Housing in Ames*

Yanzun Jiang, Siyuan Lu, Yi Tang

December 5, 2024

Table of contents

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

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 (500)

Our process begins by developing an initial multiple linear regression model (Model 1) using selected predictors. This model serves as the baseline for evaluating the relationship between the predictors and the response variable. It also provides the foundation for subsequent diagnostic checks and refinements.

After building Model 1, the two conditions (conditional mean response and conditional mean predictors) and four assumptions (linearity, uncorrelated errors, constant variance, and normality) are verified. Graphical diagnostics are included in Section A.

Variance stabilizing transformations and Box-Cox transformations are applied to the variables. These transformations ensure that the data does not violate any assumption. Following this step, a new model (Model 2) is developed using the transformed data.

Model 2 is further refined by conducting hypothesis testing for the coefficients. This step evaluates the statistical significance of them, ensuring interpretability and relevance. However, since all coefficients are significant, we would keep Model 2.

Model 2 undergoes an all-subset selection process to identify the best combination of predictors. Selection criteria for models with same size is R^2_{adj}. To choose the best of the best, R^2_{adj}, AIC, AIC_c, and BIC are used. If there are similar models, use VIF to check for multicollinearity. The resulted model is Model 3.

3 Results (750)

4 Conclusion and Limitations (350)

Appendix

A Graphs of Conditions and Assumptions Checking

A.1 Conditions

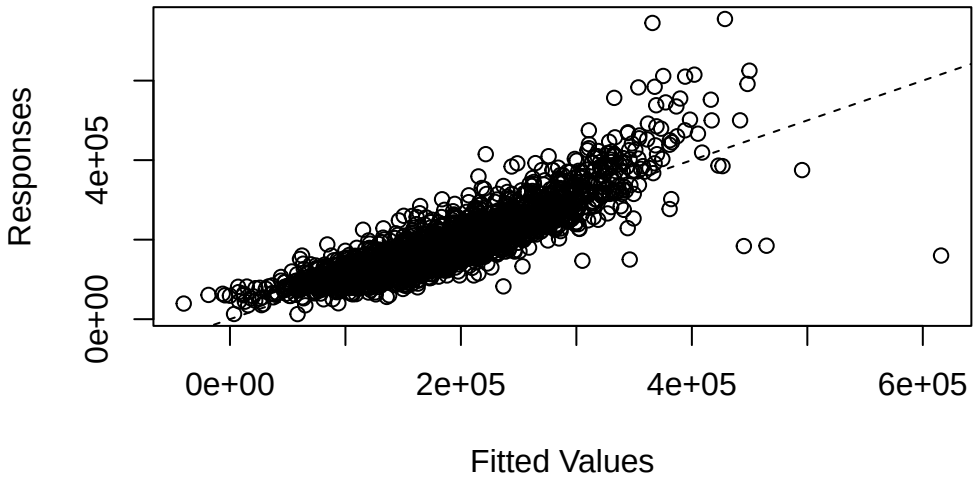

Figure 1: Responses vs. Fitted Values

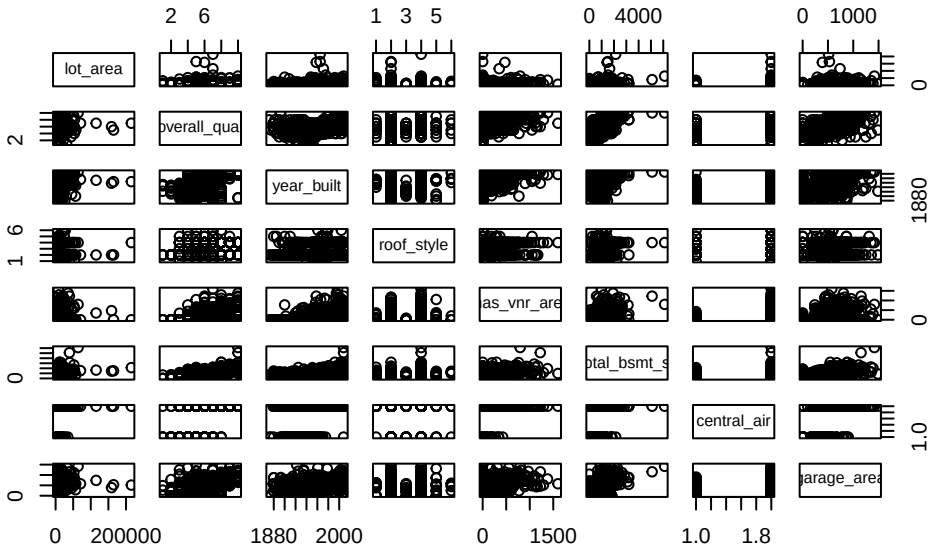

Figure 2: Pairwise Scatterplots of Predictors

We can see that both conditions are satisfied, so the residual plots would be valid.

A.2 Assumptions

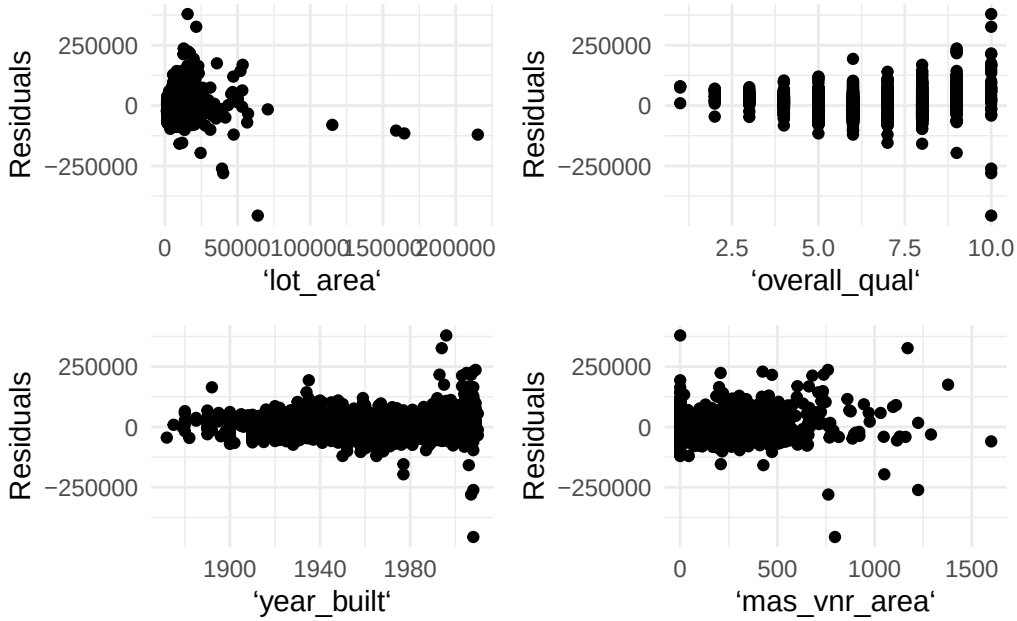

Figure 3: Residuals vs. Observation - I

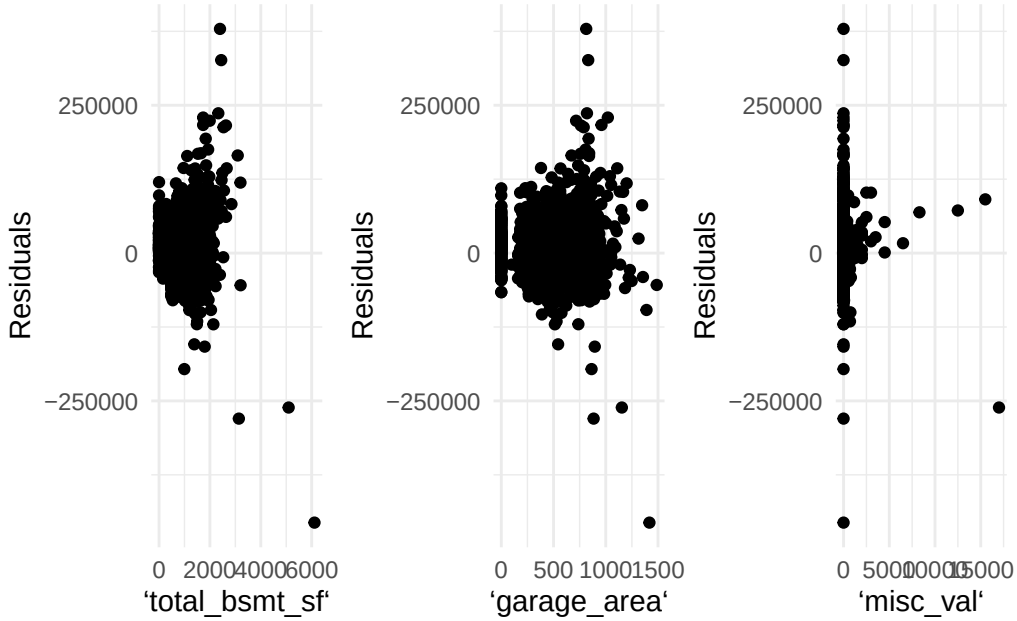

Figure 4: Residuals vs. Observation - II

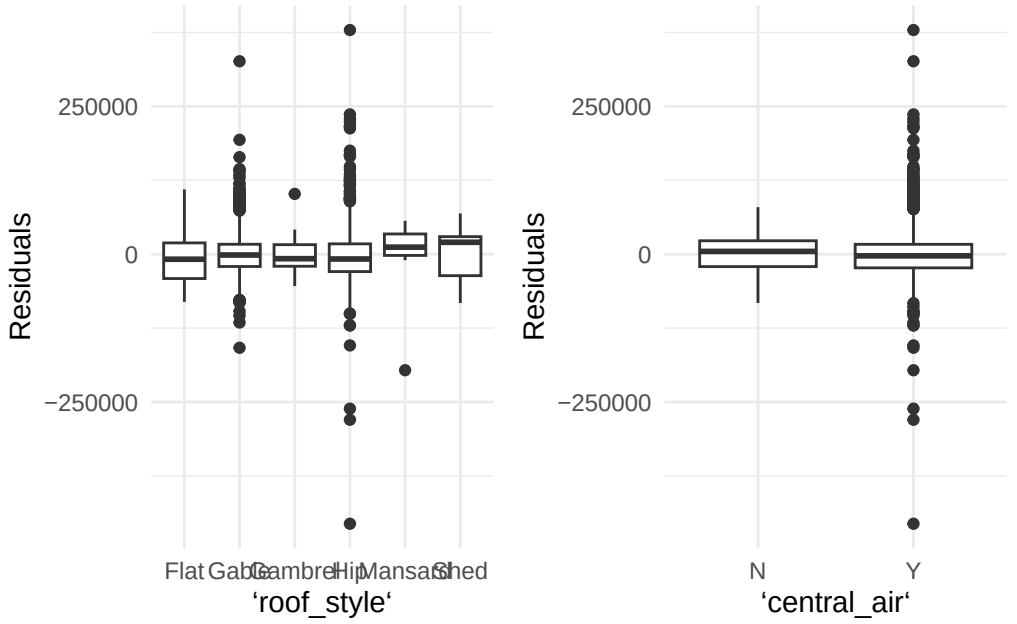

Figure 5: Residuals vs. Observation - III

We can see some violations of linearity and constant variance in `lot_area`, `overall_qual`, `mas_vnr_area`, `total_bsmt_sf`, `garage_area`, and `misc_val`.

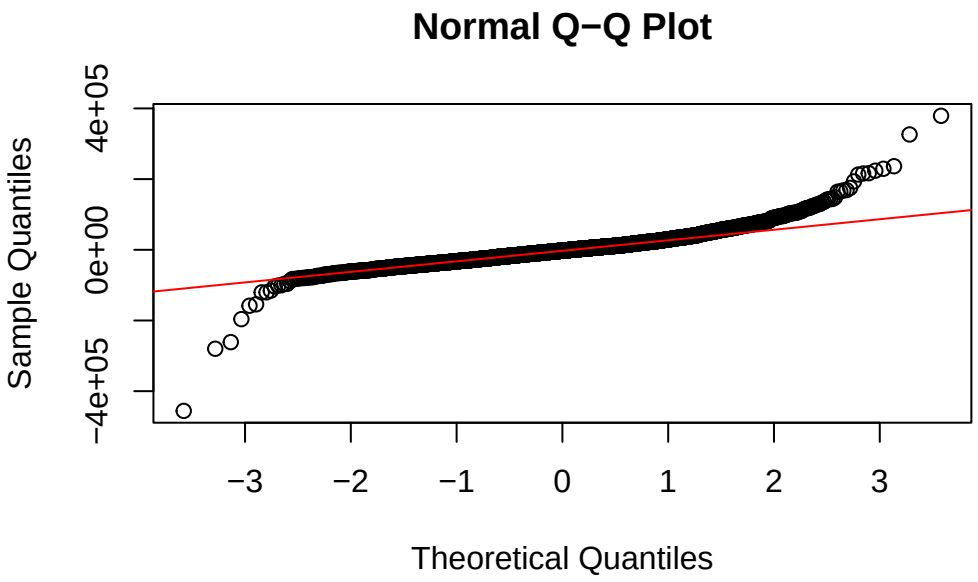

Figure 6: Q-Q Plot

Normality is not violated.

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; edited introduction; refined the model by

model selection tools; checked multicollinearity of the models; 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.
- Yi Tang: Built linear regression model; checked conditions for performing linear regression; checked extra conditions for performing multiple linear regression; showed the results of the linear regression model.

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.

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.

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