Housing in Ames*

Yanzun Jiang, Siyuan Lu, Yi Tang

October 1, 2024

Table of contents

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

1 Introduction

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
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. 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`.

`sale_price`: Priceof the house in dollars

`lot_area`: Lot size in square feet

`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 in square feet

`total_bsmt_sf`: Total square feet of basement area

`central_air`: Central air conditioning

`garage_area`: Size of garage in square feet

`misc_val`: Value of miscellaneous feature in dollars

These predictor variables all shows 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.

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 easily 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, Iowa.

3 Ethics Discussion

4 Preliminary Results

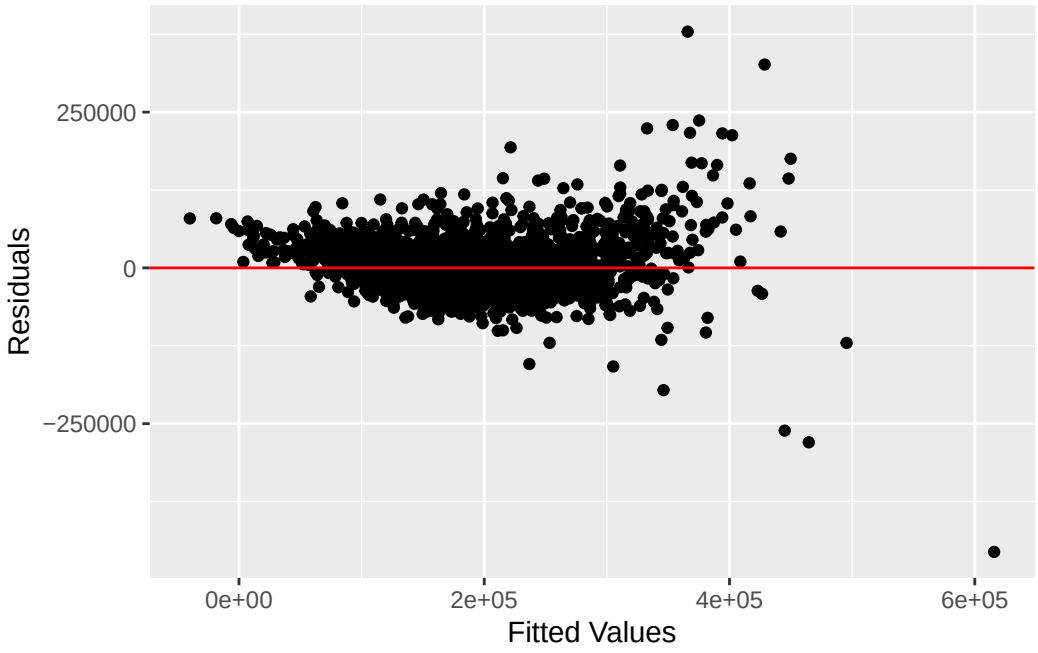

Figure 1: Residual Plot

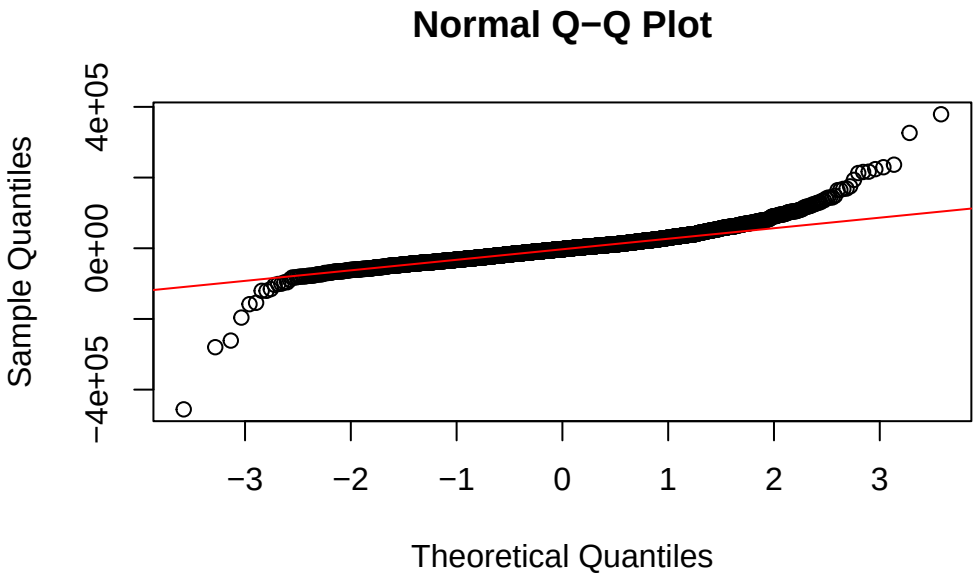

Figure 2: Q-Q Plot

A Appendix

A.1 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; 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 Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.