Housing in Ames*

Yanzun Jiang, Siyuan Lu, Yi Tang

December 5, 2024

Table of contents

1 Introduction **3**

2 Method **3**

3 Results **4**
3.1 Model 1 5
3.2 Model 2 5
3.3 Model 3 5
3.4 Model 4 5
3.5 Model 5 6
3.6 Conclusion 6

4 Conclusion and Limitations **7**
4.1 Conclusion 7
4.2 Limitation 7

Appendix **8**

A Graphs of Conditions and Assumptions Checking **8**
A.1 Conditions 8
A.2 Assumptions 10

B Ethics Discussion **14**

C Editing Demonstration **15**
C.1 Original Version of Introduction 15
C.2 Editted Version of Introduction 16

*Code and data supporting this paper is available at: https://github.com/Stary54264/Housing-in-Ames

C.3 Comments on the Process 16

D Contributions **16**

E R Packages and Dataset **17**

References **18**

1 Introduction

The research question we aim to answer is: what are the factors that influence house prices in Ames from 2006 to 2010? We would set up a linear regression model to answer this with house prices as the outcome (response). Factors that might affect the outcome (predictors) include area, quality, year of construction, facilities, etc.

By setting up the model, we can identify the factors that influence house prices, since linear regression allows us to quantify the relationship between predictors and responses, making it easier to interpret the impact of each factor alone on house prices. Our primary goal is understanding the factors that influence historical house prices, so our focus would be on description rather than prediction.

We found several peer-reviewed articles that focus on similar problems with this paper. "Influencing Factors Analysis of House Prices Based on Multiple Linear Regression" concludes that housing prices are negatively correlated with completion costs, land acquisition prices, residents' disposable income, and population density (Wang (2013)). This article provides some characteristics other than what we use that can also influence house price in national scope.

In "Dynamic Relationships Between Commodity Prices and Local Housing Market", the researchers examines the significant nonlinear relationship between agricultural commodity prices and the housing prices (Liang, Fan, and Hu (2021)). Another research, "Non-Linear Relationships Between House Size and Price", clarifies the non-linear relationship between size and price (Feng et al. (2021)). These two researches explain the non-linear relationship, between house price and other factors, providing more insights into ways that factors might affect house prices.

2 Method

The process begins by constructing the initial multiple linear regression model (Model 1) using selected predictors that are theoretically or empirically linked to the response variable. This model serves as a baseline for exploring the underlying relationships between predictors and the outcome. It provides a starting point to identify how well the chosen predictors explain the variance in the response variable.

After Model 1 is developed, it is essential to validate its reliability by checking two conditions (conditional mean response and conditional mean predictor) and four assumptions (linearity, uncorrelated errors, constant variances, and normality). Graphical tools like residual plots, and Q-Q plots are applied. This diagnostic step identifies potential issues that may compromise the model's interpretability or predictive accuracy, guiding subsequent corrections.

If diagnostic checks reveal violations, variance-stabilizing transformation and Box-Cox transformation is performed. Variance stabilizing transformation is applied to variables with many zeros, which is to take their square root. Box-Cox transformation is applied to other variables, identifing an optimal power transformation or natural logarithmic transformation for the variables. Applying these transformations ensures that the model assumptions are more closely met, enhancing its validity. After the transformation, a refined model (Model 2) is constructed to account for the adjusted data distribution.

Model 2 is refined further by testing the significance of coefficients through hypothesis testing. This step examines whether the coefficients significantly contributes to the model or if it can be excluded without sacrificing explanatory power. A non-significant coefficient could simplify the model, while a significant one may provide valuable context for interpreting the predictors. Adjustments based on these tests lead to the construction of Model 3, a more refined version that aligns closely with theoretical considerations.

The all-subset selection process begins by systematically evaluating models with the same number of predictors. At this stage, the sole evaluation criterion is R^2_{adj}, which measures how well each model explains the variance in the response variable while accounting for the number of predictors. Models with higher R^2_{adj} are preferred.

Once the best model within each size category is identified, these "best of size" models are compared to select the overall best-performing model. This step incorporates multiple criteria to balance model fit and complexity. R^2_{adj} continues to be a key measure, but additional metrics like AIC, AIC_c, and BIC are introduced. They all emphasize model fit while penalizing for complexity, with AIC_c being useful for small sample sizes, and BIC imposing a stricter penalty for model size. Models with lower criteria values are preferred. If two or more models perform similarly based on these metrics, VIF is used. VIF quantifies multicollinearity, with lower values preferred.

By the end of this process, Model 4 is identified as the best subset of predictors. It balances explanatory power and simplicity.

To refine Model 2, automated selection tools could also employed. The method we used is stepwise selection, which assess predictor subsets, streamlining the selection process. It enhance efficiency while maintaining rigor in identifying the most suitable predictors. Model 5, is the culmination of this process, balancing explanatory power and simplicity.

3 Results

In this study, we followed a systematic process to build and evaluate multiple regression models. Models 1 through 5 were developed and assessed based on several performance criteria. The models were constructed step by step, starting from a full model and progressing through various techniques to improve model performance. Below, we describe the key steps for building Models 1 through 5 and the results for each model.

3.1 Model 1

Model 1 was constructed by fitting a simple linear regression model using the initially selected predictors. At this stage, the goal was to build a baseline model to examine the relationship between the predictors and the response variable. This model was used to assess the overall fit, and the results from this initial model informed subsequent steps. This model allowed us to identify areas where assumptions may have been violated, and highlighted the need for transformations in subsequent steps. Overall, Model 1 provided the groundwork for more advanced models that followed, which involved further diagnostic checks and adjustments.

3.2 Model 2

By checking the conditions (conditional mean response and conditional mean predictors), the validity of the results of residual plots would be ensured. After checking the assumptions (linearity, uncorrelated errors, constant variances, and normality) in Model 1 using residual plots and Q-Q plots, we observed that some violations exists. Plots are available in Section A.

To stabilize the variance and improve model fit, we applied variance-stabilizing transformation with variable with zeros, and Box-Cox transformation to other variables. This transformation improved the adherence to the assumptions.

After applying the transformations, we fit another model, Model 2, and checked the assumptions again. The transformed model performed better in terms of the assumptions compared to Model 1, suggesting that the the transformations have a positive effect.

3.3 Model 3

Model 3 was built by performing hypothesis testing for each coefficient in Model 2. The goal was to evaluate the significance of each predictor and determine whether they contributed meaningfully to explaining the response variable. We set $\alpha = 0.05$, and used the function `summary()` in R to get a summary table. In the p-value column in the table, we found out that all predictors have a p-value that is smaller than 0.05, meaning all coefficients are significant.

This process led to a refined set of predictors, where only those with statistically significant relationships to the response variable remained in the model. However, since all coefficients are significant, Model 3 is the same model as Model 2.

3.4 Model 4

Model 4 was constructed using an all-subset regression approach, where the models with same number of predictors are compared using R^2_{adj}. Then, the best of bests is selected using R^2_{adj}, AIC, AIC_c, and BIC. thses values are listed in Table 1. This technique allowed us to

compare multiple models and select the one that provided the best trade-off between fit and complexity.

Table 1: Criteria for Model of Each Size

Model	R^2_{adj}	AIC	AIC_c	BIC	VIF_{max}
Model A	0.68	-8516.80	-8516.79	-8504.85	0.00
Model B	0.74	-9169.57	-9169.56	-9151.65	1.02
Model C	0.77	-9510.59	-9510.56	-9486.69	1.82
Model D	0.78	-9660.55	-9660.52	-9630.67	2.03
Model E	0.79	-9767.32	-9767.28	-9731.47	2.12
Model F	0.80	-9814.22	-9814.17	-9772.40	2.23
Model G	0.80	-9816.69	-9816.62	-9768.89	2.23

The results showed that Model F and Model G has similar R^2_{adj}. However, Model G has smaller AIC and AIC_c, while Model F has smaller BIC. Since BIC has severe penalty on more predictors, we choose Model G to be Model 4. Additionally, the maximum VIF values were all below the threshold of 5, suggesting that multicollinearity was not a concern in the chosen model. The all-subset regression approach led to a model that balanced complexity and predictive accuracy effectively. Since Model 4 is also a full model, it is the same model as Model 2.

3.5 Model 5

Finally, Model 5 was built using an automated selection tool, specifically a stepwise selection procedure, which included both forward and backward selection. The tool was set to evaluate models based on AIC, and we start with a full model. This automated process iteratively added or removed predictors to find the best model.

The stepwise selection only take one step to find the best model - the full model (Model 5), since deleting any predictor would result in a larger AIC. The automated selection procedure confirmed that the full model was indeed optimal for explaining the variance in the response variable.

3.6 Conclusion

In summary, Models 3, 4, and 5 are the full model. While each model incorporated slightly different techniques and adjustments, none of the models significantly outperformed the others. So, we are confident to conclude that the optimal model to answer the research question is just the full model with variables transformed.

4 Conclusion and Limitations

4.1 Conclusion

This study aimed to investigate the relationship between the predictors and the response variable using a series of linear models. The results is that all numerical predictors affects house prices. Based on the analysis of the final model, the linear relationship between the predictors and the response variable was confirmed, and the results support the hypothesis.

The final model we choose is Model 2 (Model 3, 4, and 5). Its formula is listed below. The result could be interpreted in this way: "For a one-unit increase in `ln(lot_area)`, we expected `ln(sale_price)` to increase by 0.187". The findings are not surprising, and they align with the literatures.

$$log(Y) = 8.57 + 0.187 \times log(X_1) + 0.280 \times X_2^{0.8} + 1.47 \times 10^{-113} \times X_3^{34} + \quad (1)$$

$$3.35 \times 10^{-3} \times X_4^{0.5} + 0.0229 \times X_5^{0.33} + 0.0348 \times X_6^{0.25} + (-1.03 \times 10^{-3}) \times X_7^{0.5} \quad (2)$$

4.2 Limitation

Despite the insights gained through this process, the study has several limitations. One significant limitation is the assumption of linearity inherent in multiple linear regression. This assumption may not hold in all cases, particularly when the true relationship between the variables is non-linear. In such cases, linear regression models may be less accurate, and alternative techniques such as non-linear regression or machine learning methods may be more appropriate. Additionally, the models evaluated in this paper were limited to a specific set of predictors and may not fully account for other potentially relevant variables. The inclusion of more variables, or the use of more complex models such as interaction terms, could improve the predictive power of the model.

Another limitation is that in this study, we did not perform a train-test split, which means the models were evaluated on the entire dataset rather than being validated on unseen data. This approach prioritizes understanding the relationships between variables and the underlying structure of the data, rather than focusing on predictive accuracy. While this method provides valuable insights into the model's fit and the significance of various predictors, it does not allow for an assessment of how well the model would generalize to new, unseen data. As a result, the primary objective here was to explore and interpret the model, rather than make predictions.

Appendix

A Graphs of Conditions and Assumptions Checking

A.1 Conditions

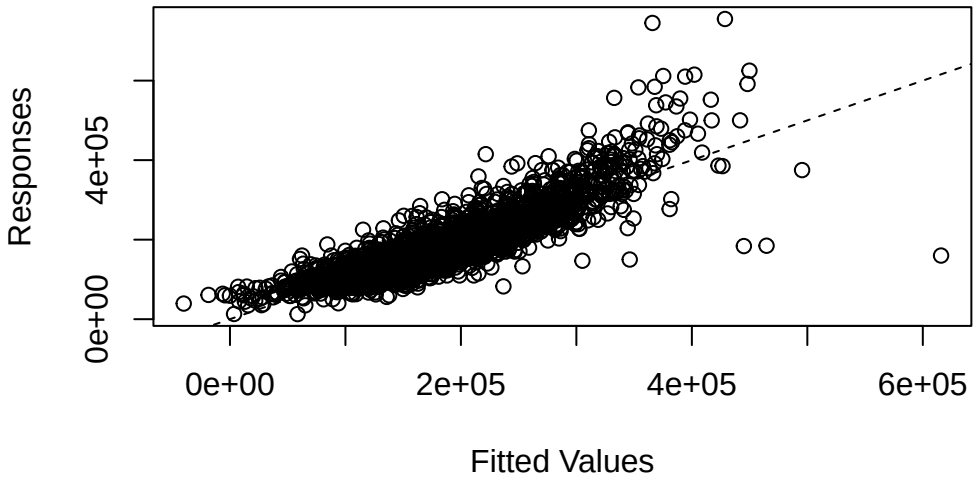

Figure 1: Responses vs. Fitted Values

We can see from Figure 1 that the pionts scattered along the diagonal, so conditional mean predictor is fulfilled.

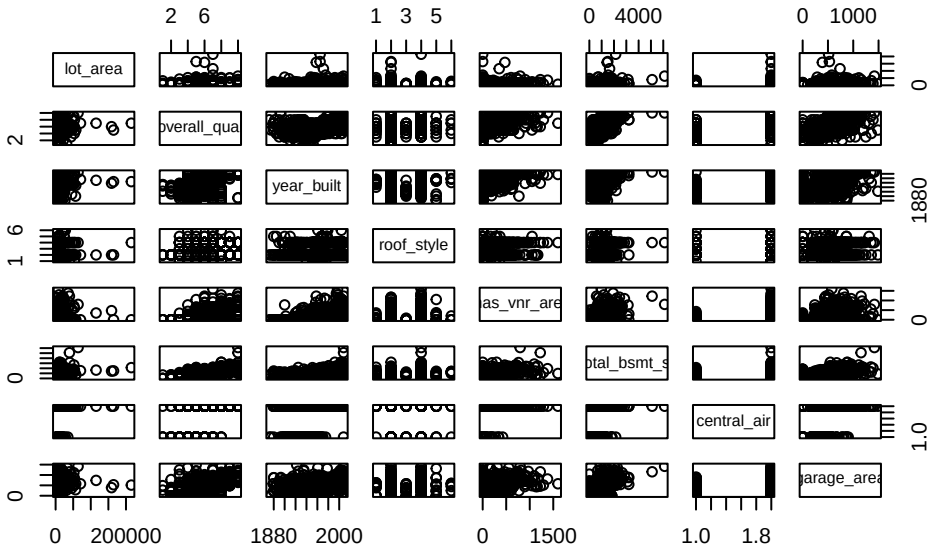

Figure 2: Pairwise Scatterplots of Predictors

We cannot see any non-linear patterns in Figure 2, so conditional mean response is fulfilled.

Since both conditions are satisfied, we are confident that the result of the residual plots would be valid.

A.2 Assumptions

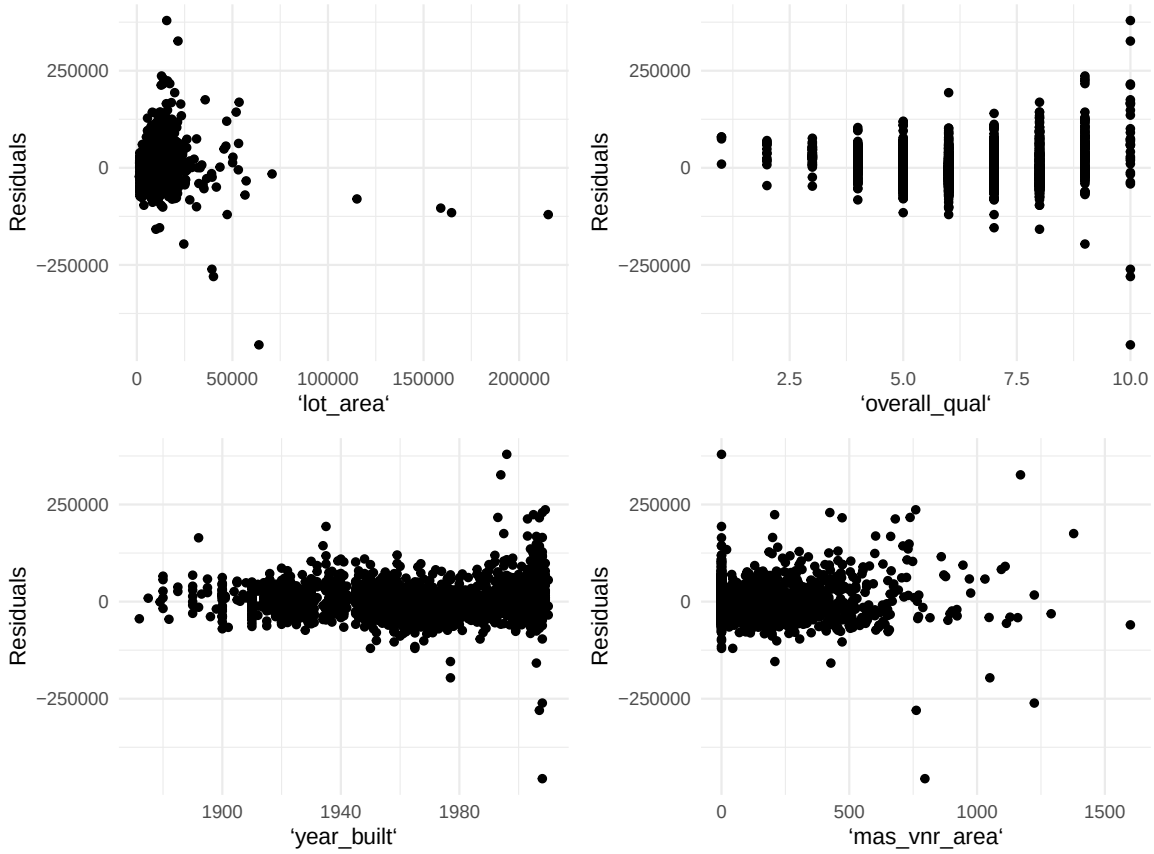

Figure 3: Residuals vs. Observation - I

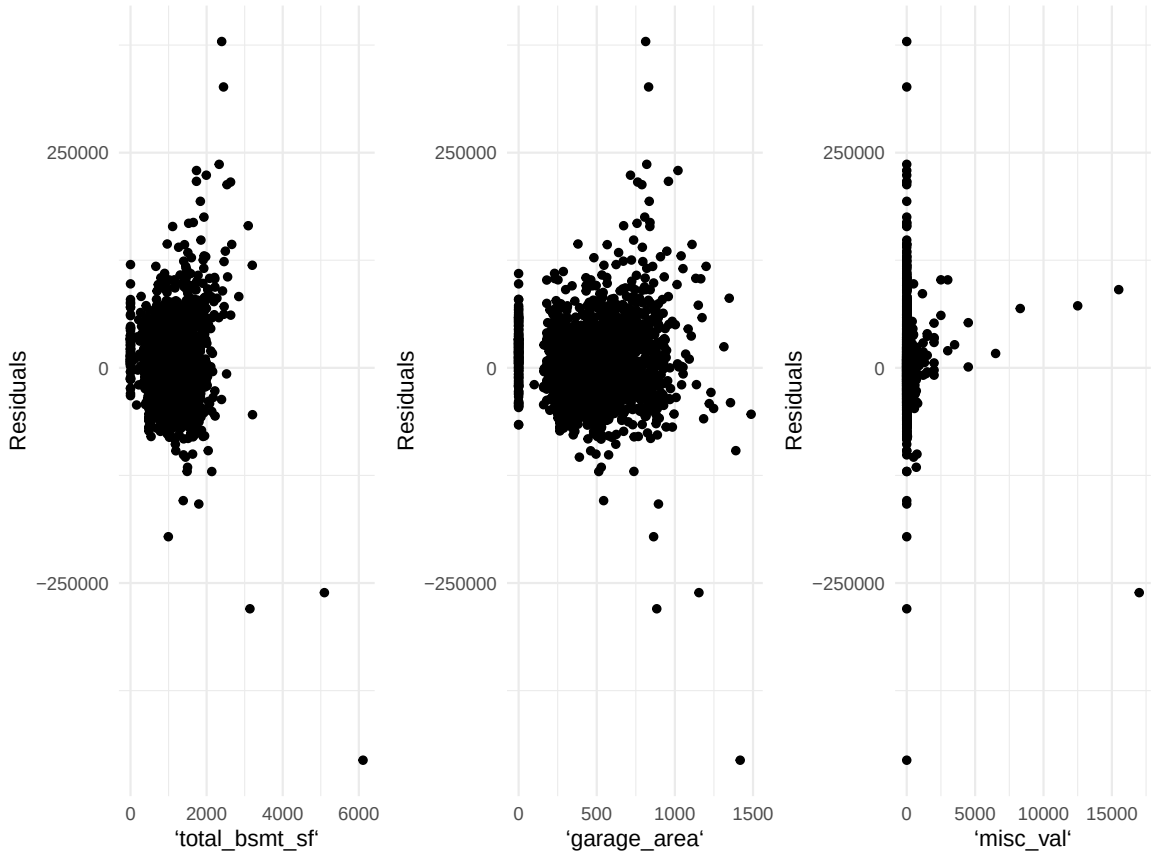

Figure 4: Residuals vs. Observation - II

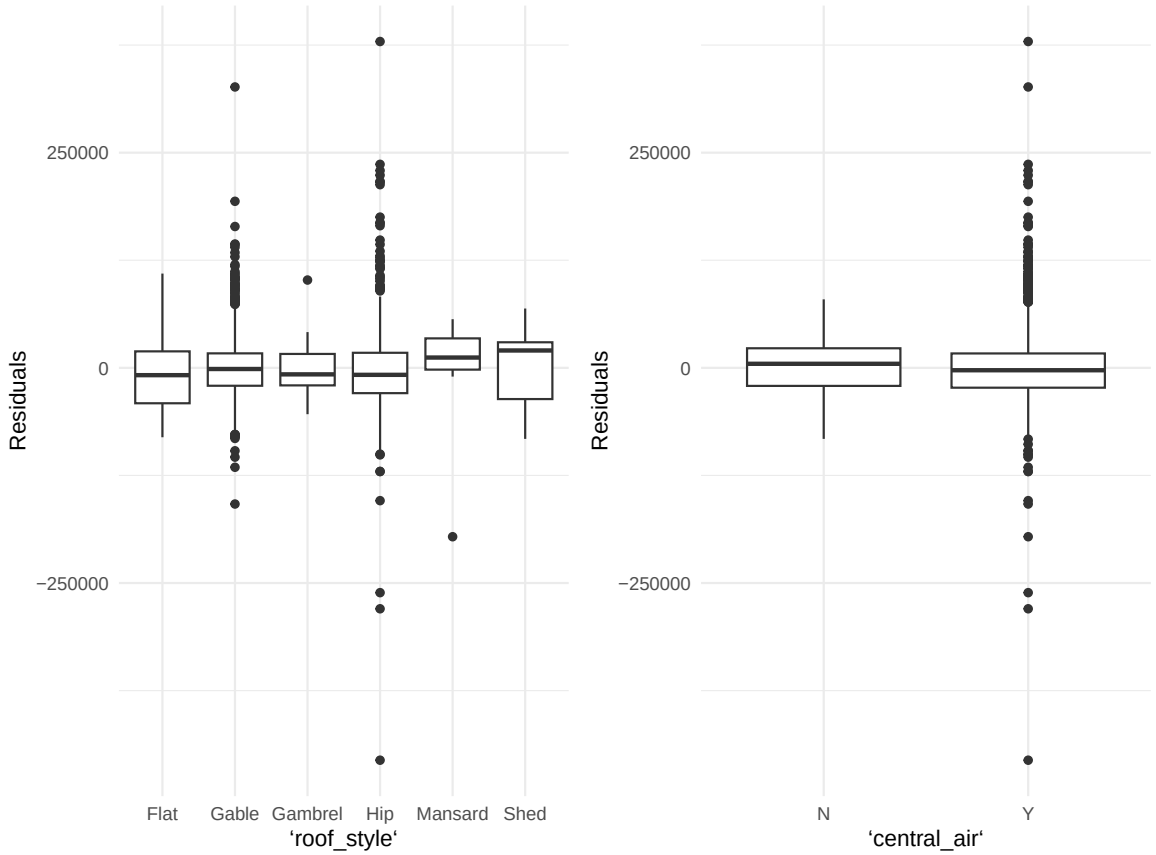

Figure 5: Residuals vs. Observation - III

We can see some violations of constant variance in Figure 3, Figure 4, and Figure 5 for variables except `year_built`, since the spread of the points becomes wider with increasing x. So, a transformation is needed for them.

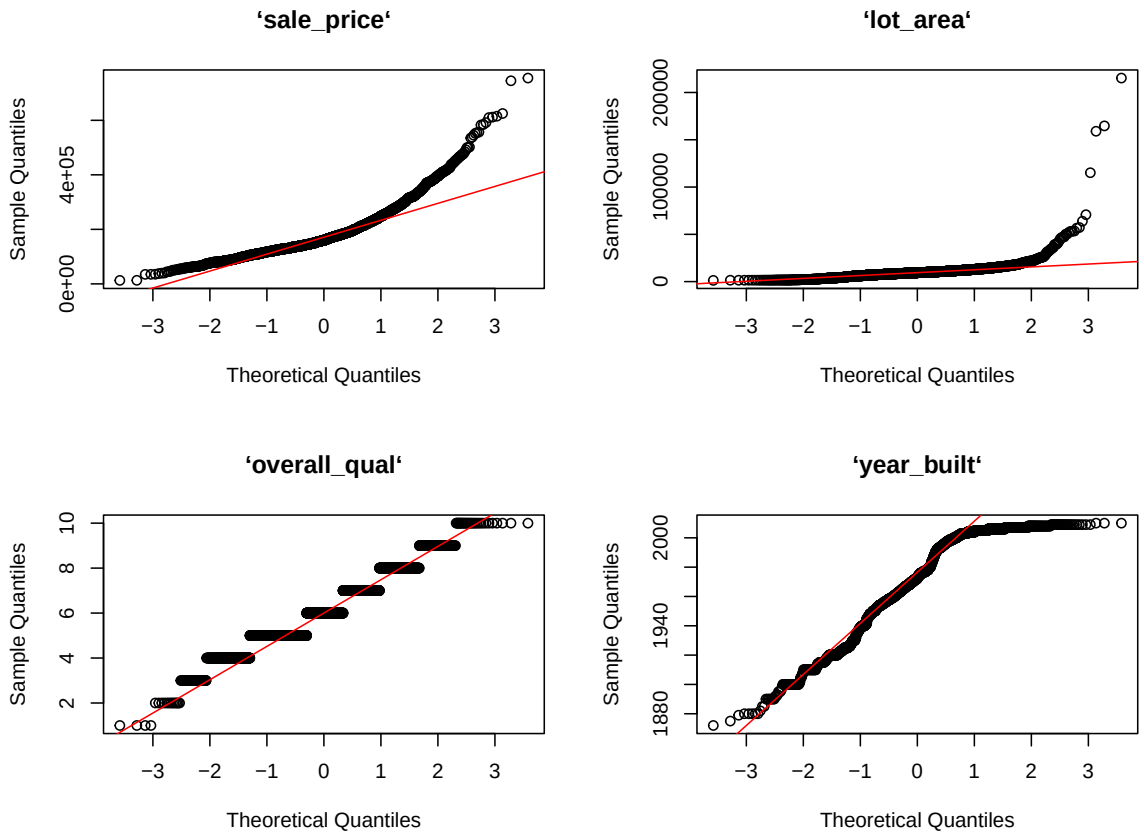

Figure 6: Q-Q Plot - I

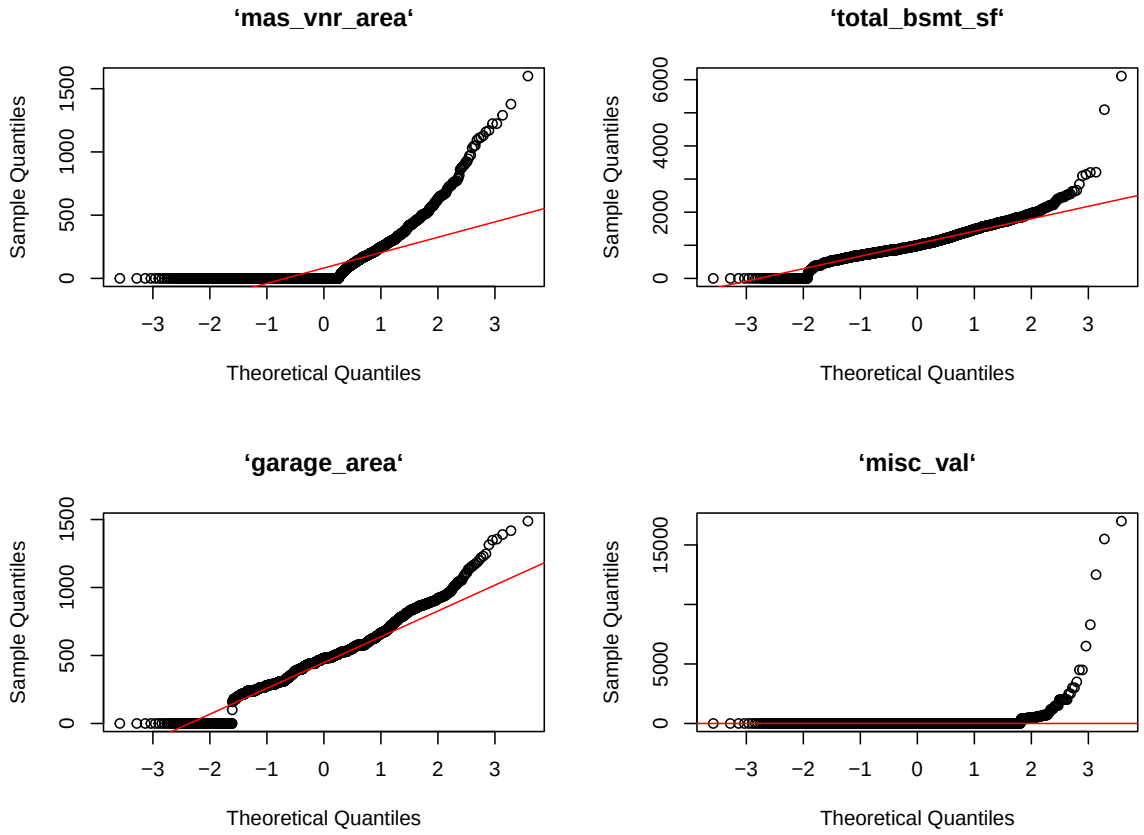

Figure 7: Q-Q Plot - II

We can see some violations of normality in Figure 6 and Figure 7 for variables except `total_bsmt_sf` and `garage_area`, since the points does not scattered along the diagonal. So, a transformation is needed for them.

B Ethics Discussion

Our data is collected from Ames City Assessor's Office (De Cock (2011)), then we cleaned the data to only keep some necessary key factors that is highly relavant to house prices. Raw and processed versions of the data from De Cock is published on Journal of Statistics Education in 2011. The cleaned data we are using includes some detailed information about housing characteristics, but does not contain personal identifiers.

The Ames housing dataset has been used widely, especially in the context of academic projects and machine learning competitions. It is often considered a modern alternative to the Boston

Housing dataset. The dataset is well-vetted and trusted by the data science community for its comprehensiveness and relevance.

The use of automated selection tools in academic research brings both opportunities and ethical challenges. Automated tools can significantly speed up research processes, but their use must be transparent, including acknowledging the specific tools and algorithms employed, as well as their limitations.

Ensuring fairness and avoiding skewed results require careful selection of training data and ongoing monitoring to detect and mitigate biases presented in AI training data. While automated tools can enhance productivity, they should not replace human judgment. By addressing these ethical considerations, we can leverage the benefits of automated selection tools while maintaining the integrity and fairness during our research practices.

C Editing Demonstration

C.1 Original Version of Introduction

The primary research question we aim to answer is: What are the key factors that significantly influence house prices in Ames from 2006 to 2010? Sale price of the house is the response variable. Predictor variables include area, overall quality index, year of construction, house facilities, and value of miscellaneous feature. By identifying and analyzing the factors that significantly influence house prices, our research can offer practical recommendations for various stakeholders, ultimately contributing to a more efficient and transparent real estate market in Ames. Similar analysis can be done to other cities to improve economic insights and investment decisions.

We will test whether there is a statistically significant linear relationship between certain property characteristics (predictors) and the sale price of houses (response) using linear regression. We use residual plots and a Q-Q plot to check the assumption. Linear regression provides coefficients that quantify the relationship between each predictor and the response variable, making it easier to interpret the impact of each factor alone on house prices. Our primary goal is to understand the impact of each predictor on house prices, so the focus should be on interpretability instead of precision.

We found several peer-reviewed articles that focus on similar problems with this paper. "Influencing Factors Analysis of House Prices Based on Multiple Linear Regression" concludes that housing prices are negatively correlated with completion costs, land acquisition prices, residents' disposable income, and population density (Wang (2013)). This article provides some characteristics other than what we use that can also influence house price in national scope.

In "Dynamic Relationships Between Commodity Prices and Local Housing Market", this research examines the significant nonlinear relationship between agricultural commodity prices

and the local housing market (Liang, Fan, and Hu (2021)). The research "Non-Linear Relationships Between House Size and Price" clarifies the nonlinear relationships between housing size and price (Feng et al. (2021)). These two researches explain another aspect, a non-linear relationship, between house price and other factors, providing more comprehensive information about house market for decision making of the developers, home purchasers, real estate appraisers, and the governments.

C.2 Editted Version of Introduction

The research question we aim to answer is: what are the factors that influence house prices in Ames from 2006 to 2010? We would set up a linear regression model to answer this with house prices as the outcome (response). Factors that might affect the outcome (predictors) include area, quality, year of construction, facilities, etc.

By setting up the model, we can identify the factors that influence house prices, since linear regression allows us to quantify the relationship between predictors and responses, making it easier to interpret the impact of each factor alone on house prices. Our primary goal is understanding the factors that influence historical house prices, so our focus would be on description rather than prediction.

We found several peer-reviewed articles that focus on similar problems with this paper. "Influencing Factors Analysis of House Prices Based on Multiple Linear Regression" concludes that housing prices are negatively correlated with completion costs, land acquisition prices, residents' disposable income, and population density (Wang (2013)). This article provides some characteristics other than what we use that can also influence house price in national scope.

In "Dynamic Relationships Between Commodity Prices and Local Housing Market", the researchers examines the significant nonlinear relationship between agricultural commodity prices and the housing prices (Liang, Fan, and Hu (2021)). Another research, "Non-Linear Relationships Between House Size and Price", clarifies the non-linear relationship between size and price (Feng et al. (2021)). These two researches explain the non-linear relationship, between house price and other factors, providing more insights into ways that factors might affect house prices.

C.3 Comments on the Process

D Contributions

Group contribution is available at https://github.com/Stary54264/Housing-in-Ames/graphs/contributors. Below is a more specific version of group contribution.

- Yanzun Jiang: Organized discussions and meetings; assigned tasks to group members; set up Github workspace for collaborating; built the model; checked conditions and assumptions; refined the model by transformation, hypothesis testing, all-subset selection, and automated selection tools; made the reference list; revised group member's work; combined group member's work together.
- Siyuan Lu: Set research question; searched and read peer-reviewed articles; introduced the project; checked data ethics; edited introduction.
- Yi Tang: Designed the poster.

E R Packages and Dataset

R (R Core Team 2023) was used to conduct this research. Packages used include `tidyverse` (Wickham et al. 2019), `knitr` (Xie 2014), `patchwork` (Pedersen 2024), `car` (Fox and Weisberg 2019), `leaps` (Fortran code by Alan Miller 2024), and `MASS` (Venables and Ripley 2002). the dataset used is Ames Housing dataset (Kuhn 2020) from De Cock (2011).

References

De Cock, Dean. 2011. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." *Journal of Statistics Education* 19 (3).

Feng, S.-T., C.-W. Peng, C.-H. Yang, and P.-W. Chen. 2021. "NON-LINEAR RELATIONSHIPS BETWEEN HOUSE SIZE AND PRICE." *International Journal of Strategic Property Management* 25 (3): 240–53. https://doi.org/10.3846/ijspm.2021.14607.

Fortran code by Alan Miller, Thomas Lumley based on. 2024. *Leaps: Regression Subset Selection*. https://CRAN.R-project.org/package=leaps.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

Kuhn, Max. 2020. *AmesHousing: The Ames Iowa Housing Data*. https://CRAN.R-project.org/package=AmesHousing.

Liang, J., Q. Fan, and Y. Hu. 2021. "Dynamic Relationships Between Commodity Prices and Local Housing Market: Evidence for Linear and Nonlinear Causality." *Applied Economics* 53 (15): 1743–55. https://doi.org/10.1080/00036846.2020.1845295.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wang, Cheng. 2013. "Influencing Factors Analysis of House Prices Based on Multiple Linear Regression." *Applied Mechanics and Materials* 415: 722–25.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.