# Analyzing Influencing Factors of House Prices in Ames by Linear Regression

## 1 Motivation and Research Question

Our research question: What are the factors that influence house prices in Ames from 2006 to 2010?

Our research can:

- Offer practical recommendations for various stakeholders
- Contribute to a more efficient and transparent real estate market
- Improve economic insights and investment decisions

## 2 Data Collection

Our Ames Housing dataset was sourced from the AmesHousing package in R, which was originally compiled by the Ames City Assessor's Office. The original dataset is reliable since it is a government official data dump from the record system.

The dataset consists of 2930 observations and 82 variables. The targeted population and some meaningful variables fit our research goal well.

## 3 Methods of Analysis

Construct the initial multiple linear regression model. (Model 1)

Check two conditions and four assumptions by residual plots, and Q-Q plots. Solve violation by variance-stabilizing and Box-Cox transformation. (Model 2)

Test the significance of coefficients and make adjustments. (Model 3)

Select the best model using R-squared adj, AIC, AICc, BIC, and VIF. (Model 4)

Stepwise selection to refine Model 2. (Model 5)

## 4 Results and Conclusions

Model 2 performed better in terms of the assumptions compared to Model 1, suggesting that the the transformations have a positive effect.
Model 3 is the same model as Model 2 since all coefficients are significant.
We choose Model G to be Model 4 because Model G has smaller AIC and AICc.
Model 4 is also a full model, the same model as Model 2.

| Model | R^2_{adj} | AIC | AIC_c | BIC | VIF_{max} |
|---|---|---|---|---|---|
| Model A | 0.68 | -8516.80 | -8516.79 | -8504.85 | 0.00 |
| Model B | 0.74 | -9169.57 | -9169.56 | -9151.65 | 1.02 |
| Model C | 0.77 | -9510.59 | -9510.56 | -9486.69 | 1.82 |
| Model D | 0.78 | -9660.55 | -9660.52 | -9630.67 | 2.03 |
| Model E | 0.79 | -9767.32 | -9767.28 | -9731.47 | 2.12 |
| Model F | 0.80 | -9814.22 | -9814.17 | -9772.40 | 2.23 |
| Model G | 0.80 | -9816.69 | -9816.62 | -9768.89 | 2.23 |

Model 5 is full model as Model 2 since deleting any predictor would result in a larger AIC.

We are confident to conclude that the optimal model to answer the research question is just the full model with variables transformed.

The final model we choose is Model 2 (3, 4, and 5). Based on the analysis of the final model, the linear relationship between the predictors and the response variable was confirmed, and the results support the hypothesis.

$$\begin{aligned} ln(Y) =& 8.57 + 0.187 \times ln(X_1) + 0.280 \times X_2^{0.8} + \\ & 1.47 \times 10^{-113} \times X_3^{34} + 3.35 \times 10^{-3} \times X_4^{0.5} + 0.0229 \times X_5^{0.33} + \\ & 0.0348 \times X_6^{0.25} + (-1.03 \times 10^{-3}) \times X_7^{0.5} + \epsilon \end{aligned}$$

**Through the final model, we can see the answer to our research question: all numerical predictors affects house prices.**

## 5 Limitations

- The assumption of linearity may not hold in all cases.
- The models use limited predictors, ignoring other potentially relevant variables.
- Did not perform a train-test split. The model cannot assess its ability to generalize to unseen data.

In future study, we can:

- Introduce interaction terms to better capture potential non-linear relationships
- Expand the predictor set
- Implement a train-test split to evaluate the model's generalizability

## 6 References


Kuhn, M. (2020). AmesHousing: The Ames Iowa Housing Data (Version 0.0.4) [R package]. Retrieved from https://CRAN.R-project.org/package=AmesHousing

De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. Journal of Statistics Education, 19(3). Taylor & Francis. https://doi.org/10.1080/10691898.2011.11889627