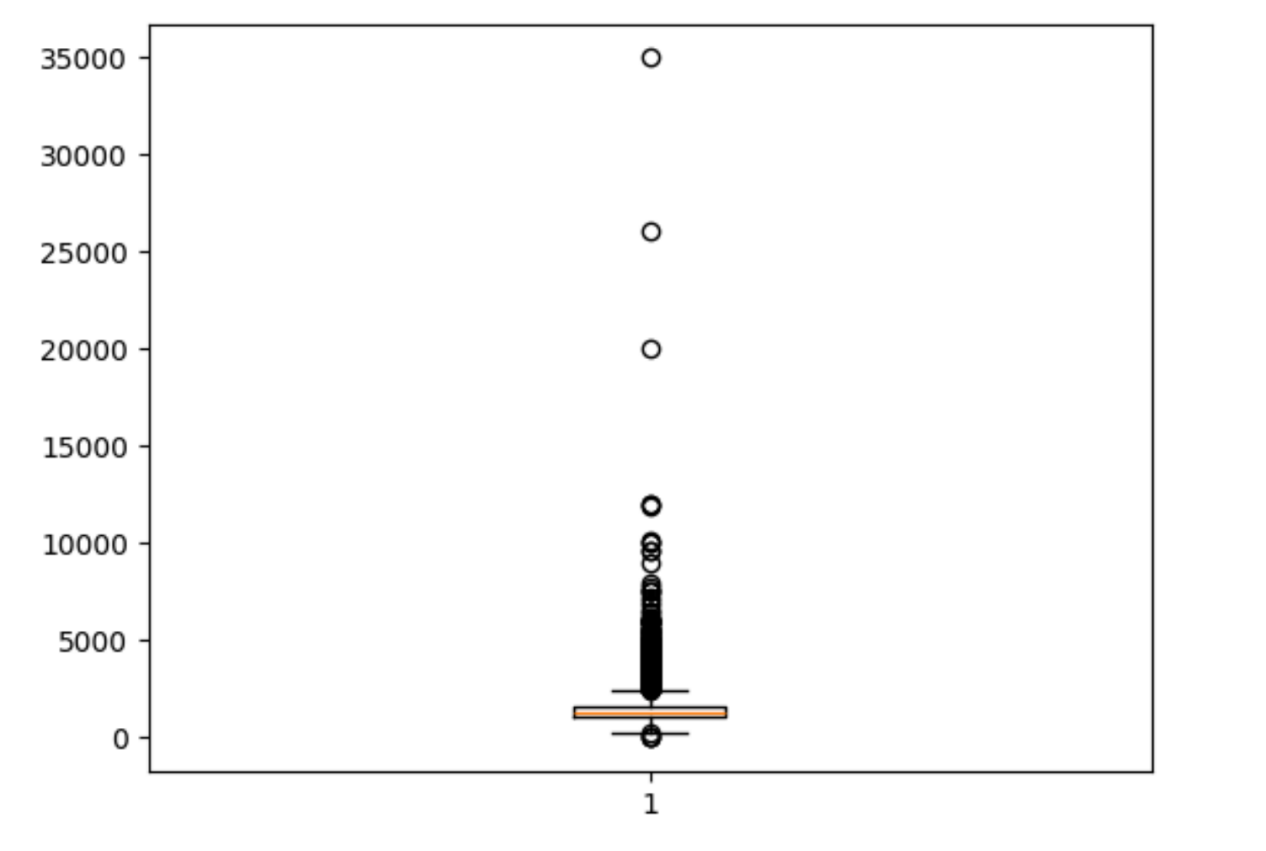
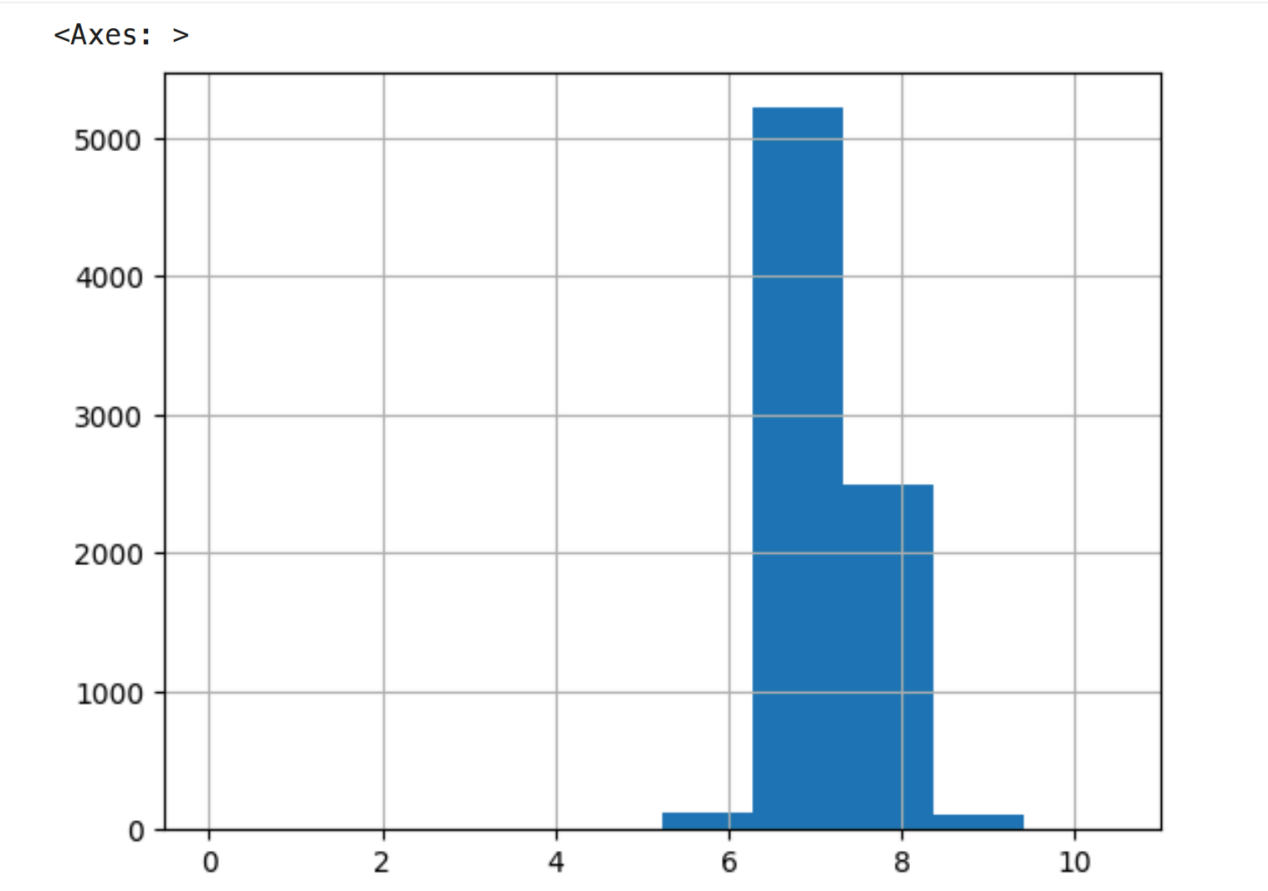
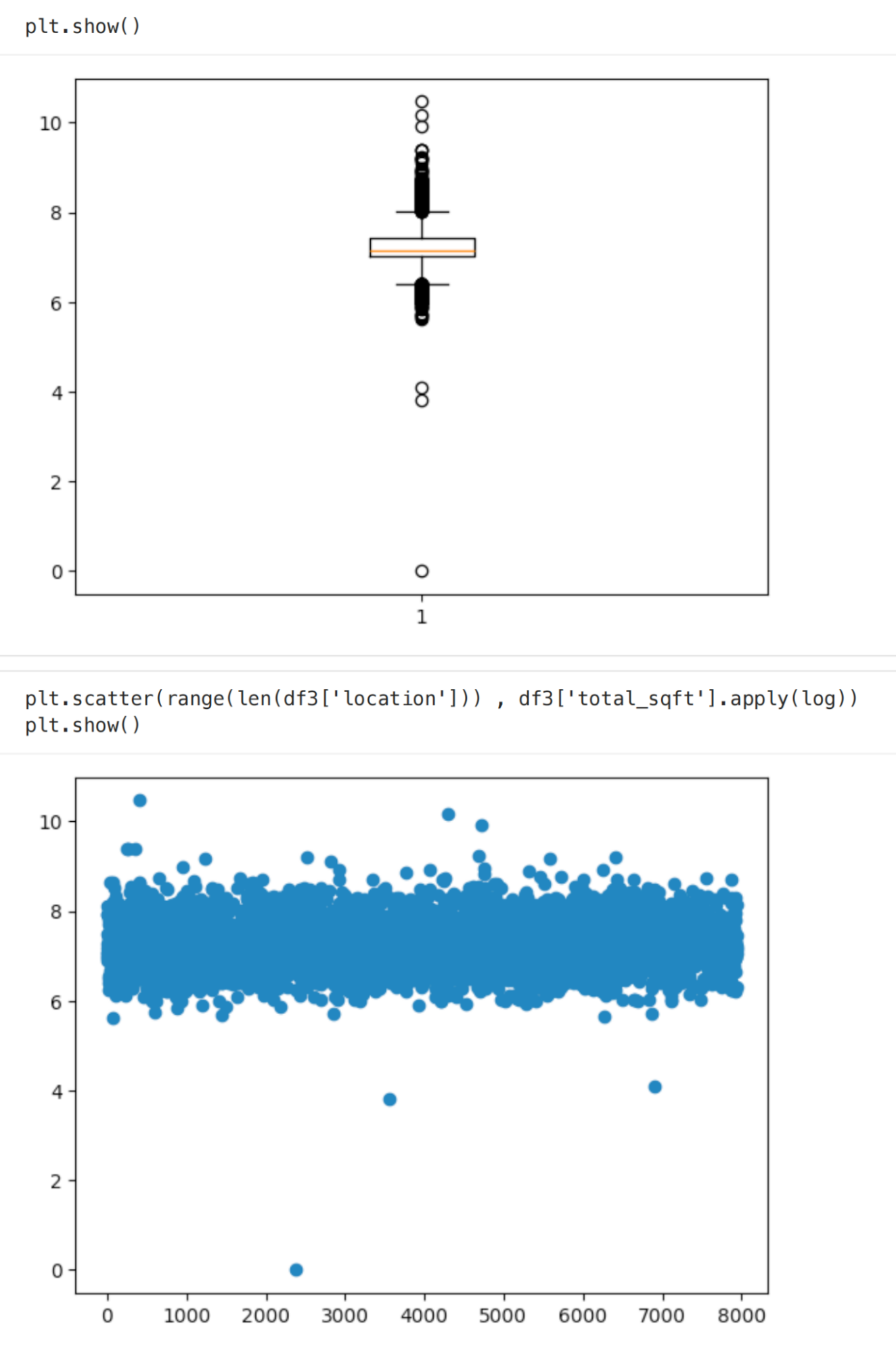
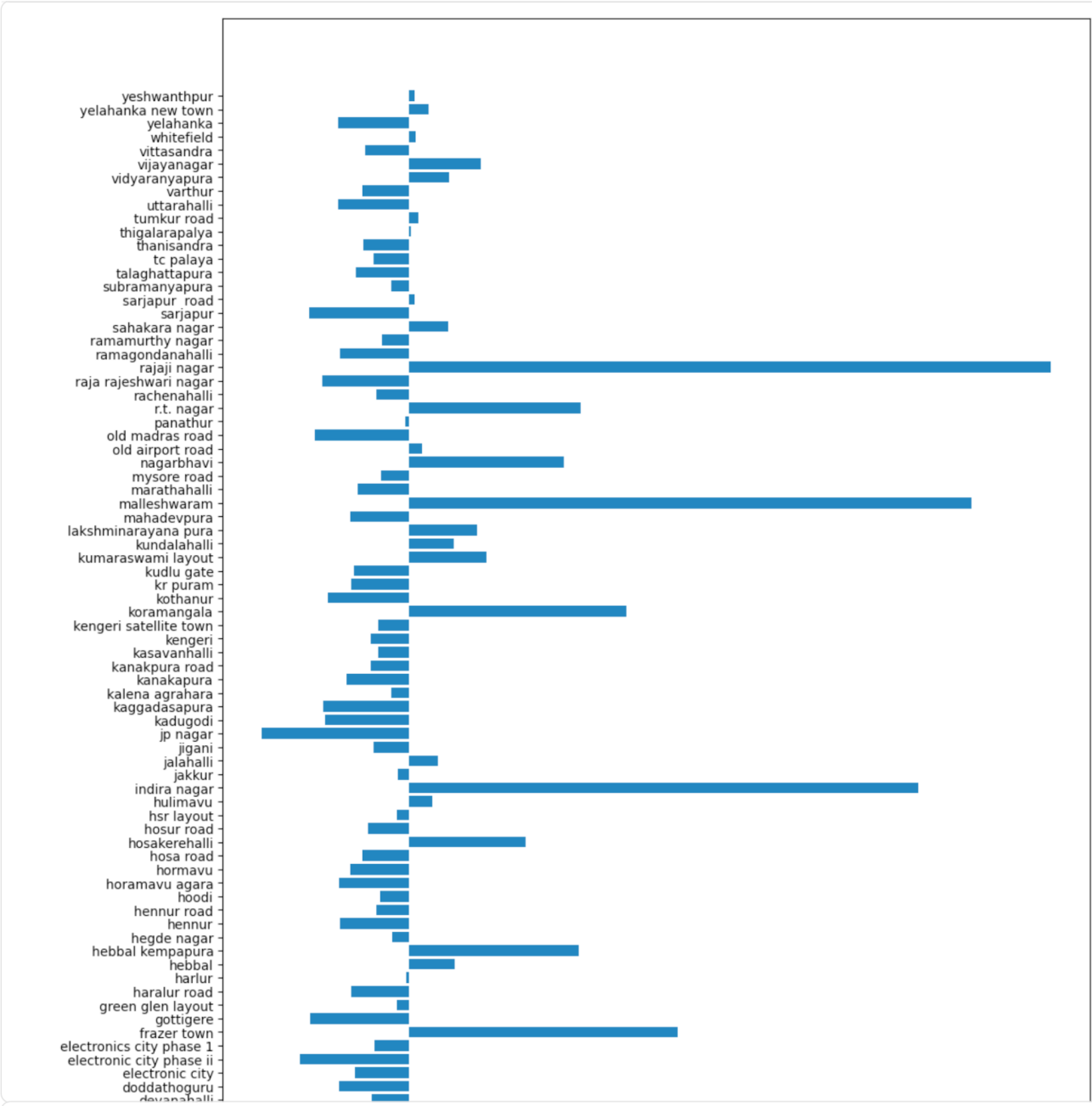
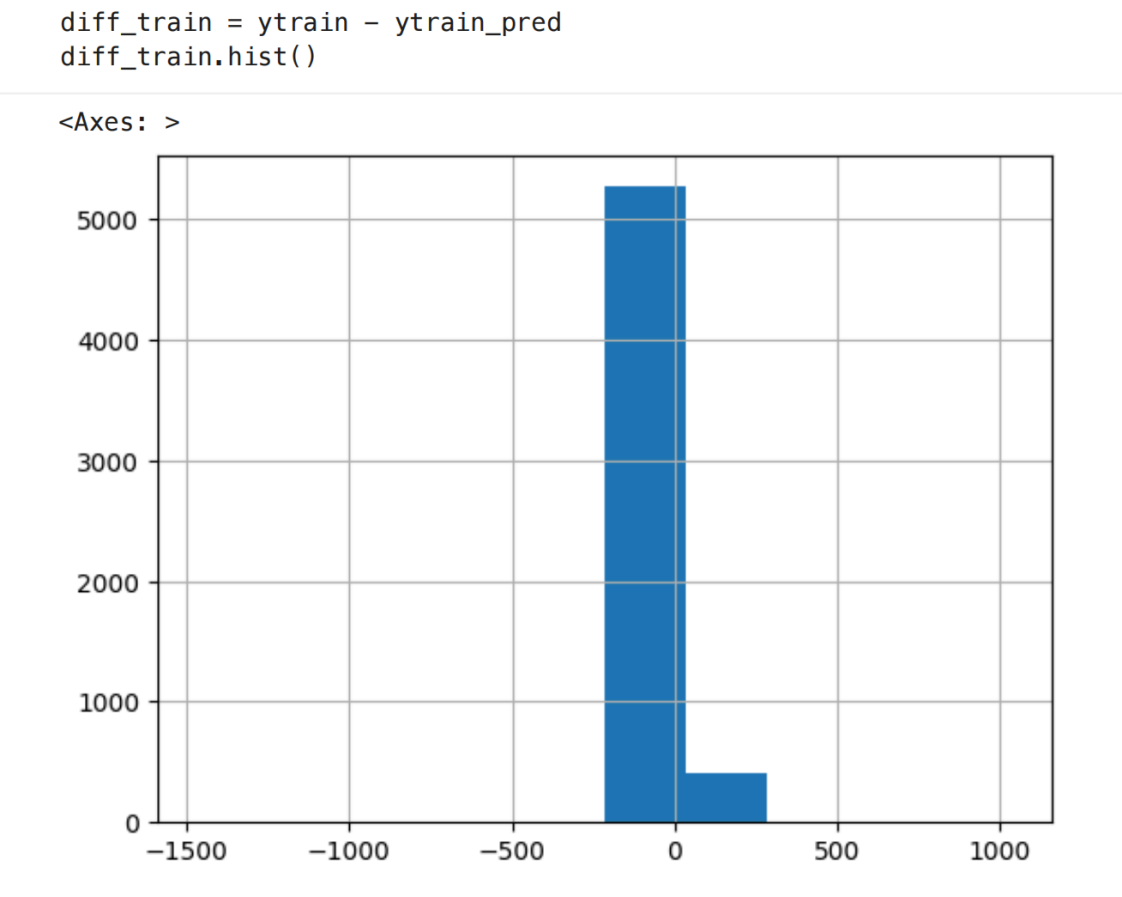
* To perform data cleaning and preprocessing on the Bengaluru house price dataset to prepare it for analysis.
* To conduct exploratory data analysis (EDA) to understand the relationships between different features and the house price.
* To build and train a linear regression model to predict house prices.
* To evaluate the performance of the model using metrics like Mean Squared Error (MSE) and Mean Absolute Error (MAE).

### **4. Methodology**

The project followed a systematic approach:

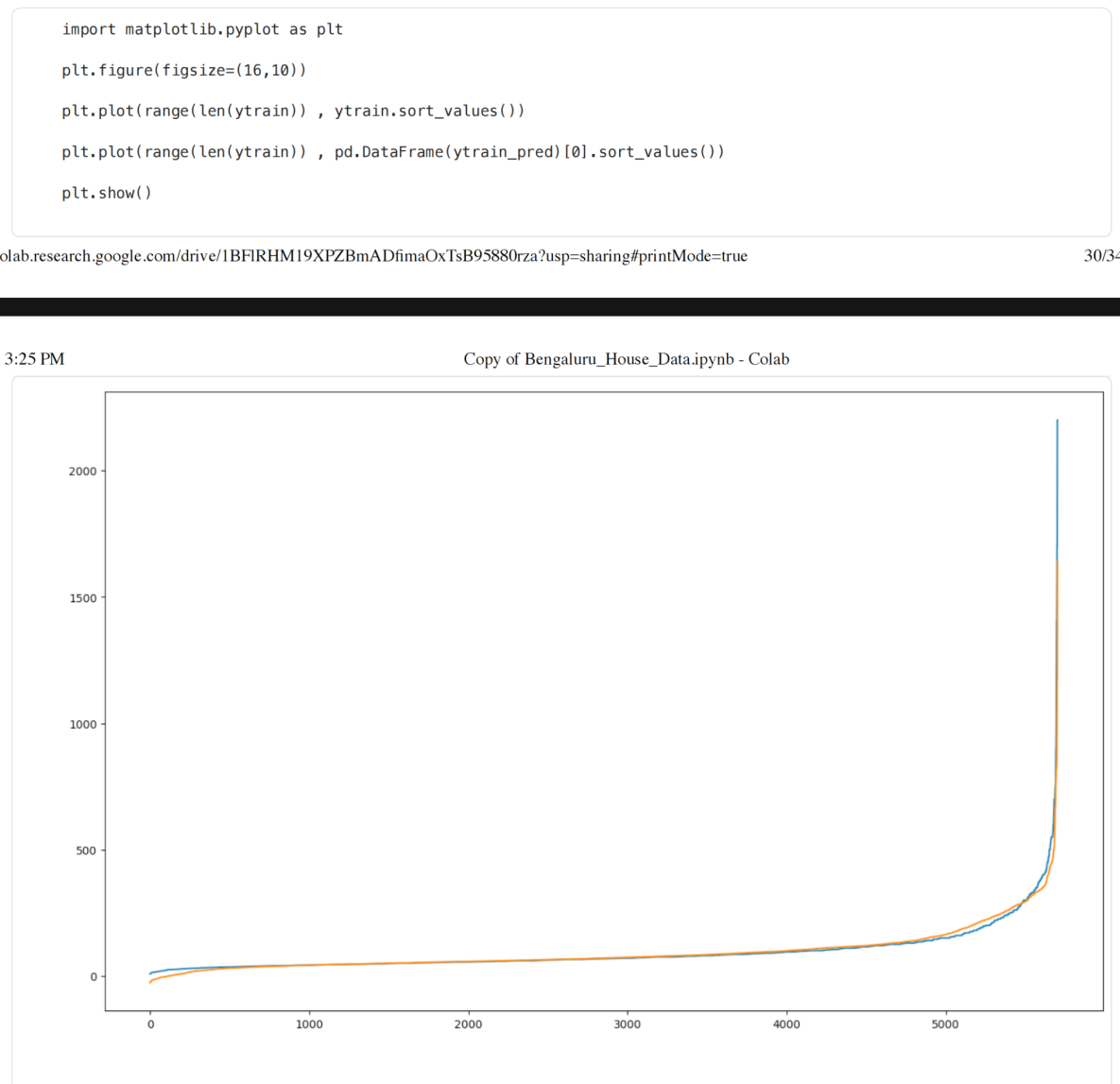
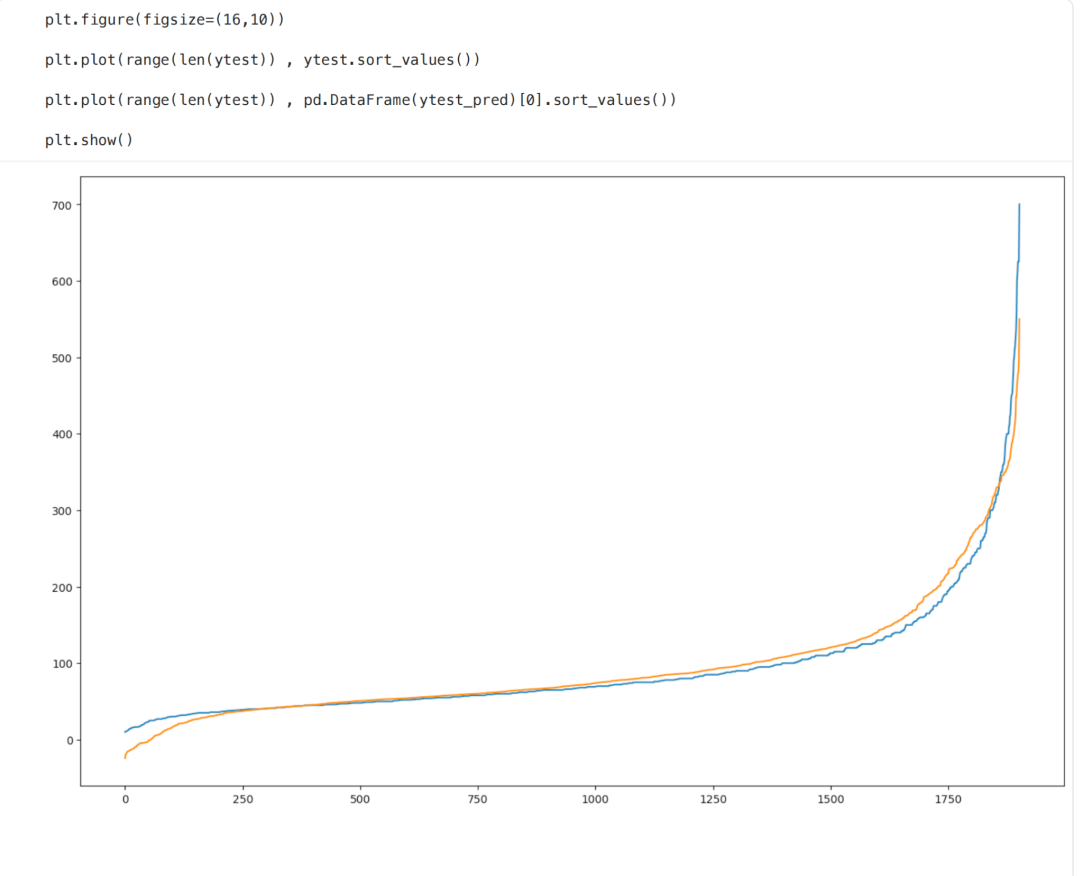
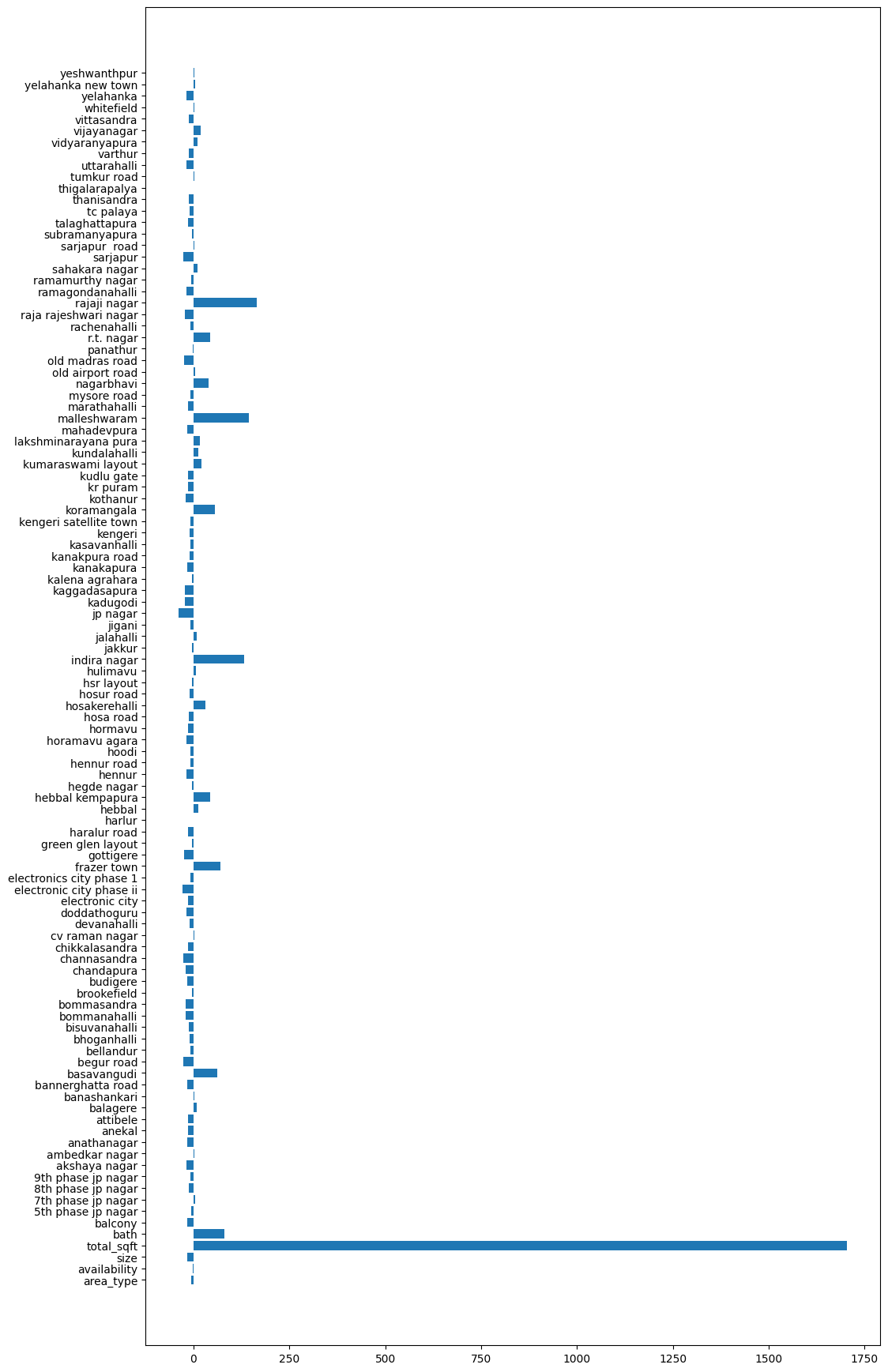
1. **Data Collection and Exploration**: The project started with the Bengaluru\_House\_Data.csv dataset, which has 13,320 rows and 9 columns. Initial exploration was done to understand the data types and check for missing values.



1. **Data Cleaning and Pre-processing**:
   * The 'society' column was dropped as it had a high number of missing values (5502 missing).
   * The 'size' column was converted to a numerical 'bhk' (bedrooms) column.
   * The  
      total\_sqft column was cleaned to handle ranges by converting them to their average numerical value.
   * Categorical features like 'area\_type' and 'availability' were simplified into fewer categories.
   * The 'location' feature was cleaned by converting to lowercase, stripping whitespace, and grouping rare locations into an 'Others' category.
   * Missing values in 'bath' and 'balcony' were imputed.
2. **Feature Engineering and Outlier Removal**:
   * A new feature, 'price\_per\_sqft' (PPS), was created to help with outlier detection.
   * Outliers were removed by filtering out data points where the PPS was beyond two standard deviations from the mean for each location.
   * 
   * 
   * 
   * 
   * 
   * 
3. **Model Building and Validation**:
   * The dataset was split into training and testing sets.
   * A  
      **Linear Regression model** was chosen for this project.
   * The data was scaled using  
      **StandardScaler** to normalize the feature ranges.
   * The model was trained on the scaled training data and its performance was evaluated on the test data using MSE and MAE.
   * 
   * 

### **5. Data Analysis and Results**

#### **Descriptive Analysis**

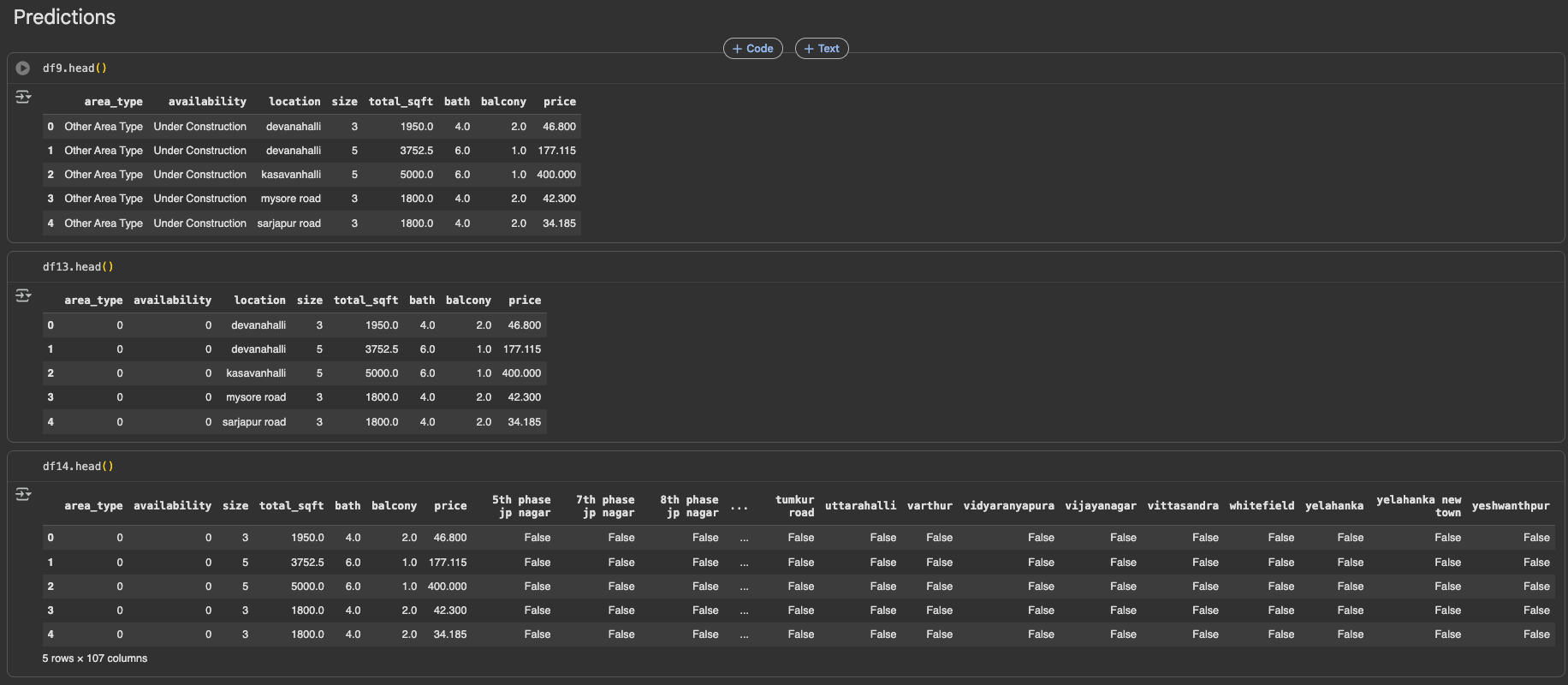
* **Data Distribution**: The distribution of total\_sqft was found to be skewed, and a log transformation was applied to normalize it for better visualization and modeling.
* **Visualizations**:
  + A histogram of the log-transformed  
     total\_sqft shows a more normal distribution.
  + A boxplot of the log-transformed  
     total\_sqft also shows a more compact distribution with fewer extreme outliers.
  + A scatter plot of the log-transformed  
     total\_sqft against the index shows the general range of values.
  + 
  + 
  + 

#### **Model Performance**

The Linear Regression model was trained and evaluated, with the following results:

|  |  |  |
| --- | --- | --- |
| Metric | Training Set | Test Set |
| **Mean Squared Error (MSE)** | 3301.67 | 3153.93 |
| **Mean Absolute Error (MAE)** | 26.56 | 28.12 |

The feature importance was visualized using a bar chart, showing the coefficients of the linear regression model for each feature. This helps in understanding which features have the most significant impact on the predicted price.



### **6. Conclusion**

This project successfully demonstrated the process of building a house price prediction model. The key takeaways are:

* **Data preprocessing is crucial**: A significant portion of the project was dedicated to cleaning and preparing the data, which is essential for building an accurate model.
* **Feature engineering can improve model performance**: Creating the 'price\_per\_sqft' feature was instrumental in identifying and removing outliers.
* **Model evaluation is key**: The model's performance was evaluated using standard metrics to understand its predictive power.

**Future Work**: To further improve the model, one could:

* Experiment with other regression models like Random Forest or Gradient Boosting.
* Perform more advanced feature engineering.
* Collect more data to improve the model's accuracy and generalizability.

### **7. APPENDICES**

#### **1. References**

* Bengaluru House Data, Kaggle
* Pandas Documentation
* Scikit-learn Documentation

#### **2. Github link for the codes developed**

[Link to Github Repository]