**1. Data Cleaning Steps**

Before training our model, we cleaned the dataset to ensure high-quality inputs. The following preprocessing steps were performed:

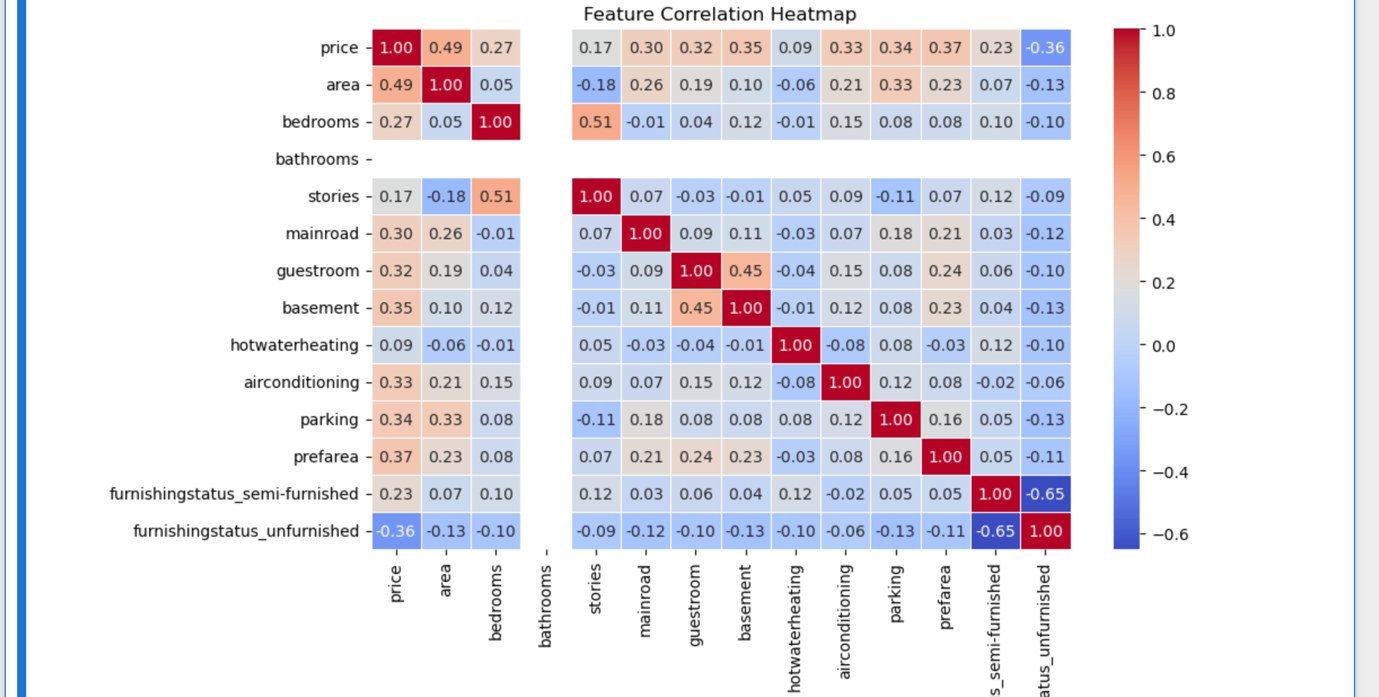
* **Handling Missing Values:**
  + No missing values in numerical and categorical columns.
* **Outlier Detection & Handling:**
  + Used **Interquartile Range (IQR)** to identify and remove extreme outliers.
* **Encoding Categorical Variables:**
  + Applied **one-hot encoding** for categorical columns such as "furnishing status".
  + Applied **label-encoding** for binary values.
* **Feature Scaling:**
  + Used **Standard Scaler** to normalize numerical features like area and price.

**Result:** The cleaned dataset was ready for exploratory data analysis (EDA).

**2. Key Insights from EDA:**

**Feature Correlation Analysis (Heatmap)**

* The **heatmap** reveals correlations between different features and house prices.
* **Strongest correlation:**
  + area (0.49 correlation with price) → Larger areas tend to have higher prices.
  + bathrooms (0.51 correlation) → More bathrooms are associated with higher property prices.
  + airconditioning, parking, basement, and guestroom also show moderate correlation with price.
* **Weak correlations:**
  + hotwaterheating and mainroad have minimal influence on house prices.
  + Some furnishing statuses (semi-furnished and unfurnished) negatively correlate with price.

****

**Scatter Plot: Area vs. Price**

* The scatter plot shows a **positive correlation** between area and price.
* While larger areas generally result in higher prices, the spread suggests **some variation**, possibly due to other features influencing price.

**Box Plot: Stories vs. Price**

* The box plot shows price distribution across different numbers of stories.
* **Higher-story houses tend to have higher median prices.**
* However, significant **overlapping of price ranges** suggests that stories alone are not a strong predictor of price.

A screenshot of a graph

Description automatically generated

**3. Model Performance Analysis**

**3.1 Train-Test Split**

We split the dataset into **80% training and 20% testing** sets.

**3.2 Model Used: Multiple Linear Regression**

We trained a linear regression model using **scikit-learn**.

**3.3 Evaluation Metrics**

The model’s performance is summarized below:

| **Metric** | **Training Set** | **Testing Set** |
| --- | --- | --- |
| R² Score | **0.571** | **0.497** |
| Mean Absolute Error (MAE) | **0.096** | **0.102** |
| Mean Squared Error (MSE) | **0.020** | **0.022** |

**Key Observations:**

* The training R² score (**0.571**) is higher than the testing R² score (**0.497**), indicating **slight underfitting**.
* **MAE and MSE are close** for training and testing, suggesting the model generalizes well but could improve.