Analysis of Characteristics that Drive Higher Selling Prices in Housing

Contribution

Every member contributed an even amount of work.

Introduction

The primary objective of this project is to investigate and identify the key factors affecting house valuation through a multiple linear regression model. This research addresses the critical question of which characteristics significantly influence higher selling prices. Such an exploration is vital for understanding the intricate dynamics of the real estate market. It aims to unravel how various elements, including the overall condition, lot size, year of construction, number of rooms above ground, and land contour, perform a linear relationship with a house's price. Housing remains a topic of considerable debate, and housing prices are a crucial indicator of the current economic climate since it is regarded as one of the essential needs of people. This analysis is intended to provide valuable insights into the determinants of property values, which are essential for both buyers and sellers to make knowledgeable decisions within the real estate market.

Lot size, which is a crucial factor of housing size, is a factor of significant concern for people to determine the price of houses. According to the article by Zhou, housing size has a notable impact on house prices, more so than policy variables (Zhou et al., 2018). In a related study, Shao identified a linear relationship between the age of a house and its price in New Taipei City, Taiwan, using a multiple linear regression model, paralleling the approach of this study (Shao, 2022). Furthermore, another article demonstrated a strong linear correlation between the number of rooms above ground and housing prices in Boston, underscoring the critical role of physical characteristics in determining the value of properties (Yilin, 2020).

Utilizing the dataset named “The Ames Iowa Housing Data”, which includes 2932 observations, we aim to gain a more comprehensive understanding of the factors affecting housing prices. This will be achieved by employing a linear regression model in the following research.

Methods

Initially, the process begins with data cleaning, where we remove missing data from the dataset then integrating the refined data into R. This is followed by dividing the dataset into two equal parts: 50% for training and 50% for testing, selected randomly. The training dataset is then utilized for subsequent analysis.

The next phase involves developing the model using the training data. This step assesses the relationship between predictors and the response variable, examining the model's linearity by evaluating uncorrelated errors, linearity violations, normality, and non-constant variance.

In the third phase, the significance level is evaluated using the training data to construct the testing model. A p-value greater than 0.05 suggests that the observed results are not merely due to chance, whereas a p-value less than 0.05 indicates a good fit of the testing model, reflecting its complexity. The fourth step involves validating the variance through an

F-test, crucial in ANOVA and regression analysis. This test checks if the data accurately represents its intended purpose. If the test statistic does not fall below the critical value at the chosen significance level, the null hypothesis is rejected, and an ANOVA F-test is conducted to compare the residual sum of squares (RSS) with the total sum of squares (SS) to confirm the presence of a linear relationship.

Following a successful F-test, where predictors are found significant, indicating a linear relationship with the response variable (such as house sale price), the partial F-test is employed. This test compares the original model with a reduced model containing only significant predictors to determine which is more effective. A small partial F-test result suggests retaining the full model over the reduced one.

In the fifth step, the data undergoes further scrutiny for Conditions 1 and 2. Condition 1 (C1) pertains to the linearity assumption, asserting a linear relationship between the predictor and the response (e.g., housing price). This is verified using plots like scatter plots, residual plots, QQ plot and box plots to ensure the conditions of the conditional mean are met, where the average response is typically a linear combination of coefficients. Condition 2 (C2) relates to the homoscedasticity assumption, evaluated by ensuring that all predictor vs. predictor scatterplots satisfy the conditions of conditional mean predictors, indicating constant variance of the error term across all levels of the independent variable. Then lastly, we find outliers and leverage points of the model.

The final step is to assess the validity of both the training and testing models, we replicate the procedures applied to the training data on the test data. This involves a thorough examination of the data and its graphical representations to analyze the coefficients. By conducting this comparative analysis, we aim to ascertain whether the final model optimally represents the underlying relationship in the data.

Results:

Description of Data

Table 1: Summary for the data of the numerical variables

	Minimum	1st quantile	Median	Mean	3rd quantile	Maximum
Sale price	13100	129425	160000	182544	215000	755000
Overall Condition	1.000	5.000	5.000	5.534	6.000	9.000
Lot Area	1300	7447	9530	10210	11494	215245
Total Rooms Above Ground	3.000	5.000	6.000	6.459	7.000	15.000
Year Remod/Add	1950	1966	1993	1985	2004	2010

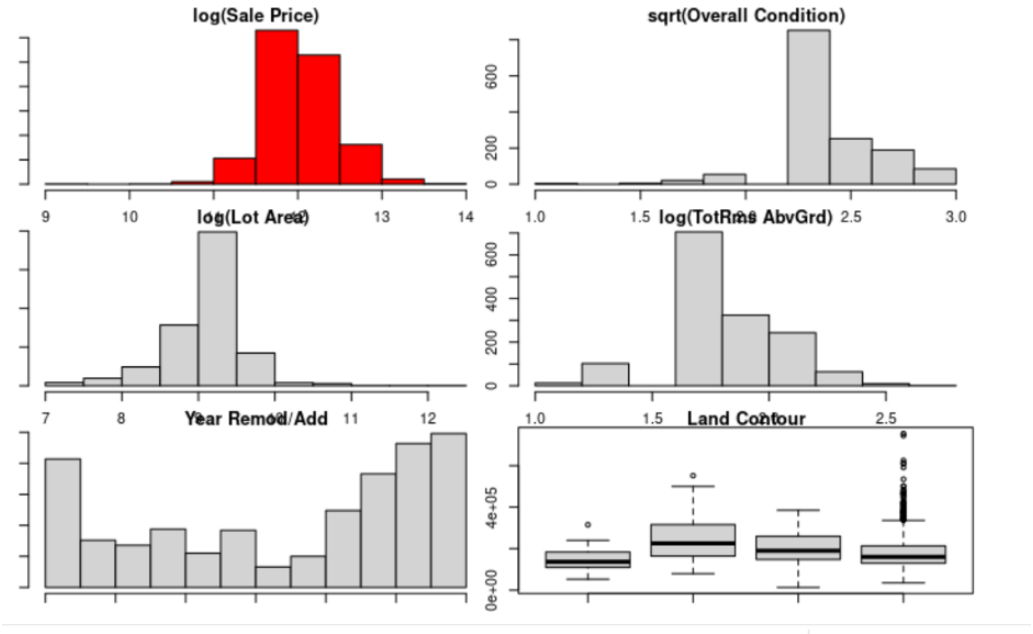


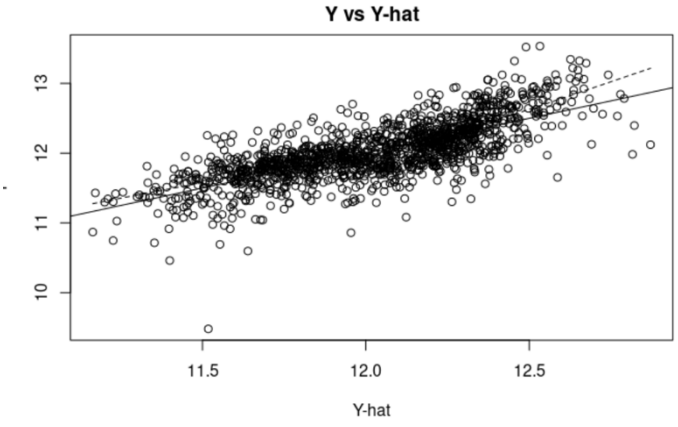
Figure 1: Histogram and scatter plots of numerical variables and boxplots of categorical variables

The histograms above display the distribution of numerical variable observations. It is evident from these histograms that the variables representing lot area and the total number of rooms above ground exhibit the most normal distribution. Additionally, the four boxplots present the scores for each category within the type of land contour, a categorical variable. However, the last boxplot is less significant due to the presence of an excessive number of outliers.

<u>Checking Final Model:</u>

1) Check if the models satisfy condition 1 and 2:

Condition 1



Condition 2

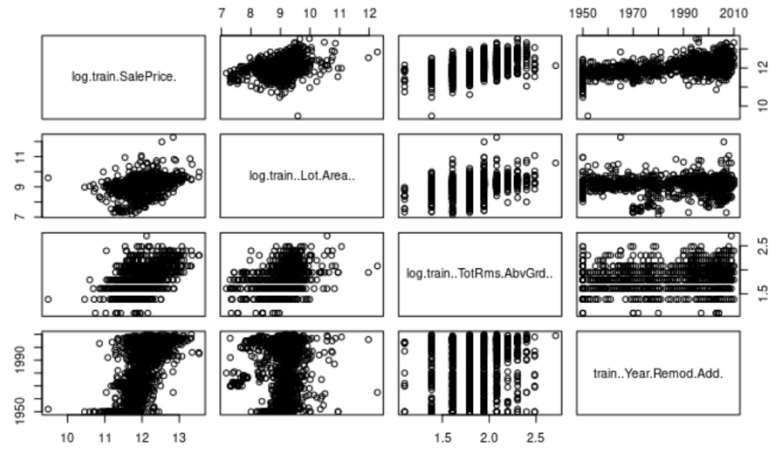


Figure 2 and 3: Plots of checking the satisfaction for condition 1 and 2

Figure 2 (the upper image) displays a scatter plot of the response variables against the fitted values, where the points are close to or on a straight line, thereby fulfilling condition 1. The lower figure, used to verify the linearity assumption, demonstrates that the points for the numerical variables align almost linearly, indicating that condition 2 is met.

2) Check if the model satisfy assumptions by residual plot and QQ plot:

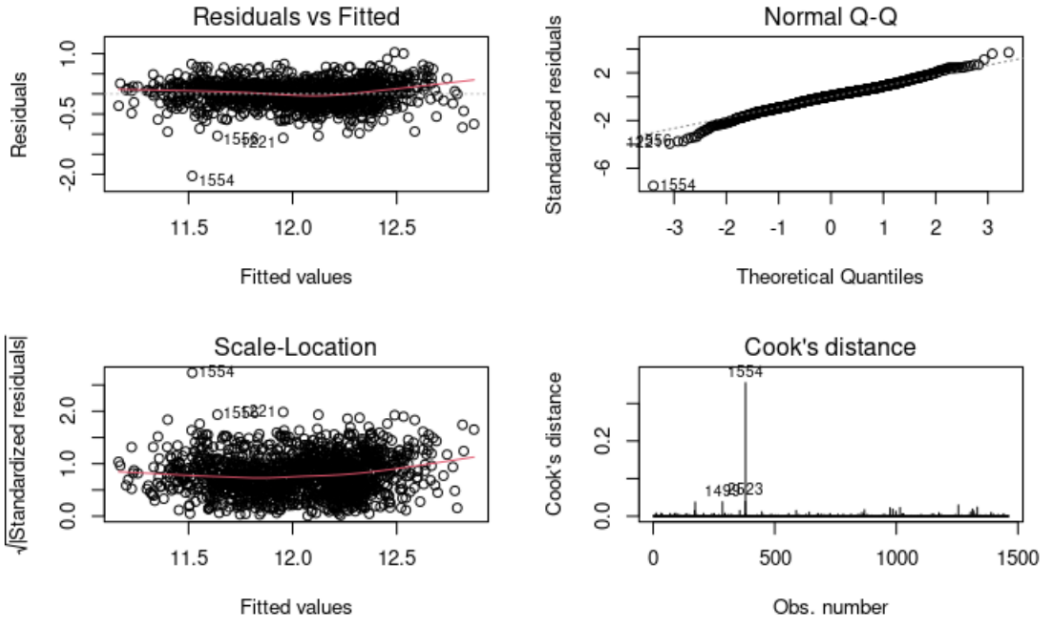


Figure 4: Residual plot and QQ plot

Analyzing the residual plot reveals no systematic, fanning, or clustered patterns. Additionally, all the points to the straight line in the QQ plot, with minimal deviations, suggests that this model meets the four key assumptions: linearity, normality, constant variance, and uncorrelated errors.

3) Checking if there are any observation that has problem in the final model

There are 2 outlier points that exist whose standard residuals are not between –4 and 4. And there are many leverage points, in which the fitted values are bigger than the value in the training set.

4) Using a partial F-test to choose the preferred model.

A new model is created, incorporating a random selection of variables, which includes part of the variables in the final model randomly. Then ANOVA formula is used to determine which model is better. Given the p-value is lower than 0.05, the final model is chosen.

Discussion:

Table 2: The coefficients of the testing model and the final model:

	Lot Area	Total Rooms Above Ground	Year Remodel	Land Contour (Hills)	Land Contour (Low)	Land Contour (Level)
Final Model	$1.644*10^{-1}$	$5.591*10^{-1}$	$1.011*10^{-2}$	$3.142*10^{-1}$	$1.255*10^{-1}$	$1.289*10^{-1}$
Testing model	0.1985	0.5012	0.0094	0.3498	0.1517	0.1133

The table presented above shows the differences in coefficients for the predictors across two models, suggesting that the final model offers an optimal fit to the data.

The Final Model

$$\text{Sale Price} = -1.070*10^{1} + 1.644*10^{-1}(\text{Lot Area}) + 5.591*10^{-1}(\text{Total Rooms Above Ground}) + 1.011*10^{-2}(\text{Year Remodel}) + 3.142*10^{-1}\text{Land Contour (Hills)} + 1.255*10^{-1}\text{Land Contour (Low)} + 1.289*10^{-1}\text{Land Contour (Level)}$$

The final model shows that when other factors remain the same, the sale price will decrease by $1,070 for each additional unit of 10,000 square feet in Lot Area, increase by $0.164 for each additional room in Total Rooms Above Ground, increase by $0.011 for each additional year since the remodel, increase by $0.314 for houses with Land Contour classified as Hills, increase by $0.125 for houses with Land Contour classified as Low, and increase by $0.129 for houses with Land Contour classified as Level. The coefficients represent the estimated changes in the sale price associated with one-unit changes in the respective variables, keeping all other variables constant. The goal of the paper is also to find out the factors that affect the price of the house, so the model is in the correct direction.

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
Overall Cond	1	1462	5.53	1.10	5	5.43	0.00	1	9
Lot Area	2	1462	10210.33	8565.90	9530	9487.50	3035.62	1300	215245
TotRms AbvGrd	3	1462	6.46	1.61	6	6.34	1.48	3	15
Year Remod/Add	4	1462	1984.83	20.51	1993	1986.32	19.27	1950	2010

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
Overall Cond	1	1468	5.59	1.13	5	5.50	0.00	1	9
Lot Area	2	1468	10085.77	7134.11	9360	9475.09	3071.95	1470	164660
TotRms AbvGrd	3	1468	6.43	1.54	6	6.33	1.48	2	14
Year Remod/Add	4	1468	1983.71	21.19	1992	1984.93	22.24	1950	2010

Limitations:

In our model, we can find lingering issues with the final model at when we are attempting to find leverage points, and outlier, although this resulted in our unperfect model. In this case the impact of this lingering issue would result in the predictive accuracy of the model, as well as real world decision making, in the case it is the price of houses. , We do not have the right to remove the points from the data, since it is unethical, and it disrupts the accuracy of the data.

Ethics

In our research project, which delves into the dynamics of housing prices, several critical ethical considerations must be meticulously addressed to uphold the integrity and reliability of our findings. Firstly, the accuracy and fairness of the data are paramount. It is essential to ensure that our data collection methods are free from biases that could skew the results. This involves a careful and objective selection of data, avoiding any tendencies that might lead to erroneous conclusions.

Particularly when examining regional house price factors, there is a risk of focusing on variables that may not be reflective of the broader housing market. In such cases, it is crucial to provide detailed, context-specific information. This approach helps in mitigating any misunderstandings and ensures that our analysis is relevant and appropriately tailored to the specificities of the regions under study.

Legal compliance in the use of data is another vital aspect. We must ensure that our data sourcing, handling, and analysis are in strict adherence to all applicable laws and regulations. This includes respecting privacy laws and ensuring that the use of data does not infringe upon the rights or cause discomfort to the data sources.

Moreover, in the context of statistical analysis, it is imperative to be forthright about the statistical methods employed, the assumptions made, and their implications for the interpretation of the results. This includes a clear exposition of how predictive models were used, the variables included, and the rationale behind these choices.

References

Shao, M. (2022). Factors Affect the House Price. *2022 14th International Conference on Computer Research and Development (ICCRD)*, 136–139. https://doi.org/10.1109/ICCRD54409.2022.9730362

Yilin, N. (2020). Linear Regression Model of House Price in Boston. *Science Discovery (Print)*, *8*(3), 52-. https://doi.org/10.11648/j.sd.20200803.12

Zhou, J., Zhang, H., Gu, Y., & Pantelous, A. A. (2018). Affordable levels of house prices using fuzzy linear regression analysis: the case of Shanghai. *Soft Computing*, *22*(16), 5407–5418. https://doi.org/10.1007/s00500-018-3090-4