

Unveiling the Impact of House Characteristics on Sale Prices in the Ames Housing Market by Multiple Linear Regression

1. Introduction

Over the past decade, the housing market has become an increasingly popular topic in the field of economics. Housing availability and affordability are fundamental issues that affect individuals, families and communities. Understanding the factors that drive sales prices in the housing market is critical to addressing these issues. In the context of the complex and dynamic environment of the Ames housing market, this study aims to provide insight into the impact of housing characteristics on sales prices by using multiple linear regression (MLR) techniques.

With the increasing focus on housing price prediction in academic research, various model-building methods have been investigated. Truong and Minh Nguyen (2019) conducted research on housing price prediction using machine learning techniques to assess their effectiveness. Their study findings demonstrated the success of machine learning in accurately predicting housing prices. In contrast, this research employs a different approach using multiple linear regression to analyze the impact of housing characteristics on sales prices in the Ames housing market. Additionally, building upon the work of Anuradha and Raga Madhuri (2019), who focused on improving house price prediction through regression techniques, this research enhances the accuracy of sales price prediction by incorporating additional variables and providing a more comprehensive understanding of the dynamics of the housing market. Furthermore, Ning-Yen Chen's study (Chen, 2022) also uses MLR to predict house prices from urban factors such as resident population, land price, and employment rate.

On the other hand, my study mainly predicts house prices from housing factors such as housing conditions and square footage.

In summary, the findings of this study are expected to provide valuable insights into the Ames housing market, shedding light on the specific house characteristics that significantly influence sale prices, thereby addressing the issue of information asymmetry. Such insights can aid sellers in accurately pricing their properties and assist buyers in making informed decisions based on the underlying factors that drive sale prices.

2. Method Selection

2.1 Variable Selection:

During the initial stage of method selection, the downloaded data was imported into R Studio. Following this, data cleaning was performed to ensure the quality of the model. Based on insights from the literature and expert knowledge, the unrelated variables were eliminated and nine independent variables were selected from them as potential predictors of house prices. To handle the large number of observations, a sample of 500 data points was randomly selected and divided into a training dataset (accounting for 70%) and a testing dataset (accounting for 30%). The training data set will be used for the subsequent analysis.

The second step involved building the model. At the beginning of model building, due to the large value of the dependent variable in data dataset, it is necessary to apply a power transformation to normalize the variables and facilitate the subsequent modeling work. Afterward, model 1 was created using nine potential variables that could affect the housing price. Based on their significance levels (below 0.05), model 1 can systematically eliminate unnecessary variables and form a more refined and optimized model 2. Then, by removing two of the variables and keeping the

remaining variables, model 3 was built and a partial F-test was performed on model 3 and model 2 to prove that these two variables are meaningful for predicting house prices, and finally, if the result of the partial F-test is less than 0.05, then model 2 is the first choice of model.

To check for multicollinearity in model 2, a VIF calculation was performed for each predictor variable. If the VIF result for each variable is less than 5, it indicates no multicollinearity issues in the model. Subsequently, a full F-test was conducted on model 2 to verify its usefulness. The null hypothesis (H_0) assumed all predictor coefficients to be 0, while the alternative hypothesis (H_a) assumed at least one coefficient to be non-zero. If the resulting p-values are all less than zero, it indicates that the model consisting of these variables is useful.

Towards the end of the model building process, the stepwise method in RStudio can be also utilized to automatically generate another house price model that minimizes the AIC.

2.2 Model Validation:

During the model validation process, as the data is divided into a training set and a test set, the data from the test set will be used to validate the model. Initially, a new model is created with the same variables as model 2. Then, all the previous steps are repeated using the data from the test dataset. Once the results are available, the accuracy and performance of the final model can be evaluated by examining the coefficients of each variable and the values obtained from the training data. In particular, higher adjusted R^2 values indicate a stronger explanatory power of the model, while smaller SSres, AIC, AICc and BIC values indicate the accuracy of the final model in estimating housing prices under various scenarios.

2.3 Model Violations and Diagnostics:

The third part focuses on assessing the assumptions of model 2, which include linearity, uncorrelated error, constant variance, and normality. To assess whether condition 1 is satisfied in this study, the data should confirm condition 1 by examining the alignment around the fitted line and the trend depicted in the graph to determine if the predicted values are close to the actual values. Additionally, condition 2 will be assessed by creating scatter plots for pairs of predictor variables to visualize their covariance and ensure the absence of significant linear relationships between them.

Once both conditions are satisfied, the residual plot can be used to detect any violations of assumptions in model 2. Specifically, a satisfactory model should first have no systematic pattern of curves. Additionally, the normality assumption can be assessed by analyzing QQ-plots, where multiple points converging into a straight line can demonstrate compliance with the assumption. Furthermore, independence and constant variance assumptions can be evaluated by examining any pattern of the residual plots. In compliance with these conditions, the residual plots should not display identifiable patterns, specifically the absence of systematic clustering and fan-shaped spreading tendencies in the distribution.

In addition, it is crucial to consider leverage points, outliers, and influential points in model 2. Leverage points refer to observations with extreme predictive values that can significantly influence the results of the model. Outliers are observations that deviate significantly from the overall model, while influential points have a significant impact on the parameters of the model. By examining these particular points, we can develop a deeper understanding of the Violations of Model 2.

3. Results Section

3.1 Description of Data:

The data is sourced from the “Ames Housing dataset”. After the selection of predictors and removal of missing values, the final dataset consists of 1460 observations of 10 variables. From this dataset, a random sample of 500 observations was chosen and divided into train and test datasets for modeling purposes.

Figure 1: Power Transformation Summary

	Estimate power	Rounded power	Lwr Bnd Wald	upper bound
Sale Price	0.0571	0.00	-0.0639	0.1782
Overall Condition	0.6895	0.50	0.4305	0.9484
Overall Quality	1.1123	1.00	0.8555	1.3690
Finished square feet	0.0809	0.08	0.0671	0.0948
Ground living area	0.0113	0.00	-0.1706	0.1933
Garage Cars	0.5564	0.56	0.5034	0.6094
Full Bathroom	0.6921	0.69	0.5885	0.7957
Month Sold	0.8562	1.00	0.6720	1.0404
Bedroom	0.9334	1.00	0.7614	1.1054
Kitchen	-25.7761	-25.78	-28.4766	-23.0756

Considering that the rounded power of the sales price variable is zero, a logarithmic transformation was applied to the sales price variable. This transformation created a variable called 'logprice,' which can be utilized in the modeling process.

Figure 2: Summary table for numerical variables for Train/Test data

Train / Test data:

Variable name	Minimum	1st quartile	Median	Mean	3rd quartile	maximum
Log (Price)	10.47 / 10.92	11.78 / 11.77	12.00 / 12.0	12.03	12.28 / 12.24	13.35 / 13.53
Overall condition	1 / 3	5 / 5	5 / 5	5.494/5.527	6 / 6	9/9
Overall quality	1 / 4	5 / 5	6 / 6	6.12 / 6.167	7 / 7	10 / 10
Finished square feet	0 / 0	0 / 0	398 / 384	459.3 /422.4	732.8 / 701.5	1646.0 / 1573.0
Ground living area	480 / 605	1132 / 1144	1452 /1491	1492 / 1537	1748 / 1794	3627 / 4316
Garage Cars	0 / 0	1 / 1	2 / 2	1.789 / 1.827	2 / 2	3 / 4
Full Bathroom	0 / 1	1 / 1	2 / 2	1.603 / 1.58	2 / 2	3 / 3
Month Sold	1 / 1	5 / 5	6 / 6	6.271 / 6.307	8 / 8	12 / 12
Bedroom	0 / 1	2 / 3	3 / 3	2.82 / 3	3 / 3	6 / 6
Kitchen	1 / 1	1 / 1	1 / 1	1.057 / 1.053	1 / 1	3 / 2

The summary table above shows basic statistics for the various numeric variables. The "Log (Price)" variable, which represents the logarithm of home prices, has a median of 12.00 and a mean of 12.03, showing a relatively moderate distribution with no extreme outliers.

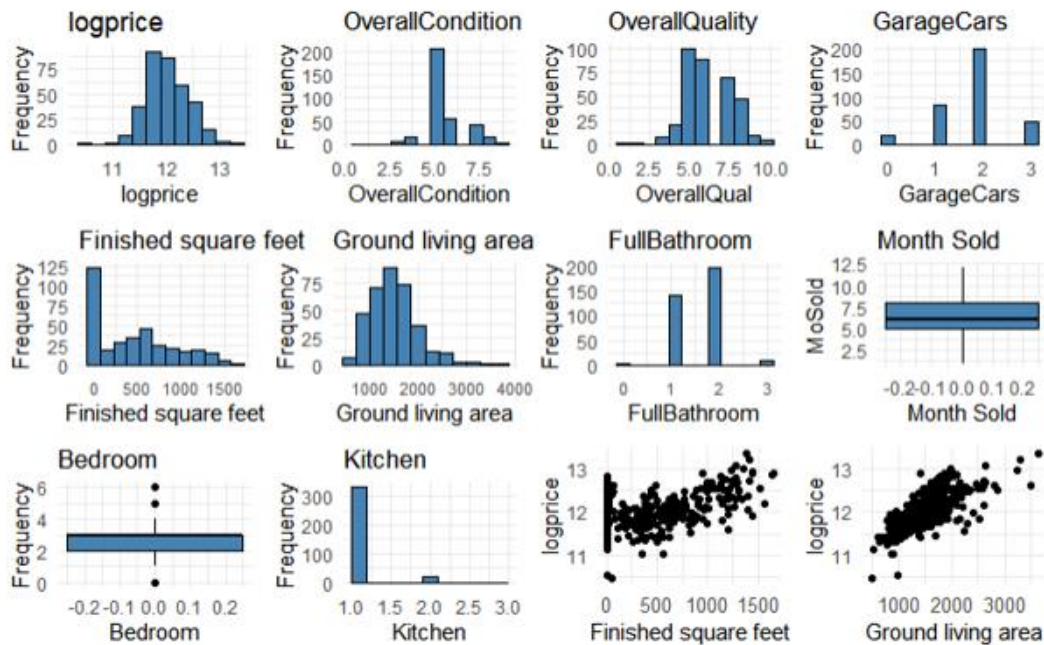


Figure 3: histogram & scatter plot & boxplot of numerical variable

[First row: Frequency vs. Logprice; Frequency vs. Overall Condition; Frequency vs. Overall Quality; Frequency vs. Garage Cars;
 Second row: Frequency vs. Finished Square feet; Frequency vs. Ground living area; Frequency vs. Full bathroom; Frequency vs. Month Sold;
 Third row: Frequency vs. Bedrooms; Frequency vs. Kitchens; Frequency vs. Full bathroom; Frequency vs. Month Sold]

The histogram above provides an initial insight into the distribution of the data for the different numerical variables. From the graph, it can be observed that the distributions of "logprice" and "ground living area" are very close to the shape of a normal distribution, and their highest frequency is at approximately 12 and 1500, respectively. On the other hand, "finished square feet" shows a right-skewed distribution, which indicate the most of the houses are smaller in size.

The two boxplots showcase the distribution of data for "Bedroom" and "Month Sold," which are also numerical variables. The boxplot for "Bedroom" displays the range and distribution of the number of bedrooms in the houses, while the boxplot for "Month Sold" depicts the distribution of house sales over different months.

Finally, two scatter plots were chosen to depict the relationship between the "price"

variable and the two numerical variables. The scatterplots clearly illustrate an upward trend, which demonstrates a positive relationship between the variables. This means that as the value of the selected numeric variable increases, the price of the house also tends to increase.

3.2 Presenting the Analysis Process and the Results:

In Step 1, based on intuition and information obtained from the literature, a careful selection process was used to remove irrelevant variables and identify nine predictors for Model 1. These predictors were selected because they were considered to be correlated with housing prices.

Model 1:

$$\begin{aligned}\widehat{Logprice} = & 10.51 + 0.01596 * [OverallCondition] + 0.1242 * [OverallQuality] + 0.0001883 \\ & * [Finishedsquarefeet] + 0.0002622 * [Groundlivingarea] + 0.1054 * [GarageCars] \\ & + 0.05124 * [FullBathroom] - 0.002598 * [MonthSold] + 0.005319 * [Bedroom] \\ & - 0.07298 * [Kitchen]\end{aligned}$$

In step 2, a rigorous analysis was performed to identify variables in model 1 with significance levels below 0.05. These variables were selected to build model 2, which was designed to refine the predictive power of the model.

Model 2:

$$\begin{aligned}\widehat{Logprice} = & 10.51 + 0.1287 * [OverallQuality] + 0.0001883 * [Finishedsquarefeet] + 0.0002647 \\ & * [Groundlivingarea] + 0.1011 * [GarageCars] + 0.04147 * [FullBathroom]\end{aligned}$$

In the third step, the most influential variables from Model 2 were carefully selected to create the third model. This selection was based on the variables with 2 or 3 stars on the right-hand side, indicating their high significance.

Model 3:

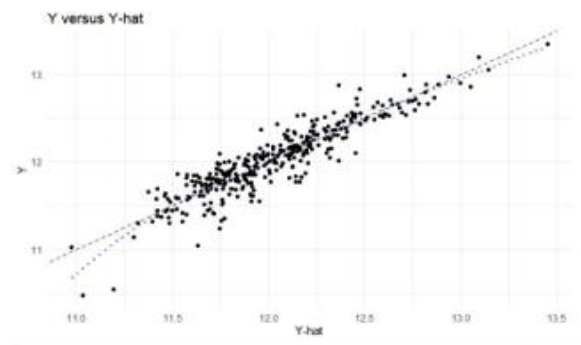
$$\begin{aligned}\widehat{Logprice} = & 10.52 + 0.1318 * [OverallQuality] + 0.0001861 * [Finishedsquarefeet] \\ & + 0.0002880 * [Groundlivingarea] + 0.1045 * [GarageCars]\end{aligned}$$

3.3 Goodness of the Final Model:

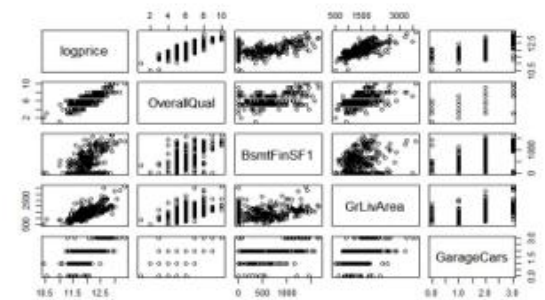
Firstly, verification was conducted to assess whether the model satisfies both Condition 1 and Condition 2.

Train model:

Condition1:

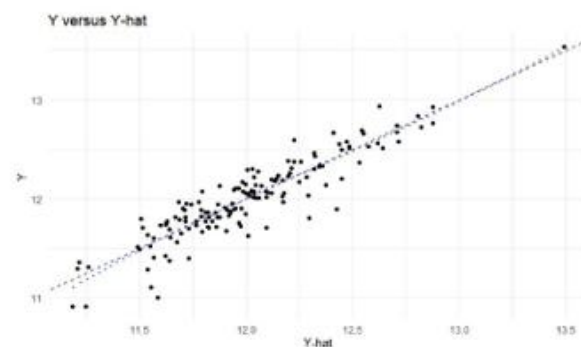


Condition2



Test model:

Condition1:



Condition2:

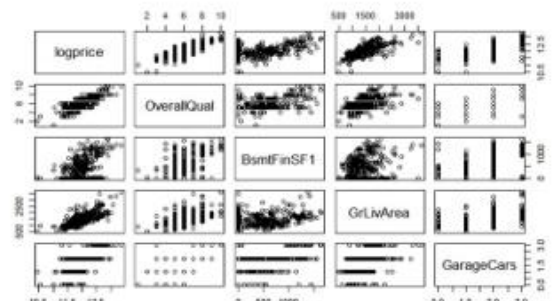


Figure 4: Diagnostic Plots for Assessing the Satisfaction of Condition 1&2 for Train & Test model

[First row: predicted vs. actual values for the training model; X1 vs. X2 vs. X3 vs. X4 for the training model

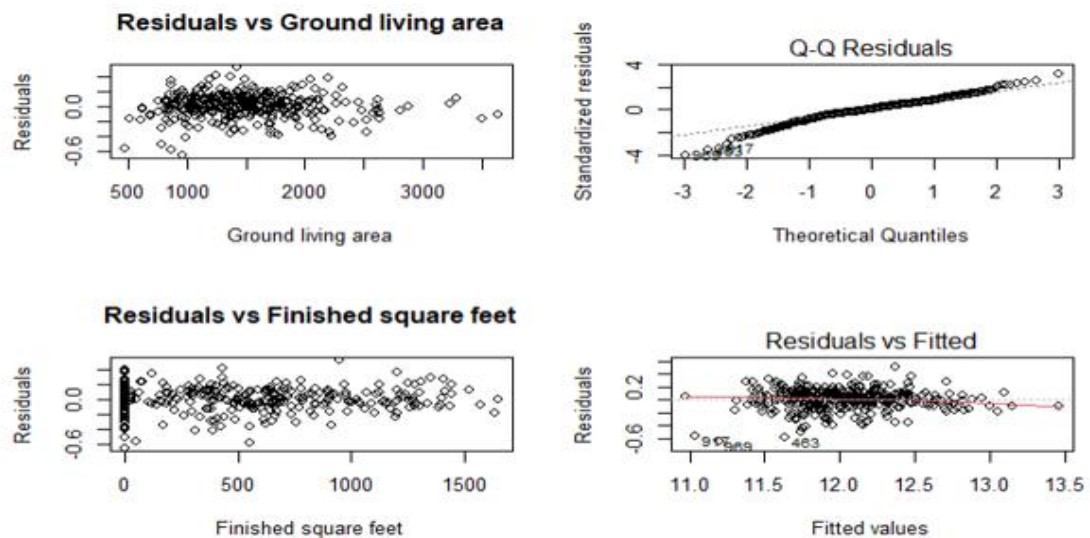
Second row: predicted vs. actual values for the test model; X1 vs. X2 vs. X3 vs. X4 for the test model]

The left side graph shows the relationship between the response and the fitted value, by observing that the points are close to or on the line, then condition 1 can be perfectly satisfied. Regarding the linearity relationship, the points in the right plot

show that the numerical variables are almost linear and there is no non-linear pattern such as cos and sin, which means that condition 2 holds.

Secondly, the model assumptions were evaluated through the analysis of residual plots and QQ-plots.

Train model:



Test model:

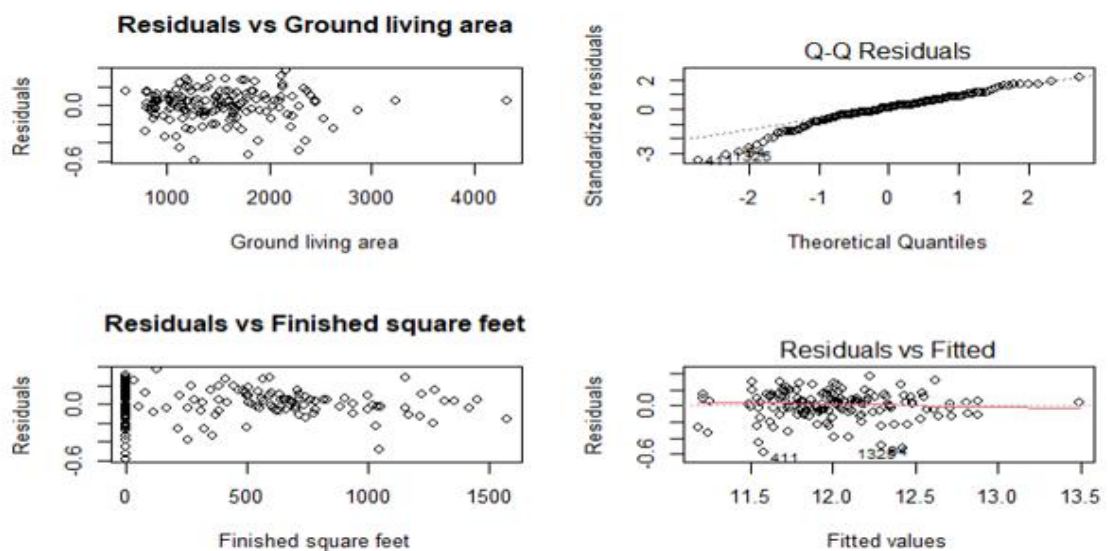


Figure5: residual plot & QO plot for Train & Test model

[First row: residual vs. Ground living area for Train model; QQ Residual Plot for Train model;
Second row: residual vs. Finished square feet for Train model; Residual vs. \hat{y} for Train model;
Third row: residual vs. Ground living area for Test model; QQ Residual Plot for Test model;
Fourth row: residual vs. Finished square feet for Test model; Residual vs. \hat{y} for Test model]

According to the residual plot on the left in Figure 5, there is no systematic pattern, cluster pattern or fan-shaped spread between this residual and fitted value or the values of the predictor variable. In addition, almost all points on the QQ-plot on the right are close to or on a straight line, except for a very few deviation points. which mean our model 3 can satisfy the 4 assumptions.

Thirdly, identification of problematic observations in Model 3 is required process as well. Based on the analysis of model 3, it can be determined that there are no problematic observations in terms of outliers and influential points, which means that the data points do not exhibit extremes that affect the results. However, there are 28 leverage points for which the fitted values exceed the corresponding values in the training set.

Finally, a partial F-test was performed to select the preferred model. By randomly selecting a subset of variables from model 3, a new model was created, called model 4. Then, the ANOVA formula was used again to compare the model 3 and model 4, and because the p-value was less than 0.05, model 3 was selected.

4. Discussion Section


	Train model	Test model
Overall Quality	0.132	0.135
Finished square feet	0.000186	0.000196
Ground living area	0.000288	0.000203
Garage Cars	0.105	0.164
VIF for Overall Quality	1.872639	1.796140
VIF for Finished square feet	1.069862	1.057714
VIF for Ground living area	1.487363	1.745126
VIF for Garage Cars	1.603288	1.400050
Adjusted R^2	0.8379265  (Ctrl) ▾	0.8280718
AIC	-1265.12	-529.07
AICc	-1264.8795488	-528.4911017
BIC	-1237.9754281	-507.0066002
SSRes	9.2113719	4.1793924

Figure 6: Comparison of Coefficients and performance for Train & Test Models

4.1 Final Model Interpretation and Importance:

$$\widehat{\text{Logprice}} = 10.52 + 0.1318 * [\text{OverallQuality}] + 0.0001861 * [\text{Finishedsquarefeet}] \\ + 0.0002880 * [\text{Groundlivingarea}] + 0.1045 * [\text{GarageCars}]$$

The final model shows that a one unit increase in overall quality rating is associated with a 13.18% increase in average price, holding all other factors constant. This finding has important implications for real-life housing situations. For example, the 13.18% price increase due to an increase in overall quality rating underscores the importance buyers place on quality when making home purchase decisions. This means that investing in improving the quality of a home, for example through renovations or better building materials, can provide sellers with a substantial return on their investment. Similarly, for each additional square foot of finished area, we

observe a price increase of 0.0186%. This finding suggests that larger finished areas contribute positively to the overall value of a property. Buyers are willing to pay a premium for additional living space, recognizing its potential for enhanced comfort and functionality. Other variables vary in a similar way. The purpose of this paper is to identify the factors that influence housing prices and how these factors affect housing prices, so the model is on the right track for us to apply.

4.2 Limitations of the Analysis:

Based on the results obtained, it can be seen that there are some differences in the coefficients of the garage cars between the train model and the test model. One possible contributing factor could be the inherent differences in the data between the train and test sets (by Figure2). In addition, the slight decrease in R^2 represents the model prediction performance is not enough perfect in the test, which I think may be due to the existence of some other variables that have an impact on price, such as garage area and garage type. Because of this, future analyses could benefit from the inclusion of these variables to address these minor issues and potentially improve the model performance.

Reference List

- House Price Prediction Using Regression Techniques: A Comparative Study.* (n.d.). House Price Prediction Using Regression Techniques: A Comparative Study | IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8882834?casa_token=A43kZl5qRjIAA AAA:adTsjZJsYQNrVLJanohq2Lu8IMTUY0spssbnmzBLxk8lCMLP5KKZEmMv5KF0PlvTINvRYw
- H., & Chen, N. (2022, May 5). *House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis*. House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis. <https://doi.org/10.1155/2022/9590704>
- Housing Price Prediction via Improved Machine Learning Techniques.* (2020, July 27). Housing Price Prediction via Improved Machine Learning Techniques - ScienceDirect. <https://doi.org/10.1016/j.procs.2020.06.111>
- House Prices - Advanced Regression Techniques | Kaggle. (n.d.). House Prices - Advanced Regression Techniques | Kaggle. <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data?select=train.csv>
- Kaushal, A., & Shankar, A. (2021, April 27). House Price Prediction Using Multiple Linear Regression. House Price Prediction Using Multiple Linear Regression by Anirudh Kaushal, Achyut Shankar:: SSRN. <https://doi.org/10.2139/ssrn.3833734>
- H., & Zhang, Q. (2021, October 29). Housing Price Prediction Based on Multiple Linear Regression. Housing Price Prediction Based on Multiple Linear Regression. <https://doi.org/10.1155/2021/7678931>
- Nelder, J. A. (1994, December 1). The statistics of linear models: back to basics - Statistics and Computing. SpringerLink. <https://doi.org/10.1007/BF00156745>