

Housing in Ames*

Yanzun Jiang, Siyuan Lu, Yi Tang

October 1, 2024

Table of contents

1	Introduction	2
2	Data Description	2
3	Ethics Discussion	3
4	Preliminary Results	3
A	Appendix	4
A.1	Contributions	4
	References	5

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

1 Introduction

The R Core Team (2023)

2 Data Description

Table 1: Preview of Data

sale_price	lot_area	overall_qual	year_built	roof_style	mas_vnr	total_bsmt_sf	central_air	garage_area	misc_val
215000	31770	6	1960	Hip	112	1080	Y	528	0
105000	11622	5	1961	Gable	0	882	Y	730	0
172000	14267	6	1958	Hip	108	1329	Y	312	12500
244000	11160	7	1968	Hip	0	2110	Y	522	0
189900	13830	5	1997	Gable	0	928	Y	482	0
195500	9978	6	1998	Gable	20	926	Y	470	0

The Ames Housing dataset (Table 1) 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. It was cleaned using `tidyverse` package (Wickham et al. (2019)). 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`.

#TODO: explain variables

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

The variability in `lot_area` and the distribution of the `year_built` variable indicate interesting trends that warrant further exploration. By analyzing these variables, this study aims to provide insights into how specific property characteristics affect housing prices in Ames, Iowa.

3 Ethics Discussion

4 Preliminary Results

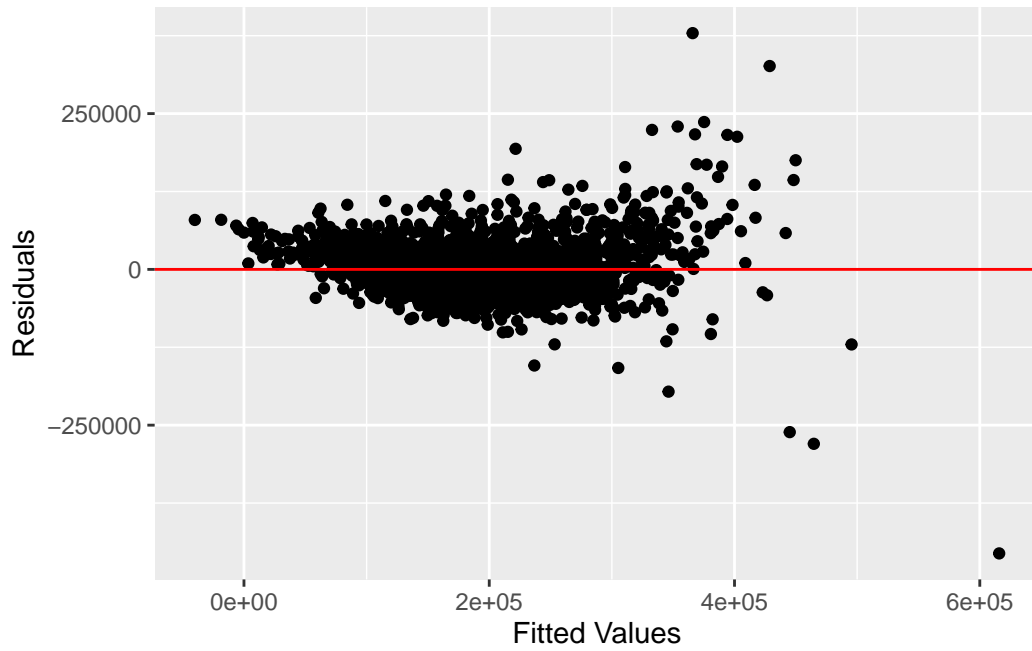


Figure 1: Residual Plot

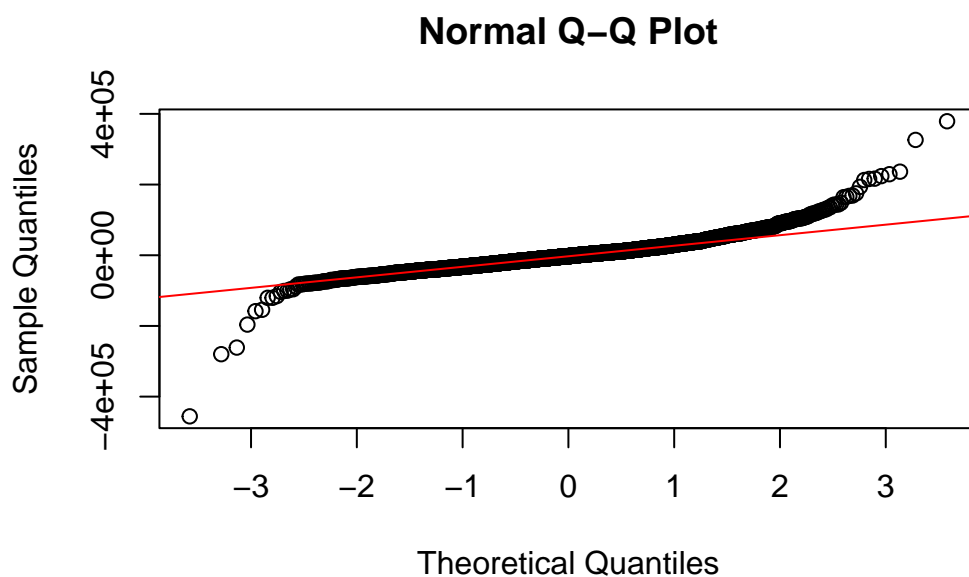


Figure 2: Q-Q Plot

A Appendix

A.1 Contributions

Yanzun Jiang: Organized discussions and meetings; set up Github workspace for collaborating; downloaded data for setting up the linear regression model; cleaned data to make further analysis easier; completed Section 2 in the proposal; made the reference list; revised and combined group member's work together.

Siyuan Lu:

Yi Tang: Built linear regression model to predict house sale prices by using five key predictors in cleaned data. It assisted to understand the relationship between variables and ensure data meets key assumptions for statistical validity.

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).
- Kuhn, Max. 2020. *AmesHousing: The Ames Iowa Housing Data*. <https://CRAN.R-project.org/package=AmesHousing>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Golemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.