Housing in Ames*

Yanzun Jiang, Siyuan Lu, Yi Tang

October 1, 2024

Table of contents

1 Introduction **2**

2 Data Description **3**

3 Ethics Discussion **3**

4 Preliminary Results **4**

A Appendix **5**

A.1 Contributions 5

References **6**

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

1 Introduction

Understanding the relationships between variables that influence house prices can provide deeper insights for various stakeholders, including homeowners, potential buyers, real estate agents, and policymakers, to make better decisions. Accurate prediction of house prices helps to make informed decisions about buying, selling and investing in real estate. In addition, it helps to assess the economic health of an area and plan for future developments.

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 lot size, overall quality, year of construction, roof style, masonry veneer area, total basement area, central air conditioning, garage size, and value of miscellaneous feature. The hypothesis we state is that there is a statistically significant positive relationship between certain property characteristics (predictors) and the sale price of houses (response). This hypothesis will be tested using linear regression which is appropriate when there is a linear relationship between the response variable and predictor variables. We use scatter plots and correlation coefficients 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. Specifically, It will provide 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/accuracy.

(**influencing?**) analyzed the influence factors for market supply and demand of commodity housing, and established price influence factors analysis model based on multiple linear regression. The article studies influencing factors of house price using the same method, linear regression, as we do, providing an example of setting a linear regression on factors and house price. In Zhou et al. (2021), authors develop a novel fuzzy linear regression framework using symmetric and asymmetric trapezoidal fuzzy numbers for determining the relationship of particular (non-) policy factors with the house prices. The article shows the relationship between policy factors and house price, which demonstrates a different aspect of our research question. In the real estate industry, the price of property plays a crucial role in economic growth. The research Zaki et al. (2022) attempts to predict the price of a house using MLTs. Here, the price of the property is predicted using Extreme Gradient (XG) boosting algorithm and hedonic regression pricing. Both XGBoost and hedonic pricing models use 13 variables as inputs to predict house prices. The article uses an alternative regression method to study the relationship between multiple factors and house price.

In Section 2, we introduce the source and overview of our dataset. Variables used in regression is explained in detail. Then, in Section 3 we talk about the data ethics. In Section 4, we address the preliminary result plots and descriptions.

2 Data Description

Table 1: Preview of Data

sale_price	lot_area	overall_qual	year_built	roof_style	mas_vnr_area	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

Our data is collected from (**amehousing?**), then we cleaned the data to keep some keey factors we want. Raw and processed versions of the data from De Cock published on Journal

of Statistics Education in 2011. More detailed information about source of data is described in **?@sec-data.Stakeholders** of our analysis including homeowners, potential buyers, real estate agents, and policymakers.

4 Preliminary Results

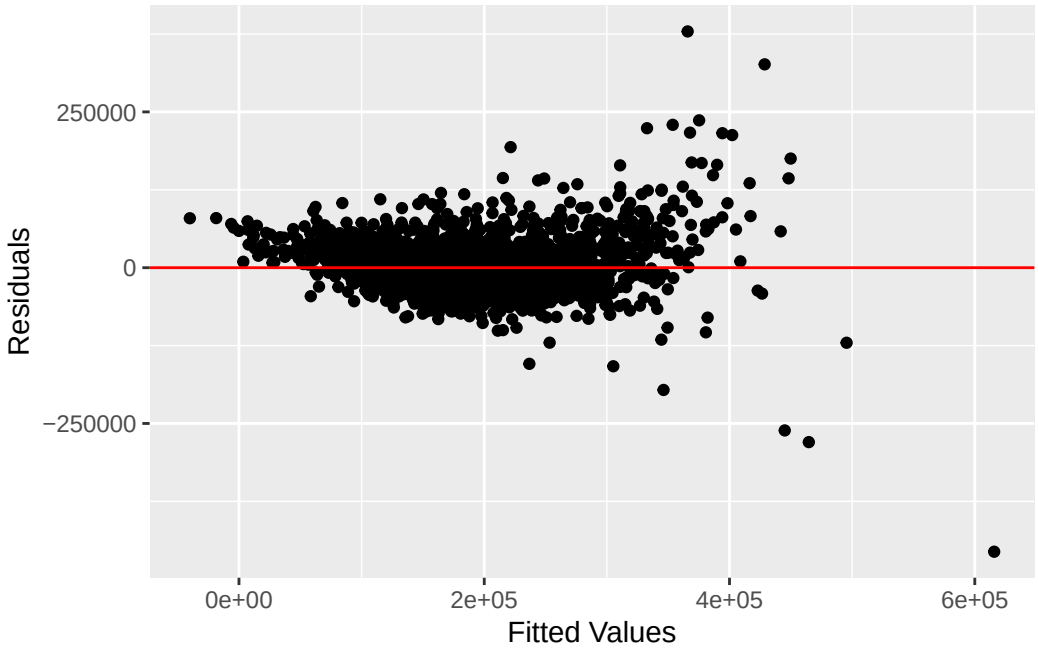

Figure 1: Residual Plot

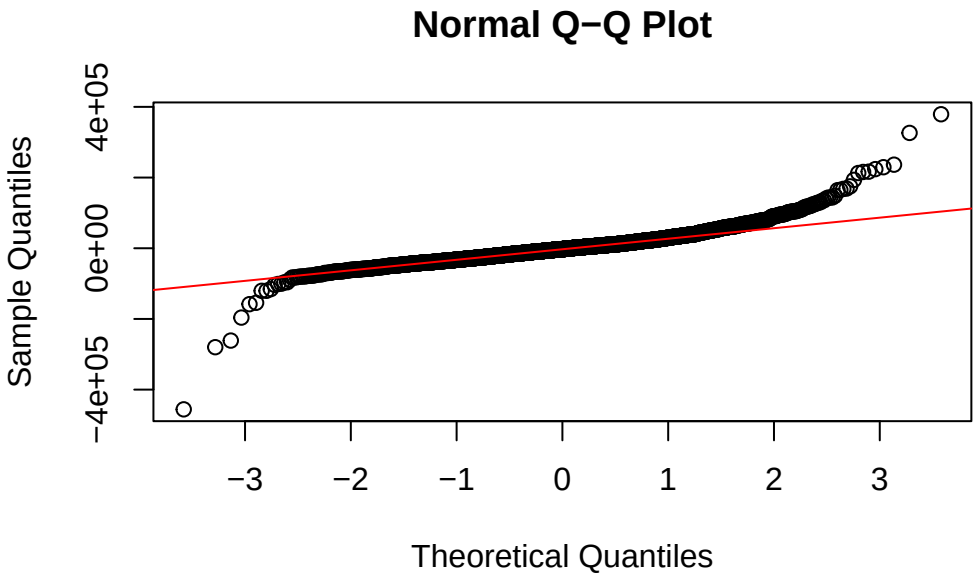

Figure 2: Q-Q Plot

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

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* 34 (27). https://doi.org/10.1002/cpe.7342.

Zhou, Jian, Yixuan Shen, Athanasios A. Pantelous, and Hui Zhang. 2021. "The Range of Uncertainty on the Property Market Pricing: The Case of the City of Shanghai." *Finance Research Letters* 40: 101720. https://doi.org/10.1016/j.frl.2020.101720.