Predict Sale Price Using Multiple Linear Regression (November 2023)

Abstract—**This paper delves into the intriguing world of real estate, offering a fresh perspective on predicting the housing sale prices. Multiple linear regression has been used on the dataset to unravel the complexities of the housing market. Accuracy in the sale price prediction of a house is paramount for initiating buying and selling transactions, and valuation of property, and provides valuable information to investors and owners. An in-depth data exploration which involves exploring variables, different patterns and correlations between various variables is done. The final model indicates that only some necessary features influence the sale price of the house and necessary evaluation methods have been used to explain the efficiency and robustness of the model in predicting housing sale price.**

Keywords—House Sale Price Prediction, Multiple Linear Regression, Accuracy

I. INTRODUCTION

The real estate industry is a fast-growing and profitable industry. Not just buildings, apartments or commercial spaces, house sales have always been popular within the real estate business. Houses are one of the basic necessities or needs according to renowned psychologist Abraham Maslow. House prices can be influenced by various factors like Total Square feet, location, overall condition and many other factors. Prediction of house sale prices helps real estate maximize profits and help homebuyers make decisions and plan their investments which in turn saves a lot of time and money [1]. A supervised learning algorithm called Multiple Linear Regression predicts house sale prices using various factors. Multiple Linear Regression consists of a single dependent variable and at least one independent variable [2]. The Multiple Linear Regression model is used for predicting house sale prices using various predictors, where the dependent variable must be continuous.

II. METHODOLOGY

A. Overview

The framework used in this data analysis using statistics is the CRISP-DM methodology. The CRISP-DM (Cross Industry Standard for Data Mining) [3] is a framework or model for data mining that contains six phases. The order of phases is not strict.

Business Understanding: The objectives and goals of the project are understood from a business perspective. A problem definition and initial plan are designed.

Data Understanding: This phase involves data gathering and activities to familiarize with the data takes place. The data quality is accessed and basic insights are drawn.

Data Preparation: This phase can be performed many times as it prepares the raw data into a final dataset that can be used for analysis. This phase includes activities like variable selections, cleaning of data, transformations of variables and creating new attributes.

Modelling: One or more modelling techniques are selected and implemented on the dataset that has been prepared and cleaned.

Evaluation: One or more models that have performed well will be evaluated and checked if it fulfils business objectives.

Deployment: Model creation is not the final step of the project. More often, the user carries out the deployment steps for use in the business.

The CRISP-DM framework was fully implemented except deployment phase.

B. Business Understanding

The main challenge is to forecast the selling price of a house. The dependent variable i.e., the sale price is being predicted based on the independent variables or factors.

C. Data Understanding

The first step in this phase is to analyze the dimensions of the data that has been made available by a real estate company. There are a total of 2413 rows and 18 columns. The dataset includes the following variables: -

Variable	Type	Unit Measure
1. Lot Frontage	Numerical Continuous	Feet
2. Lot Area	Numerical Continuous	Square Feet
3. Building Type	Categorical	No Unit Measure
4. House Style	Categorical	No Unit Measure
5. Overall Condition	Categorical Ordinal	No Unit Measure
6. Year Built	Numerical Discrete	Year
7. External Condition	Categorical Ordinal	No Unit Measure
8. Total Basement Square Feet	Numerical Continuous	Square Feet
9. First Floor Square Feet	Numerical Continuous	Square Feet
10. Second Floor Square Feet	Numerical Continuous	Square Feet
11. Full Bath	Numerical Discrete	Number of Bathrooms
12. Half Bath	Numerical Discrete	Number of Half Bathrooms

13. Bedroom Above Ground	Numerical Discrete	Number of Bedrooms above ground
14. Kitchen Above Ground	Numerical Discrete	Number of Kitchens above ground
15. Fireplaces	Numerical Discrete	Number of Fireplaces
16. Longitude	Numerical Continuous	Degrees
17. Latitude	Numerical Continuous	Degrees
18. Sale Price	Numerical Continuous	Currency E.g.: Dollars

Table 1: Variables

Descriptive statistics were calculated as shown below to understand the data better.

Variable	Mean	Median	Min	Max
1	55.46	60.00	0.00	313.00
2	10060	9360	1300	215245
6	1969	1971	1872	2010
8	1023	970	0	3206
9	1134	1060	334	3820
10	339.2	0	0	1872
11	1.539	2	0	4
12	0.378	0	0	2
13	2.855	3	0	6
14	1.04	1	0	3
15	0.603	1	0	4
16	-93.64	-93.64	-93.69	-93.58
17	42.03	42.03	41.99	42.06
18	175568	159000	35000	755000

Table 2: Descriptive Statistics

Descriptive statistics helps organize and summarize information from the dataset.

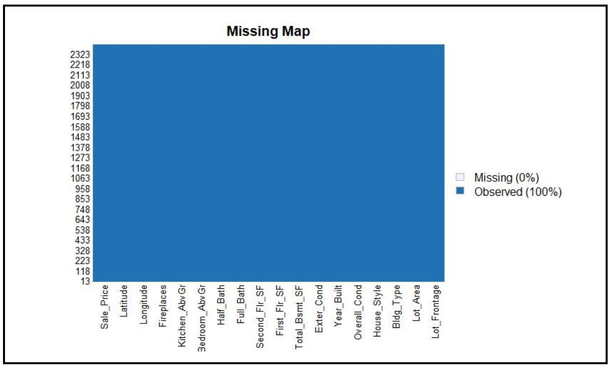


Figure 1: Amelia Missing map

Using Amelia package in RStudio, missing values if any are checked. There are no missing values.

Variable numbers #1, #2, #9, #10 and #18 are right skewed as seen in the histograms.

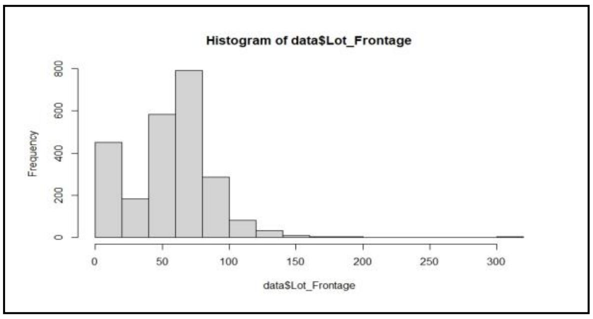


Figure 2: Histogram – Lot_Frontage

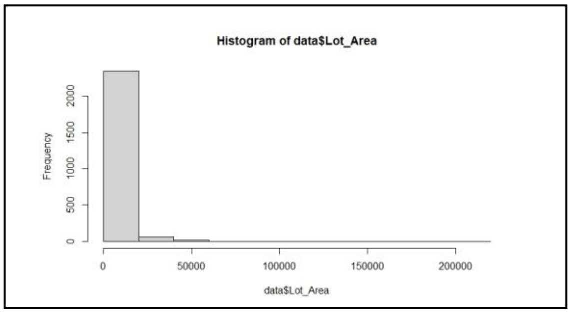


Figure 3: Histogram – Lot_Area

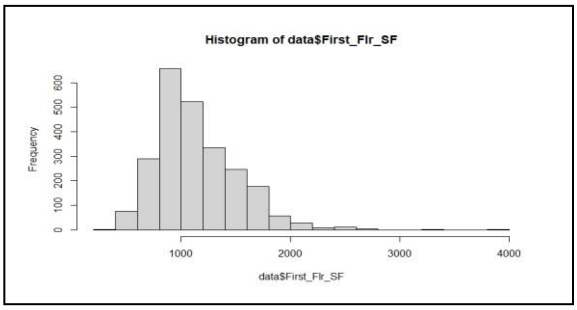


Figure 4: Histogram – First_Floor_SF

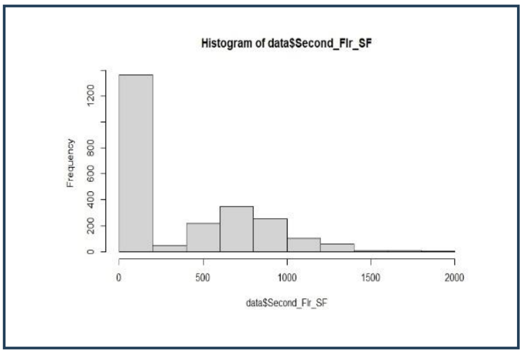


Figure 5: Histogram – Second_Floor_SF

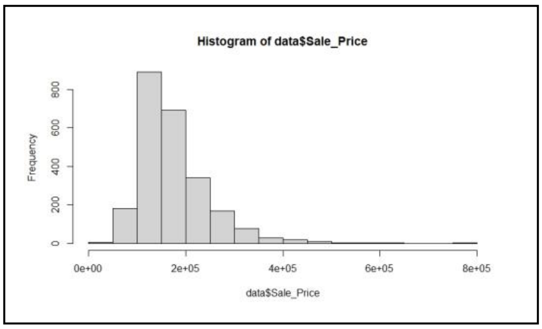


Figure 6: Histogram – Sale_Price

D. Data Preparation

Variables #1, #2, #9, #10, #18 are all right-skewed. So, these variables have been normalized using transformations. The main purpose of transformations is to make one the orders of magnitude and measurement units [4]. A map has also been plotted based on variables #16 and #17 with respect to variable #18 where variable #18 has been split into two labels i.e., Low and High based on the median value of variable #18.

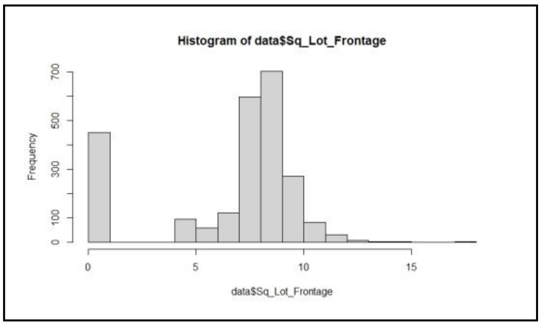


Figure 7: Histogram – Sq_Lot_Frontage

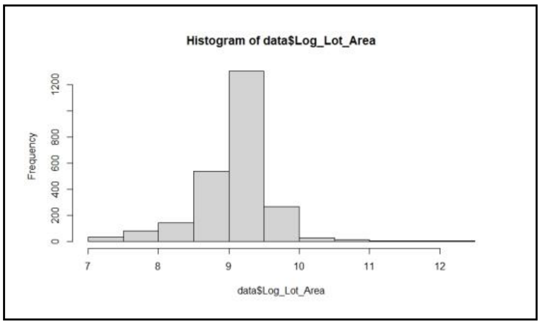


Figure 8: Histogram – Log_Lot_Area

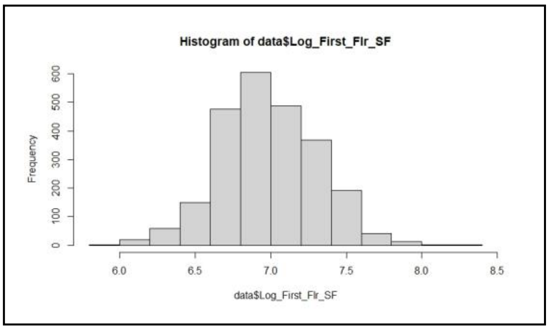


Figure 9: Histogram-Log_First_Flr_SF

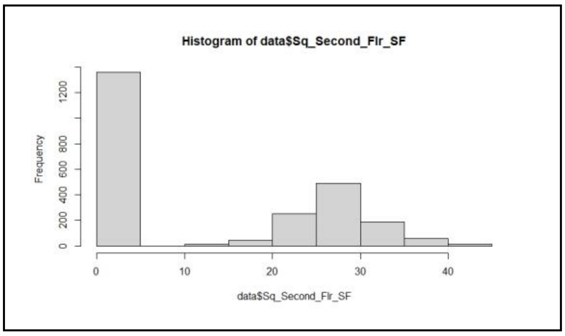


Figure 10: Histogram – Sq_Second_Flr_SF

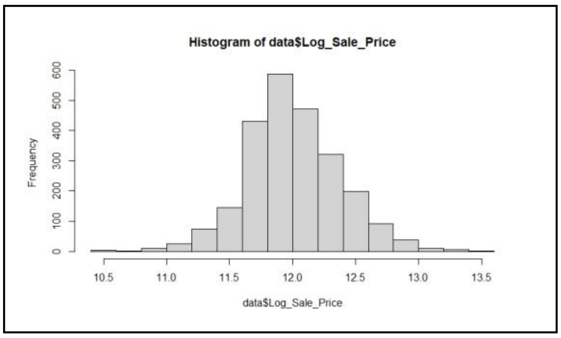


Figure 11: Histogram – Log_Sale_Price

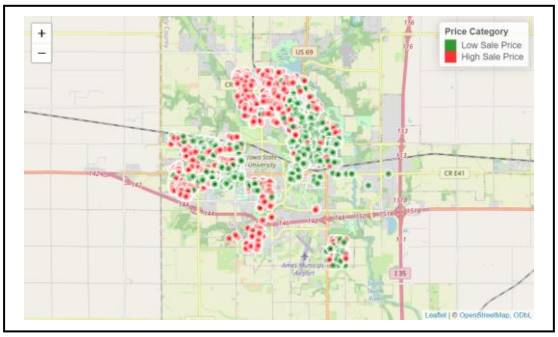


Figure 12: Map- Distribution of houses based on price

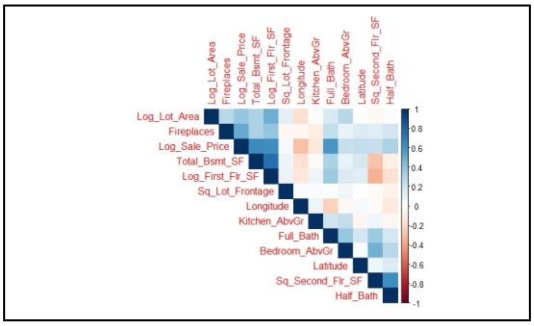
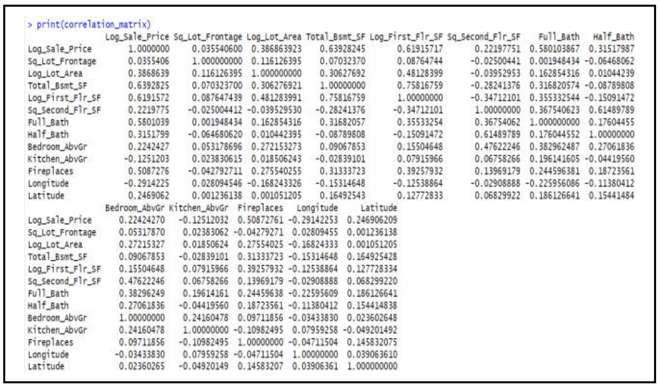


Figure 13: Correlation matrix heat map

Outliers have been detected in dependent variables like variable #18 and also in independent variables like variables #1, #2, #9 and #10. However, on outlier exploration, the data is not incorrect and is right to incorporate these outliers in the prediction of house sale price. The next step in this process is to calculate the correlation matrix. It is seen that when square root transformation is applied to variable #1, the correlation matrix shows that it is not a strong predictor. However, it is included in the multiple linear regression model as it is an important factor while purchasing a house. It can be removed if the business need requires it to be omitted.



```
> print(correlation_matrix)
                Log_Sale_Price Sq_Lot_Frontage Log_Lot_Area Total_Bsmt_SF Log_First_Flr_SF Sq_Second_Flr_SF    Full_Bath   Half_Bath
Log_Sale_Price       1.0000000     0.035540600  0.386863923    0.63928245       0.61915717       0.22197751 0.580103867  0.31517987
Sq_Lot_Frontage      0.0355406     1.000000000  0.116126395    0.07032370       0.08764744      -0.02500441 0.001948434 -0.06468062
Log_Lot_Area         0.3868639     0.116126395  1.000000000    0.30627692       0.48128399      -0.03952953 0.162854316  0.01044239
Total_Bsmt_SF        0.6392825     0.070323700  0.306276921    1.00000000       0.75816759      -0.28241376 0.316820574 -0.08789808
Log_First_Flr_SF     0.6191572     0.087647439  0.481283991    0.75816759       1.00000000      -0.34712101 0.355332544 -0.15091472
Sq_Second_Flr_SF     0.2219775    -0.025004412 -0.039529530   -0.28241376      -0.34712101       1.00000000 0.367540623  0.61489789
Full_Bath            0.5801039     0.001948434  0.162854316    0.31682057       0.35533254       0.36754062 1.000000000  0.17604455
Half_Bath            0.3151799    -0.064680620  0.010442395   -0.08789808      -0.15091472       0.61489789 0.176044552  1.00000000
Bedroom_AbvGr        0.2242427     0.053178696  0.272153273    0.09067853       0.15504648       0.47622246 0.382962487  0.27061836
Kitchen_AbvGr       -0.1251203     0.023830615  0.018506243   -0.02839101       0.07915966       0.06758266 0.196141605 -0.04419560
Fireplaces           0.5087276    -0.042792711  0.275540255    0.31333723       0.39257932       0.13969179 0.244596381  0.18723561
Longitude           -0.2914225     0.028094546 -0.168243326   -0.15314648      -0.12538864      -0.02908888 -0.225956086 -0.11380412
Latitude             0.2469062     0.001236138  0.001051205    0.16492543       0.12772833       0.06829922 0.186126641  0.15441484
                Bedroom_AbvGr Kitchen_AbvGr  Fireplaces   Longitude     Latitude
Log_Sale_Price     0.22424270   -0.12512032  0.50872761 -0.29142253  0.246906209
Sq_Lot_Frontage    0.05317870    0.02383062 -0.04279271  0.02809455  0.001236138
Log_Lot_Area       0.27215327    0.01850624  0.27554025 -0.16824333  0.001051205
Total_Bsmt_SF      0.09067853   -0.02839101  0.31333723 -0.15314648  0.164925428
Log_First_Flr_SF   0.15504648    0.07915966  0.39257932 -0.12538864  0.127728334
Sq_Second_Flr_SF   0.47622246    0.06758266  0.13969179 -0.02908888  0.068299220
Full_Bath          0.38296249    0.19614161  0.24459638 -0.22595609  0.186126641
Half_Bath          0.27061836   -0.04419560  0.18723561 -0.11380412  0.154414838
Bedroom_AbvGr      1.00000000    0.24160478  0.09711856 -0.03433830  0.023602648
Kitchen_AbvGr      0.24160478    1.00000000 -0.10982495  0.07959258 -0.049201492
Fireplaces         0.09711856   -0.10982495  1.00000000 -0.04711504  0.145832075
Longitude         -0.03433830    0.07959258 -0.04711504  1.00000000  0.039063610
Latitude           0.02360265   -0.04920149  0.14583207  0.03906361  1.000000000
```

Figure 14: Correlation matrix values

Domain expertise plays a major role here. The correlation matrix gives insights such as strong positive correlations of Log_Sale_Price with Log_First_Flr_SF, variables #8, #11 and #15. A very weak correlation was found between Log_Sale_Price and Sq_Lot_Frontage. Variable #16 and Log_Sale_Price have a weak negative correlation. Variables Sq_Lot_Frontage, #14 and #17 have low correlations with Log_Sale_Price. However, it will be included in the model for predicting housing sale prices.

E. *Modelling*

The seed is set to ensure the same random values are generated every time the code is run. The seed is set to 22238590. The dataset is split into train values and test values. All the variables are considered in the multiple linear regression model to predict the sale price. This method is called the Backward Elimination method. All variables have been included and none have been eliminated. Fitting all variables in the model yields a model with good accuracy. Gauss-Markov Assumptions are key conditions that ensure the ordinary least squares estimator will produce the optimal linear unbiased estimators of coefficients.

- Parameters Linearity
- No perfect Multicollinearity
- Random Sample from population
- Homoscedasticity

The train$fitted.values give the predicted sale price on a logarithmic scale. These values are used for diagnostic processes and model evaluation. To see how well the model performed, predictions will be made on test data. No perfect multicollinearity is not met. However, the parameters are linear, random sampling and homoscedasticity have been met.

Condition	Test	Pass Value
Homoscedasticity	Plot of residual vs predicted value	Randomness should be shown with no particular funnel pattern
No autocorrelation between errors	Durbin Watson test	A value of approximately 2 is best.
No multicollinearity between independent variables	Using Correlation matrix	Closer the number is to 1 or -1 stronger correlation. If no correlation then value is 0.
Data points not influential	Cook's distance	Values <1 and close to 0.

Table 3: Assumptions in Multiple Linear Regression

It is an ideal multiple linear regression model if all the assumptions are met.

- Gauss Markov Assumptions
- Other Assumptions
- Highest R-squared value
- Lowest Standard error

F. *Evaluation*

The ultimate model will be evaluated against the business objective to ensure it aligns with the goals of the business and the context stated in the problem. This can be achieved by analyzing the model output using various statistical measures.

III. Model OutPut and Evaluation

The multiple linear regression model performs well on the training data with an R-squared value of 0.8844. This implies 88.44% of the variation. The Durbin- Watson test value is 1.96 which can be rounded off to 2. Below is the P-P plot illustrating the model residuals.

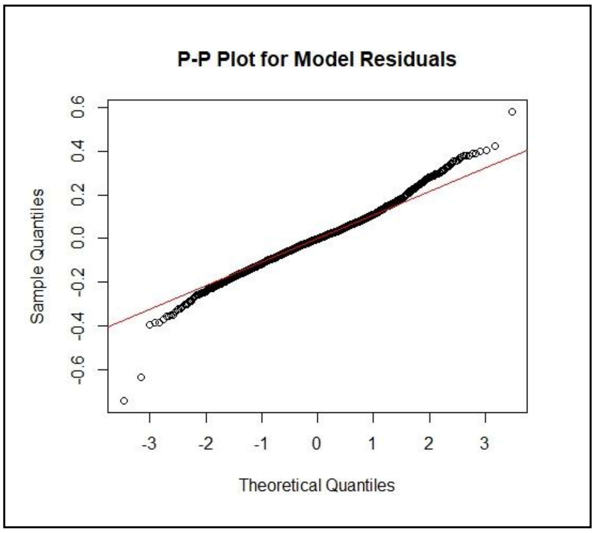


Figure 15: P-P Plot for Model Residuals

From Figure 15, it can be interpreted that the points almost closely follow the diagonal line. Hence, it can be concluded that the residuals are normally distributed. The points in the middle of the distribution should follow the diagonal line closely. Slight deviation at the ends is common due to variability in sampling.

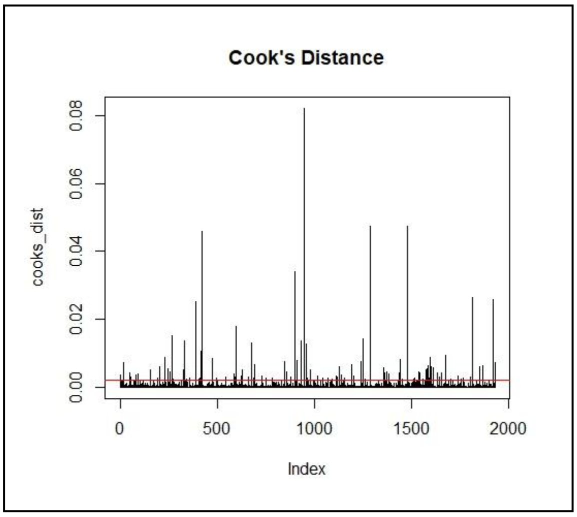


Figure 16: Cook's Distance

It can be observed from Figure 16 that the Cook's Distance is high. These points are influential and have a large influence on the regression model. These influential points are not errors or outliers but these are legitimate points.

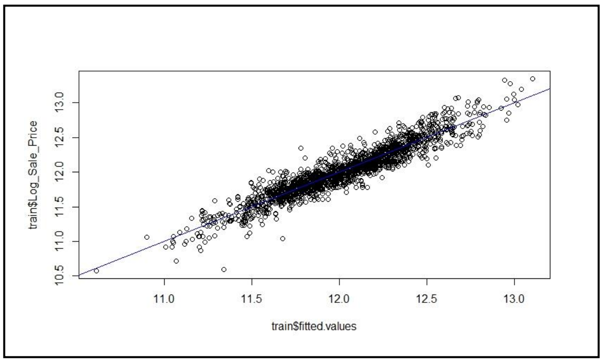


Figure 17: Scatterplot of fitted vs actual values

Figure 17 above shows the scatterplot comparing the fitted values vs the actual transformed values. This plot checks the fit of the multiple linear regression model. The points are closely clustered around the line of perfect fit. This indicates that the model has good predictive accuracy.

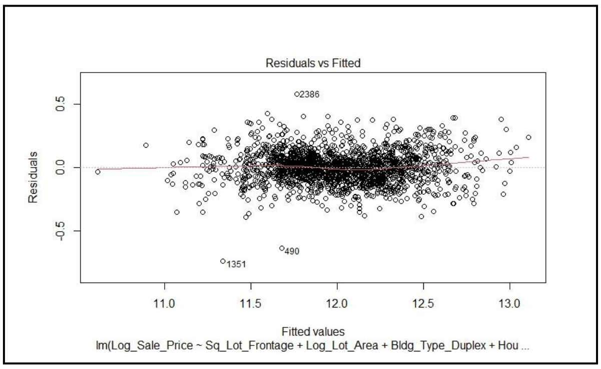


Figure 18: Residual plot

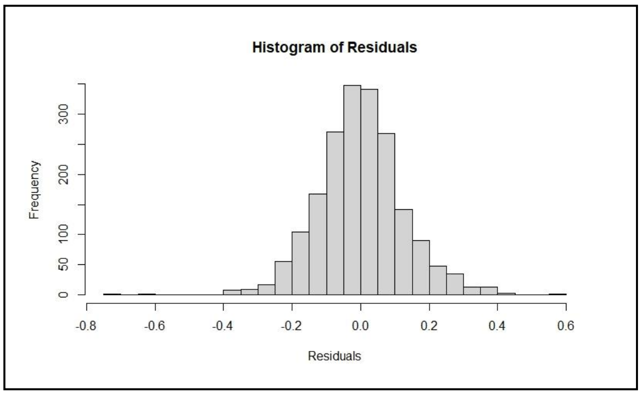


Figure 19: Histogram of Residuals

The figure 18 shows the Residual Plot to check for non-linearity and homoscedasticity. Since the residuals are scattered randomly and no clear pattern is observed, it can be inferred that the model fits the data quite well.

The histogram in Figure 19 above is roughly symmetrically distributed around zero. There is a clear peak and the histogram is bell-shaped. There is slight skewness as it has been discussed previously about not meeting all the Gauss-Markov Assumptions.

Now, the model will be tested using the test data to evaluate the performance. The test R-squared value stands at 0.8924 which implies that 89.24% of variability in Log_Sale_Price can be predicted from all independent predictors. This indicates that the model fits data well.

The Mean Absolute Error (MAE) is 0.1013, the Mean Squared Error (MSE) is 0.0170 and the Root Mean Squared Error (RMSE) is 0.1306. The low values of MAE, MSE and RMSE indicate that predictions are close to actual values. The summary of the model shows that most of the predictors are significant. However, there are a few insignificant predictors which can be removed if the business requires it to be removed.

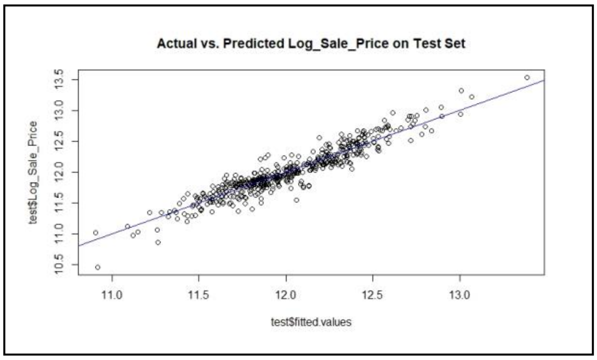


Figure 20: Scatterplot of actual vs predicted on test set

The figure above displays the scatter plot of actual vs predicted Log_Sale_Price on the test dataset. The majority of prediction points are close to the actual values since most of the data points are close to the diagonal line of perfect fit.

IV. Conclusion

The business problem statement was to predict the sale price of houses. The regression model implemented as shown above shows that the performance is quite good. Therefore, future work has to be whether to include predictors that are statistically less significant based on the domain knowledge and business requirement.

V. References

[1] M. R. Putri, I. G. Wijaya, F. P. Praja, A. Hadi, and F. Hamami, “The comparison study of regression models (multiple linear regression, ridge, lasso, random forest, and polynomial regression) for house price prediction in West Nusa tenggara,” *2023 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)*, 2023.

[2] K. Iraoui, R. Moustabchir, H. Charifi, M. Ouattab, and A. Chirmata, “Ozone concentrations predicting in Agadir City (Morocco) using the multiple linear regression (forward regression analysis),” *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2023.

[3] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, and R. Wirth, ‘The CRISP-DM Process Model’. CRISP-DM Consortium, 2000 1999.

[4] Sławomir Pasikowski, “Normalization Transformations of Data in the Procedure of Statistical Methods Used in Education Research,” Lubelski Rocznik Pedagogiczny, vol. 41, no. 4, pp. 91–101, Dec. 2022.