Analysis for Statistics:

Investigating Factors Influencing Housing Prices through Multiple Linear Regression

(November 2023)

Etinosa Eghaghe
School of Computing
National College Of Ireland
x23138548@student.ncirl.ie

Abstract **-** In today's world, understanding the factors that impact housing price in the real estate sector is paramount. This statistical analysis focuses on understanding the factor influencing housing sale price. Using the R programming language and the robust methodology of multiple linear regression (MLR) as a statistical tool to analyse, we look at the impart of various independent variables on our dependent variable sale_price. We aspire to extract meaningful insights that not only explain present market trends but also offer a predictive framework for future demand housing price. The final MLR algorithm effectively predict and analyse the sale price, the has dataset with 18 variable, Gauss-Markov assumptions was meet. That means OLS estimates are BLUE (best linear unbiased model) and other assumption of MLR. while further sophisticated statistical analysis can still be used to enhance the method.

Keywords— Multiple Linear Regression, Gauss-Markov, OLS, BLUE

1. INTRODUCTION

The purpose of this analysis is to investigate the factors influencing housing prices using Multiple Linear Regression (MLR) to build a model that aids in determining the relationship between sale_price and some other factors in the dataset given for this analysis, that can influence housing prices. and understanding these relationships is difficult for various stakeholders. The final model should meet the Gauss Markov assumptions of MLR.[1] There should be linearity between the predictor and other response, Relationships should not be overfitted, there should be no multicollinearity, residuals should be normal distributed and homoscedasticity.

Multiple Linear Regression refers to a statistical technique that uses many explanatory variables to predict the result of a response variable [1]-[2] Many studies have utilized MLR to explore the determinants of housing prices. For instance, examined the impact. This analysis aims to contribute to the existing body of knowledge by considering a set of relevant that collectively contribute to the determination of housing prices.

2. METHODOLOGY

2.1 Overview

A comprehensive analysis and modelling process undertaken on a housing dataset. The objective is to predict housing prices using multiple linear regression. The analysis involved several key stages, including data exploration, visualization, preprocessing, development, evaluation of predictive models, and conclusions.

2.2 Data Exploration

Dataset Loading The initial steps involve setting the working directory and loading the dataset from the housing_dat.csv file. Upon loading the dataset,

Descriptive Statistics: are method implemented to summarize and view relevant features of a dataset. Which include measures of mean, median, mode, range, variance, standard deviation, and graphical representation such as histogram.[3]-[4] was performed by examining the data structure through the following practices.

- **str(housing_dat):** This function to display an overview the dataset structure,
- **summary(housing_dat):** This function shows summary statistics for each variable, such as mean, median and quartiles.
- **dim(housing_dat):** is used to view the dimension of the dataset rows (2413), column (18)
- **sapply(housing_dat):** This function type of variables in the dataset, which contains numerical and categorical variables.

Figure 1a and 1b shows the dataset structure, type, distribution, mean, median, mode e.tc

Lot_Frontage	Lot_Area	Bldg_Type	House_Style	Overall_Cond	Year_Built	Exter_Cond
"integer"	"integer"	"factor"	"factor"	"factor"	"integer"	"factor"
Total_Bsmt_SF	**First_Flr_SF**	**Second_Flr_SF**	**Full_Bath**	**Half_Bath**	**Bedroom_AbvGr**	**Kitchen_AbvGr**
"integer"	"integer"	"integer"	"integer"	"integer"	"integer"	"integer"
Fireplaces	Longitude	Latitude	Sale_Price			
"integer"	"numeric"	"numeric"	"integer"			

Figure 1a: sapply()

```
 Lot_Frontage       Lot_Area         Bldg_Type              House_Style
Min.   :  0.00   Min.   :  1300   Duplex  :  78   One_Story       :1189
1st Qu.: 37.00   1st Qu.:  7390   OneFam  :2002   Two_Story       : 726
Median : 60.00   Median :  9360   Twnhs   :  93   One_and_Half_Fin: 270
Mean   : 55.46   Mean   : 10060   TwnhsE  : 188   SLvl            : 115
3rd Qu.: 77.00   3rd Qu.: 11404   TwoFmCon:  52   SFoyer          :  66
Max.   :313.00   Max.   :215245                   Two_and_Half_Unf:  22
                                                  (Other)         :  25
      Overall_Cond    Year_Built      Exter_Cond   Total_Bsmt_SF   First_Flr_SF
Average      :1282   Min.   :1872   Excellent:  11   Min.   :   0   Min.   : 334
Above_Average: 474   1st Qu.:1953   Fair     :  53   1st Qu.: 784   1st Qu.: 866
Good         : 352   Median :1971   Good     : 266   Median : 970   Median :1060
Very_Good    : 139   Mean   :1969   Poor     :   2   Mean   :1023   Mean   :1134
Below_Average:  82   3rd Qu.:1998   Typical  :2081   3rd Qu.:1246   3rd Qu.:1350
Excellent    :  39   Max.   :2010                    Max.   :3206   Max.   :3820
(Other)      :  45
Second_Flr_SF      Full_Bath       Half_Bath     Bedroom_AbvGr   Kitchen_AbvGr
Min.   :   0.0   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.00
1st Qu.:   0.0   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:2.000   1st Qu.:1.00
Median :   0.0   Median :2.000   Median :0.000   Median :3.000   Median :1.00
Mean   : 339.2   Mean   :1.539   Mean   :0.378   Mean   :2.855   Mean   :1.04
3rd Qu.: 704.0   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:1.00
Max.   :1872.0   Max.   :4.000   Max.   :2.000   Max.   :6.000   Max.   :3.00

   Fireplaces      Longitude        Latitude       Sale_Price
Min.   :0.000   Min.   :-93.69   Min.   :41.99   Min.   : 35000
1st Qu.:0.000   1st Qu.:-93.66   1st Qu.:42.02   1st Qu.:129500
Median :1.000   Median :-93.64   Median :42.03   Median :159000
Mean   :0.603   Mean   :-93.64   Mean   :42.03   Mean   :175568
3rd Qu.:1.000   3rd Qu.:-93.62   3rd Qu.:42.05   3rd Qu.:206900
Max.   :4.000   Max.   :-93.58   Max.   :42.06   Max.   :755000
```

Figure 1b: Summary

2.3 Visualization

Histogram: This allows us to visualize each variable for a better understanding of the data distribution and potential patterns. Geom_histogram is used for numerical variables while geom_bar for categorical variable distribution view the dependent variable sale price is almost normal so the next step can be taken. ASshow below in figure 2a and 2b, respectively.

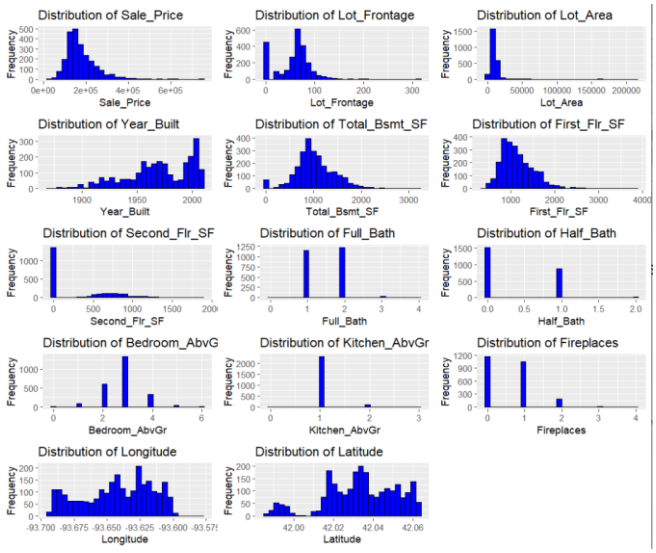

Figure 2a: numeric variable

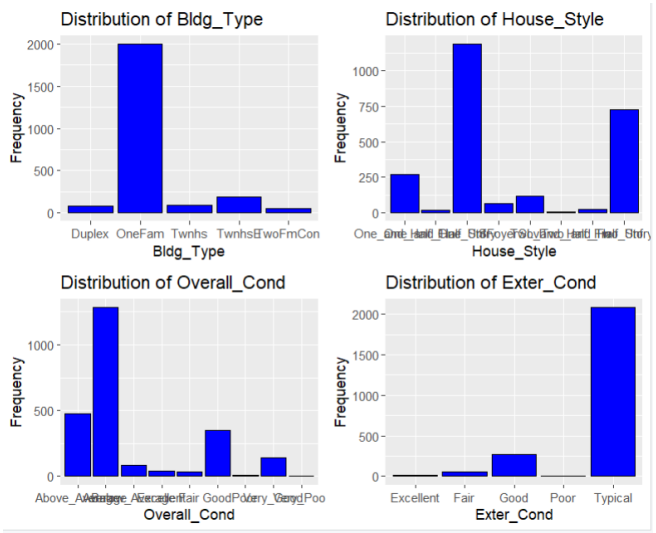

Figure 2b: Categorical variable

3. DATA PREPROCESSING

3.1 Handling Missing Values: In this stage missing values was check because in data processing handling missing value is important in other to ensure the accuracy and reliability of subsequent analyses.[3] Data can be missing due to various reason which include error during data collection or simply the absence of information for certain observations, a count of missing values in each column was displayed. In R, the 'is.na ()' function is mostly used to identify missing values but with the dataset used for this analysis there was no missing values. below is a subset.

.

Variable	Lot_Frontage	Lot_Area	Bldg_Type	House_Style	Overall_Cond	Year_Built
Missing Values	0	0	0	0	0	0

Figure 3: no missing values

3.2Encoding Categorical Variables: next was to encode categorical variable as MLR algorithm requires numeric input. Which represents qualitative data. Categorical variables were encoded using one-hot encoding to make them suitable for modeling. And after one-hot encoding the original categorical variables (Blog_Type,House_Style Overall_Cond, Exter_Cond) were removed from the dataset.

Variable	Data Type
Bldg_Type	"factor"
House_Style	"factor"
Overall_Cond	"factor"
Exter_Cond	"factor"

Categorical Variable type

3.3 Outlier Detection and Correction: Is a crucial phase to ensure that statistical models are duly influenced by extreme values. Outliers are data points that significantly differ from most of the data and can have a substantial impact on the results of statistical analyses.[1] Identifying and addressing outliers helps improve the robustness and reliability of the models. boxplots were utilized to visualize and correct outliers using the interquartile range (IQR) method.

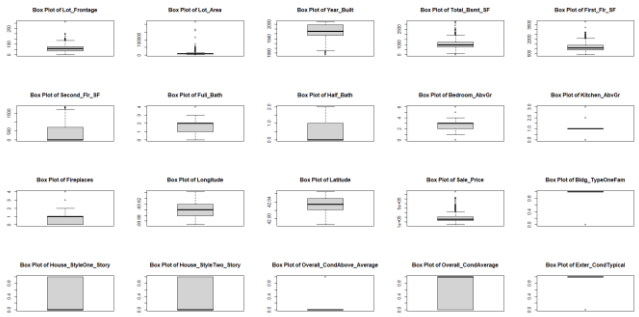

Figure 4a: Before IQR Test

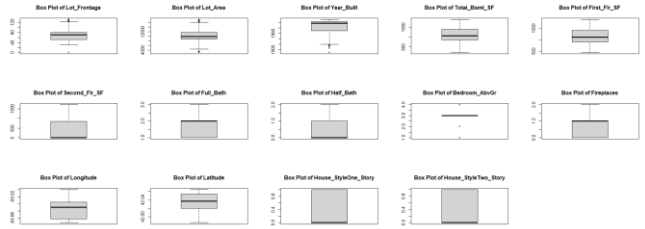

Figure 4b: After IQR Test

3.4 Feature Selection: This involves choosing a subset of features that are relevant. to retain the most informative and influential features. This leads to eliminating redundant or less important ones, it is difficult due to many reasons. But it improves model performance, reduces overfitting, Computational Efficiency and Enhanced Interpretability, for this analysis unnecessary columns without range were removed from the dataset to streamline the modeling process. Example in of variable with no range in figure 5

```
Overall_CondVery_Poor Exter_CondExcellent Exter_CondFair Exter_CondGood Exter_CondPoor
Min.   :0             Min.   :0           Min.   :0      Min.   :0      Min.   :0
1st Qu.:0             1st Qu.:0           1st Qu.:0      1st Qu.:0      1st Qu.:0
Median :0             Median :0           Median :0      Median :0      Median :0
Mean   :0             Mean   :0           Mean   :0      Mean   :0      Mean   :0
3rd Qu.:0             3rd Qu.:0           3rd Qu.:0      3rd Qu.:0      3rd Qu.:0
Max.   :0             Max.   :0           Max.   :0      Max.   :0      Max.   :0
```

Figure 5: sample of variables with no range

3.5 Standardization: Also called normalization used to scale numeric variables in a dataset for them to have similar scale, this make sure that variable is on common scale to prevent certain variables from disproportionately influencing the model training process. It transforms each numerical variable in the dataset so that it has a mean of 0 and standard deviation of 1.in this analysis selected numerical features were standardized using min-max normalization to ensure consistent scales.

.

$$Z = (X - mean(X)) / std(X)$$

4. MULTIPLE LINEAR REGRESSION

4.1 Model Development: The dataset was split into training and testing sets, by setting the seed to generate randomly. This is a statistical method to estimate the accuracy of the model performance when deploy on data that it was not train with. This separation helps ensure that the model is tested on data it has never seen during training, providing a more realistic assessment of its generalization performance. Training data is used to train the model and adjust it, while Test sets is used to evaluate the performance of the model. For this analysis we split the dataset into two parts on percentage of 80% to training the model and 20% for testing the model performance.

4.2 Model1: The first MLR model was performed using sale price as the dependent variable and selected independent variables model summary is generated to provide a comprehensive overview of this performance. Key statistics below show the residual standard error, R-squared, adjusted R-squared, F-statistic, and P-value. These show the goodness of fit to training data of the model.

- The model's performance was assessed using summary statistics,
- variance inflation factor (VIF) for multicollinearity to see the linearity.
- non-constant variance test, and autocorrelation test (Durbin-Watson).
- histogram of residuals provided insights into model fit.

Statistic	Value
Residual Standard Error	24820
Degrees of Freedom	653
Multiple R-squared	0.8593
Adjusted R-squared	0.8563
F-Statistic	284.8
F-Statistic Degrees of Freedom (numerator, denominator)	14, 653
F-Statistic p-value	< 2.2e-16

summary (model1):

```
Lot_Frontage                Lot_Area                Year_Built
    1.057179                1.245813                 2.900095
Total_Bsmt_SF               First_Flr_SF            Second_Flr_SF
    8.706442                9.270068                10.038313
Full_Bath                   Half_Bath               Bedroom_AbvGr
    3.181022                3.727887                 1.615522
Fireplaces                  Longitude               Latitude
    1.457523                1.470257                 1.232130
House_StyleOne_Story House_StyleTwo_Story
    4.556522                7.524044
```

VIF (model1): check for multicollinearity

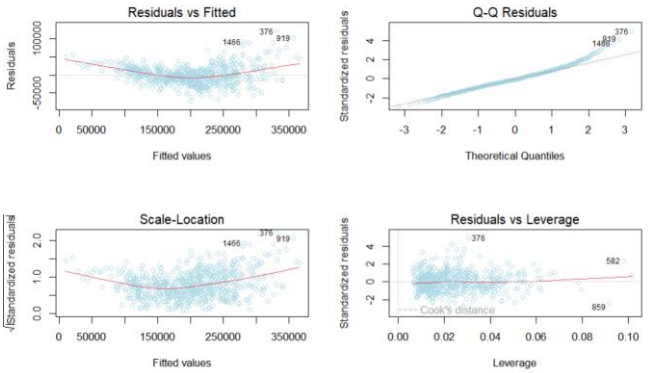

Figure 7: plot(model1): relationship b/w variables

Test	Value
Non-constant Variance Score Test	
Variance Formula	~ fitted.values
Chi-square	108.7061
Degrees of Freedom (Df)	1
p-value	< 2.22e-16

ncvTest(model1)

Lag	Autocorrelation	D-W Statistic	p-value
1	0.01024435	1.975368	0.768

DurbinWatsonTest(model1): Autocorrelation

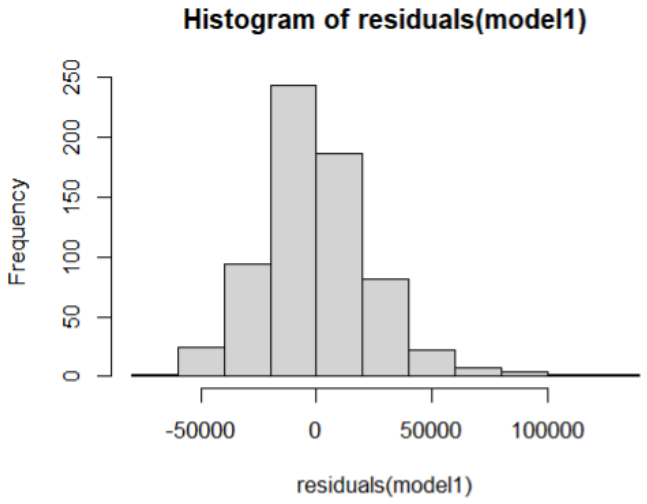

Figure 8: Residual

4.3 Model Improvement: To address multicollinearity, a second model was developed by removing variables with high VIF to achieve the assumption.

Metric	Value
Residual Standard Error	39740
Degrees of Freedom (DF)	657
Multiple R-squared	0.6372
Adjusted R-squared	0.6316
F-statistic	115.4
F-statistic DF	10 and 657
p-value	< 2.2e-16

summary(model2):

```
     Lot_Frontage         Lot_Area       Year_Built 
         1.040507         1.160178         2.691735 
       Full_Bath        Half_Bath     Bedroom_AbvGr 
         2.644829         2.289870         1.373976 
      Fireplaces        Longitude         Latitude 
         1.254906         1.462952         1.227350 
House_StyleOne_Story 
         2.384529 
```

VIF (model2): check for multicollinearity

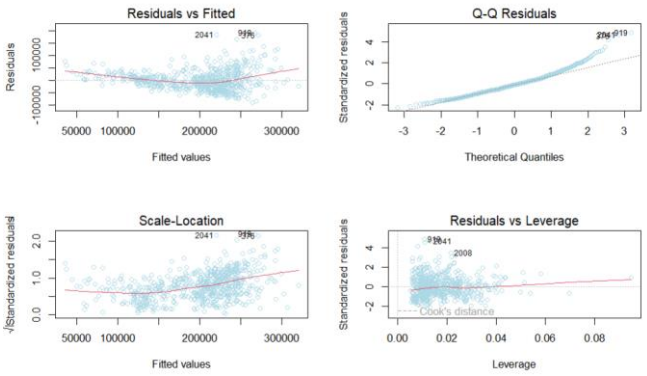

Figure 9: plot(model2): relationship b/w variables

Test	Value
Test Type	Non-constant Variance
Variance Formula	~ fitted.values
Chi-square	115.7961
Degrees of Freedom (Df)	1
p-value	< 2.22e-16

ncvTest(model2): Non-constant Variance Score Test

Lag	Autocorrelation	D-W Statistic	p-value
1	-0.03033418	2.057995	0.44

durbinWatsonTest(model2): Autocorrelation

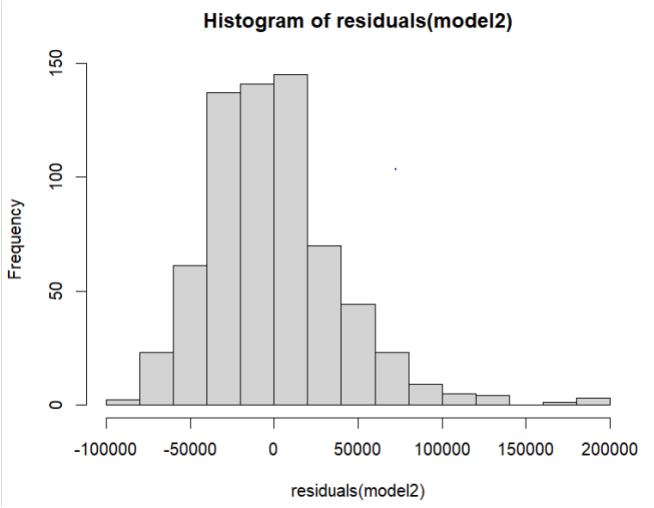

Figure 10: Residuals

4.4 Final Model: A log transformation was applied to the response variable to get a better model that meet Gauss-Markow assumptions. This final model showed an adjusted R-squared of 0.742 and a residual standard error of 0.1686. The non-constant variance test indicated significant heteroscedasticity (p = 0.0021), and the Durbin-Watson test showed no significant autocorrelation (p = 0.574). This is better than the second model.

Metric	Value
Residual Standard Error	0.1686
Degrees of Freedom (DF)	657
Multiple R-squared	0.7458
Adjusted R-squared	0.742
F-statistic	192.8
F-statistic DF	10 and 657
p-value	< 2.2e-16

summary (final_model)

```
     Lot_Frontage         Lot_Area       Year_Built        Full_Bath 
         1.040507         1.160178         2.691735         2.644829 
       Half_Bath     Bedroom_AbvGr       Fireplaces        Longitude 
         2.289870         1.373976         1.254906         1.462952 
        Latitude House_StyleOne_Story 
         1.227350         2.384529 
```

VIF (final_model): multicollinearity

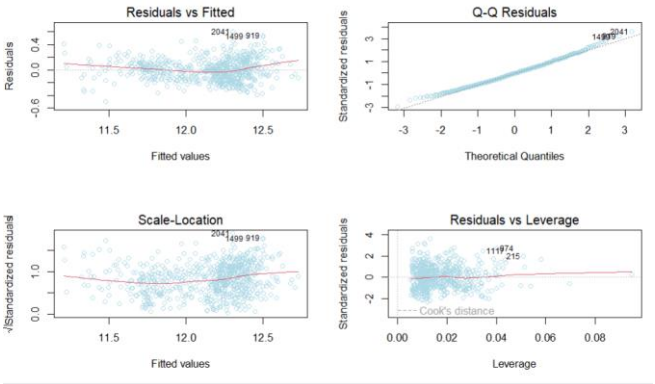

Figure 11: plot(final_model): relationship b/w variables

Test	Value
Non-constant Variance Score Test	
Variance Formula	~ fitted.values
Chi-square	9.451764
Degrees of Freedom (Df)	1
p-value	0.0021095

ncvTest(final_model): Non-constant Variance Score Test

Lag	Autocorrelation	D-W Statistic	p-value
1	-0.02299099	2.043147	0.574

durbinWatsonTest(final_model): Autocorrelation

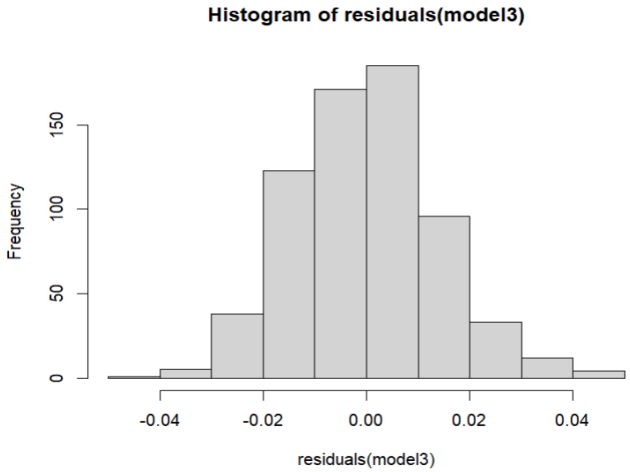

Figure 12: Residual

5. EVALUATION

The predictor capability of the final model was evaluated using a test dataset, and cross-validation result are obtained to assess the model performance the table below show the result

Metric	Value
RMSE	0.1011789
Rsquared	0.9092361
MAE	0.0797689

6. CONCLUSION

The analysis successfully preprocesses the housing dataset, identifies influential variables, and builds a regression model for predicting housing prices. Further refinement and validation may be required for real-world applications, but this report serves as a comprehensive guide for the initial steps in data analysis and regression modeling.

7. REFERENCE

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

[2] J. P. Hoffmann, *Linear Regression Models: Applications in R* (Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences). CRC Press, 2021.

[3] A. Field, J. Miles, and Z. Field, *Discovering Statistics Using R*. SAGE Publications, 2012.

[4] A. Roberts and J. M. Roberts, *Multiple Regression: A Practical Introduction*. SAGE Publications, 2020.