Investigating Factors Influencing Housing Prices: A Multiple Linear Regression Approach

Muddassir Ahmed
School of Computing
National College of Ireland
Dublin, Ireland
x23138688@student.ncirl.ie

***Abstract*—This research implements a Multiple Linear Regression approach to investigate the intricate factors influencing housing prices. The linear regression has two types: simple regression and multiple regression (MLR). Multiple liner regression is statistical approach which use regression model for complex relationship between two or more independent variables and one dependent variable. Using a comprehensive dataset, which is comprising on 18 variables and 2413 housing entries, the study searches into data analysis, preparation, modeling, and diagnostic evaluation to unveil significant understanding into the dynamics of housing pricing. visualisations and statistical tests applied, than results checked with Gauss-Markov assumptions.**

***Index Terms*—Multiple Linear Regression**

I. Introduction

Multiple Liner Regression (MLR) is used to check relationship between two variables. The linear regression has two types: simple regression and multiple regression (MLR). We predict dependent variable or the output, MLR [3] . For example, Govt and private Real estate organisations maintainig data of houses, which contains number of rooms, number of washrooms, kitchen size, number of floors and other variables. [1] These variables can use to predict price of houses. It would be interesting that sale price will be measure by different variables. We can check outliers, houses that should really sell for more, given their location and characteristics [4] This study endeavors to explore the nuanced factors that impact housing prices through the application of Multiple Linear Regression. The dataset is bound to 18 variables, providing a strong foundation for the investigation. [2].

II. Data Analysis

A. Descriptive Statistics

The introductory analysis involves a rigorous examination of basic dataset characteristics, including count, mean, minimum, and maximum values. This provides an initial understanding of the dataset's distribution.

B. Visualization

Visualization technique is consist of scatter plots, box plots, and pie charts, are employed to mmt the relationships between variables and the distribution of categorical data. These visualizations offer a comprehensive overview of the dataset's inherent patterns.

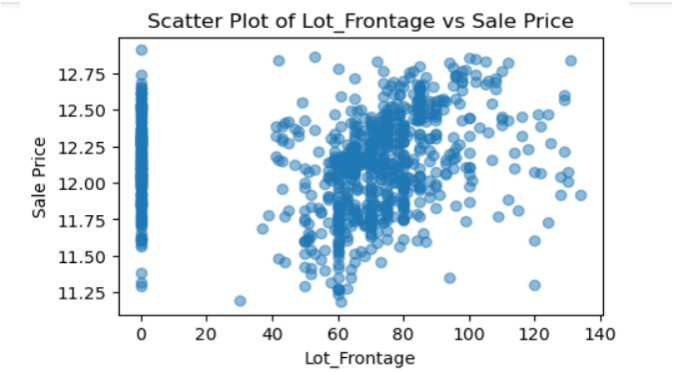

Fig. 1. Image

Below picture shows sale price increase with respect of year. In year 2000 price of houses increase 20 million.

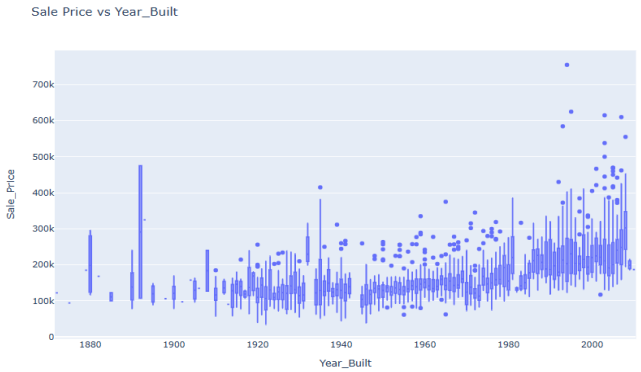

Fig. 2. Image

53.1 percent houses condition is average and 19.6 percent is above average. Overall distribution is shows in figure.

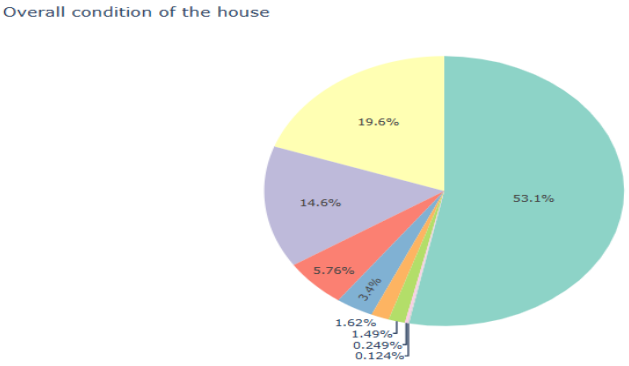

Fig. 3. Image

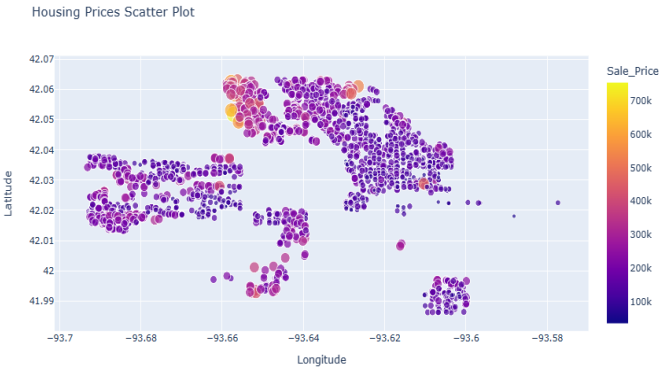

Fig. 4. Image

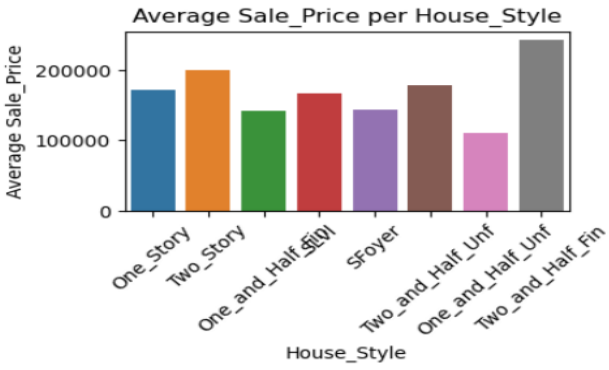

Fig. 5. Image

C. Levels of Measurement

An assessment of the level of measurement for each variable is crucial. This step distinguishes between nominal, ordinal, interval, and ratio variables, providing a nuanced understanding of the dataset's characteristics.

```
Levels of Measurement:
Lot_Frontage: Interval or Ratio
Lot_Area: Interval or Ratio
Bldg_Type: Nominal
House_Style: Nominal
Overall_Cond: Nominal
Year_Built: Interval or Ratio
Exter_Cond: Nominal
Total_Bsmt_SF: Interval or Ratio
First_Flr_SF: Interval or Ratio
Second_Flr_SF: Interval or Ratio
Full_Bath: Ordinal
Half_Bath: Ordinal
Bedroom_AbvGr: Ordinal
Kitchen_AbvGr: Ordinal
Fireplaces: Ordinal
Longitude: Interval or Ratio
Latitude: Interval or Ratio
Sale_Price: Interval or Ratio
```

Fig. 6. Image

III. Data Preparation

A. Categorical Encoding

The encoding of categorical variables is undertaken, utilizing label encoding techniques. Concurrently, a meticulous check for null values is executed, ensuring the dataset's integrity.

```
Count of Null Values for Each Variable:
Lot_Frontage              0
Lot_Area                  0
Year_Built                0
Total_Bsmt_SF             0
First_Flr_SF              0
Second_Flr_SF             0
Full_Bath                 0
Half_Bath                 0
Bedroom_AbvGr             0
Kitchen_AbvGr             0
Fireplaces                0
Longitude                 0
Latitude                  0
Sale_Price                0
Overall_Cond_encoded      0
Bldg_Type_encoded         0
Exter_Cond_encoded        0
House_Style_encoded       0
dtype: int64
```

Fig. 7. Image

B. Skewness and Kurtosis

Skewness and kurtosis are computed for each feature, and subsequent transformations are applied to address extreme skewness values. Log and Box-Cox transformations are deployed to achieve normality.

Result shows before applying Skewness and Kurtosis

	Skewness	Kurtosis
Lot_Frontage	-0.081114	1.165690
Lot_Area	13.393921	270.466068
Year_Built	-0.586724	-0.438653
Total_Bsmt_SF	0.456386	1.716119
First_Flr_SF	1.043301	2.256534
Second_Flr_SF	0.804118	-0.559404
Full_Bath	0.245013	-0.570307
Half_Bath	0.663992	-1.162667
Bedroom_AbvGr	0.184159	1.471746
Kitchen_AbvGr	4.681574	21.974672
Fireplaces	0.739075	0.138247
Longitude	-0.337513	-0.972557
Latitude	-0.507778	-0.083129
Sale_Price	1.745358	5.825764
Overall_Cond_encoded	1.309029	0.365022
Bldg_Type_encoded	2.216017	4.754053
Exter_Cond_encoded	-2.406885	4.640708
House_Style_encoded	0.498645	-1.568070

Fig. 8. Image

Defined acceptable threshold to values and identified unacceptable skewness. Applied log transforming only features which has extreme sknewness values.
Result shows skewness after transformation.

```
Kitchen_AbvGr            4.261726
Bldg_Type_encoded        0.985095
Overall_Cond_encoded     0.542142
Sale_Price               0.155859
First_Flr_SF             0.024600
Lot_Area                -0.513180
dtype: float64
```

Fig. 9. Image

Result of computing Skewness and kurtosis of the features again to check after applying transformation

	Skewness	Kurtosis
Lot_Frontage	-0.081114	1.165690
Lot_Area	-0.513180	3.793270
Year_Built	-0.586724	-0.438653
Total_Bsmt_SF	0.456386	1.716119
First_Flr_SF	0.024600	0.030996
Second_Flr_SF	0.804118	-0.559404
Full_Bath	0.245013	-0.570307
Half_Bath	0.663992	-1.162667
Bedroom_AbvGr	0.184159	1.471746
Kitchen_AbvGr	4.261726	21.846437
Fireplaces	0.739075	0.138247
Longitude	-0.337513	-0.972557
Latitude	-0.507778	-0.083129
Sale_Price	0.155859	0.675376
Overall_Cond_encoded	0.542142	-0.603059
Bldg_Type_encoded	0.985095	3.501112
Exter_Cond_encoded	-2.406885	4.640708
House_Style_encoded	0.498645	-1.568070
transformed_feature	-2.105647	2.436031

Fig. 10. Image

C. Outlier Handling

The identification and removal of outliers through box plots are essential for refining the dataset. A systematic analysis is conducted to determine how outlier removal affects the distribution of the dataset.

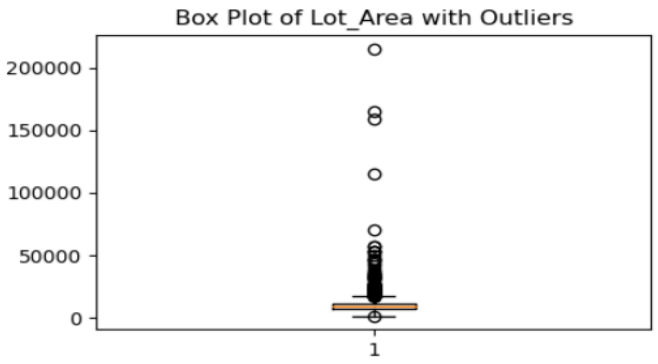

Fig. 11. Image

Each variable has below mentioned outliers.

```
Lot_Frontage, Outliers: 16
Lot_Area, Outliers: 223
Year_Built, Outliers: 10
Total_Bsmt_SF, Outliers: 116
First_Flr_SF, Outliers: 11
Second_Flr_SF, Outliers: 5
Full_Bath, Outliers: 4
Half_Bath, Outliers: 0
Bedroom_AbvGr, Outliers: 61
Kitchen_AbvGr, Outliers: 98
Fireplaces, Outliers: 11
Longitude, Outliers: 0
Latitude, Outliers: 0
Sale_Price, Outliers: 46
Overall_Cond_encoded, Outliers: 974
Bldg_Type_encoded, Outliers: 411
Exter_Cond_encoded, Outliers: 332
House_Style_encoded, Outliers: 0
transformed_feature, Outliers: 332
```

Fig. 12. Image

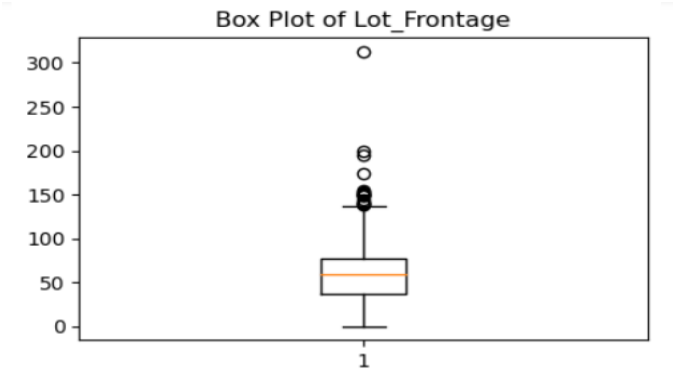

Fig. 13. Image

D. Collinearity

The presence of collinearity is assessed through the calculation of Variance Inflation Factors (VIF). Features with high collinearity are found and removed to improve the stability of further investigations.

	Variable	VIF
0	Lot_Frontage	1.057380
1	Lot_Area	1.260352
2	Year_Built	4.651591
3	Total_Bsmt_SF	7.223903
4	First_Flr_SF	7.747448
5	Second_Flr_SF	9.185824
6	Full_Bath	3.179241
7	Half_Bath	3.539582
8	Bedroom_AbvGr	1.691223
9	Fireplaces	1.492917
10	Longitude	1.428484
11	Latitude	1.215659
12	Sale_Price	9.318871
13	Overall_Cond_encoded	1.195664
14	Bldg_Type_encoded	0.000000
15	Exter_Cond_encoded	0.000000
16	House_Style_encoded	5.703537
17	transformed_feature	0.000000

Fig. 14. Image

Created correlation matrix for the features

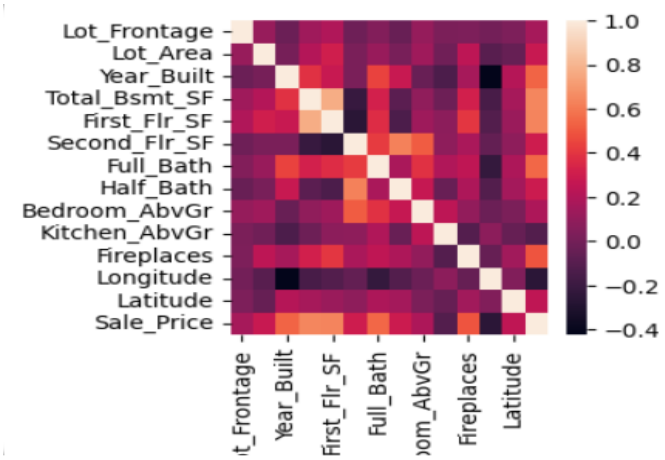

Fig. 15. Image

IV. Modeling

A. Feature Selection

The correlation between four predictor variables and housing prices was used to select them. The purpose of this deliberate choice is to improve the predictive power of the model. Based on correlation matrix we have choose Year built, Full bath, Total Bsmt SF, Fireplaces variables for predictor.

B. Data Splitting

The training and testing sets of the dataset are divided, making it easier to assess how well the model generalizes.

C. Analysis of Variance (ANOVA)

ANOVA tests confirm the statistical significance of selected features in explaining variance in housing prices, offering critical insights into the predictive power of the chosen variables.

```
Feature: Year_Built, F-Statistic: 5803326.458295206, p-value: 0.0
Feature: Full_Bath, F-Statistic: 288140.60993254377, p-value: 0.0
Feature: Total_Bsmt_SF, F-Statistic: 9819.368716716332, p-value: 0.0
Feature: Fireplaces, F-Statistic: 264256.21363024315, p-value: 0.0
```

Fig. 16. Image

Now, calculating Critical value to compare it with p-value to assess if the independent variables has a significant effect on the dependent variable.
Number of independant variables are four.
Level of alpha is 0.05
Result: Critical Value: 2.3818997423603165

D. Principal Component Analysis (PCA)

I have applied the PCA to Standardize the features and also preserving the original column names then I get the variance ratio
Result: [0.49847314 0.23618105 0.19547792]

	PC_Year_Built	PC_Full_Bath	PC_Total_Bsmt_SF
0	-0.688721	-0.108374	-0.936544
1	-0.818594	-0.070899	-0.773649
2	-0.388378	-0.218637	-1.345757
3	-0.542061	-0.273688	-1.289900
4	0.298906	0.953369	0.554108
...	...	...	...
892	-0.445513	0.434994	-0.396539
893	1.561946	-0.775815	0.292475
894	1.283840	-0.873624	0.396050
895	0.578849	1.313800	0.810675
896	-0.675018	0.030690	-0.756750

897 rows × 3 columns

Fig. 17. Image

E. Linear Regression

Multiple linear regression was employed to model the relationship between selected features and housing prices. The model's statistical metrics, such as Coefficients, R-squared, Adjusted R-square, Residuals, F-statistic and P-value indicated its overall performance.
Result show in below diagram.

```
Coefficients:
const     12.115230
x1        -0.208344
x2        -0.000271
x3         0.013965
dtype: float64
R-squared: 0.7864297336854311
Adjusted R-squared: 0.785354715566398
Residuals:
739       0.068015
1406     -0.247348
734      -0.022121
124       0.066186
733       0.042089
             ...
4        -0.091412
274       0.241088
1327      0.263113
905      -0.013580
2037      0.279239
Length: 600, dtype: float64
F-statistic: 731.5502127469191
p-value: 2.762430310240747e-199
```

Fig. 18. Image

V. Interpretation

The coefficients and statistics derived from the Multiple Linear Regression model are thoroughly interpreted to gain insights into the strength and direction of relationships between predictor variables and housing prices.

VI. Diagnostics

A. Gauss-Markov Assumptions

We verified the model against the Gauss-Markov assumptions, including linearity, normality of residuals, homoscedasticity, and independence of residuals.

B. Linearity Graph

Scatter plot used to check linearity. Results shows in below diagram

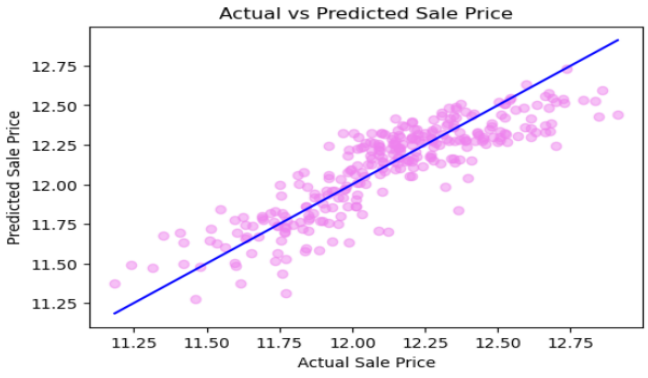

Fig. 19. Image

C. RESET Test

The RESET test confirmed the linearity assumption, ensuring the model's appropriateness for predicting housing prices.
Result: statistic=7.138328594320825
p-value=0.028179393441924546

D. Normality of Residuals Graph

Q-Q plot shows normality of residuals. Results shows diagram

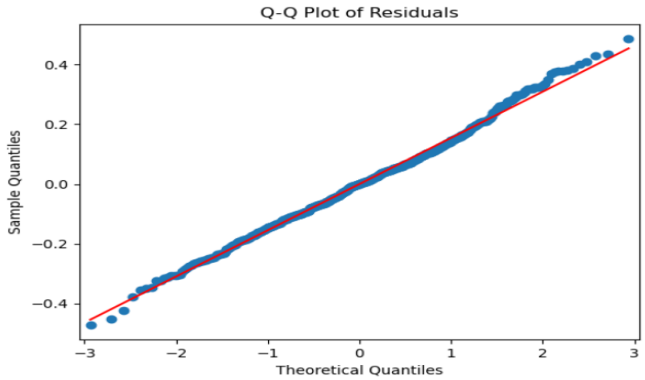

Fig. 20. Image

Shapiro-Wilk tests validated the normality of residuals, a crucial assumption for accurate predictions.
Result: statistic=0.9956386089324951
p-value=0.09212029725313187
Histogram shows residuals

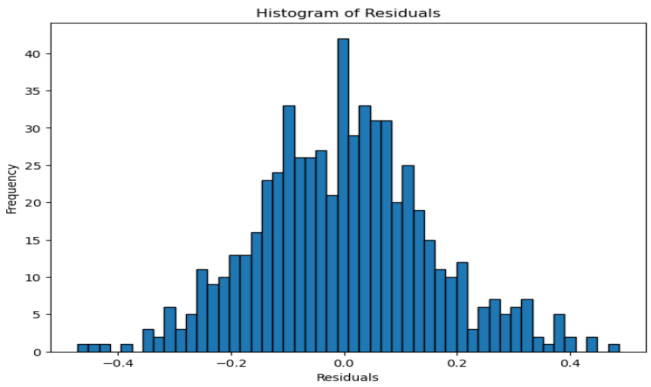

Fig. 21. Image

E. Homoscedasticity

Scatter plot use to show result.

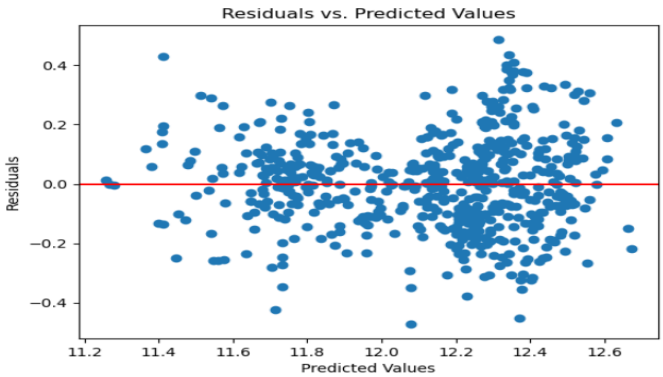

Fig. 22. Image

The Breusch-Pagan test confirmed homoscedasticity, indicating consistent variance across predicted values.

Breusch-Pagan Test Results: LM Statistic: 31.44375508533588,
P-value: 6.854704042547495e-07

F. Independence of Residuals

We ensured the independence of residuals, crucial for unbiased predictions and reliable model performance

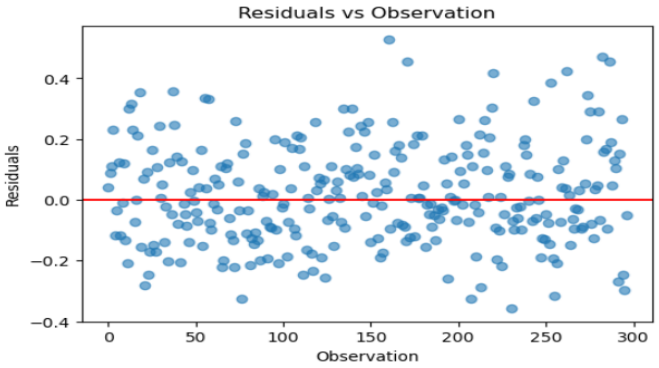

Fig. 23. Image

G. Evaluation

Diagnostic tests, such as the RESET test, Q-Q plots, and histograms, are employed to validate the model's adherence to assumptions. Strong F-statistics and high R-squared values support the robustness of the model's fit.

VII. Conclusion

This paper presented a real estate dataset describing the sale price of houses. Dataset consist of 2413 houses data and total 18 variables. In this paper, Multiple linear regression modal use to analyse the sale price in dataset. The goal was to train data to predict the sale price of houses in testing dataset. A summary of the research's findings emphasizes how well the model explains the variance in housing prices. The diagnostic evaluations add significant value to the field of housing price analysis by bolstering confidence in the model's predictive abilities.

References

[1] Azad Abdulhafedh. Incorporating multiple linear regression in predicting the house prices using a big real estate dataset with 80 independent variables. *Open Access Library Journal*, 9(1):1–21, 2022.

[2] P Bruce and A Bruce. Practical statistics for data scientists. o'reilly media. *Inc., Sebastopol, CA*, 2017.

[3] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[4] Daniela Witten and Gareth James. *An introduction to statistical learning with applications in R*. springer publication, 2013.