Investigating Factors That Influence Housing Prices Using Multiple Linear Regression

Temitope Oladimeji – x23187204

National College of Ireland

MSc in Data Analytics

Statistics for Data Analytics

CA – Semester 1 2023/2024

Abstract — This study investigates variables influencing a target variable through a comprehensive analysis of a set of predictor variables. The model was built based on the OLS (ordinary least squares) estimates and its predictive capabilities were evaluated using a selected portion of the dataset. The final linear regression model with influential predictors were investigated based on the significance of its β-value associated with the t-test. Other metrics were evaluated to capture a more accurate model while ensuring the Gauss-Markov assumptions were met.

1. INTRODUCTION

In regression analysis, the aim is often to identify explanatory variables related to a response variable. The simplest regression model assumes a linear relationship between a single predictor variable and the target variable [1]. This can be described by the following formula:

$$Y = \beta o + \beta 1X1 + \varepsilon$$

where $\beta o + \beta 1$ represent the intercept and slope terms in the linear model. These parameters are usually called the regression coefficients [1]. The goal is to select the model parameters such that the difference between actual response values and predicted values are minimized.

Multiple linear regression uses more than one predictor variable to estimate the value of the target variable. The x-variable co-efficient is the amount by which the target variable (y) changes if the x variable increases by on, and the values of all other x variables in the model remain unchanged. This can be described by the following formula:

$$Y = \beta o + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \varepsilon$$

An MLR model can be implemented by investigators to predict a person's height from skeletal remains [2].

This paper details the steps followed to build a linear regression model that analyses predictor variables that have a major influence on the target variable based on the interpretation of its parameters.

2. METHODOLOGY

2.1. Overview

The KDD (knowledge discovery in database) methodology is regarded as an automatic exploratory data analysis of large databases [3]. It is an iterative process that is less business driven than its counterpart CRISP-DM. One key aspect of KDD is its division of processes, each with its own benefits and drawbacks [3].

i. Data Selection – data or a subset of data that will be used for analysis is selected.

ii. Data Pre-processing – encompasses data cleaning, data transformation and feature selection.

iii. Data Mining – most suitable DM techniques are applied to the data to discover patterns and relationships.

iv. Evaluation – the quality of the model is assessed and its patterns are interpreted.

v. Result Exploitation – through visualisation tools, the discovered knowledge is reported [2].

2.2. Data Understanding

The dataset has 18 variables, consisting of 2413 observations. Table 1 shows a summary of the variables in the dataset:

	Variable name	Type
1	Lot_Frontage	int
2	Lot_Area	int
3	Bldg_Type	factor
4	House_Style	factor
5	Overall_Cond	Factor
6	Year_Built	int
7	Exter_Cond	factor
8	Total_Bsmt_SF	int

9	First_Flr_SF	int
10	Second_Flr_SF	int
11	Full_Bath	int
12	Half_Bath	int
13	Bedroom_AbvGr	int
14	Kitchen_AbvGr	int
15	Fireplaces	int
16	Longitude	num
17	Latitude	num
18	Sale_Price	int

Table 1: Dataset variables

2.3. Data Pre-processing

After understanding the variables in the dataset, it is worth pre-processing to effectively capture any underlying patterns, and implement domain knowledge to analyse which predictor variables are worth evaluating the model with. The pre-processing stage is generally to ensure the data is fit enough to satisfy its intended use.

The pre-processing stage initially involves visualisations of distributions to further review the data prior to any cleaning. Figure 1 shows the distribution of the first four numerical variables:

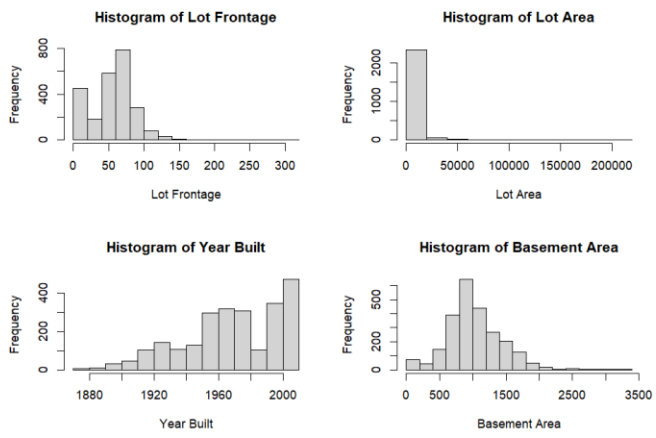

Figure 1: Visual distribution of numerical variables

It is worth noting that categorical variables should also be visualized, however its plot usually differs to that of a numerical variable, as shown in figure 2:

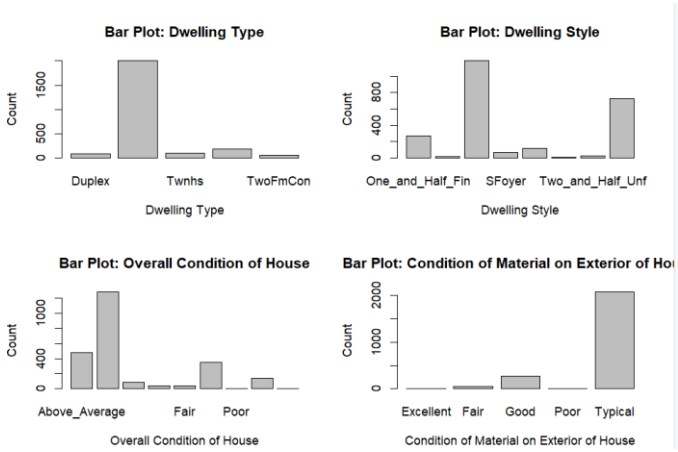

Figure 2:Visual distribution of categorical variables

2.3.1. Data Cleaning

Data cleaning is a part of the pre-processing stage that involves handling missing values, identifying & removing outliers, and resolving inconsistencies. The handling of missing values is especially important, as ignoring them can lead to bias and complications in analysing the data [2]. Imputation is commonly used for quantitative data [2], and has the benefit of preserving the sample size instead of discarding the columns or rows with missing values.

In this study, the imputation was done after encoding the categorical variables, subsequently binding new "child" columns to the existing dataset, and dropping their parent variables. Dropping them ensured the data was reduced, and only important variables were used for further analysis. The dataset only had one variable which had missing values, and this was handled by imputing with the mean.

Noise removal is another step to be considered while cleaning the data. This step involves handling outliers that are likely to be noise. It is important to remove noise because it affects the interpretations of the data, models built and the evaluations [2]. Figure 3 below shows an example of a box plot, with the outliers being represented as small circular points:

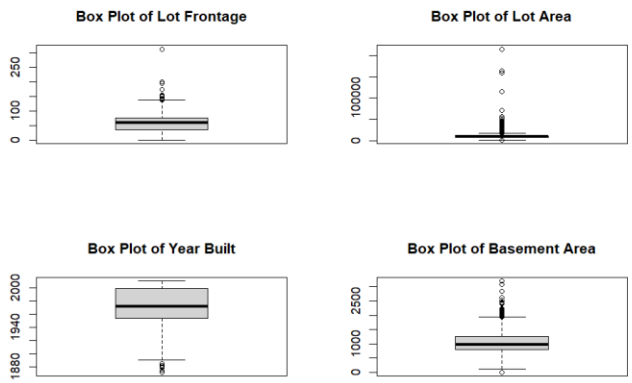

Figure 3: Box Plot (Before IQR test)

Statistical analysis was used to remove the outliers. There are various tests such as z-scores and IQR, but the latter was used as it is useful for non-normal distribution and is a fundamental component of box-and-whisker plots, providing a visual representation of the data spread, as seen in figure 3.

The box-and-whisker plot helps identify outliers easily. For the IQR (inter-quartile range) test, the upper and lower bound have to be specified and any data point that falls beyond these bounds would be detected as outliers, and removed.

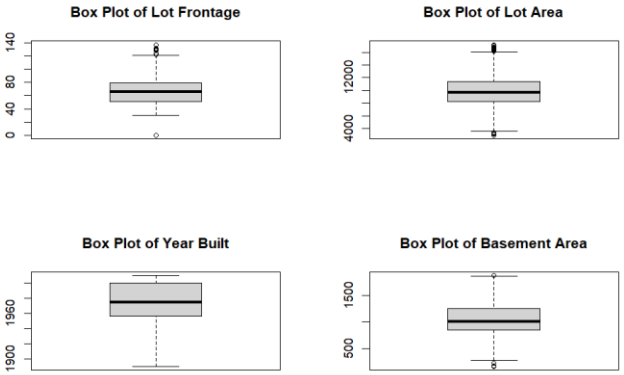

Figure 4: Box Plot (After IQR test)

2.3.2. Data Transformation

Data Transformation converts raw data into a suitable format for analysis and modelling. It is especially useful where the values are vastly different in scale [4].

Normalisation scales the data to a specific range e.g. (0 to 1). Standardisation gives the data a mean of 0 and a standard deviation of 1. Min-max normalization (linear transformation of original data) is one of the ways of normalizing data, which is used in this analysis. Other methods include: z-score normalization, decimal normalization, and unit vector normalization.

```
Total_Bsmt_SF   First_Flr_SF    Second_Flr_SF
Min.   : 160   Min.   : 432.0   Min.   :   0.0
1st Qu.: 847   1st Qu.: 887.5   1st Qu.:   0.0
Median :1008   Median :1064.0   Median :   0.0
Mean   :1063   Mean   :1122.8   Mean   : 356.1
3rd Qu.:1256   3rd Qu.:1335.0   3rd Qu.: 768.0
Max.   :1884   Max.   :1929.0   Max.   :1721.0
```

Figure 5: Before Min-max normalization

```
Total_Bsmt_SF     First_Flr_SF     Second_Flr_SF
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.3985   1st Qu.:0.3043   1st Qu.:0.0000
Median :0.4919   Median :0.4222   Median :0.0000
Mean   :0.5240   Mean   :0.4614   Mean   :0.2069
3rd Qu.:0.6357   3rd Qu.:0.6032   3rd Qu.:0.4463
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

Figure 6: After Min-max normalization

2.4. Data Modelling

Multiple linear regression was the model used to build the model. The goal is to find the line that best fits a set of data points – a approach based on the least squares method which estimates the parameters of a linear regression model [5], by minimizing the sum of the squared residuals.

To avoid biased ordinary least squares (OLS) estimators, a set of assumptions must be met. These assumptions are known as the Gauss-Markov assumptions, and adhering to them ensures that OLS estimators are BLUE (best linear unbiased estimators):

i. The errors should be independent of each other.
ii. Homoscedasticity – variance of errors should be constant across all levels of the independent variables.
iii. Errors should be normally distributed.
iv. Absence of multicollinearity between independent variables.
v. No autocorrelation between the errors.
vi. No influential data points.
vii. Zero mean of residuals.

Assumption	**Test**	**P-Value**	**Conclusion**
Independence of Errors	Durbin-Watson Test	$p1$	$p1 \geq 0.05$ (Acceptance of independence)

Assumption	Test	P-Value	Conclusion
Homoscedasticity	Breusch-Pagan Test (or others)	$p2$	$p2 \geq 0.05$ (Acceptance of homoscedasticity)
Normality of Errors	Shapiro-Wilk Test (or others)	$p3$	$p3 \geq 0.05$ (Acceptance of normality)
Absence of Multicollinearity	Variance Inflation Factor (VIF)	$p4$	$p4 \geq 0.05$ (Acceptance of no multicollinearity)
No Autocorrelation	Ljung-Box Test (or others)	$p5$	$p5 \geq 0.05$ (Acceptance of no autocorrelation)
No Influential Data Points	Cook's Distance, Influence Plot	-	Visual inspection
Zero Mean of Residuals	Mean of Residuals	-	Should be approximately zero

Table 1: Gauss-Markov Assumptions

The assessment of regression-based models are usually done by investigating if every assumption has been met, and if quantitative metrics such as the standard error and adjusted R^2 value is appropriate. Typically, a model would be chosen as the best if:

- Gauss-Markov assumptions are met
- It has a very low standard error
- It has a high adjusted R^2 value, and
- It performs well on unseen data

The performance of a model is usually evaluated on a different sample than the one used to develop either. This prevents bias when rewarding a model [6].

3. Evaluation

A fundamental practice in statistical modelling is splitting the data into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate the performance of the model – that is, it assesses how well the model generalizes to new, unseen data.

Another key factor in data splitting is identifying overfitting. Overfitting occurs when a model performs too well on the training data but poorly on the testing data. By evaluating the model on a testing set, a more accurate estimate of its performance is observed, and it results in an unbiased analysis of different models.

In this study, the dataset (post-processing) had 1295 observations of 16 variables. A 70:30 split ensured the training data had 906 observations and the testing data had 389 observations.

4. Model Analysis

4.1. Model 1

The first model built used the relevant independent variables. These predictor variables were carefully chosen using domain knowledge and statistical analysis such as variables with sufficient range and a feasible correlation with the dependent variable.

Figure 5 below shows the necessary plots for the model:

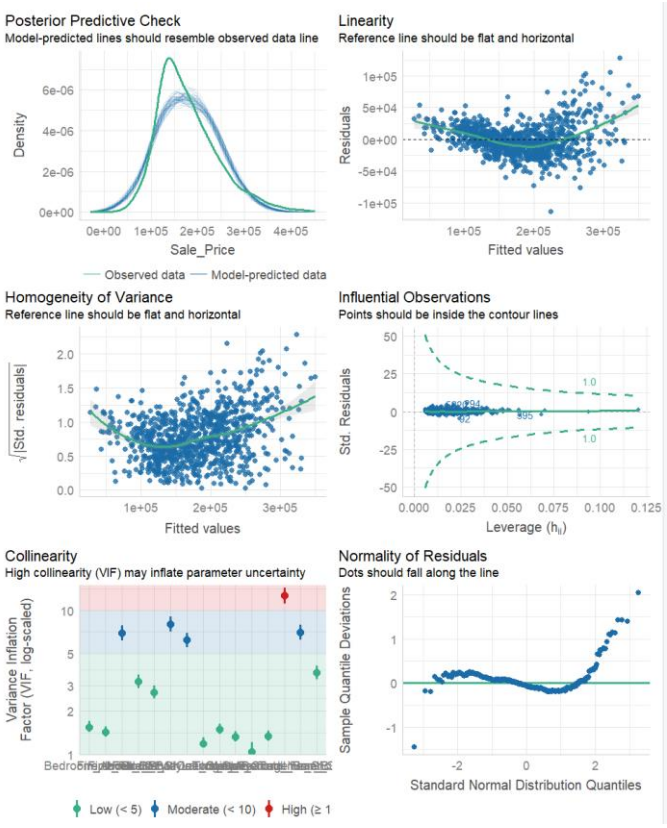

Figure 5: Model 1 Visualization

The plot gives a summary of how the model was distributed in line with the assumptions. Table 2 gives a more detailed statistical analysis of the model:

Test	Value	Interpretation
Adjusted R-squared	0.8497	The proportion of variance in the DV explained by the model, adjusted for the number of predictors.
Residual Error	24880	
F-test p-value	< 2.2e-16	P-value associated with the F-statistic. Small p-value indicates overall model significance.
Durbin-Watson Test	0.404	Tests for autocorrelation in the residuals. P-value > 0.05 suggests no significant autocorrelation.
NCV Test	Chi-square = 224.7722, Df = 1, p = < 2.22e-16	Tests for homoscedasticity. Small p-value suggests non-constant variance.

Table 2: Model 1 Summary Stats

From the table and plot above, the first model evidently violated multiple tests including the homoscedasticity test indicating that the variance is non-constant, normality of residuals, and linearity. One method of addressing the heteroscedasticity issue is to transform the target variable '*Sale_Price*'. This was done for Model 2.

The test for multicollinearity was also violated, as figure 6 proves.

```
Low Correlation

                Term  VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
         Lot_Frontage 1.05 [ 1.01,  1.21]         1.02      0.96     [0.83, 0.99]
             Lot_Area 1.32 [ 1.23,  1.44]         1.15      0.76     [0.69, 0.81]
           Year_Built 3.70 [ 3.32,  4.14]         1.92      0.27     [0.24, 0.30]
            Full_Bath 3.20 [ 2.88,  3.58]         1.79      0.31     [0.28, 0.35]
            Half_Bath 2.70 [ 2.44,  3.00]         1.64      0.37     [0.33, 0.41]
        Bedroom_AbvGr 1.55 [ 1.43,  1.70]         1.24      0.65     [0.59, 0.70]
           Fireplaces 1.43 [ 1.32,  1.56]         1.19      0.70     [0.64, 0.76]
            Longitude 1.49 [ 1.38,  1.63]         1.22      0.67     [0.61, 0.73]
             Latitude 1.19 [ 1.12,  1.30]         1.09      0.84     [0.77, 0.89]
Overall_Cond_numeric 1.34 [ 1.25,  1.47]         1.16      0.75     [0.68, 0.80]

Moderate Correlation

                Term  VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
       Total_Bsmt_SF 7.04 [ 6.26,  7.93]         2.65      0.14     [0.13, 0.16]
        First_Flr_SF 6.94 [ 6.17,  7.82]         2.63      0.14     [0.13, 0.16]
House_StyleOne_Story 8.04 [ 7.15,  9.08]         2.84      0.12     [0.11, 0.14]
House_StyleTwo_Story 6.21 [ 5.53,  6.99]         2.49      0.16     [0.14, 0.18]

High Correlation

         Term   VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
Second_Flr_SF 12.62 [11.17, 14.28]         3.55      0.08     [0.07, 0.09]
```

Figure 6: Model 1 – Multicollinearity test

4.2. Model 2

To address heteroscedasticity, the target variable was log-transformed. Square root transformation is another potential solution. The independent variables remained unchanged so a violation of the multicollinearity test would still be expected.

Test	Value	Interpretation
Adjusted R-squared	0.89	The proportion of variance in the DV explained by the model, adjusted for the number of predictors.
Residual Error	0.1167	
F-test p-value	< 2.2e-16	P-value associated with the F-statistic. Small p-value indicates overall model significance.
Durbin-Watson Test	0.65	Tests for autocorrelation in the residuals. P-value > 0.05 suggests no significant autocorrelation.
NCV Test	Chi-square = 5.257754, Df = 1, p = 0.021849	Tests for homoscedasticity. Small p-value suggests non-constant variance.

Table 3: Model 2 Summary Stats

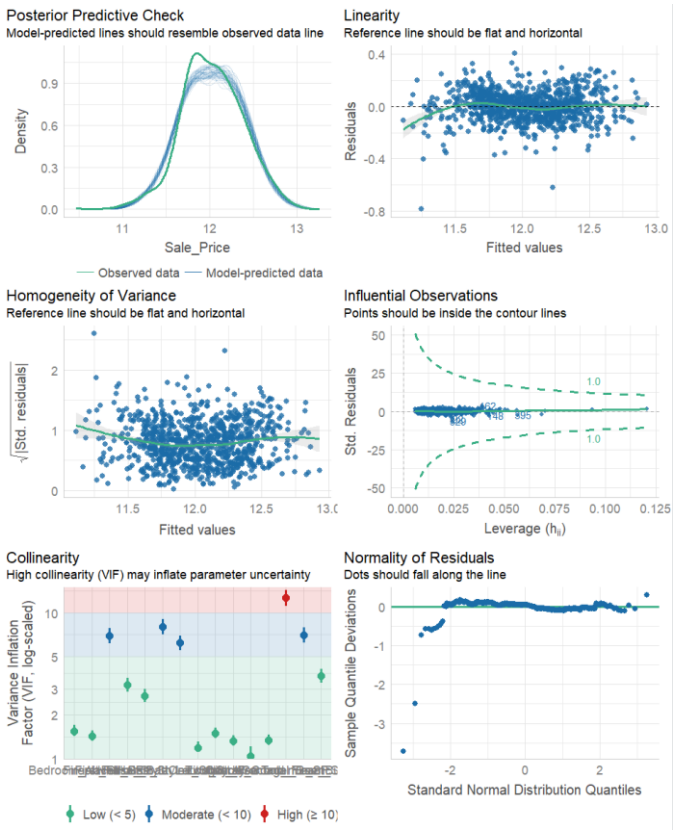

Figure 7: Model 2 Visualization

Log-transforming the target variable appeared to have a slight positive effect on the variance distribution, however the small p-value from the test still suggests a violation is present.

The model also had far lower residual errors than the first model indicating a positive progress. This is also backed up by an increase in the R^2, but correlation was still present.

Evaluating the collinearity test suggested variables with a low VIF value are statistically significant. Model 3 addresses this issue by dropping the variables with VIF > 5, while still keeping the log-transformation of the target variable.

4.3. Model 3

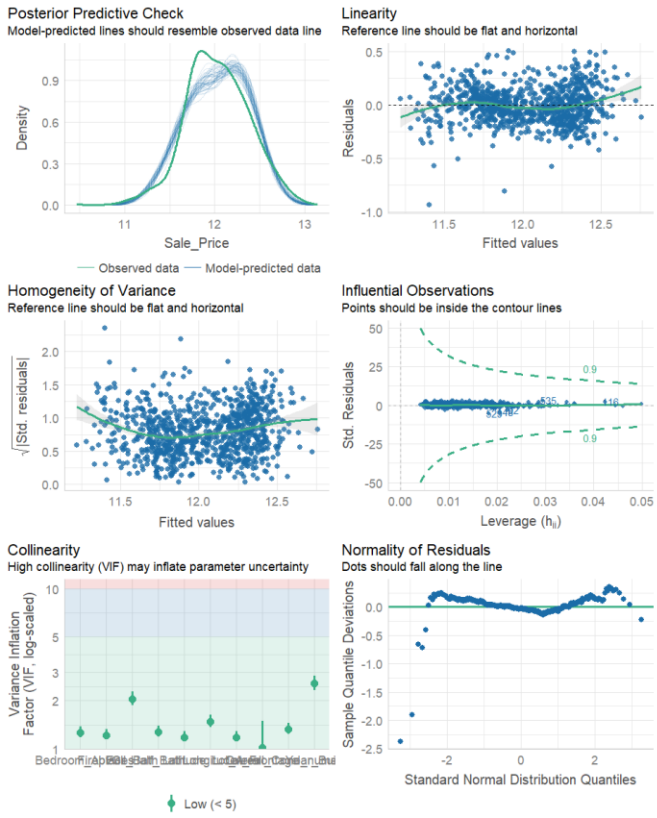

Figure 8: Model 3 Visualization

The third model was built with the following variables which had low VIF.

Low Correlation

Term	VIF	VIF 95% CI	Increased SE	Tolerance	Tolerance 95% CI
Lot_Frontage	1.02	[1.00, 1.49]	1.01	0.98	[0.67, 1.00]
Lot_Area	1.18	[1.11, 1.29]	1.08	0.85	[0.78, 0.90]
Year_Built	2.58	[2.33, 2.87]	1.61	0.39	[0.35, 0.43]
Full_Bath	2.06	[1.87, 2.28]	1.43	0.49	[0.44, 0.53]
Half_Bath	1.27	[1.19, 1.39]	1.13	0.79	[0.72, 0.84]
Bedroom_AbvGr	1.26	[1.18, 1.38]	1.12	0.79	[0.73, 0.85]
Fireplaces	1.22	[1.14, 1.33]	1.10	0.82	[0.75, 0.87]
Longitude	1.48	[1.37, 1.63]	1.22	0.67	[0.61, 0.73]
Latitude	1.18	[1.11, 1.29]	1.08	0.85	[0.78, 0.90]
Overall_Cond_numeric	1.33	[1.24, 1.45]	1.15	0.75	[0.69, 0.81]

Figure 9: Model 3 – Multicollinearity test

Test	Value	Interpretation
Adjusted R-squared	0.7645	The proportion of variance in the DV explained by the model, adjusted for the number of predictors.
Residual Error	0.1708	
F-test p-value	< 2.2e-16	P-value associated with the F-statistic. Small p-value indicates overall model significance.
Durbin-Watson Test	0.764	Tests for autocorrelation in the residuals. P-value > 0.05 suggests no significant autocorrelation.
NCV Test	Chi-square = 0.2482488, Df = 1, p = 0.61831	Tests for homoscedasticity. Error variance appears to be homoscedastic.

Table 4: Model 3 Summary Stats

As seen from the table, the final model met all the linear regression assumptions. Although its linearity and variance homogeneity both seem skewed a bit more than those of the second model, its validity is justified by meeting all the criteria. Also, the variance explained 76.45% of the model, which would still be considered a good model.

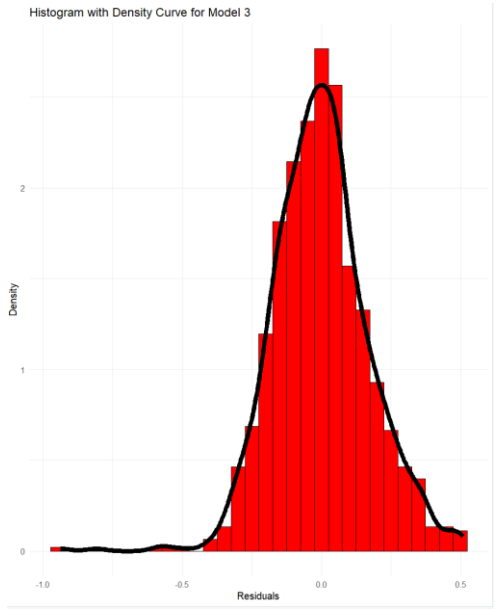

Figure 10: Model 3 – Distribution of Errors (Residual Plot)

5. Model Evaluation

To evaluate the predictive capability of the third model, it was assessed on the 30% test dataset. The predicted and actual values were then plotted against each other via a scatter plot, where linearity suggests the model captures underlying linear relationships in the data, indicating a good fit. Figure 11 shows a visual representation of this alignment.

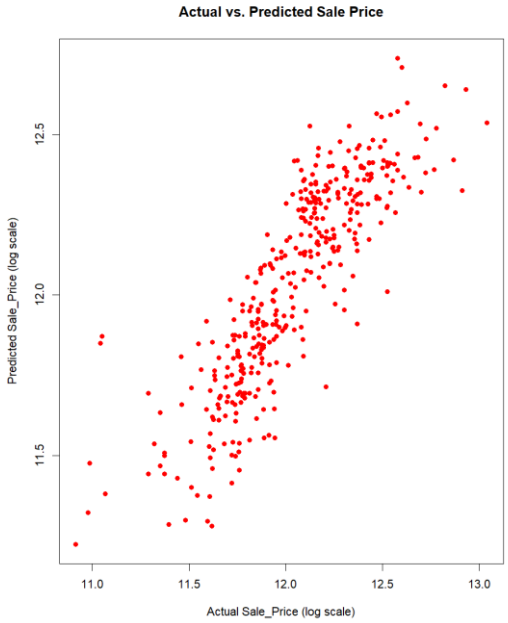

Figure 11: Predictive capability of model 3

5.1. Cross-Validation

To complement this visual inspection, a 10-fold cross-validation was conducted using the model.

The model's performance was assessed, revealing a root mean squared error (RMSE) of 0.1172726, an R^2 value of 0.8914042, and a mean absolute error (MAE) of 0.08929857. The low values of the RMSE and MAE, along with a high R^2 value, indicates that the model has good predictive accuracy and explains a significant proportion of the variance in the target variable.

Cross-validation was performed to further assess the model's performance on a random, unseen, split of data. A high r-squared value suggests that the model is good at making predictions, and implies that the chosen predictors provide meaningful information about the target variable, which is the purpose of this analysis.

6. Conclusion

In conclusion, this analysis has provided valuable insights into the factors influencing housing prices. The pre-processing approach, involving exploratory data analysis, data cleaning, transformation, and model building, has offered a detailed understanding of the dataset. The final linear regression model, validated using diagnostic tests, and with a 76.5% predictive accuracy, shows a reasonable fit to the data. The Gauss-Markov assumptions have also been tested and met, adding to the model's reliability.

Improvements can be made on the model. Firstly, feature engineering could be implemented to better select relevant predictors that could help in capturing the model's linearity. This could be done via correlation matrix to assess the strength of relationships between independent variables and the target variable. It also helps in identifying highly correlated variables which is one of the assumptions to be met.

As KDD is an iterative process, it is also worth considering reassessing feature importance and continually refine the model to ensure the model's relevance and improve on its predictive accuracy.

7. References

[1] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis* (Springer Series in Statistics). Springer New York, 2013.

[2] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis* (Wiley Series in Probability and Statistics). Wiley, 2013.

[3] S. García, J. Luengo, and F. Herrera, *Data Pre-processing in Data Mining* (Intelligent Systems Reference Library). Springer International Publishing, 2014.

[4] D. J. Olive, *Linear Regression*. Springer International Publishing, 2017.

[5] X. Yan and X. Su, *Linear Regression Analysis: Theory And Computing*. World Scientific Publishing Company, 2009.

[6] M. J. Zaki and W. Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.