STAT 4355.001 APPLIED LINEAR MODELS

FINAL PROJECT REPORT

GROUP 15 : HOUSE PRICE PREDICTION

Elshad Jafarli ♦ Pamela Kamdefwere ♦ Natthiya Sae Ngow

Chuan-Fa Tang

TABLE OF CONTENTS

PART I

INTRODUCTION TO THE DATASET

In this project, we explore a house price prediction dataset sourced from Kaggle, focusing on real estate markets in Sydney and Melbourne in Australia. The dataset contains 2,000 observations and 9 variables, covering houses built between 1900 and 2023. It includes a variety of features that may influence house prices, consisting of 6 numerical and 3 categorical variables.

ANALYSIS GOAL

The goal of this project is to develop predictive models that estimate house prices based on a range of structural and locational features. By applying statistical analysis, we aim to identify the variables that most strongly influence property values. Since many of these features exhibit a linear relationship with house prices, we focus on using linear regression models to determine which variables are the most and least significant predictors.

DATASET OVERVIEW

The datasets of 2,000 entries with 9 keys variables (6 numerical variables and 3 categorical variables) including the target variable *Price*. The *Id* column is excluded from analysis, as it is used solely as a identifier for each house (Ranging from 1 to 2,000 houses)

DATA VARIABLES

9 Keys Variables:

Feature	Data Type	Description
Area	int	Size of the house in square feet (Ranges from 500 to 5000 sqft)
Bedrooms	int	Number of bedrooms (From 1 to 5)
Bathrooms	int	Number of bathrooms (From 1 to 4)
Floors	int	Number of floors (From 1 to 3)
YearBuilt	int	Year the house was constructed (Built from 1900 to 2023)
Location	chr	Categorical: Downtown, Suburban, Urban, or Rural
Condition	chr	Categorical: Poor, Fair, Good, Excellent
Garage	chr	Binary: Yes or No
Price	int	Target variable : House sale price ($50,000–$1,000,000)

The dataset consists of two data types :

1. Numerical variables : Area, Bedroom, Bathrooms, Floors, YearBuilt Price.
2. Categorical variables : Location, Condition, Garage.

The target variable is Price.

EXPLORATORY DATA ANALYSIS (EDA)

We began by cleaning the dataset, checking for missing values, outliers, and duplicates. The data was found to be complete with no missing values.

Missing values	
Feature	Count
id	0
Area	0
Bedrooms	0
Bathrooms	0
Floors	0
YearBuilt	0
Location	0
Condition	0
Garage	0
Price	0

Descriptive Statistics and Observations :

House Price Summary Table : summary(house_data)

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Area	501.0	1653.0	2833.0	2786.0	3888.0	4999.0
Bedrooms	1.0	2.0	3.0	3.003	4.0	5.0
Bathrooms	1.0	2.0	3.0	2.553	4.0	4.0
Floors	1.0	1.0	2.0	1.994	3.0	3.0
YearBuilt	1900.0	1930.0	1961.0	1961.0	1993.0	2023.0
Price	50005.0	300098.0	539254.0	537677.0	780086.0	999656.0

The house price summary table shows that the average house area is approximately 2,786 square feet. On average, homes have 3 bedrooms and 2.5 bathrooms, with most having 2 floors. The mean year built is 1961, with properties ranging from 1900 to 2023. The average house price is approximately $537,677 and the most expensive home is priced nearly $1,000,000.

Top_10_Most_Expensive_Houses_Summary : summary(top10)

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Area	566	875.2	2703.5	2259.5	3070.5	4340
Bedrooms	1	1.25	3.5	2.8	4	4
Bathrooms	1	1.25	2	2.2	2.75	4
Floors	1	3	3	2.7	3	3
YearBuilt	1910	1940	1959	1961	1980	2012
Price	996357	996992	997472	997747	998117	999656

We sorted the data by descending price and selected the top 10 most expensive homes for closer analysis. The summary table reveals that the average area is approximately 2,260 square feet, which is smaller than the dataset's overall average, indicating that larger size is not always associated with higher prices. Some of these high-priced homes are relatively compact, likely due to factors such as location.

The average number of bedrooms is around 2.8, showing that the most expensive homes do not necessarily have more bedrooms. The average number of bathrooms is about 2.2, reflecting that a modest number of bathrooms is common among top-priced properties. This suggests that bathroom count alone is not a strong determinant of price at the high end of the market.

All of the top homes have 3 floors, indicating that multi-level living is typical among the most expensive properties in the dataset.

The average year built is 1961, with homes ranging from 1910 to 2012. This wide range shows that newer construction is not always more valuable. Lastly, the price range is very narrow, falling between $996,357 and $999,656.

Top 10 Most Expensive Houses Price

Ranking	Id	Area	Bedrooms	Bathrooms	Floors	YearBuilt	Location	Condition	Garage	Price
1	1005	3099	1	2	3	1997	Suburban	Good	No	$ 999,656
2	1788	566	4	1	1	1945	Urban	Fair	Yes	$ 999,453
3	1007	736	3	2	3	1939	Downtown	Good	Yes	$ 998,128
4	1553	2860	2	2	3	1910	Downtown	Fair	No	$ 998,084
5	242	2985	4	3	3	1969	Urban	Poor	Yes	$ 997,719
6	95	4340	1	1	3	1983	Urban	Poor	Yes	$ 997,226
7	1728	3597	4	2	3	2012	Downtown	Fair	Yes	$ 997,176
8	844	1293	4	4	3	1958	Urban	Poor	Yes	$ 996,931
9	1864	572	1	1	3	1960	Suburban	Fair	Yes	$ 996,740
10	21	2547	4	4	2	1935	Suburban	Good	Yes	$ 996,357

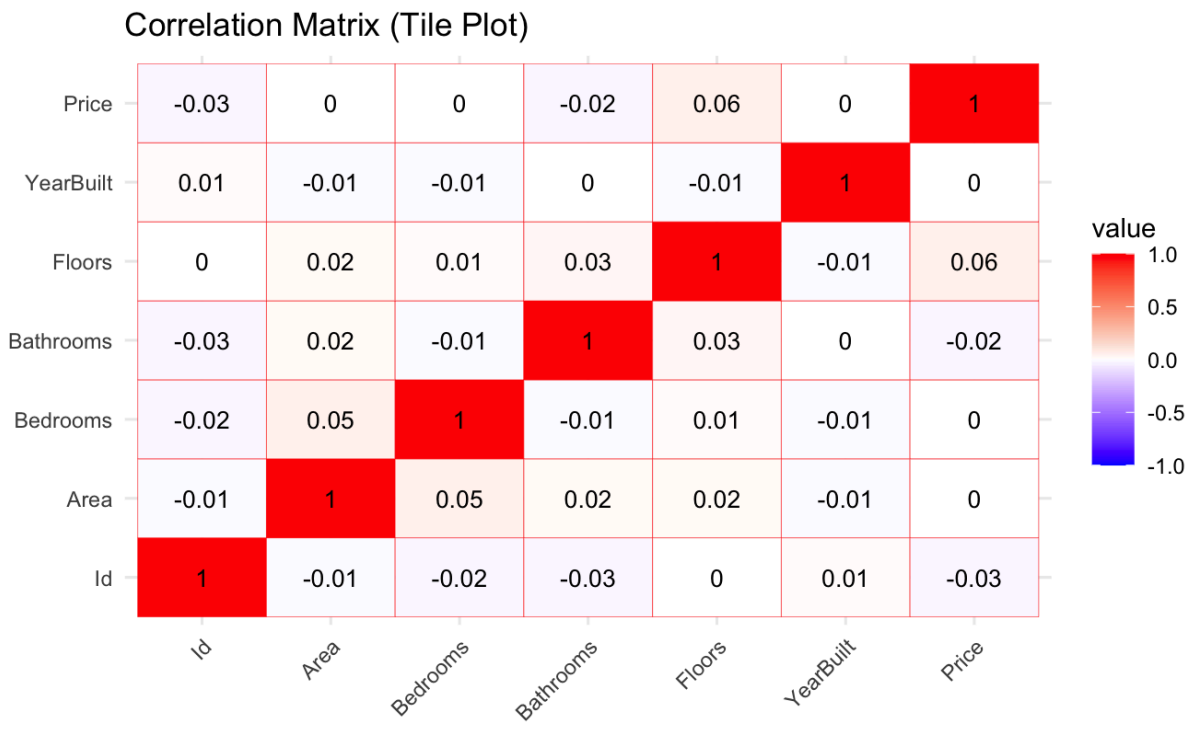
The top 10 most expensive houses price table show some interesting patterns. The living areas range from 566 to 4,340 square feet, and the most expensive house has an area of 3,099 square feet , indicating that larger size does not always correlate with higher price. Most of these homes have 2 to 3 bathrooms and 3 floors, indicating that multi-level living is a common feature in high-end properties.

Interestingly, the most expensive home was built in 1997, is located in a Suburban area, and does not have a garage. This suggests that garage availability and

location alone may not always be strongly associated with higher prices, particularly in the luxury segment.

Correlation Matrix :

We examined the correlations among the numerical features in the dataset to identify potential linear relationships. The results are summarized in the correlation matrix shown below.

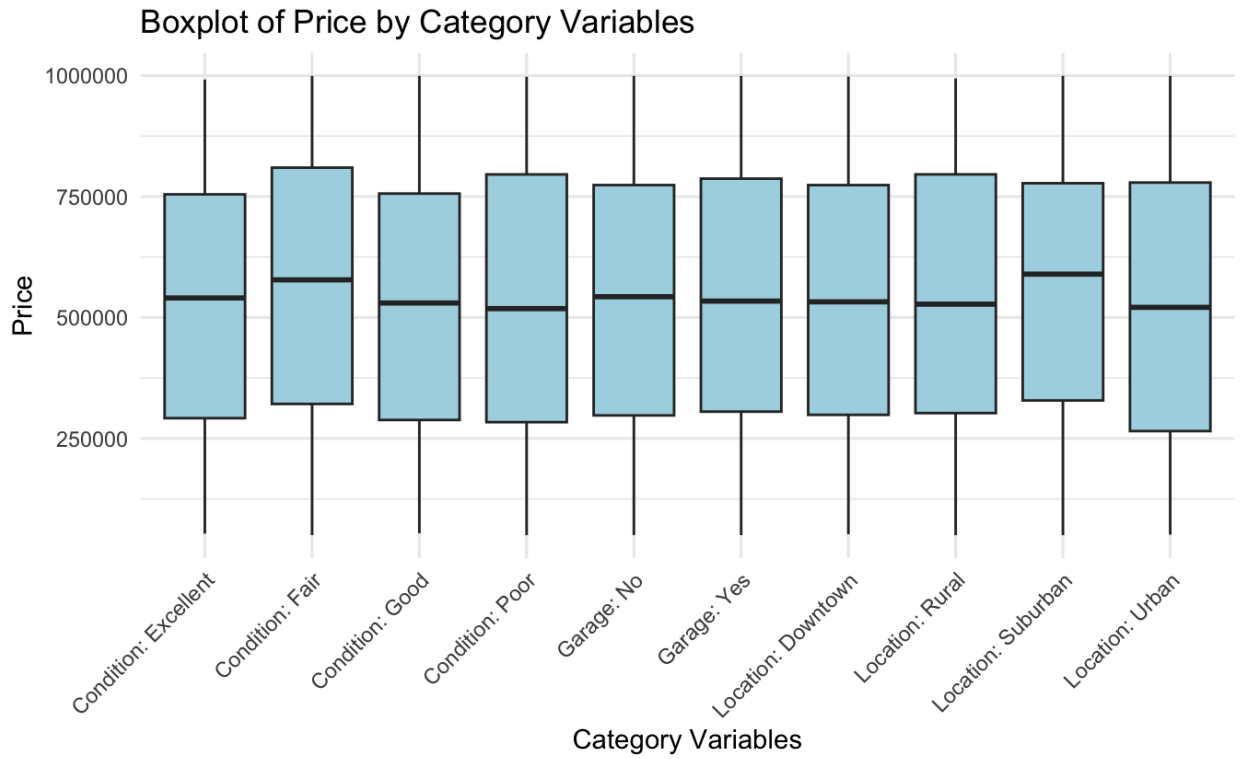


The correlation matrix (Tile Plot) shows that the most correlated numeric variable with Price is Floors, with a correlation of +0.06. While this suggests a very weak positive trend , indicating that homes with more floors may be slightly more expensive. All other features such as Area, Bedrooms, Bathrooms, YearBuilt are off-diagonal values close to zero, indicating very weak correlations among the variables. This suggests that the numerical variables are mostly independent of each other, and no strong linear correlation exists between any numeric

predictor and Price. These results support the idea that categorical variables such as Condition, Garage, and Location may be more important in predicting house price.

Boxplot Analysis :

We visualized the price distribution across categorical variables including Condition, Garage, and Location. The results are summarized in the box plot shown below.



The boxplot shows that homes with a garage tend to have slightly higher prices, but there is significant overlap with those without a garage. For Condition, there is minimal variation in price across the categories “Excellent,” “Good,” “Fair,” and “Poor,” indicating that condition alone does not have a strong impact on price. Location shows a slightly higher median price for Downtown and Suburban areas, but the distributions still overlap considerably. In conclusion, while Garage, Condition, and Location each have some influence on price, none of

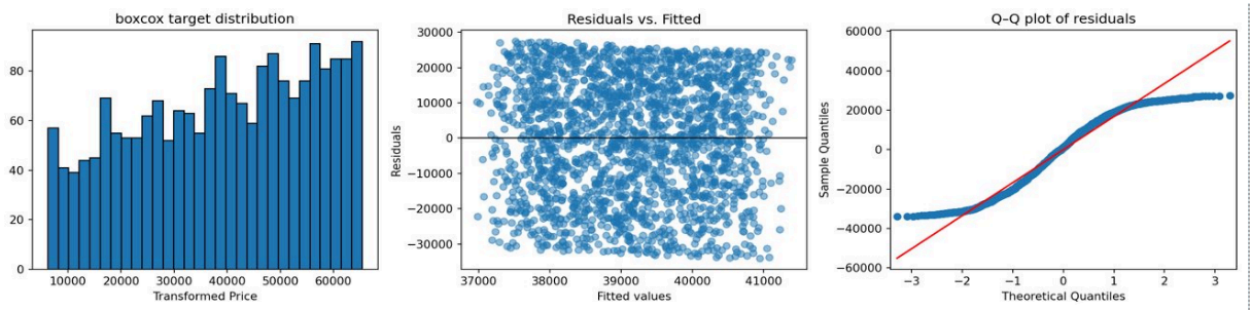
them demonstrate a clear or dominant separation, indicating that no single categorical variable independently drives house price in this dataset.

PART II

TRANSFORMATION :

The residual analysis on raw predictors didn't yield satisfying results so we have to approach methods of transformations for these predictors to see if any pattern emerges in the residuals and possibly if R^2 improves by doing these transformations.

Box–Cox TRANSFORMATION:

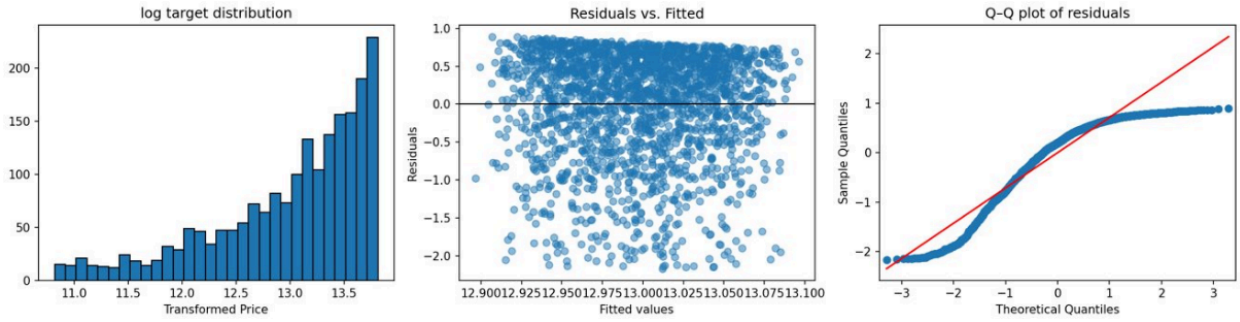



```
=== Transformation: boxcox ===
  • Box-Cox λ = 0.7851
  • R² = 0.0037
  • RMSE = 16752.3373
  • Coefficients:
      Area        : -0.0078
      Bedrooms    : -28.2439
      Bathrooms   : -265.2050
      Floors      :  1164.8482
      YearBuilt   :  1.9714
      Location    : -201.2442
      Condition   : -115.1329
      Garage      :  167.9504
      Intercept: 34433.5186
```

The Box–Cox transform (lambda ≈ 0.785) yielded a more symmetric but still non normal target distribution, and fitting a linear model on the standardized

features produced low predictive performance (R^2≈0.0037) with a high RMSE (~16752), indicating a poor overall fit for the dataset. Residuals vs fitted plot shows no obvious pattern or heteroscedasticity but the Q–Q plot has heavy tails on both ends, confirming deviations from normality. Coefficient estimates on the transformed scale are small for Area (–0.0078), YearBuilt (+1.97), and Garage (167.95), somewhat negative for Bedrooms (–28.24), Bathrooms (–265.21), Location (–201.24), Condition (–115.13), and a positive effect of Floors (1164.85), while intercept sits at around 34433. In summary, the Box–Cox adjustment improved symmetry but didn't enhance the model performance or normalize the residuals to a sufficient point.

LOGARITHMIC TRANSFORMATION:

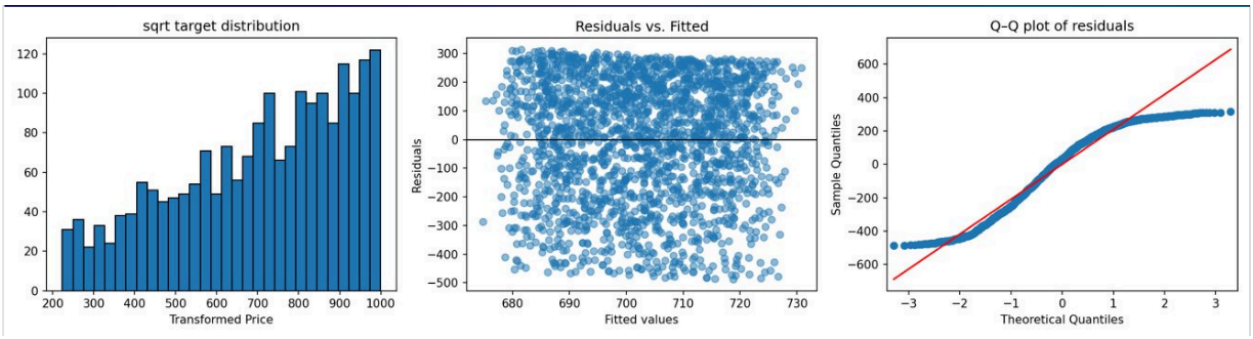



```
=== Transformation: log ===
  • R² = 0.0034
  • RMSE = 0.7126
  • Coefficients:
      Area        : -0.0000
      Bedrooms    :  0.0026
      Bathrooms   : -0.0107
      Floors      :  0.0449
      YearBuilt   :  0.0001
      Location    : -0.0110
      Condition   : -0.0008
      Garage      :  0.0229
      Intercept: 12.8270
```

The log-transform made it more noticable for the right-skew in the target, with the transformed prices still heavily skewed and cluttered at higher values.Linear model on standardized predictors achieved an R^2≈0.0034 and RMSE≈0.7126, showing almost no predictive power. The residuals vs fitted plot reveals increasing heteroscedasticity as fitted values rise and the Q–Q plot shows an S

shaped curve with heavy tails on both ends confirming substantial spread from normality. Coefficient estimates remain very small (Area ≈ –0.0, Bathrooms ≈ –0.0107, Floors ≈ 0.0449, Garage ≈ 0.0229) and the intercept sits at around 12.83. Despite reducing the scale, log-transform didn't do much to improve model performance

SQUARE ROOT TRANSFORMATION:




```
=== Transformation: sqrt ===
  • R² = 0.0036
  • RMSE = 209.2881
  • Coefficients:
      Area        : -0.0005
      Bedrooms    :  0.0335
      Bathrooms   : -3.2508
      Floors      :  14.3193
      YearBuilt   :  0.0216
      Location    : -2.8027
      Condition   : -0.9564
      Garage      :  3.4605
      Intercept: 648.7487
```

The square root transform modestly reduced the skew and resulted in a slightly right leaning but more balanced target distribution also a R^2≈0.0036 with RMSE≈209.29, still indicating minimal predictive power and hardly better than previous transformations. Residuals vs fitted plot is randomly scattered without clear patterns, while Q–Q plot's heavy tails on both ends form a slight S curve, showing better normality than log transform but still not sufficient. On transformed scale, coefficients are small: Area (−0.0005), Bedrooms (0.0335), Bathrooms (−3.2508), Floors (14.3193), YearBuilt (0.0216), Location (−2.8027), Condition (−0.9564), Garage (3.4605), with an intercept of 648.75. Overall, square

root transformation improved symmetry more than the log, but did not practically enhance model performance.

PART III

MODEL IMPROVEMENT TECHNIQUES

Focus:
Improving and evaluating a multiple linear regression model to predict house prices.

Method:
Using a cleaned version of the dataset and applying transformations and evaluation methods to improve accuracy and interpretability.

AIM:
To build a reliable model, asses prediction performance and propose future improvements

MODEL OVERVIEW

• Multiple linear regression model
• statistical method used to explain the relationship between one dependent variable (response) and several independent variables (predictors)
• Our dependent variable was the log-transformed house price
• helps normalize the data making the model assumptions more valid

MODEL DETAILS

• Response variable: LogPrice (natural logarithm of the Price column)
• Predictor variables: All other numeric columns, such as:
• bedrooms
• bathrooms
• sqft_living
• saft_lot
• floors

• year_built, etc.

```
##
## Call:
## lm(formula = LogPrice ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98155 -0.13441  0.06253  0.19874  0.27007
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.167e+01  3.325e-01  35.088   <2e-16 ***
## Id          -1.092e-05  1.054e-05  -1.037    0.300
## Area        -4.341e-06  4.677e-06  -0.928    0.353
## Bedrooms     6.183e-03  4.241e-03   1.458    0.145
## Bathrooms   -2.465e-04  5.457e-03  -0.045    0.964
## Floors       1.553e-03  7.484e-03   0.208    0.836
## YearBuilt    4.938e-06  1.687e-04   0.029    0.977
## Price        2.456e-06  2.193e-08 111.980   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2424 on 1592 degrees of freedom
```



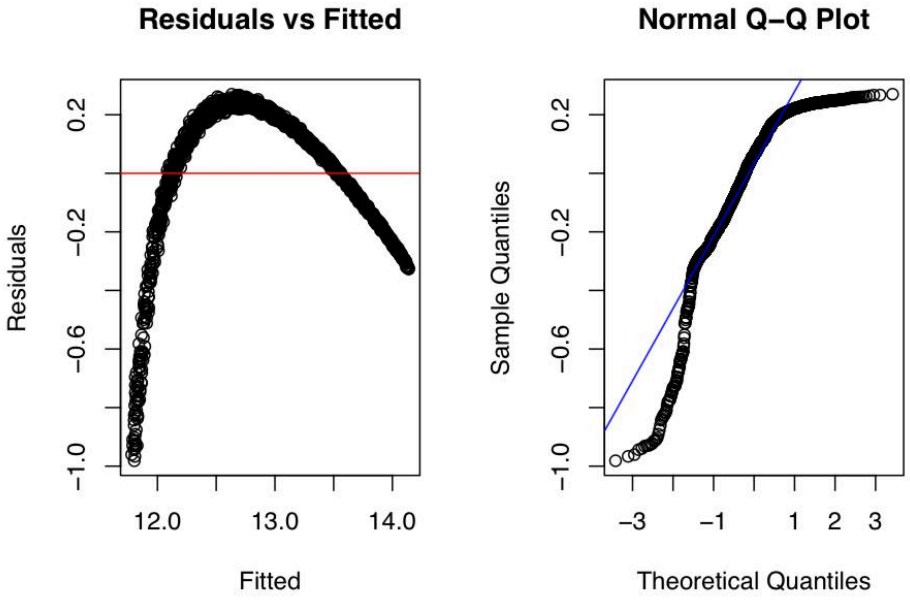
```
## Model Evaluation:
cat("RMSE:", round(rmse, 3), "\n")

## RMSE: 0.221
cat("MAE:", round(mae, 3), "\n")

## MAE: 0.178
cat("R-squared:", round(r2, 4), "\n")

## R-squared: 0.8971
```

We tested the model on unseen test data and evaluated its prediction performance using:

Root Mean Squared Error (RMSE):
Measures the average magnitude of the prediction errors.

Mean Absolute Error (MAE):
Measures average absolute difference between predicted and actual values.

R-squared (R'):
Measures how much of the variation in house prices is explained by the model.



DIAGNOSTIC ANALYSIS

To check whether our model met the assumptions of linear regression, we created diagnostic plots:

RESIDUALS VS FITTED PLOT

The plot showed that residuals were randomly scattered around 0, indicating a good fit and no strong patterns (which would violate linearity or constant variance assumptions).

Q-Q PLOT

• Plotted quantiles of residuals against a normal distribution.
• Residuals mostly followed the line, supporting the assumption of normally distributed errors.
Multicollinearity Check
• We computed a correlation heatmap and VIF (Variance Inflation Factor) to detect multicollinearity:
• High correlations between predictors (e.g., sqft_living and saft_above) suggest redundancy.
• Variables with VIF > 5 or 10 were flagged as potentially problematic.

TO FURTHER ENHANCE THE MODEL, WE SUGGEST

• Feature Selection: Use stepwise regression or regularization to select the most informative variables.
• Nonlinear Modeling: Try Random Forests, Decision Trees, or Gradient Boosting if relationships are not linear.
• Data Enrichment: Add new variables such as neighborhood rating, school quality, or crime statistics.
• Standardization: Scale features like square footage and year built to improve numerical stability.

Appendix

Presentation Slides : Attached as Group15.pptx

Proposal Document : Attached as Group15-Final Proposal.pdf

R Code :

- Part I : Attached as PartI.rmd.
- Part II : Attached as Code.R
- Part III : Attached as PartIII.pdf

Team Roles:

- Natthiya Sae Ngow - Proposal , Part I , Presentation and Final Report
- Elshad Jafarli - Proposal , Part II , Presentation and Final Report
- Pamela Kamdefwere - Proposal , Part III , Presentation and Final Report

Reference :

- House Price Prediction : https://www.kaggle.com/datasets/zafarali27/house-price-prediction-dataset/data