FRONT PAGE

Business Understanding -

The goal of this data analytics project is to understand what factors influence house prices and to develop a model that can help predict prices based on property features. This is relevant for real estate agencies, property developers, and individual investors who need to evaluate property values and make pricing decisions based on key features such as size, amenities, and location preferences.

The dataset consists of 545 house listings and 13 variables, including price, physical attributes, such as area, bedrooms, and bathrooms, and lifestyle or infrastructure-related features, such as air conditioning, furnishing status.

Price ranges –

- Minimum – 1,750,000
- Maximum – 13,300,000

Categorical Variables –

- Main road
- Guestrooms
- Basement
  Hot water heating
- air conditioning
- Prefarea
- Furnishing status

Data Understanding -

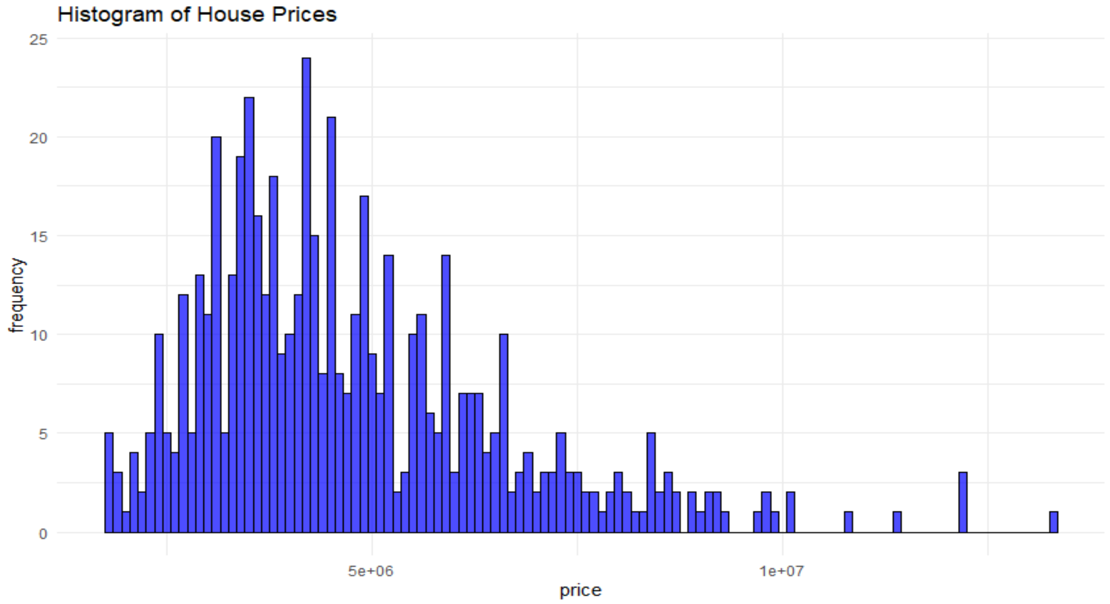


Figure 1: Histogram of House Prices

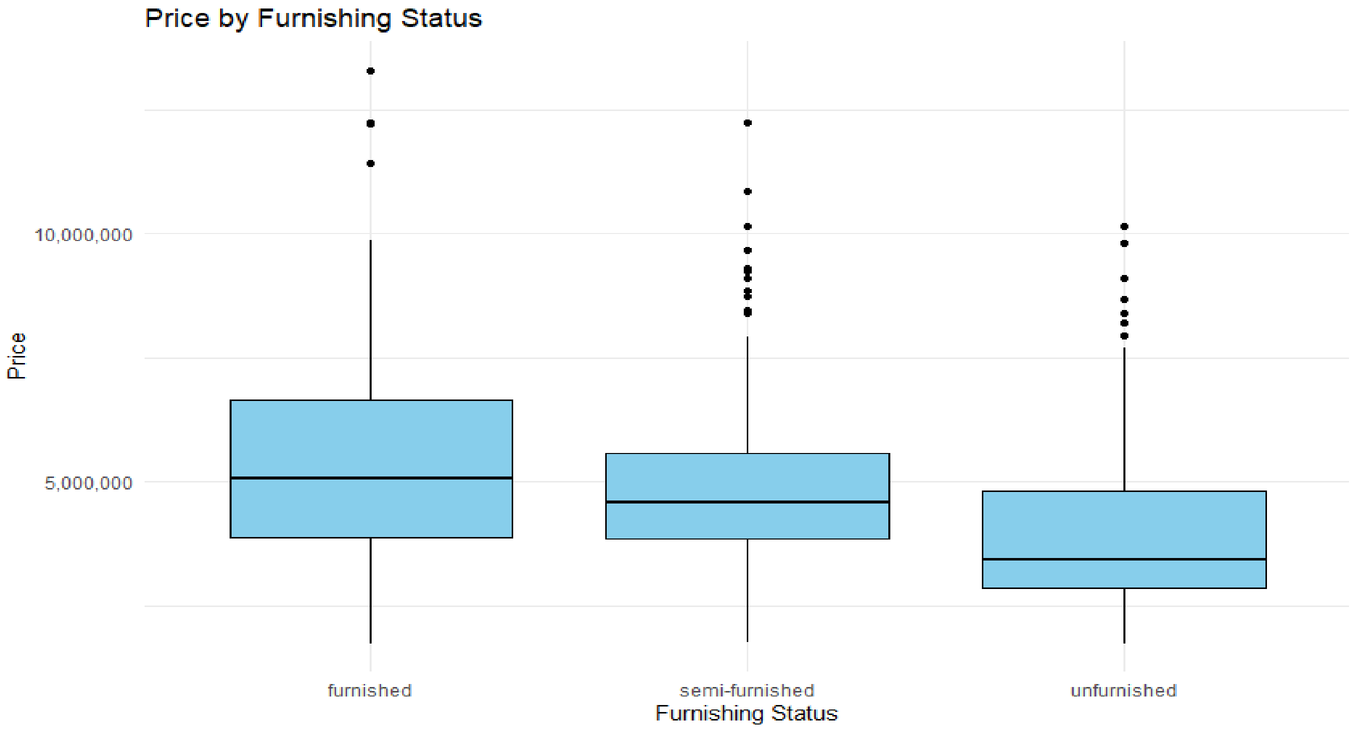


Figure 2: Boxplot of Price by Furnishing Status - Houses in which are furnished cost more than those without.

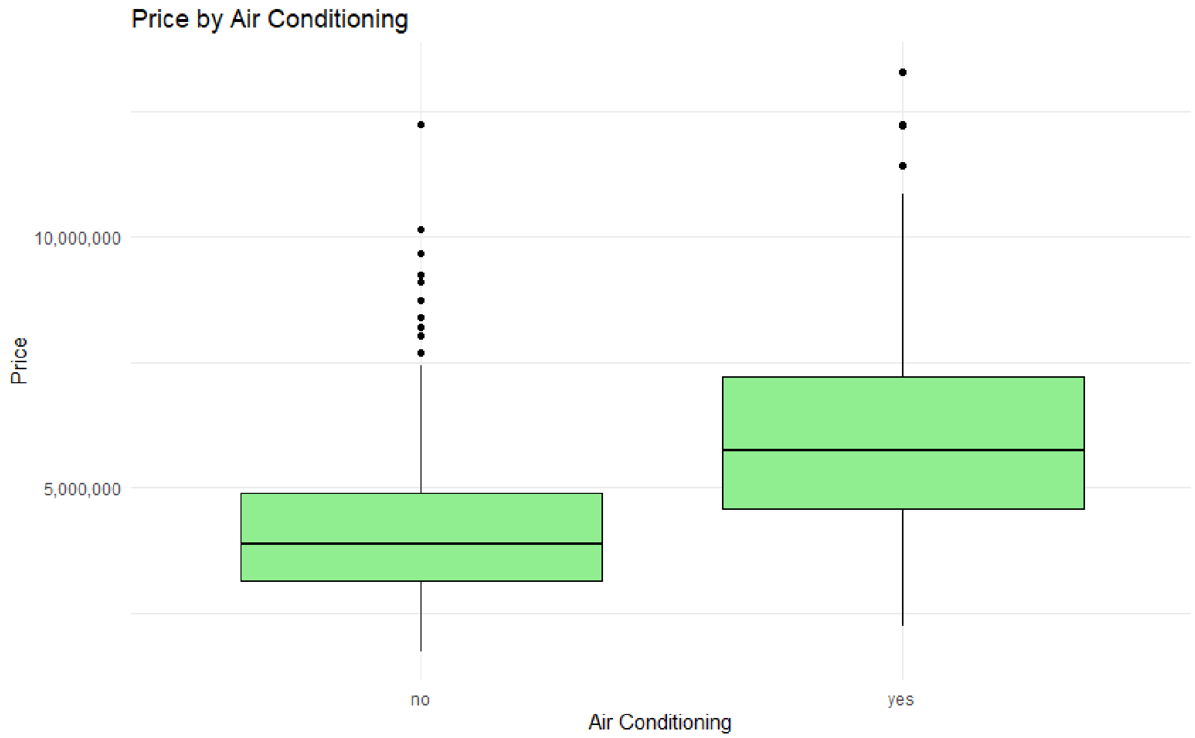


Figure 3: Boxplot of Price by Air Conditioning - Houses with A/C tend to cost more than those without it.

Figure 5

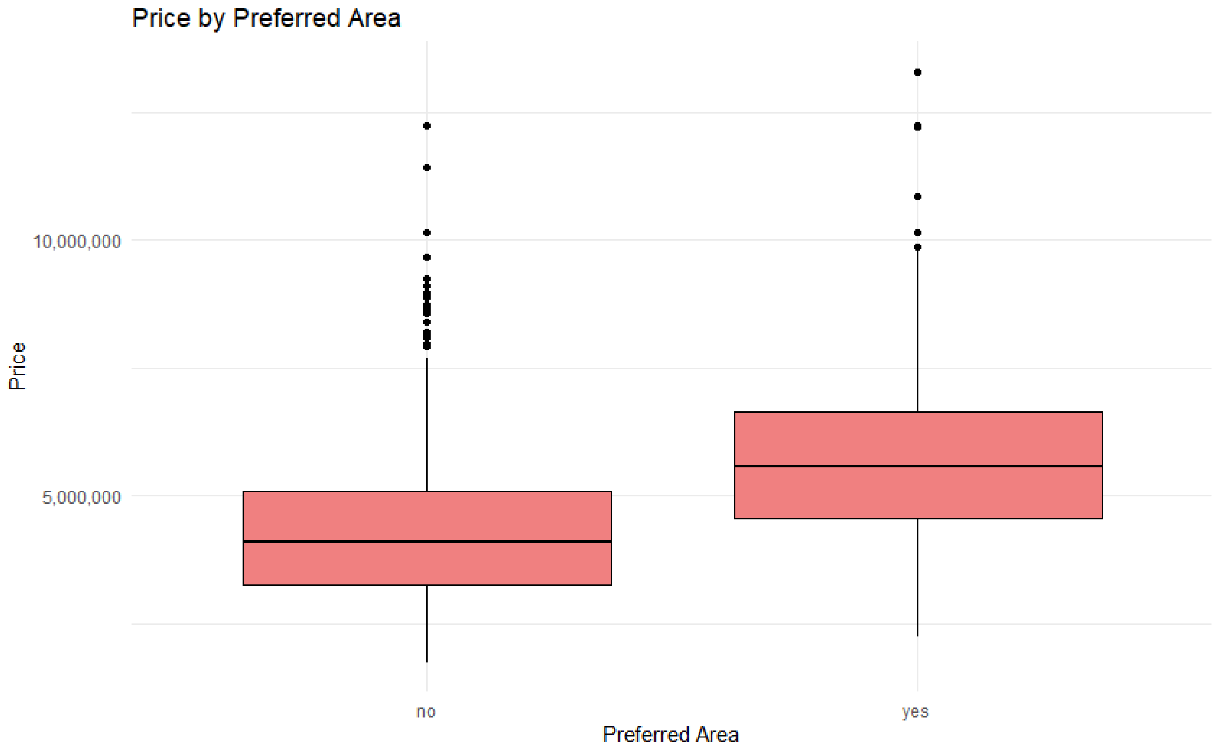


Figure 4: Boxplot of Price by Preferred Area – Home sin preferred areas clearly sell at a premium.

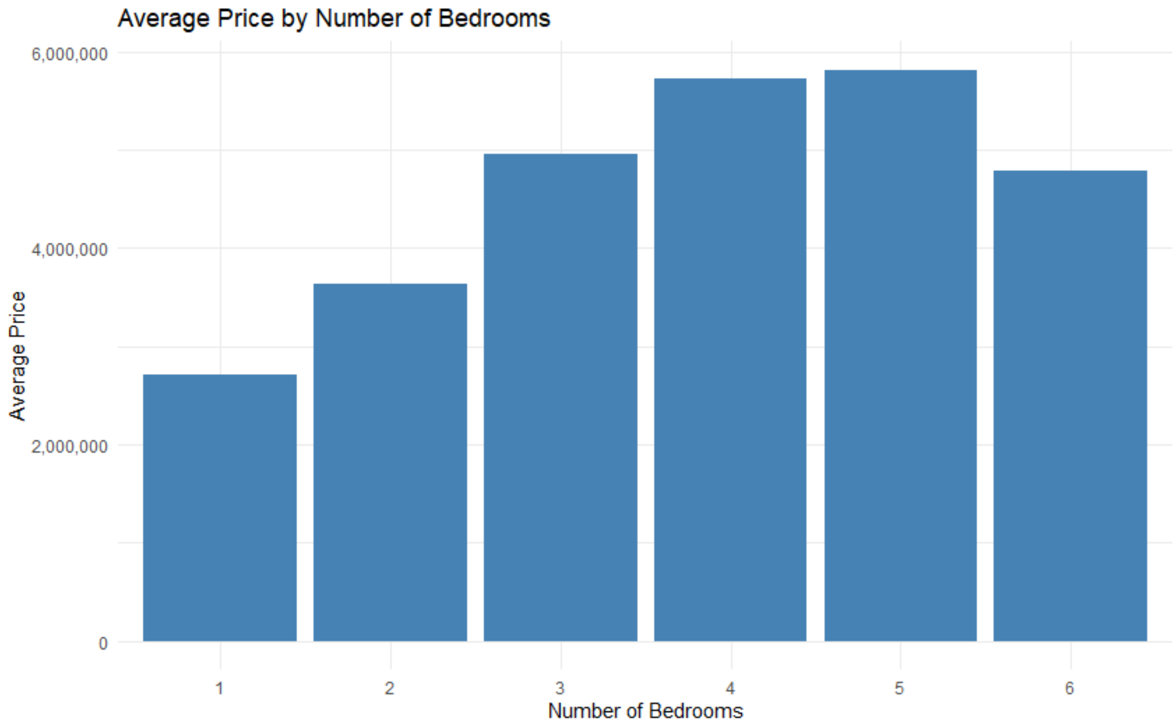


Figure 5: Average Price by Number of Bedrooms.

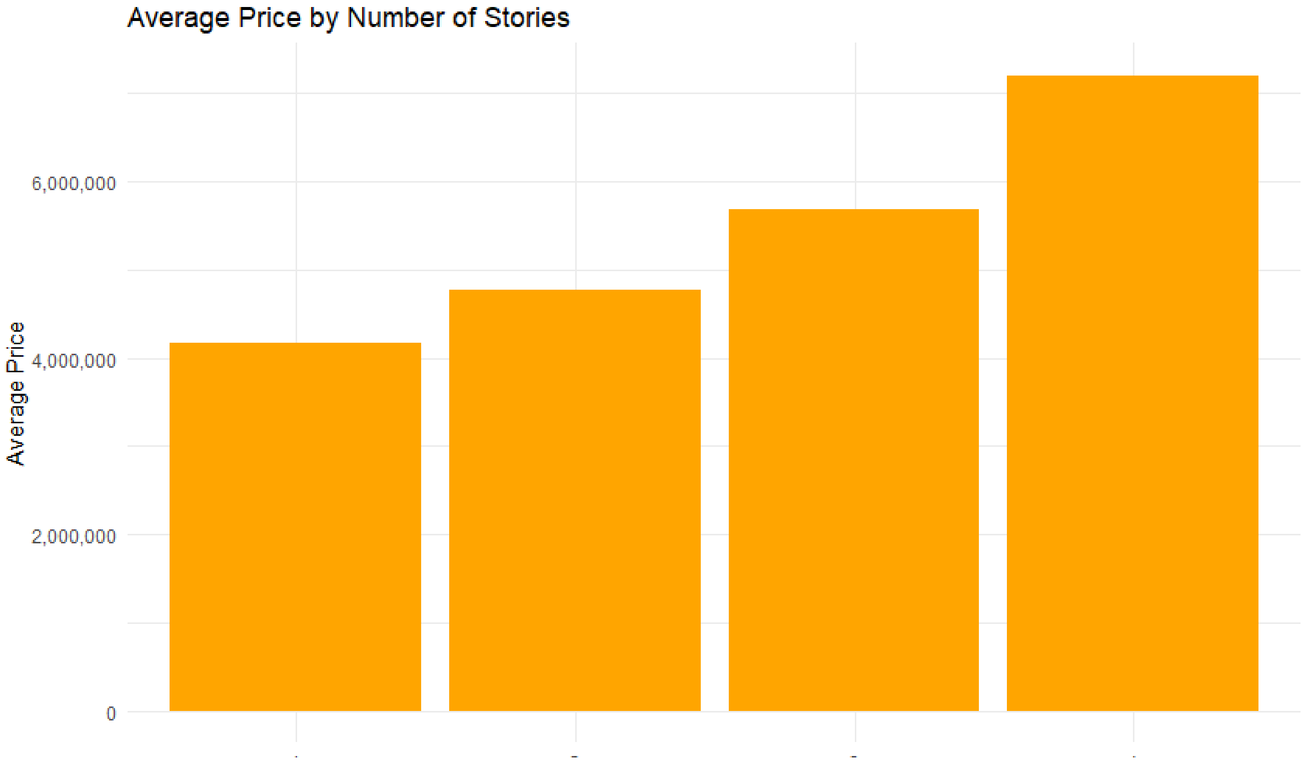
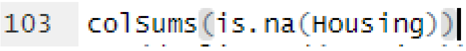


Figure 6: Average Price by Number of Stories.

Data Preparation –

Before modelling. The dataset was reviewed and prepared to ensure quality and consistency:

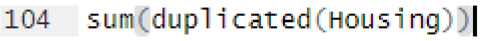
Missing values –



```
colSums(is.na(Housing))
```

The dataset was checked using the above line of code. No missing values were found, so no imputation was necessary.

Duplicates –

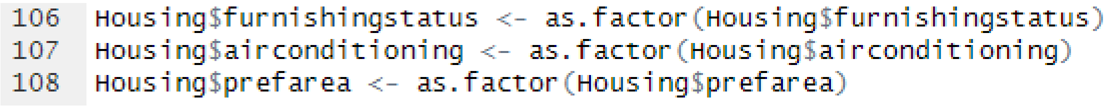


```
sum(duplicated(Housing))
```

Duplicate rows were identified using the above line of code and confirmed to be absent.

Data Types –

Categorical variables –



```
Housing$furnishingstatus <- as.factor(Housing$furnishingstatus)
Housing$airconditioning <- as.factor(Housing$airconditioning)
Housing$prefarea <- as.factor(Housing$prefarea)
```

Categorical variables were explicitly converted to factors using the above function. This was necessary for regression modelling.

Outliers –

Extreme values were visually inspected through boxplots. While a few high-price outliers were present, they were not removed, as they appeared legitimate and relevant to real estate pricing.

Transformations –

To improve model fit and address skewness in the price variable, a log transformation was applied. This resulted in a more normal distribution of residuals and better homoscedasticity in the regression model.

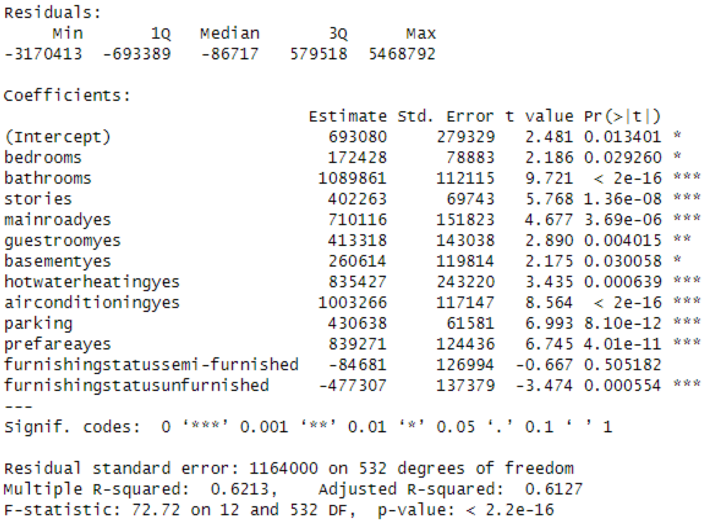
Feature Encoding –

R automatically handled dummy variable creation for categorical features in the regression model. For example, furnishing status was split into reference fully furnished and two dummy categories, semi-furnished, and unfurnished.

Modelling -

Model Fit –

- Multiple R Squared – 0.6213 – figures suggests 62.1% of the variation in house prices.
- Adjusted R Squared – 0.6127 – figures shows that predictors add real value.
- F-Statistic – 72.72, p < 2.2e-16 – figure shows that the model is statistically significant overall – predicts price better than chance.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-3170413  -693389   -86717   579518  5468792 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      693080     279329   2.481 0.013401 *  
bedrooms                         172428      78883   2.186 0.029260 *  
bathrooms                       1089861     112115   9.721  < 2e-16 ***
stories                          402263      69743   5.768 1.36e-08 ***
mainroadyes                      710116     151823   4.677 3.69e-06 ***
guestroomyes                     413318     143038   2.890 0.004015 ** 
basementyes                      260614     119814   2.175 0.030058 *  
hotwaterheatingyes               835427     243220   3.435 0.000639 ***
airconditioningyes              1003266     117147   8.564  < 2e-16 ***
parking                          430638      61581   6.993 8.10e-12 ***
prefareayes                      839271     124436   6.745 4.01e-11 ***
furnishingstatussemi-furnished   -84681     126994  -0.667 0.505182    
furnishingstatusunfurnished     -477307     137379  -3.474 0.000554 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1164000 on 532 degrees of freedom
Multiple R-squared:  0.6213,	Adjusted R-squared:  0.6127 
F-statistic: 72.72 on 12 and 532 DF,  p-value: < 2.2e-16
```

Figure 7: Output Summary from Standard Linear Regression Model

This output displays the estimated coefficients, standard errors, t-values, and p-values for each predictor in the model predicting house prices on the original scale. Key variables such as bathrooms, air conditioning, and preferred area show strong, statistically significant effects. The adjusted R-squared of 0.6127 indicates that about 61% of the variance in housing prices is explained by the model.

Coefficients from Standard Linear Regression Model – (Not Logged)

Predictor	Estimate	p-value	Interpretation
Intercept	693,080	0.013	Baseline price when all other predictors = 0 (not very meaningful alone)
Bedrooms	+172,428	0.029	Each extra bedroom increases price by approximately $172,000
Bathrooms	+1,089,861	< 0.0001	Huge, significant impact — bathrooms greatly increase value
Stories	+402,263	< 0.0001	More stories add significant value
Main road (Yes)	+710,116	< 0.0001	Homes on main roads are more expensive
Guest room (Yes)	+413,318	0.004	Guestroom adds substantial value
Basement (Yes)	+260,614	0.030	Basement contributes positively to price
Hot water heating (Yes)	+835,427	0.0006	Strong positive effect on price
Air conditioning (Yes)	+1,003,266	< 0.0001	Very strong, highly significant — A/C is a major price driver
Parking (per space)	+430,638	< 0.0001	Parking spaces significantly increase home value
Preferred area (Yes)	+839,271	< 0.0001	Homes in preferred areas are valued much higher
Furnishing: Semi-furnished	–84,681	0.505	Not statistically significant
Furnishing: Unfurnished	–477,307	0.0006	Significantly lower prices than fully furnished homes

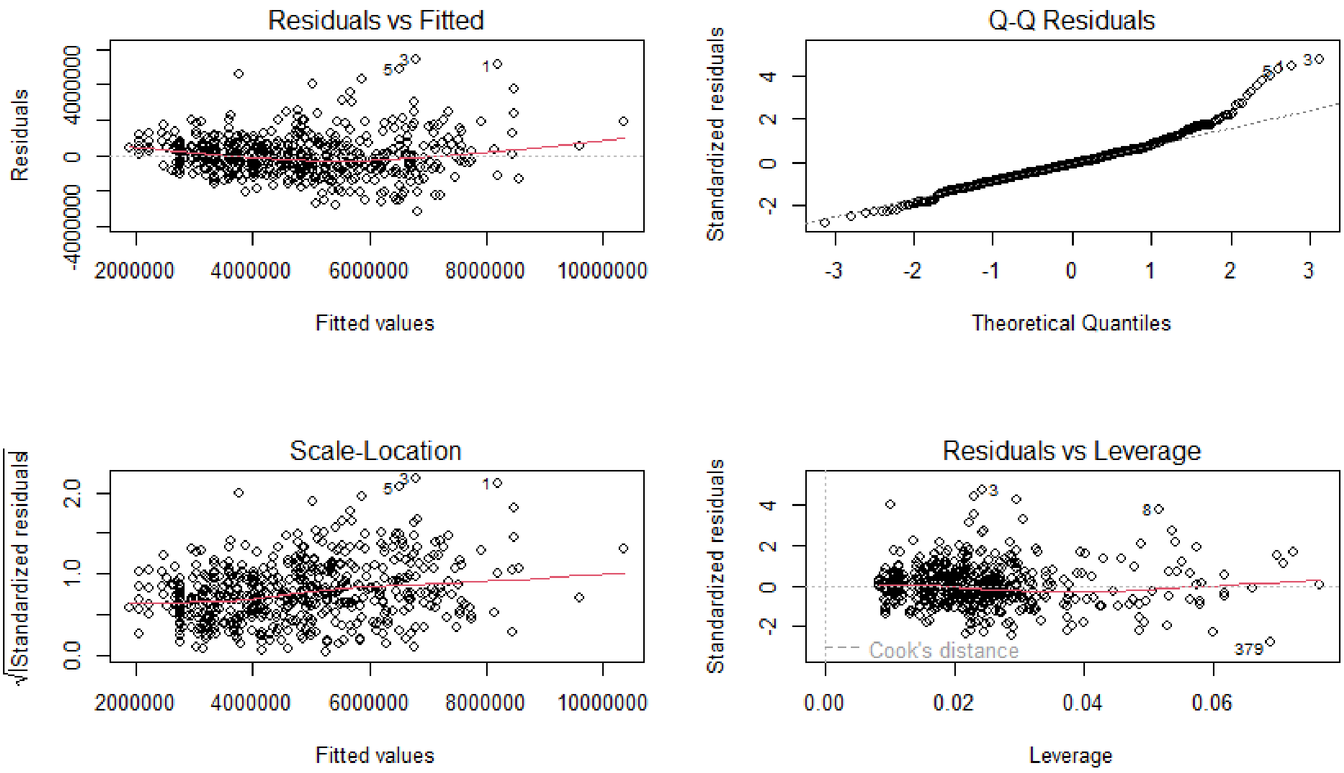
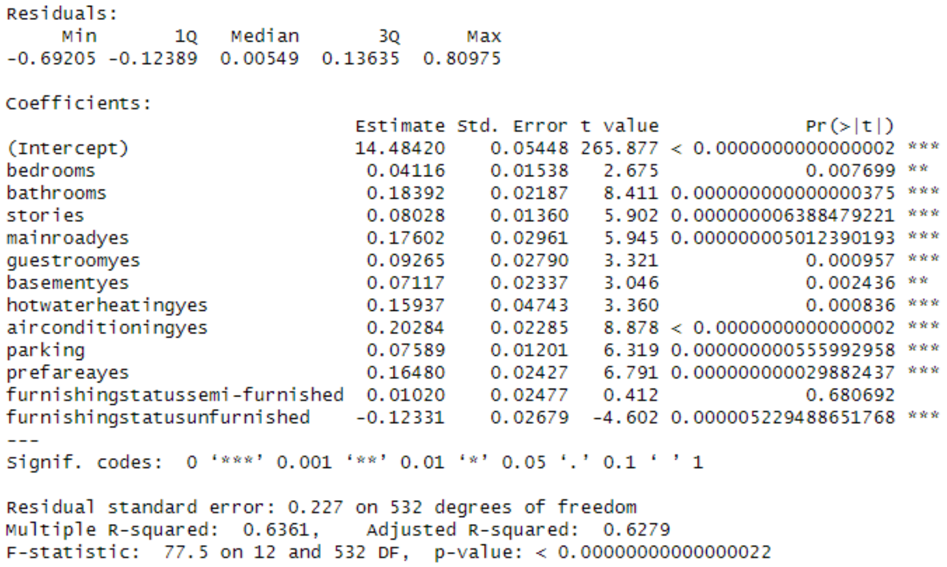


Figure 8: Residual Diagnostics for Linear Regression Model – Where fitted values range from 2 million to 10 million – unlogged.

These plots check the assumptions of linear regression. The Residuals vs Fitted plot shows non-constant variance, and the Q-Q plot shows moderate deviation from normality. These issues motivate the log transformation of the target variable.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.69205 -0.12389  0.00549  0.13635  0.80975 

Coefficients:
                              Estimate Std. Error t value             Pr(>|t|)    
(Intercept)                   14.48420    0.05448 265.877 < 0.0000000000000002 ***
bedrooms                       0.04116    0.01538   2.675             0.007699 ** 
bathrooms                      0.18392    0.02187   8.411 0.000000000000000375 ***
stories                        0.08028    0.01360   5.902 0.000000006388479221 ***
mainroadyes                    0.17602    0.02961   5.945 0.000000005012390193 ***
guestroomyes                   0.09265    0.02790   3.321             0.000957 ***
basementyes                    0.07117    0.02337   3.046             0.002436 ** 
hotwaterheatingyes             0.15937    0.04743   3.360             0.000836 ***
airconditioningyes             0.20284    0.02285   8.878 < 0.0000000000000002 ***
parking                        0.07589    0.01201   6.319 0.000000000555992958 ***
prefareayes                    0.16480    0.02427   6.791 0.000000000029882437 ***
furnishingstatussemi-furnished 0.01020    0.02477   0.412             0.680692    
furnishingstatusunfurnished   -0.12331    0.02679  -4.602 0.000005229488651768 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.227 on 532 degrees of freedom
Multiple R-squared:  0.6361,	Adjusted R-squared:  0.6279 
F-statistic:  77.5 on 12 and 532 DF,  p-value: < 0.00000000000000022
```

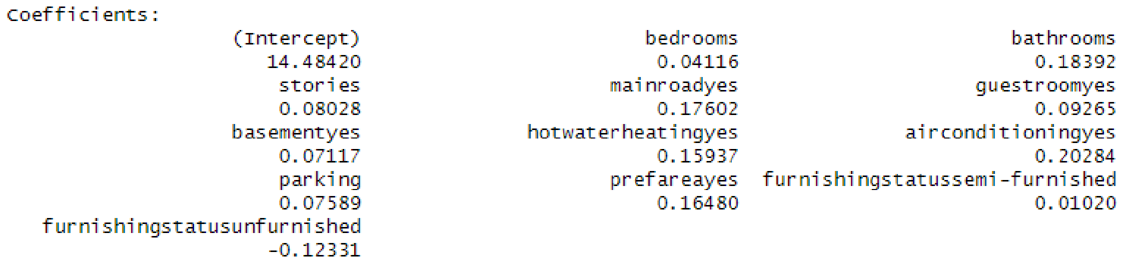
Figure 9: Output Summary from Log-Transformed Regression Model -

This version of the model uses the natural logarithm of price as the response variable to address skewness and heteroscedasticity. The adjusted R-squared improves slightly to 0.6279, indicating a better fit. Variables like air conditioning, bathrooms, and preferred areas remain highly significant. Coefficients in this model are interpreted as approximate percentage changes in price.

Coefficients from Log-Transformed Regression Model (Logged)

Variable	Estimate	Meaning (Approximate % Change in Price)
Bedrooms	0.041	+4.1% per additional bedroom
Bathrooms	0.184	+18.4% per additional bathroom
Stories	0.080	+8.0% per extra story
Main road (Yes)	0.176	+17.6% if located on a main road
Guest room (Yes)	0.093	+9.3% increase in price if the home has a guest room
Basement (Yes)	0.071	+7.1% increase with a basement
Hot water heating (Yes)	0.159	+15.9% increase if the house has hot water heating
Air conditioning (Yes)	0.203	+20.3% increase in price if A/C is present

Variable	Estimate	Meaning (Approximate % Change in Price)
Parking (per space)	0.076	+7.6% per parking space
Preferred area (Yes)	0.165	+16.5% if the house is in a preferred area
Furnishing: Semi-furnished	0.010	Not statistically significant (p = 0.681)
Furnishing: Unfurnished	–0.123	–12.3% cheaper compared to fully furnished homes (significant)



```
Coefficients:
                    (Intercept)                        bedrooms                       bathrooms 
                       14.48420                         0.04116                         0.18392 
                        stories                     mainroadyes                    guestroomyes 
                        0.08028                         0.17602                         0.09265 
                    basementyes              hotwaterheatingyes              airconditioningyes 
                        0.07117                         0.15937                         0.20284 
                        parking                     prefareayes  furnishingstatussemi-furnished 
                        0.07589                         0.16480                         0.01020 
   furnishingstatusunfurnished 
                       -0.12331 
```

Figure 10: Coefficients Extracted from Log-Transformed Regression Model.

<u>Interpretation of Coefficients from Log Model – (Plain Summary)</u>

Variable	Estimate	Interpretation
(Intercept)	14.484	Base log(price) when all other variables are 0
Bedrooms	0.041	+4.1% price per additional bedroom
Bathrooms	0.184	+18.4% price per bathroom
Stories	0.080	+8.0% price per extra story
Main road (yes)	0.176	+17.6% if the house is on a main road
Guest room (yes)	0.093	+9.3% if a guest room is available
Basement (yes)	0.071	+7.1% if there's a basement
Hot water heating (yes)	0.159	+15.9% if hot water heating is installed
Air conditioning (yes)	0.203	+20.3% with air conditioning

Variable	Estimate	Interpretation
Parking (per space)	0.076	+7.6% per parking space
Preferred area (yes)	0.165	+16.5% if the house is in a preferred area
Furnishing: Semi-furnished	0.010	Negligible effect; not statistically significant
Furnishing: Unfurnished	–0.123	–12.3% lower price compared to fully furnished

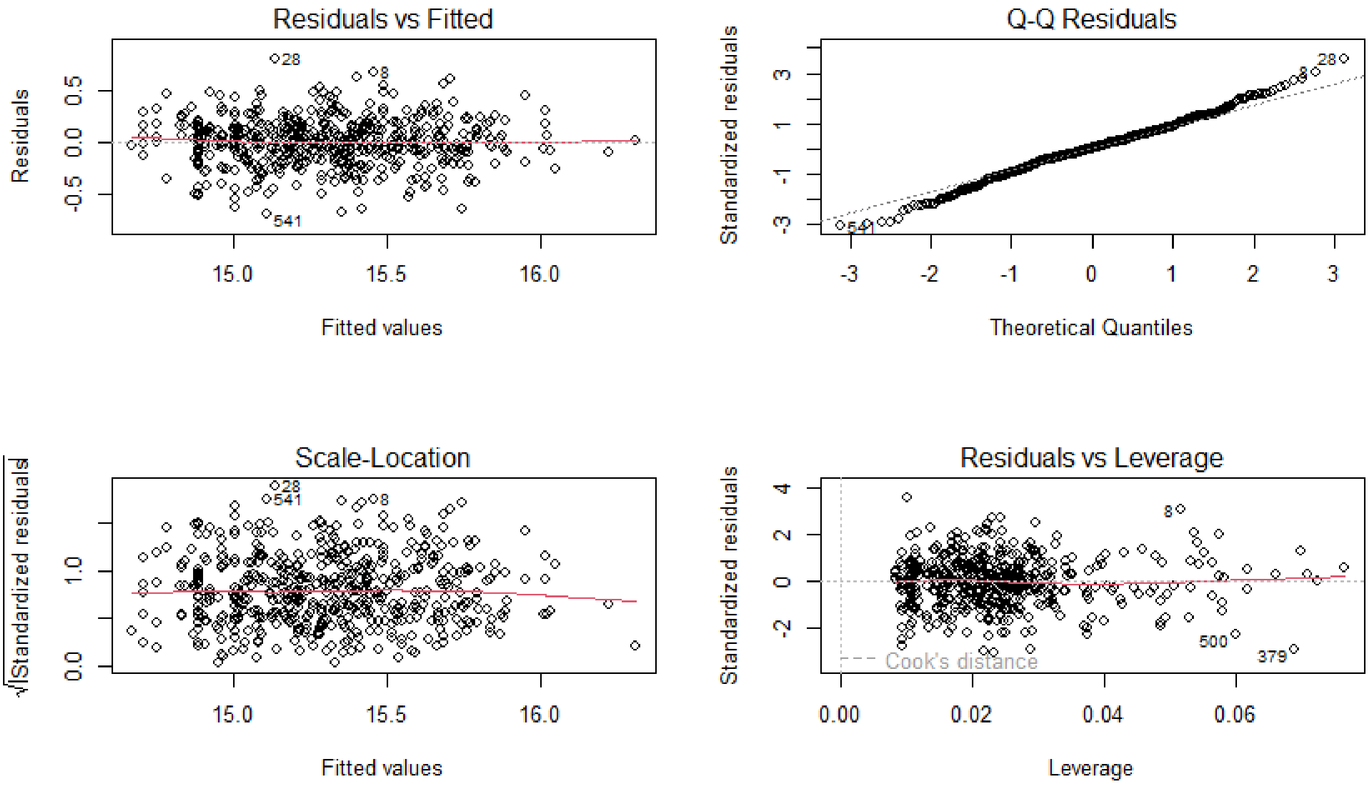


Figure 11: Residual Diagnostic Plot for Log-Transformed Regression Model. Where fitted values range from 15 to 16 – log price scale.

After applying a log transformation to the price, residuals display improved behavior. The variance appears more constant, and residuals are closer to normally distributed. No highly influential points are observed.

Evaluation -

The final log-transformed regression model successfully explained approximately 63% of the variation in house prices. Key influencing factors included air conditioning, bathrooms, and location preferences. The model was evaluated through residual diagnostics and showed improved fit and assumption satisfaction after log transformation.

From an ethical perspective, if deployed in real-world housing markets, predictive models must be used with care. Pricing algorithms should be transparent and avoid reinforcing socioeconomic or geographic bias. Fair access to housing data and protection of sensitive personal or locational data are essential, especially under regulations such as GDPR.