Housing in Ames*

Yanzun Jiang, Siyuan Lu, Yi Tang

November 24, 2024

Table of contents

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

1 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. The hypothesis we state is that there is a statistically significant linear relationship between certain property characteristics (predictors) and the sale price of houses (response). This hypothesis will be tested using linear regression, which would be appropriate in this case. 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 on house prices: estimates of how much the sale price is expected to change with a one-unit change in each predictor variable, holding other variables constant. Our primary goal is to understand the impact of each predictor on house prices, so the focus should be on interpretability instead of precision or accuracy.

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 housing completion costs, land acquisition prices, urban residents' disposable income, and urban population density (Wang (2013)). This article provides some characteristics other than what we use that can also influence house price in national scope.

In "Flood Risk as a Price-Setting Factor in the Market Value of Real Property", the analyzed market consider flood risk almost indifferent with house price compare to other factors in the analysis. (Cupal (2015)). This article offers a great example of our research since it uses multiple linear regression with some similar predictors as ours.

The research "House Price Prediction Using Hedonic Pricing Model and Machine Learning Techniques" shows that XGBoosting has higher accuracy in comparison to hedonic pricing model in prediction of property price. (Zaki et al. (2022)). This article provides an alternative method to study the relationship between multiple factors and house price.

2 Data Description

Table 1: Preview of Data (First Half)

sale_price	lot_area	overall_qual	year_built	roof_style
215000	31770	6	1960	Hip
105000	11622	5	1961	Gable

Table 1: Preview of Data (First Half)

sale_price	lot_area	overall_qual	year_built	roof_style
172000	14267	6	1958	Hip
244000	11160	7	1968	Hip
189900	13830	5	1997	Gable
195500	9978	6	1998	Gable

Table 2: Preview of Data (Second Half)

mas_vnr_area	total_bsmt_sf	central_air	garage_area	misc_val
112	1080	Y	528	0
0	882	Y	730	0
108	1329	Y	312	12500
0	2110	Y	522	0
0	928	Y	482	0
20	926	Y	470	0

The Ames Housing dataset (Table 1, Table 2) was sourced from the `AmesHousing` package (Kuhn (2020)) in R (R Core Team (2023)). It was originally compiled by the Ames City Assessor's Office through a comprehensive data dump of property tax records from 2006 to 2010, and it aimed to document residential property sales (De Cock (2011)). The dataset was initially designed for property tax assessments and general valuation, focusing on property characteristics such as lot area, the year built, and sale price. In contrast, this research aims to analyze how various property features influence house prices in Ames.

The dataset consists of 2930 observations and 82 variables relevant to understanding housing market dynamics. After cleaning, we selected 1 response variable, `sale_price`, and 9 predictor variables: `lot_area`, `overall_qual`, `year_built`, `roof_style`, `mas_vnr_area`, `total_bsmt_sf`, `central_air`, `garage_area`, and `misc_val`.

- `sale_price`: Price of the house
- `lot_area`: Lot size
- `overall_qual`: Rates the overall material and finish of the house
- `year_built`: Original construction date
- `roof_style`: Type of roof
- `mas_vnr_area`: Masonry veneer area
- `total_bsmt_sf`: Total area of basement

- `central_air`: Central air conditioning
- `garage_area`: Size of garage
- `misc_val`: Value of miscellaneous feature

These predictor variables all show the quality of the house, which affects the price of the house directly. So, we believe there is a linear relationship between these predictor variables and the response variable.

In this analysis, we will use these packages in R: `tidyverse` (Wickham et al. (2019)), `knitr` (Xie (2014)), and `patchwork` (Pedersen (2024)).

Table 3: Summarize Table of Numerical Data

	Mean	Standard_Deviation	Median
sale_price	180425.31	79811.03	160000
lot_area	10143.13	7898.24	9434
overall_qual	6.09	1.41	6
year_built	1971.13	30.22	1973
mas_vnr_area	101.97	179.15	0
total_bsmt_sf	1050.52	440.66	990
garage_area	472.34	215.23	479
misc_val	51.07	568.76	0

From the summary table (Table 3), we can see that `mas_vnr_area` and `misc_val` might be right-skewed since their mean is a lot greater than their median. An interesting point is that the standard deviation of `misc_val` is quite large, which indicate that houses in Ames might differs significantly in miscellaneous features. By analyzing these variables, we aim to provide insights into how specific property characteristics affect housing prices in Ames.

3 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. More detailed information about source of data is described in Section 2. The cleaned data we are using includes some detailed information about housing characteristics, but does not contain personal identifiers.

The analysis can provide deeper insights for stakeholders, including homeowners, potential buyers, real estate agents, and policymakers, to make better decisions about buying, selling

and investing in real estate. The Ames housing dataset has gained popularity, 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.

4 Preliminary Results

4.1 Method

```
#| echo: false
#| warning: false
#| message: false

# Convert variables to factors
data$central_air <- as.factor(data$central_air)
data$roof_style <- as.factor(data$roof_style)

# Fit the linear model
model1 <- lm(sale_price ~
               lot_area + overall_qual + year_built + roof_style +
               mas_vnr_area + total_bsmt_sf + central_air + garage_area +
               misc_val, data = data
)

#summary(model1)

# Extract variables with  p < 0.001 (***) from 14 variables

model2 <- lm(sale_price ~
                lot_area + overall_qual + year_built + mas_vnr_area +
                total_bsmt_sf + garage_area + misc_val, data = data)

#summary(model2)

# Perform stepwise selection starting with model1
model3 <- stepAIC(model1,
                  scope = list(lower = ~1,  # Minimum model (intercept-only)
                               upper = ~ lot_area + overall_qual + year_built +
                                       roof_style + mas_vnr_area +
                                       total_bsmt_sf + central_air +
```

```
                                        garage_area + misc_val),  # Full model
                  direction = "both",  # Both forward and backward selection
                  k = 2,  # Use AIC criterion with k = 2
                  trace = TRUE)  # Display stepwise process
```

```
Start:  AIC=61515.99
sale_price ~ lot_area + overall_qual + year_built + roof_style +
    mas_vnr_area + total_bsmt_sf + central_air + garage_area +
    misc_val

                Df  Sum of Sq        RSS   AIC
- central_air    1 1.7367e+09 4.5259e+12 61515
<none>                        4.5242e+12 61516
- roof_style     5 3.8960e+10 4.5631e+12 61531
- misc_val       1 4.3253e+10 4.5674e+12 61542
- year_built     1 5.3062e+10 4.5772e+12 61548
- mas_vnr_area   1 1.9237e+11 4.7165e+12 61635
- lot_area       1 2.4866e+11 4.7728e+12 61669
- total_bsmt_sf  1 2.7405e+11 4.7982e+12 61685
- garage_area    1 3.1028e+11 4.8344e+12 61707
- overall_qual   1 2.2202e+12 6.7443e+12 62674

Step:  AIC=61515.11
sale_price ~ lot_area + overall_qual + year_built + roof_style +
    mas_vnr_area + total_bsmt_sf + garage_area + misc_val

                Df  Sum of Sq        RSS   AIC
<none>                        4.5259e+12 61515
+ central_air    1 1.7367e+09 4.5242e+12 61516
- roof_style     5 3.9256e+10 4.5651e+12 61530
- misc_val       1 4.3398e+10 4.5693e+12 61541
- year_built     1 5.1446e+10 4.5773e+12 61546
- mas_vnr_area   1 1.9342e+11 4.7193e+12 61635
- lot_area       1 2.4812e+11 4.7740e+12 61668
- total_bsmt_sf  1 2.7302e+11 4.7989e+12 61683
- garage_area    1 3.0854e+11 4.8344e+12 61705
- overall_qual   1 2.2192e+12 6.7451e+12 62672
```

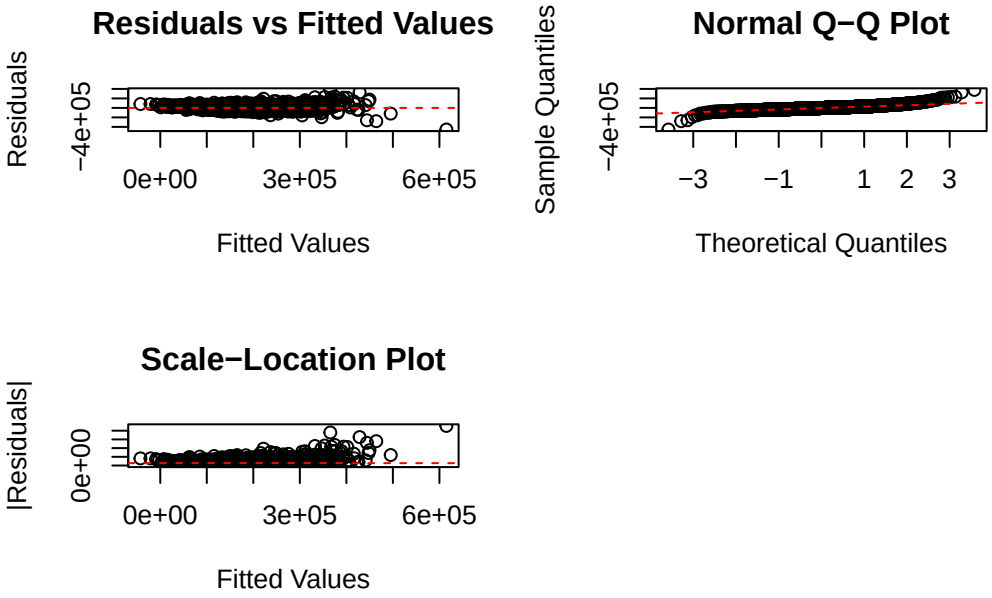
```
#summary(model3)
```

Model 3, derived through stepwise selection (stepAIC), is the most optimal model compared to Models 1 and 2. It achieves a similar adjusted R-squared (0.7543) to Model 1 but with

fewer predictors and a lower AIC (61515.11), indicating a better balance between explanatory power and simplicity. While Model 1 includes non-significant variables, and Model 2 uses only the most significant ones, Model 3 refines the selection to include only the predictors that optimize the model's performance, making it more efficient and interpretable.

```
    Durbin-Watson test

data:  model3
DW = 1.5393, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```



The diagnostic analysis shows that the assumptions of linearity and constant variance are not violated, as evidenced by the residuals vs. fitted plot and scale-location plot. However, the Durbin-Watson test indicates significant positive autocorrelation ($p < 2.2e\text{-}16$), violating the uncorrelated errors assumption. Additionally, the Q-Q plot reveals some deviations in the tails, suggesting a potential violation of the normality assumption. Addressing the autocorrelation issue and considering adjustments for normality may improve the model's validity.

```
# 1. Addressing Autocorrelation by Including Lagged Residuals
# Generate lagged residuals
residuals_lagged <- c(NA, head(residuals(model3), -1))  # Lagged residuals
data$residuals_lagged <- residuals_lagged
```

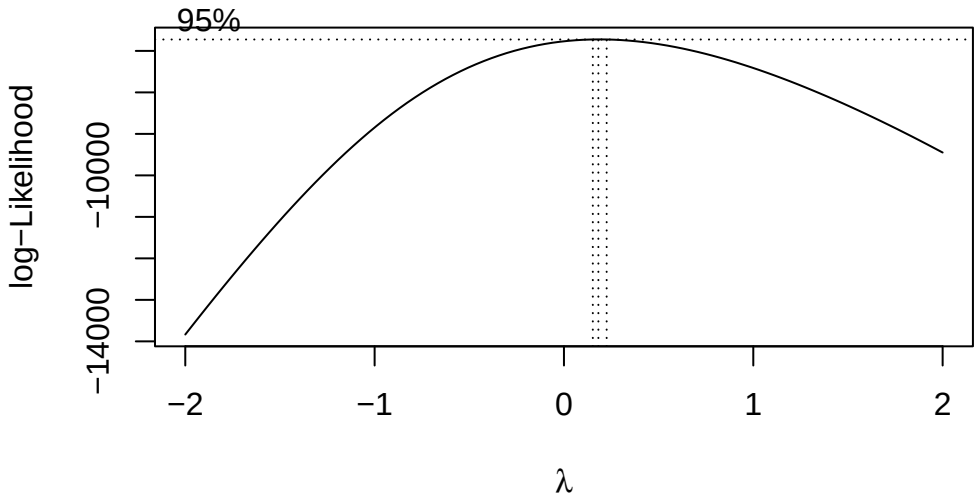
```r
# Refit model by adding lagged residuals
model4 <- lm(sale_price ~ lot_area + overall_qual + year_built + mas_vnr_area +
               total_bsmt_sf + garage_area + misc_val + residuals_lagged,
             data = data, na.action = na.exclude)

# Check for autocorrelation again
dw_test_model4 <- dwtest(model4)
print(dw_test_model4)  # Improved DW indicates less autocorrelation
```

```
    Durbin-Watson test

data:  model4
DW = 2.0154, p-value = 0.6462
alternative hypothesis: true autocorrelation is greater than 0
```

```r
# 2. Addressing Normality with Box-Cox Transformation
# Box-Cox Transformation to find the best lambda
boxcox_result <- boxcox(model3, lambda = seq(-2, 2, by = 0.1))  # Test range of lambda
```



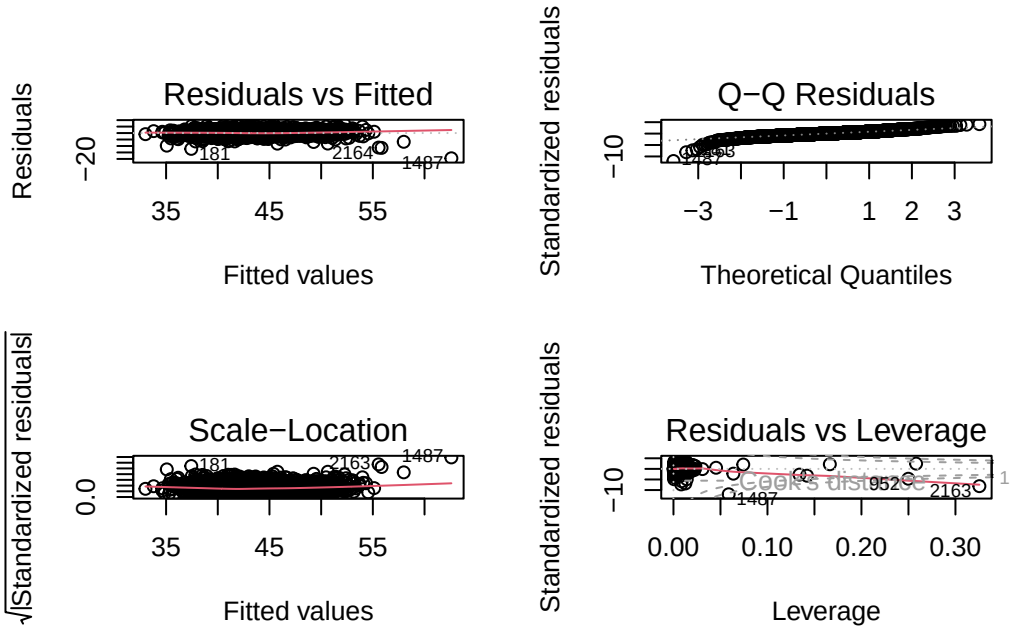
```r
lambda_optimal <- boxcox_result$x[which.max(boxcox_result$y)]  # Optimal lambda
print(lambda_optimal)
```

```
[1] 0.1818182
```

```r
# Apply the Box-Cox transformation to the response variable
if (lambda_optimal == 0) {
  data$sale_price_transformed <- log(data$sale_price)
} else {
  data$sale_price_transformed <- (data$sale_price^lambda_optimal - 1) / lambda_optimal
}

# Refit the model with transformed response
model5 <- lm(sale_price_transformed ~ lot_area + overall_qual + year_built +
               mas_vnr_area + total_bsmt_sf + garage_area + misc_val,
             data = data)

# Diagnostic plots for the new model
par(mfrow = c(2, 2))
plot(model5)
```

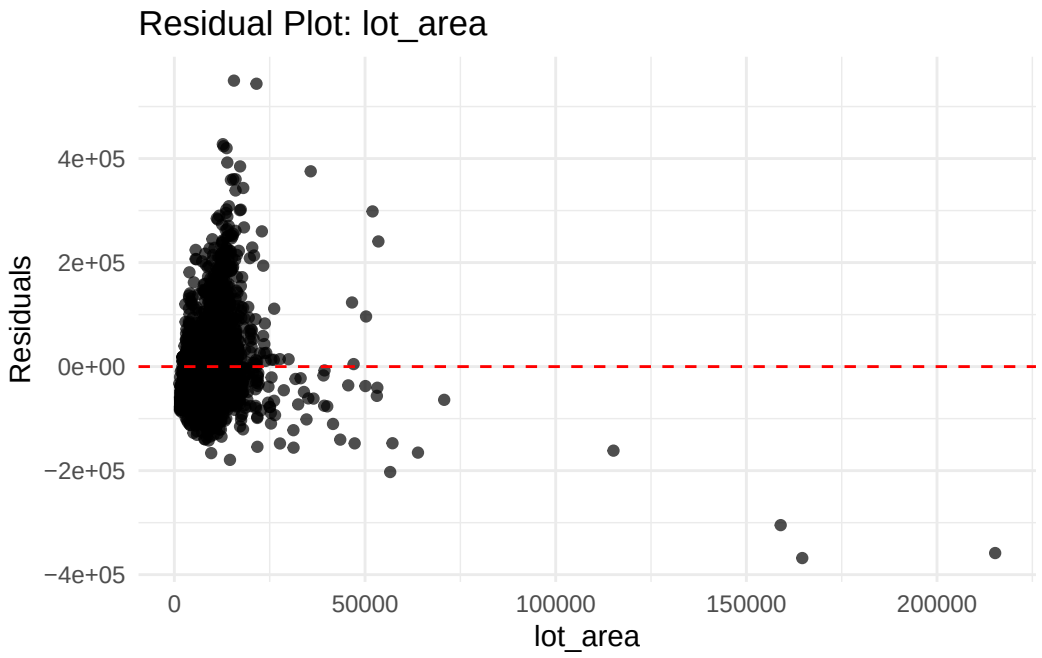


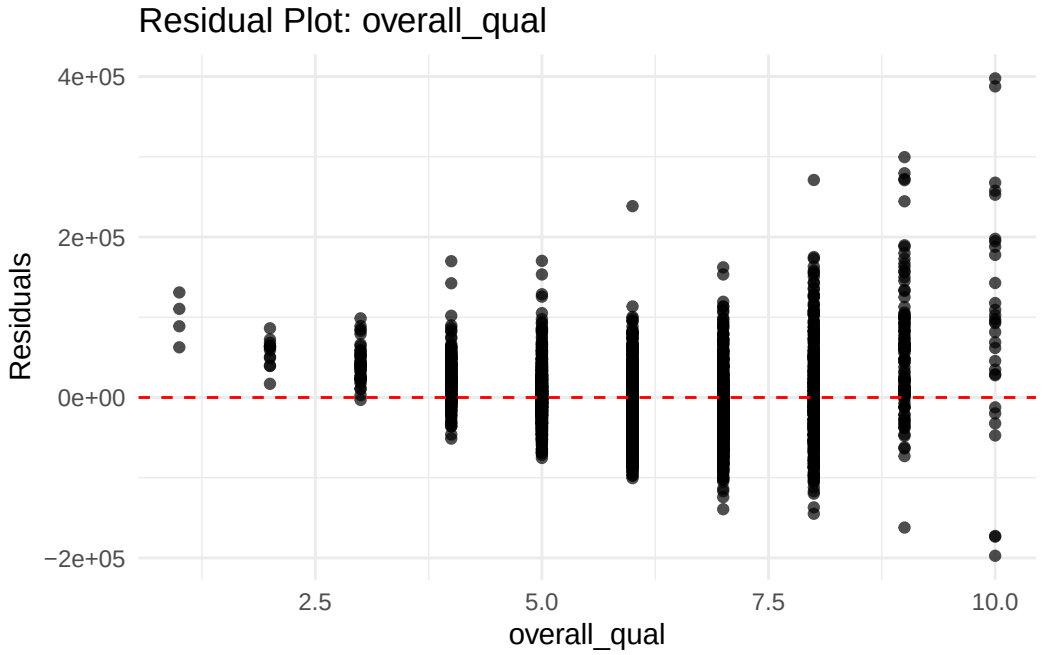
```r
par(mfrow = c(1, 1))
```

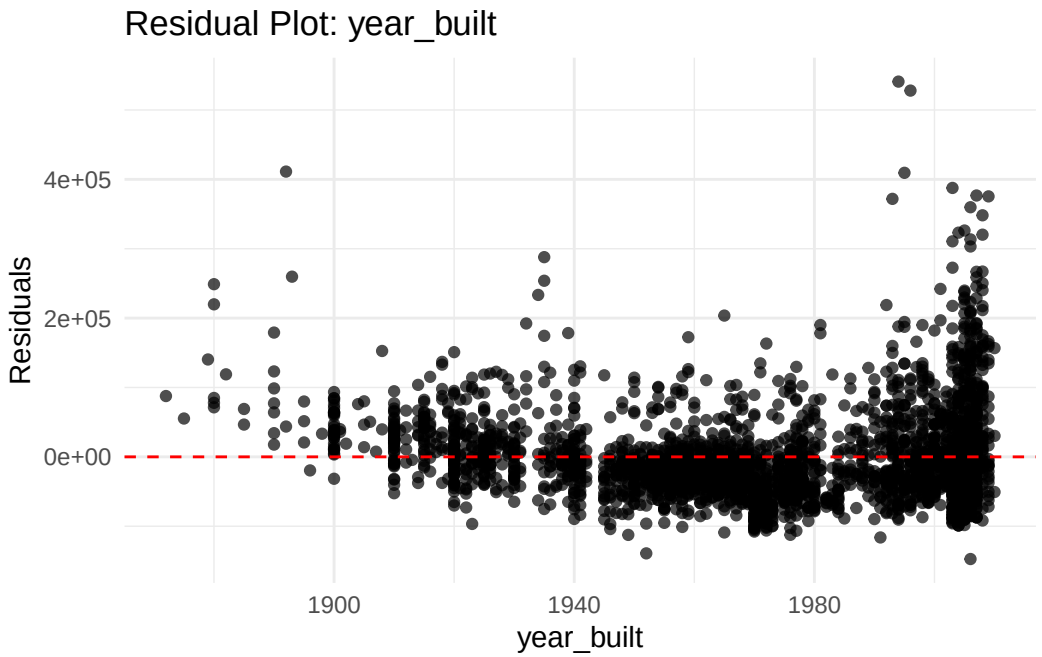
To address violations of assumptions, a lagged residual term was added to mitigate autocorrelation, improving the Durbin-Watson statistic. Additionally, a Box-Cox transformation was applied to normalize the response variable, with an optimal determined through analysis. These adjustments ensure the model aligns better with linear regression assumptions and produces more reliable inferences.

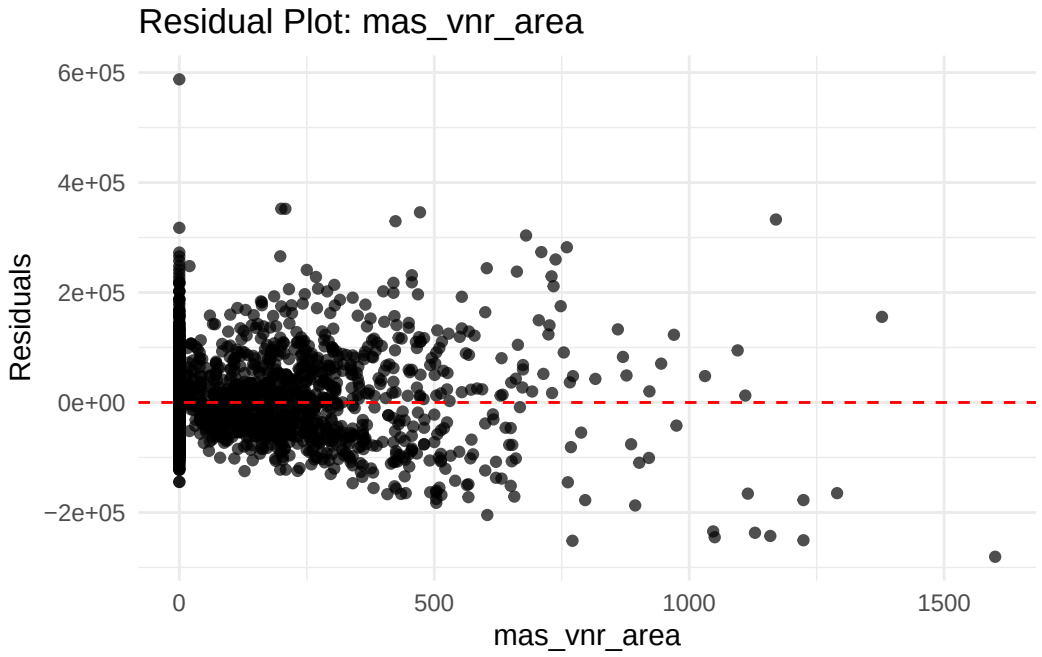
A Appendix

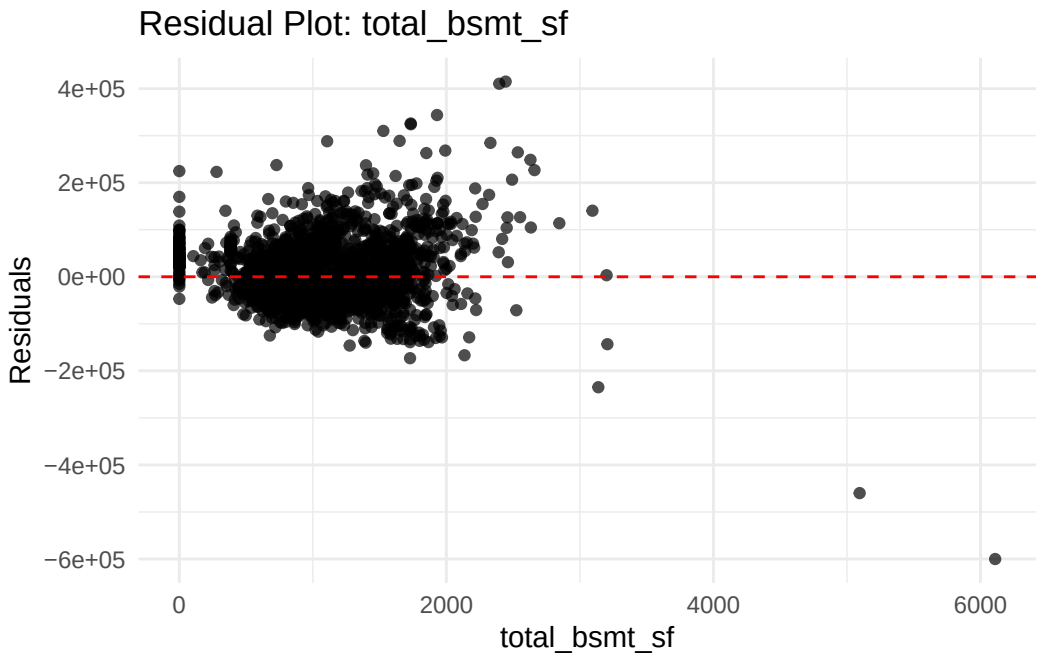
A.1 Residual plot of predictors

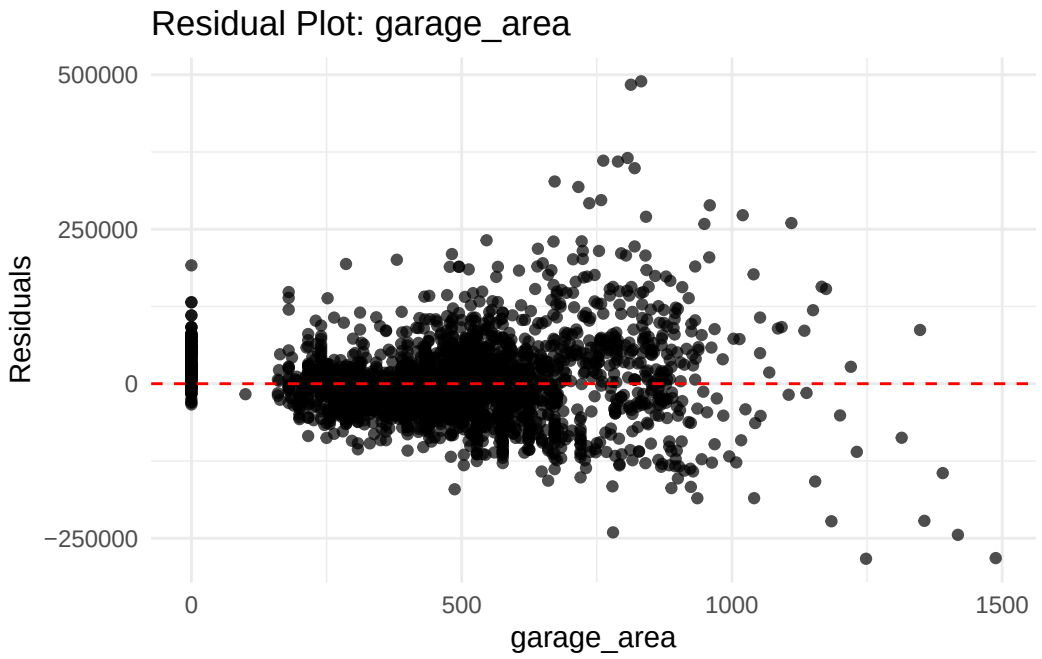


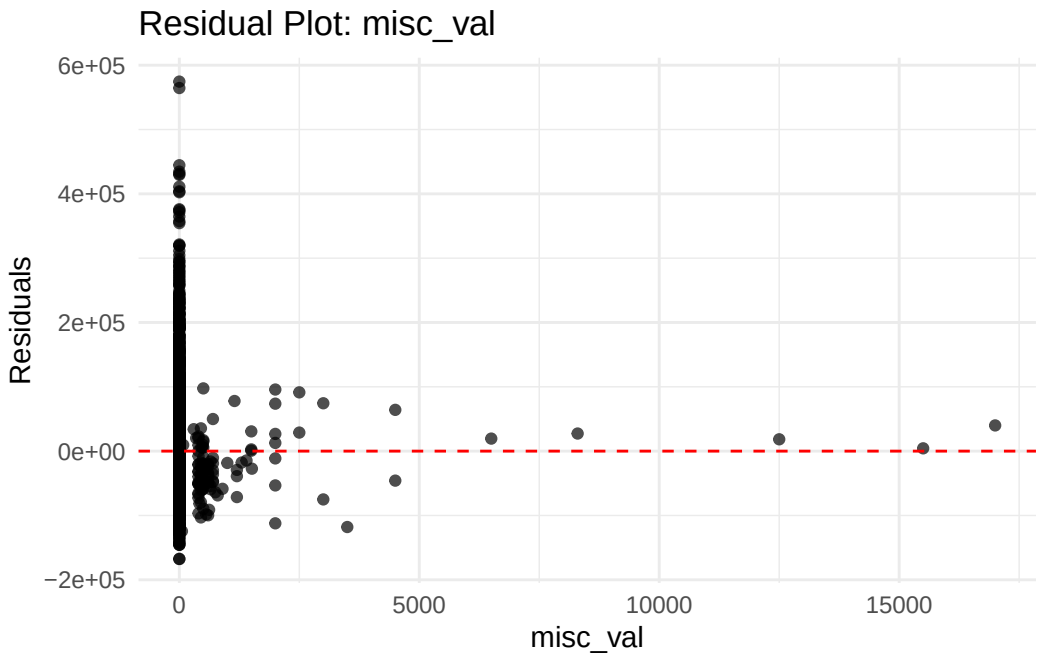



Residual Plot: year_built
Residuals
4e+05
2e+05
0e+00
1900
1940
1980
year_built


Residual Plot: mas_vnr_area
Residuals
6e+05
4e+05
2e+05
0e+00
−2e+05
0
500
1000
1500
mas_vnr_area


Residual Plot: total_bsmt_sf
Residuals
4e+05
2e+05
0e+00
−2e+05
−4e+05
−6e+05
0
2000
4000
6000
total_bsmt_sf


Residual Plot: garage_area
Residuals
500000
250000
0
−250000
0
500
1000
1500
garage_area



A.2 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; downloaded data for setting up the linear regression model; cleaned data to make further analysis easier; introduced the dataset; made the summary table; created file for R code; made the reference list; revised group member's work; combined group member's work together.
- Siyuan Lu: Set research question and hypothesis; 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

Cupal, Martin. 2015. "Flood Risk as a Price-Setting Factor in the Market Value of Real Property." *Procedia Economics and Finance* 23: 658–64.

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).

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

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/.

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.

Zaki, John, Anand Nayyar, Surjeet Dalal, and Zainab H Ali. 2022. "House Price Prediction Using Hedonic Pricing Model and Machine Learning Techniques." *Concurrency and Computation: Practice and Experience* 34 (27): e7342.