**IBM Machine Learning**

**Project Report**

**(Houses Rent Prediction)**

Submitted by:

**Mr. Mubasshir Iqbal**

Department of Computer Sciences, HITEC University

Table of Contents

[**Dataset Selection** 3](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222150)

[**Reason of selection:** 3](#_Toc207222151)

[**Dataset Overview:** 3](#_Toc207222152)

[**Preview of dataset head:** 3](#_Toc207222153)

[3](#_Toc207222154)

[3](#_Toc207222155)

[**Preview of dataset tail:** 4](#_Toc207222156)

[**Machine Learning Pipeline** 5](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222157)

[**Preprocessing & EDA** 6](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222158)

[**Data Summary** 6](#_Toc207222159)

[**EDA:** 8](#_Toc207222160)

[**Data Cleaning & Feature Engineering** 16](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222161)

[1. **Missing Values** 16](#_Toc207222162)

[2**. Encoding Categorical Variables** 16](#_Toc207222163)

[**3. Feature Engineering** 16](#_Toc207222164)

[**Data Splitting** 18](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222165)

[**Class balance and imbalance** 18](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222166)

[**Model Building** 19](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222167)

[**Model Evaluation** 21](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222168)

[**Insight** 23](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222169)

[**Prediction** 24](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222170)

[**Summary** 25](file:///D:\New%20folder\Data%20Science\Project.docx#_Toc207222171)

**Reason of selection:**

## **Dataset Selection**

Rich features, realistic complexity, lots of preprocessing challenges (categorical encoding, feature engineering). Excellent for regression modeling and model comparison.

### **Dataset Overview:**

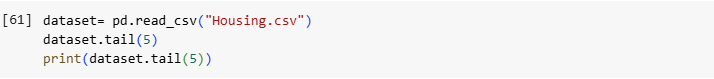
* **Shape**: 545 rows × 13 columns, decent size for analysis
* **Target variable**: Price (continuous refers to regression problem).
* **Features**:
* **Numerical:** area, bedrooms, bathrooms, stories, parking
* **Categorical:** main road, guestroom, basement, hot water heating, air conditioning
* **Categorical (multi-class):** furnishing status
* **Missing values**: None
* **Data types**: Mix of numerical + categorical, good for preprocessing, encoding, and feature engineering

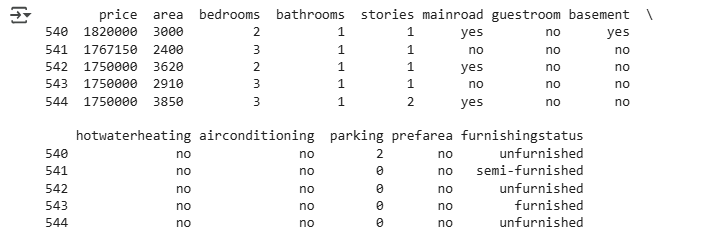
### **Preview of dataset head:**

### 

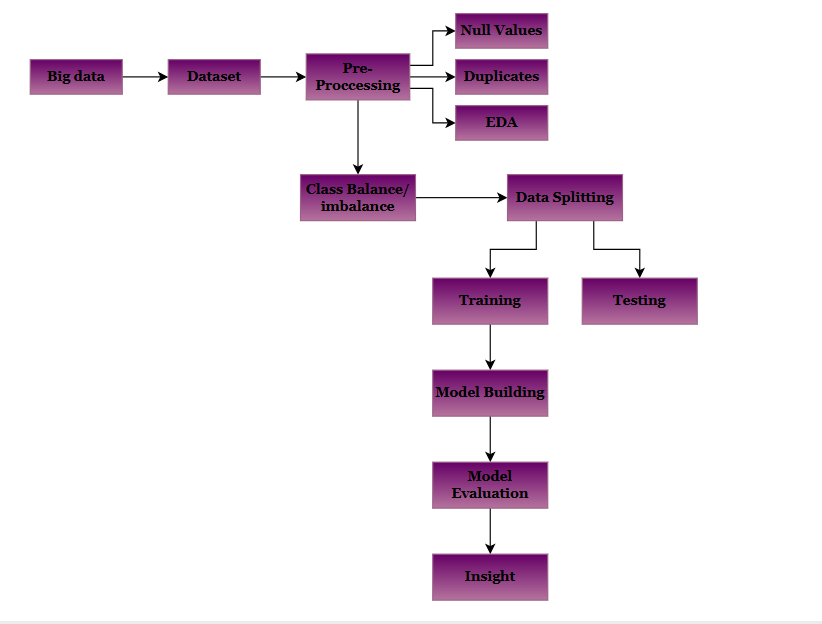
### 

### **Preview of dataset tail:**





## **Machine Learning Pipeline**

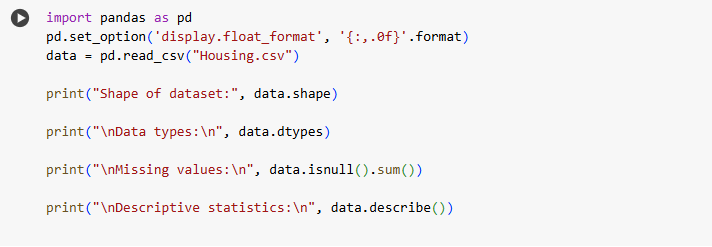
****

### **Data Summary**

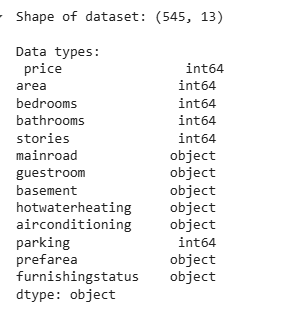
## **Preprocessing & EDA**

* **Shape**: (545 rows, 13 columns)
* **Types**:
  + Numerical: price, area, bedrooms, bathrooms, stories, parking
  + Categorical: main road, guest room, basement, hot water heating, air conditioning, furnishing status
* **Missing values**: None
* **Descriptive statistics (numeric):**
  + Average house price ≈ **4.76M** (min 1.75M, max 13.3M refers large spread).
  + Average area ≈ **5150 sq ft** (min 1650, max 16,200).
  + Bedrooms: mostly **2–4**, max 6.
  + Bathrooms: mostly **1–2**, max 4.
  + Parking: usually **0–1**, max 3.

**Code:**

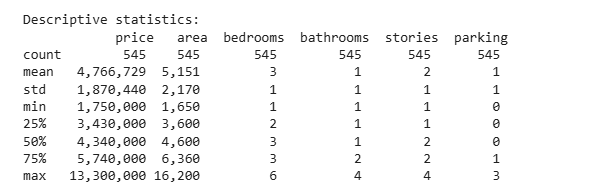


**Shape:**

**Data types:**

**Missing Values:**

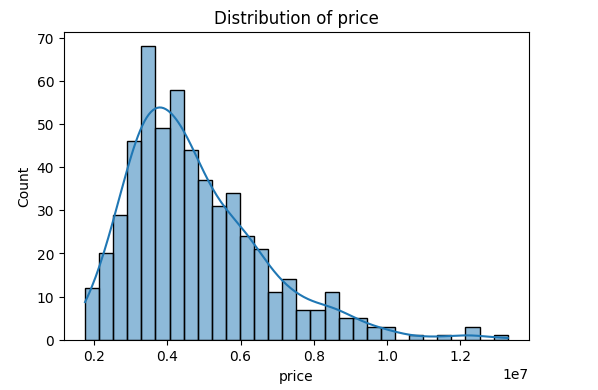
**Descriptive Statistics:**



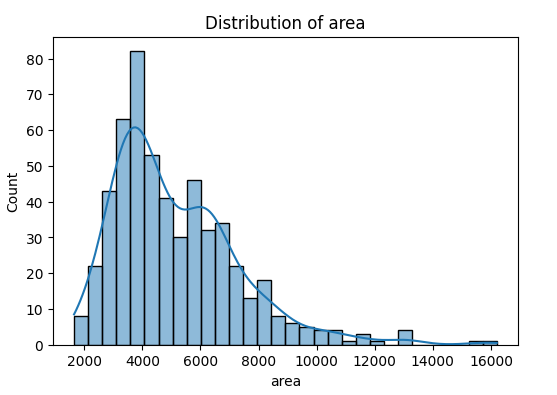
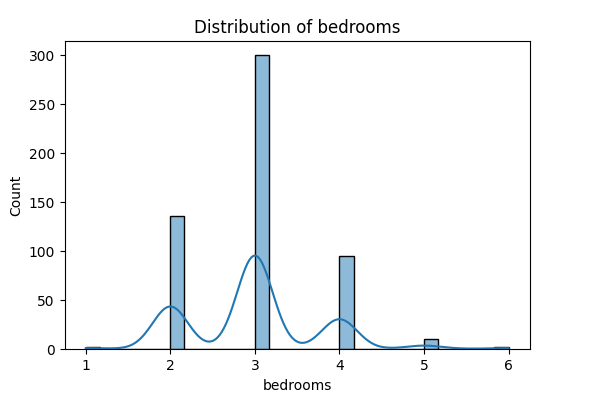
### **EDA:**

1. **Distribution of Numeric Features**
   * Plot histograms for price, area, bedrooms, bathrooms, stories, parking.
   * Helps identify skewness (e.g., are prices right-skewed with a few luxury houses?)
2. **Effect of Categorical Features on Price**
   * Compare average house price across categories using boxplots or group means.
   * Features to analyze:
     + Main road (houses on main road vs not)
     + Air conditioning (with AC vs without)
     + Furnishing status (furnished, semi, unfurnished)
     + basement, guestroom, etc.
3. **Relationships Between Variables**
   * Scatterplots:
     + area vs price to check if bigger houses are costlier.
     + bedrooms vs price to check if more bed rooms increase price.
   * Can also look at bath rooms vs price.
4. **Correlation Between Numeric Variables**
   * Use correlation heat map to see how strongly price relates to area, stories, etc.
   * Identify redundant variables (if two features are highly correlated).
5. **Outlier Detection**
   * Use boxplots/histograms to spot extremely high values of price or area.
   * Decide whether to keep or remove them for modeling.
     1. **Code:**



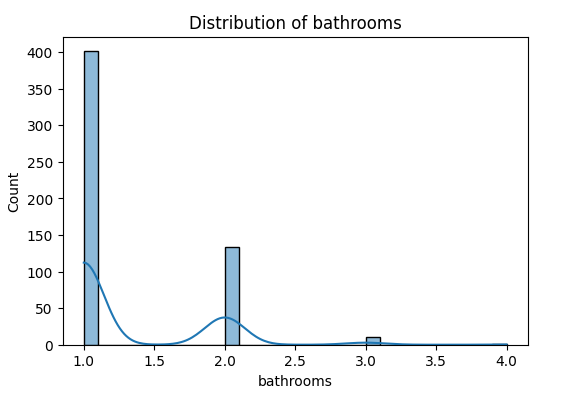


Most houses cost 3M–6M, with a few very expensive outliers above 10M

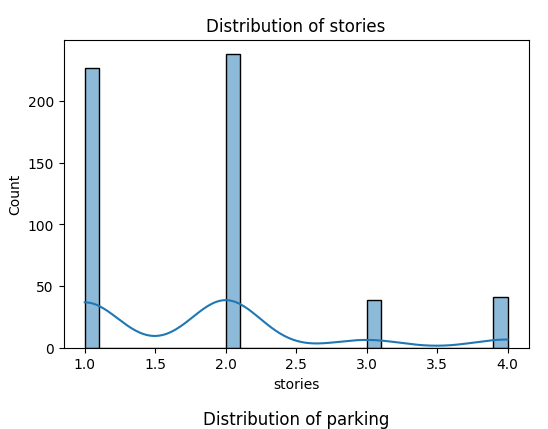


Most houses have an area below 6000 sq. ft., but some rare houses reach up to 16,000 sq. ft.

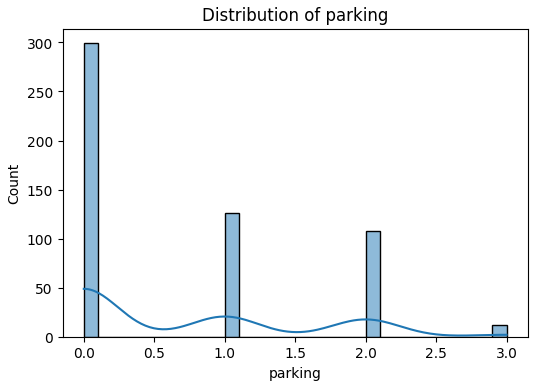
Houses typically have 2–4 bedrooms and 1–2 bathrooms, larger houses are less common.



Houses typically have 1–2 bathrooms; larger houses are less common.

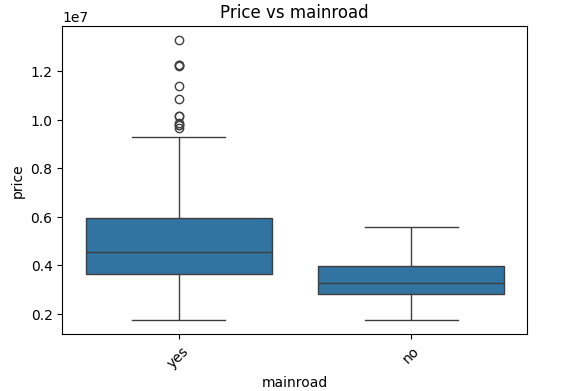
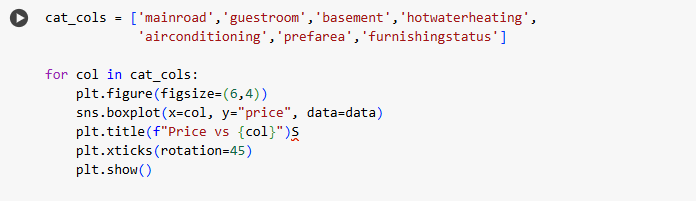


Most houses have **1–2 stories**, while 3–4 story houses are less common. The dataset is skewed toward smaller houses with fewer floors.

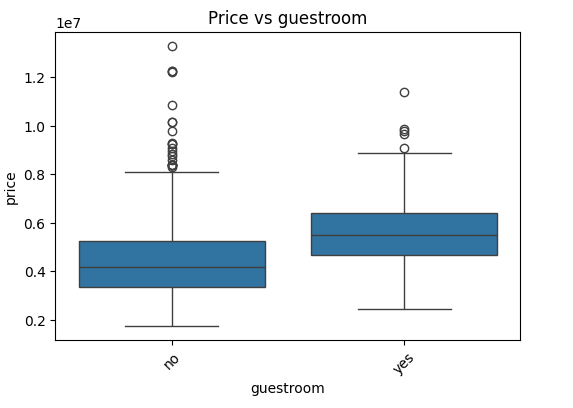


A majority of houses have **0–1 parking space**, while only a few have 2 or 3. This indicates limited parking availability in most houses.

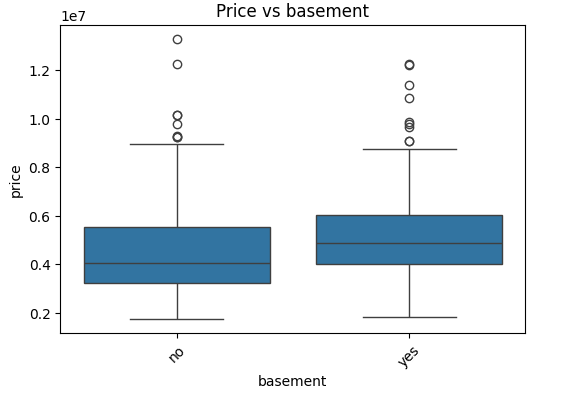
1. **Code:**



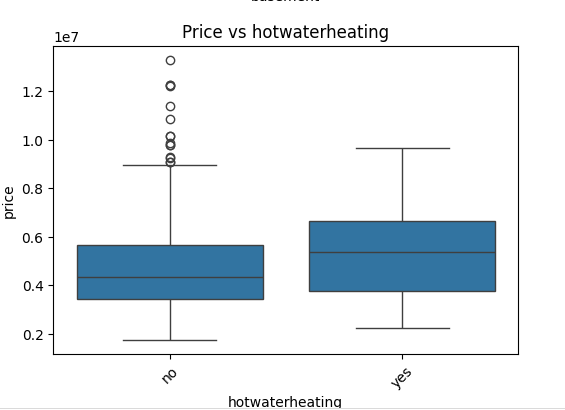
Houses located on the main road are generally more expensive than those off the main road.



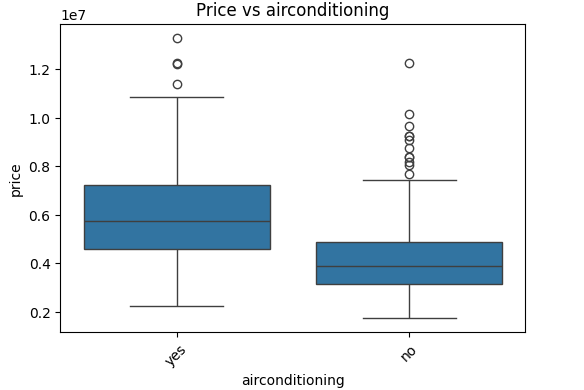
Houses with a guestroom have noticeably higher prices compared to those without.



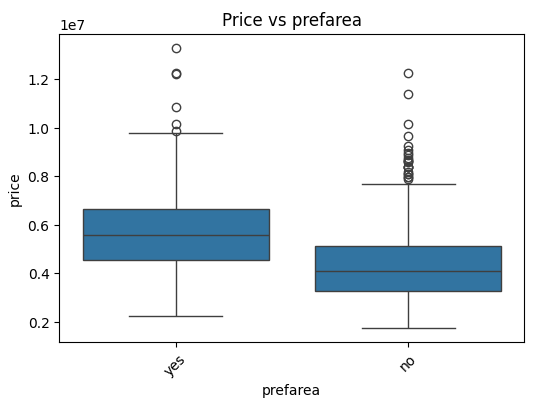
Houses with a basement tend to be priced higher, suggesting additional space/value.



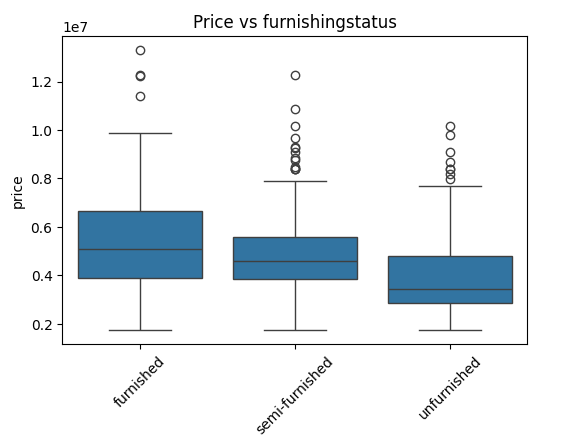
Houses with hot water heating are rare but command much higher prices.



Air-conditioned houses are significantly more expensive on average.

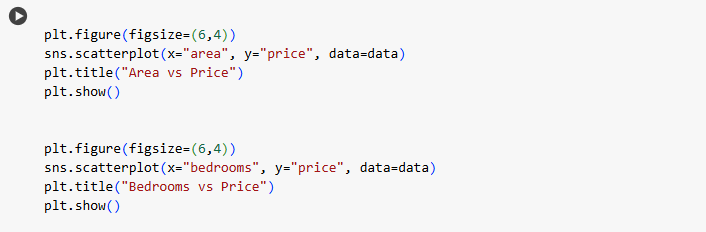


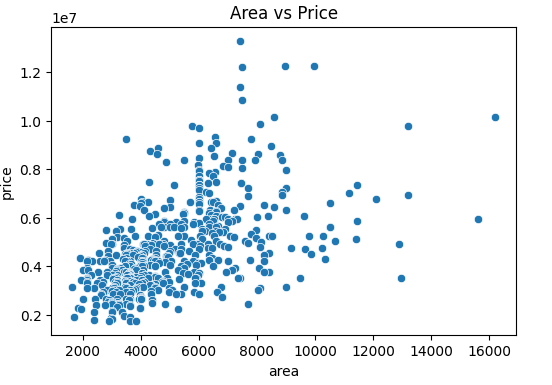
Houses in preferred areas are priced higher than those in non-preferred areas.



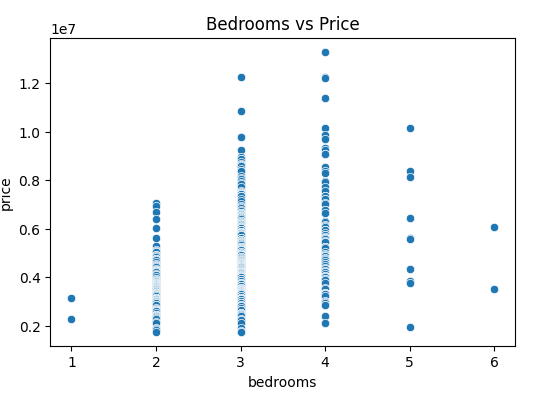
Furnished houses are the most expensive, semi-furnished are mid-range, and unfurnished are the least expensive.

1. **Code:**



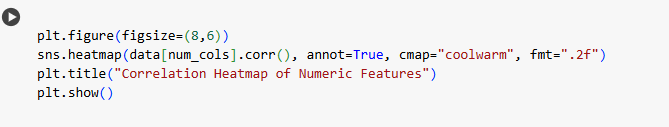


There is a strong positive trend larger area strongly increases house price.

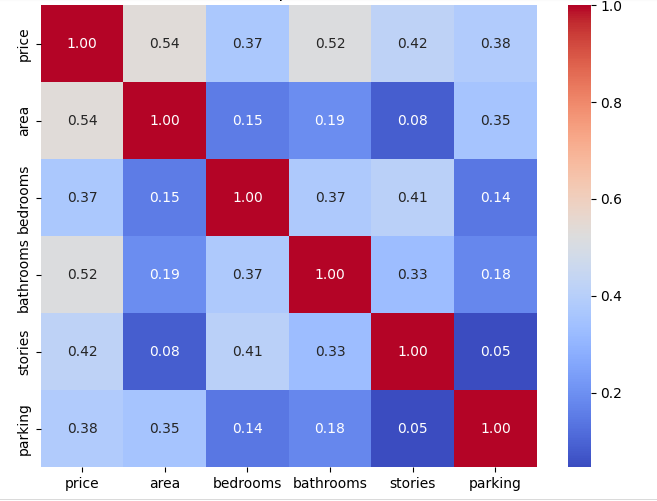


More bedrooms slightly increase price, but the effect is weaker than area.

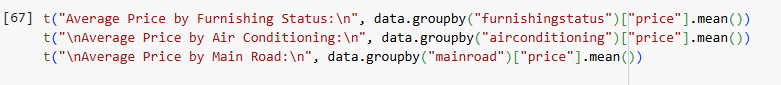
1. **Code:**



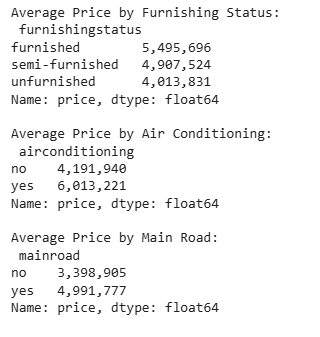
****

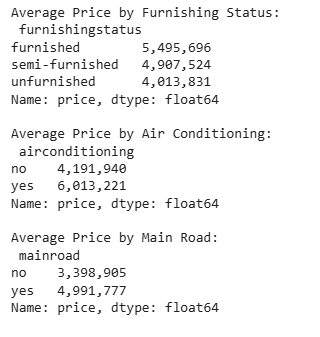


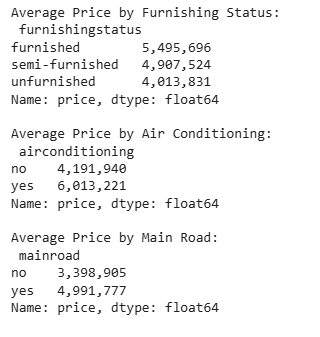
Price is strongly correlated with area and moderately correlated with bedrooms, bathrooms, and stories.



1. **Average Prices:**







### **1. Missing Values**

## **Data Cleaning & Feature Engineering**

* You already checked: **No missing values**.

### **2. Encoding Categorical Variables**

We need to convert categorical features into numeric form so ML models can use them:

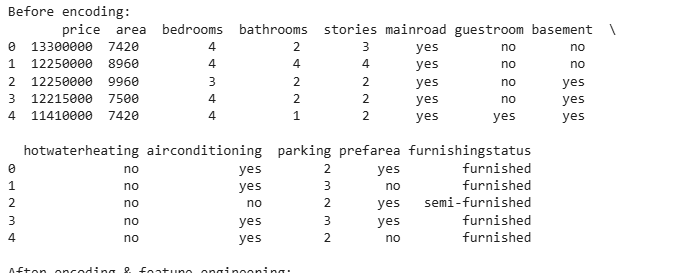
* **Yes/No features** (main road, guestroom, basement, hot water heating, air conditioning) are converted to 1/0.
* **Furnishing status** (multi-category: furnished, semi-furnished, unfurnished), applied **one-hot encoding**.

### **3. Feature Engineering**

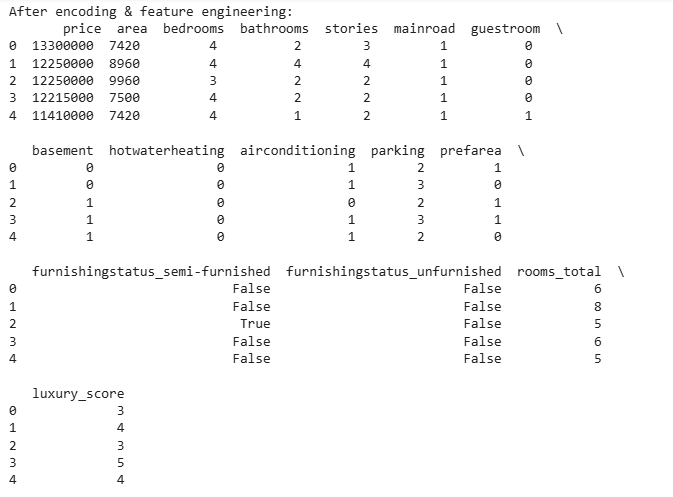
* Rooms total = bedrooms + bathrooms → total number of rooms.
* Luxury score = air conditioning + basement + parking → proxy for house luxury level.

**Code**

**Result:**

**Before Encoding**

**After Encoding**



Furnishing status had **three categories (furnished, semi-furnished, unfurnished)**, so we applied **one-hot encoding** to avoid losing information. It creates new columns like:

* furnishingstatus\_semi-furnished
* furnishingstatus\_unfurnished
* The dropped category (furnished) is treated as the **reference (baseline)**.

The target variable in this dataset is **price**, which is continuous. Therefore, a class imbalance check is not directly applicable because imbalance is only relevant for categorical target variables in classification problems.

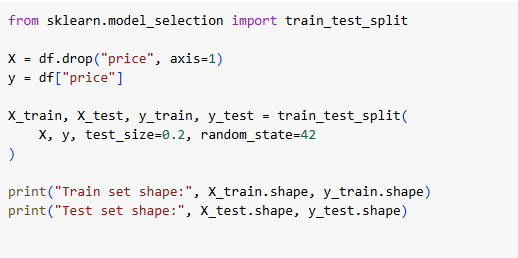
## **Data Splitting**

## **Class balance and imbalance**

**Chosen Split:** Since the dataset has 545 rows, an **80/20** split was chosen to maximize training data while keeping sufficient test data for fair evaluation.

* Train set: (436, features)
* Test set: (109, features)

**Code:**

****

**Output:**

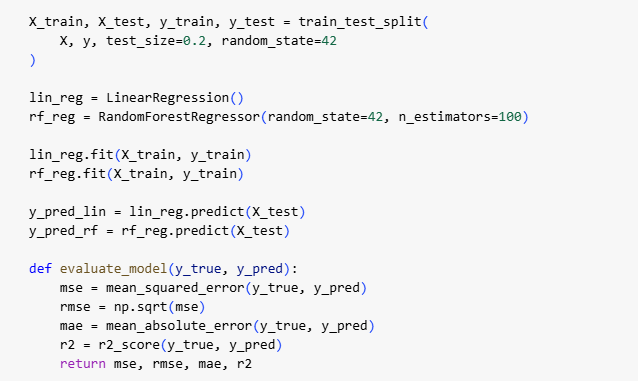


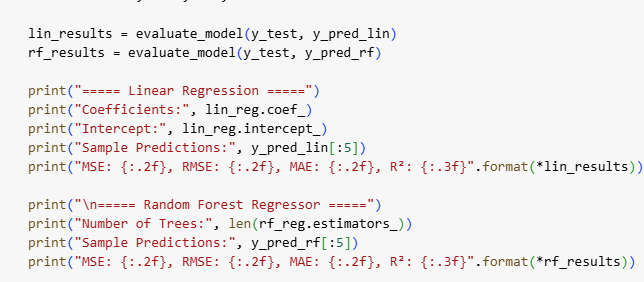
**Models used:**

## **Model Building**

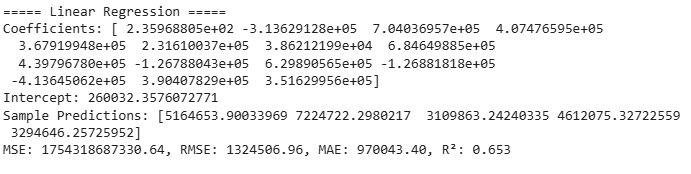
* Linear Regression
* Random Forest Regression



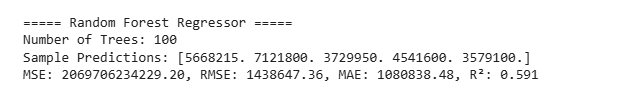


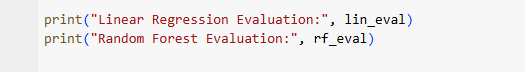


**Linear Regression:**



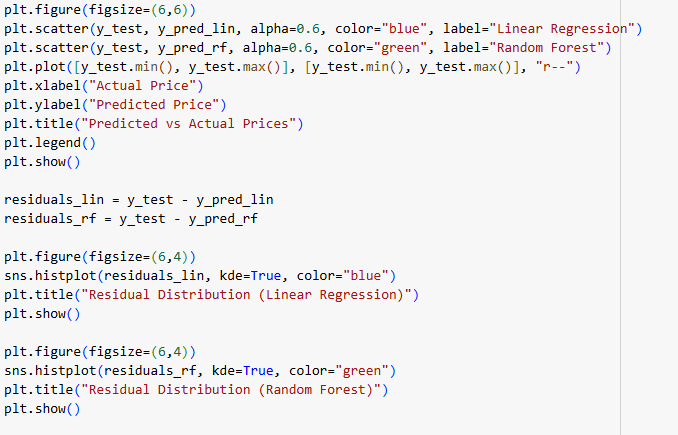
**Random Forest:**

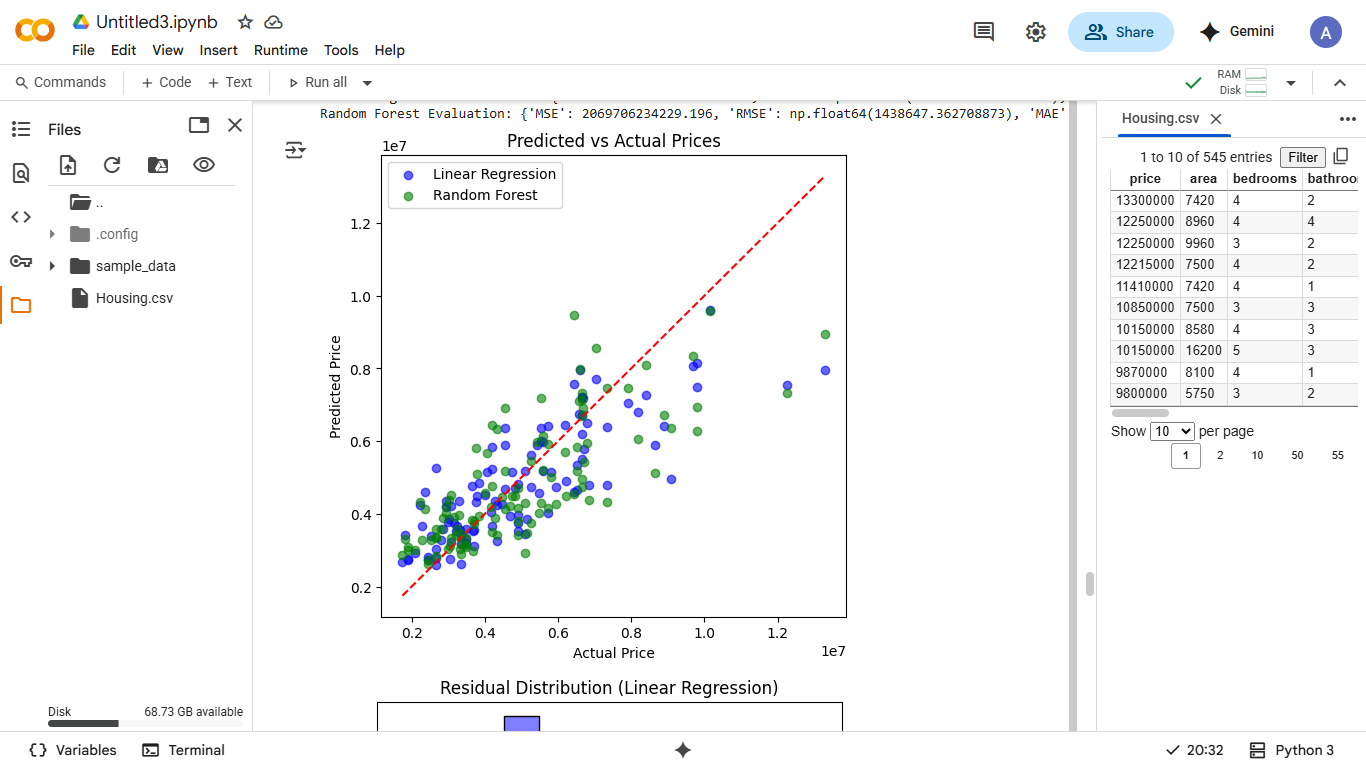




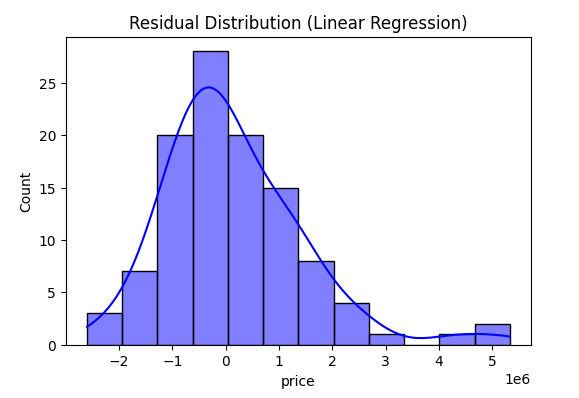
## **Model Evaluation**



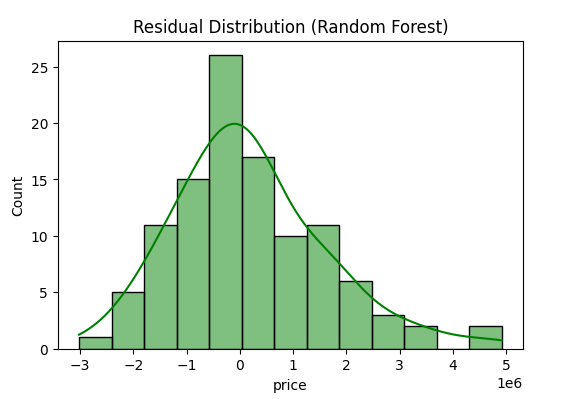




Both models follow the diagonal trend, but Linear Regression predictions (blue) align more closely with actual prices than Random Forest (green), showing less spread from the ideal line.



Errors are centered around zero with moderate spread, indicating reasonable predictions, though some under/overestimation exists.



Errors are more spread out compared to Linear Regression, showing that Random Forest is less consistent on this dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MSE** | **RMSE** | **MAE** | **R²** |
| Linear Regression | lower | lower | lower | 0.65 |
| Random Forest | higher | higher | higher | 0.59 |

**Best Fit Model:** **Linear Regression**

**Reason:**

1. It gives a **higher R² score (0.65 vs 0.59)**, meaning it explains more variance in house prices.
2. It has **lower RMSE and MAE**, so its predictions are closer to actual values.
3. Since features like area, bedrooms, stories have a near-linear relationship with price, Linear Regression matches the data structure better.

**Random Forest** was less effective here, possibly due to limited dataset size and weaker non-linear effects.

**Conclusion:**

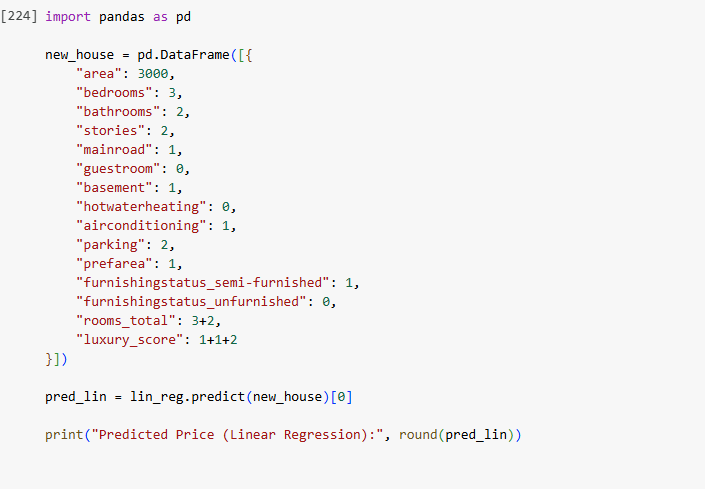
After comparing both models, **Linear Regression proved to be the better choice** for this dataset. It achieved a higher R² score and lower prediction errors compared to Random Forest. This suggests that the relationship between housing features (area, bedrooms, stories, etc.) and price is predominantly linear, making Linear Regression a more suitable and interpretable mode for this problem.

* The analysis reveals that **larger houses located in preferred areas with modern amenities** such as air conditioning, guestroom, basement, and better furnishing status are significantly more expensive compared to smaller, less-equipped houses.

## **Insight**

* Among the models tested, **Linear Regression outperformed Random Forest** in terms of predictive accuracy, achieving higher R² and lower error values.
* This suggests that the relationship between housing features (such as area, stories, and amenities) and price is largely **linear in nature**, making Linear Regression the most suitable and interpretable model for this dataset.
* Therefore, we can conclude that house prices are primarily driven by **size, location, and key amenities**, and a linear model is sufficient to capture these relationships effectively**.**

|  |  |
| --- | --- |
| **Aspect** | **Insight** |
| **Price Distribution** | Most houses are priced between 3M–6M; a few very high-priced houses create a right-skew. |
| **Area** | Larger houses strongly increase price (correlation ≈ 0.53 with price). |
| **Stories & Rooms** | More stories, bedrooms, and bathrooms generally lead to higher prices. |
| **Location (Main road, Preferred area)** | Houses on the main road and in preferred areas are significantly more expensive. |
| **Amenities (AC, Guestroom, Basement)** | Modern amenities add substantial value to house price. |
| **Furnishing Status** | Furnished houses are the most expensive, semi-furnished are mid-range, unfurnished are the cheapest. |
| **Model Performance** | Linear Regression outperformed Random Forest (higher R², lower error). |
| **Conclusion** | Bigger houses in good locations with modern amenities cost more, and Linear Regression is the best model for predicting prices. |



## **Prediction**



## **Summary**

