Analyzing the Impact of Heating Quality and Central Air Conditioning on House Sale Prices Using Multiple Linear Regression

STA302H1 Final Project Part 3 Report

Mohammad Danish Malik & Yutong Han

December 6, 2024

Contributions

Mohammad Danish Malik: Danish is responsible for the **Introduction**, **Methods**, **Conclusions & Limitations**, **Ethical Discussion**, **Bibliography** and **Appendix Section** of the report. He also dealt with the **Editing Demonstration** and **proofread the Poster**.

Yutong Han: Yutong is responsible for the **R coding aspect** of the project as well as the **Results** Section of the Report. She was also **primarily responsible for the Poster**.

Introduction

The housing market is influenced by various factors, and understanding the drivers of house prices is essential for real estate professionals, policymakers, and homebuyers. This study examines the impact of **heating quality and central air conditioning on house sale prices**, using the **Ames Housing Dataset**. The research question is: **How do heating quality and central air conditioning (predictors) affect house prices (response), while accounting for overall quality, living area and basement area (predictors)?**

Heating quality and air conditioning are increasingly important as buyers prioritize comfort and energy efficiency. **Hahn et al. (2018)** found that energy-efficient heating systems significantly raised house prices in Germany, highlighting the relevance of heating quality. Similarly, **Cağlayan and Arikan (2011)** demonstrated that features like heating systems and living area strongly influenced housing prices in Istanbul, reinforcing their inclusion as predictors. **Anselin and Lozano-Gracia (2008)** emphasized the role of internal home features and environmental factors in determining house prices in California. From the results of these studies we hypothesize that, controlling for other characteristics, houses with higher heating quality and central air conditioning will have higher sale prices.

Linear regression is appropriate for this study as it quantifies the relationship between predictors and house prices while controlling for other factors. This approach allows us to estimate how much heating quality and air conditioning contribute to sale prices in a clear, interpretable manner. The focus is on **interpretability** to aid stakeholders in understanding how these features influence property value. This study builds on existing research by analyzing these predictors in a new context and identifying their specific contributions to house pricing.

Methods

We tried answering our research question in **R** using **Multiple Linear Regression**. The Ames Housing Dataset was reviewed initially to ensure there were no missing values for the predictors mentioned in the introduction section. An initial model was fitted using these predictors. Categorical predictors were encoded to ensure appropriate use in the model. This initial model served as the foundation for subsequent analysis and refinement.

To validate the MLR model, the **conditional mean response** was assessed by examining a scatterplot of the response variable (sale price) versus fitted values. A systematic deviation from randomness in this plot would indicate a violation of linearity between the predictors and the response. The **conditional mean predictor condition** was checked through pairwise scatterplots of predictors to ensure their relationships were approximately linear.

Multicollinearity was assessed using **Variance Inflation Factors (VIF)**. A VIF value exceeding the threshold of 5 would indicate significant multicollinearity, requiring corrective measures such as removing or combining predictors, to ensure stable coefficient estimates.

Once these conditions were verified, linear regression assumptions were checked systematically:

- **Uncorrelated Errors:** Residual independence was assessed using a residual vs fitted plot. A random scatter around zero would confirm independence, while patterns or trends would suggest serial correlation.

- **Linearity:** Residual vs predictor plots were used to identify non-linear relationships. A random distribution around zero would confirm linearity, while systematic deviations would indicate a need for Box-Cox transformations.

- **Homoscedasticity:** Residual vs fitted plots were also used to check for constant variance. A funnel-like pattern would suggest heteroscedasticity, which could be addressed using variance-stabilizing transformations.

- **Normality:** A Q-Q plot was used to verify whether residuals followed a normal distribution. Deviations from the diagonal line in this plot would prompt the use of Box-Cox transformations to improve normality.

Leverage and influential points were identified using **leverage values, Cook's Distance, and DFBETAs** to ensure model stability and reliability. Observations flagged by these metrics were reviewed for potential errors or unusual characteristics. Points causing substantial changes in coefficients, residual patterns, or overall model fit were carefully evaluated, with their inclusion or exclusion based on their impact on the model.

ANOVA tests were conducted to test the overall significance of the transformed model ($p < 0.05$). Predictors that did not contribute significantly to reducing residual variability were considered for exclusion to maintain a parsimonious model. These tests guided the selection of meaningful predictors while avoiding unnecessary complexity.

Individual **t-tests** were used to evaluate the significance of each predictor, with p-values below 0.05 indicating statistical relevance. Non-significant predictors were considered for removal unless their inclusion was theoretically justified or necessary to maintain adherence to model assumptions.

Adjusted R^2 was reviewed to assess the balance between explanatory power and model simplicity. Values closer to 1 were preferred, provided they did not result from overfitting due to unnecessary predictors.

A **cyclical approach** was adopted throughout the analysis. If violations of assumptions or issues with model diagnostics were found, corrective measures were applied, and the model-building process restarted from the initial step. This iterative process ensured the development of a robust final model.

A simplified version of this process is shown in **Figure 1** below.

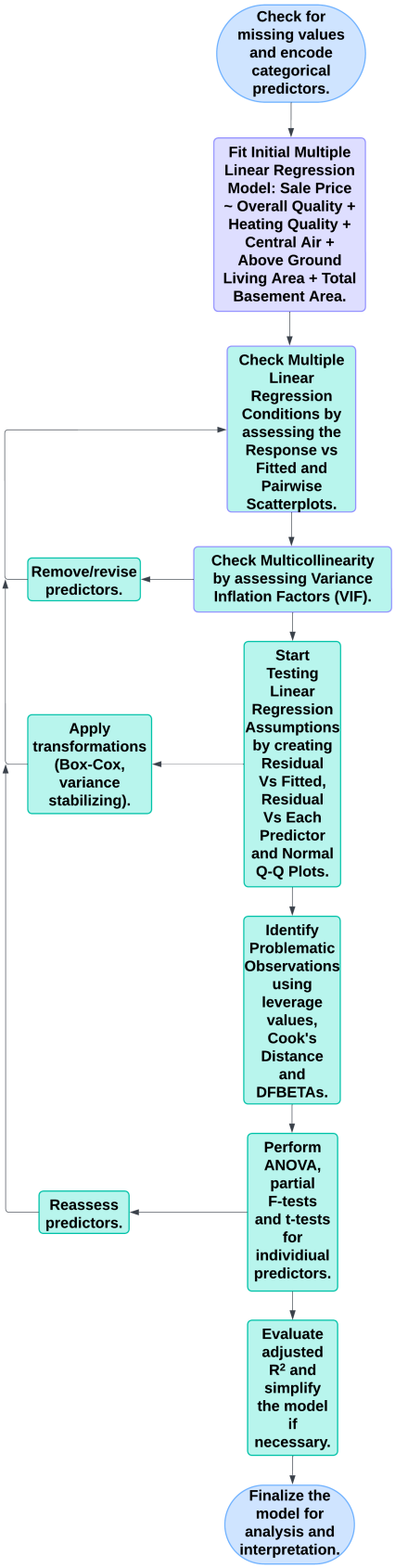

Figure 1: A simplified flowchart of our model building process

Analysis and Results

The analysis began with an initial multiple linear regression (MLR) model aimed at understanding the relationship between house prices (SalePrice) and selected predictors: **Heating Quality (categorical), Central Air Conditioning (binary), Overall Quality (ordinal), Above Ground Living Area (continuous), and Total Basement Area (continuous)**. These predictors were chosen based on **theoretical relevance to the research question and encoded as needed for inclusion in the model**. Fitting the initial model we ended up with the following model:

$$SalePrice = \beta_0 + \beta_1 \times Overall.Qual2 + \beta_2 \times Overall.Qual3 + \beta_3 \times Overall.Qual4 + \beta_4 \times Overall.Qual5 + \beta_5 \times Overall.Qual6 + \beta_6 \times Overall.Qual7 + \beta_7 \times Overall.Qual8$$
$$+\beta_8 \times Overall.Qual9 + \beta_9 \times Overall.Qual10 + \beta_{10} \times Heating.QCgd + \beta_{11} \times Heating.QCTA + \beta_{12} \times Heating.QCFa + \beta_{13} \times Heating.QCPo + \beta_{14} \times Central.AirY + \beta_{15} \times Gr.Liv.Area + \beta_{16} \times Total.Bsmt.SF$$

Initial Model Diagnostics: MLR Conditions and Multicollinearity

Following the fitting of the initial model, the Multiple Linear Regression (MLR) conditions were assessed. The conditional mean response was validated through a scatterplot of residuals vs. fitted values. This plot (part of **Figure 2** below) showed curved patterns, indicating **potential violations of linearity**. The conditional mean predictor condition was checked through **pairwise scatterplots of predictors**, which revealed approximately linear relationships between numerical predictors. Multicollinearity was assessed using **Variance Inflation Factors (VIF)**, with **all values falling below 5, confirming no significant multicollinearity**.

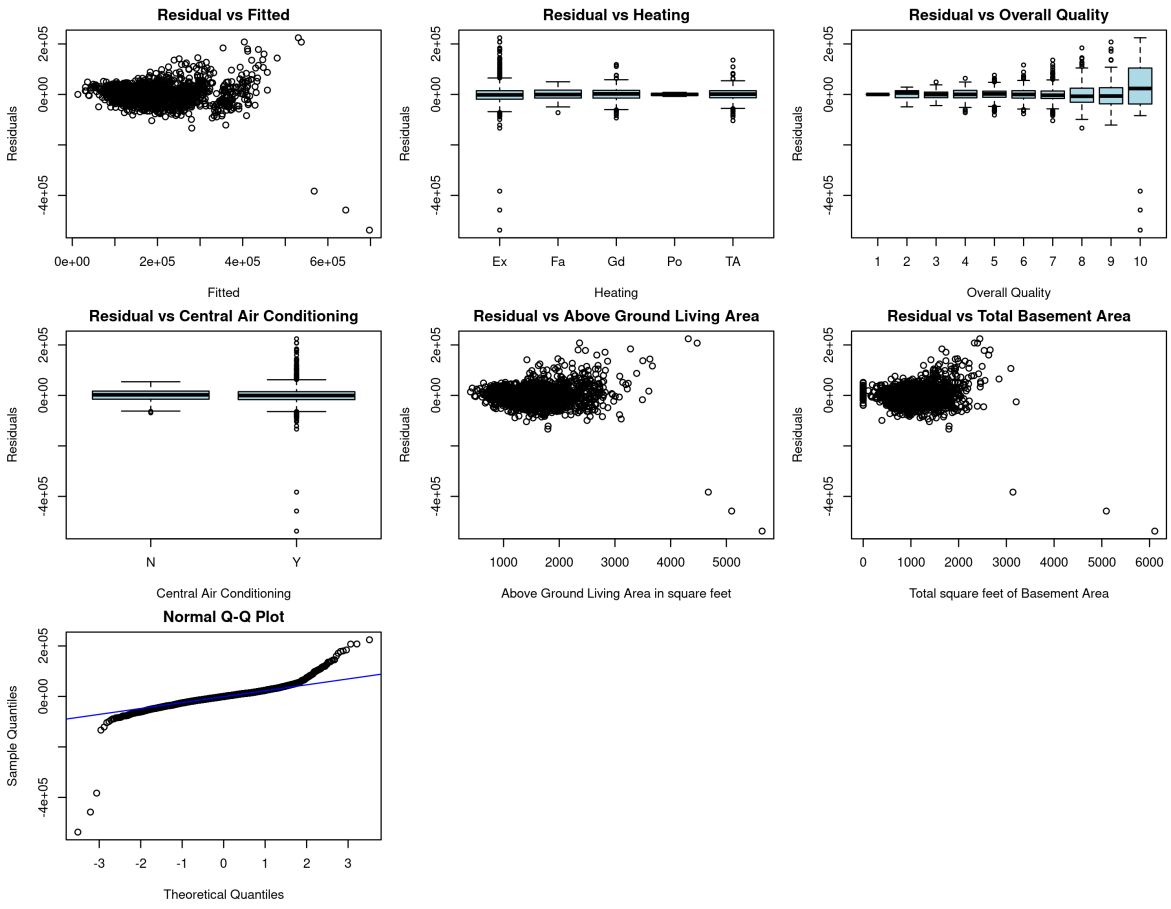

Figure 2: Diagnostic Plots for Initial Model

Assumption Testing and Transformation

The initial model's diagnostics (Figure 2 above) revealed **violations of key linear regression assumptions**. The residuals vs. fitted plot displayed curved patterns, indicating **non-linearity**. The residuals vs. predictors plots showed signs of **heteroscedasticity**, and the Q-Q plot indicated **deviations from normality**, particularly in the tails. Crucially, however, the plots did not show clustering at different places which suggested that the **Uncorrelated Errors Assumption had not been violated**. The potential violations of linearity, heteroscedasticity and normality prompted remedial measures.

To address these violations, a **Box-Cox transformation was applied**, which suggested a $\lambda \approx 0.141$ value. This has been included in the appendix of this report as Figure 4.

Consequently, **a quarter-power transformation** ($SalePrice^{\frac{1}{4}}$) **was applied for interpretability and to improve adherence to assumptions**. Post-transformation diagnostics (**Figure 3 below**) showed significant improvement. Residuals appeared randomly distributed around zero, **addressing non-linearity and heteroscedasticity concerns**. The Q-Q plot demonstrated **better alignment with normality**, though slight deviations in the tails persisted.

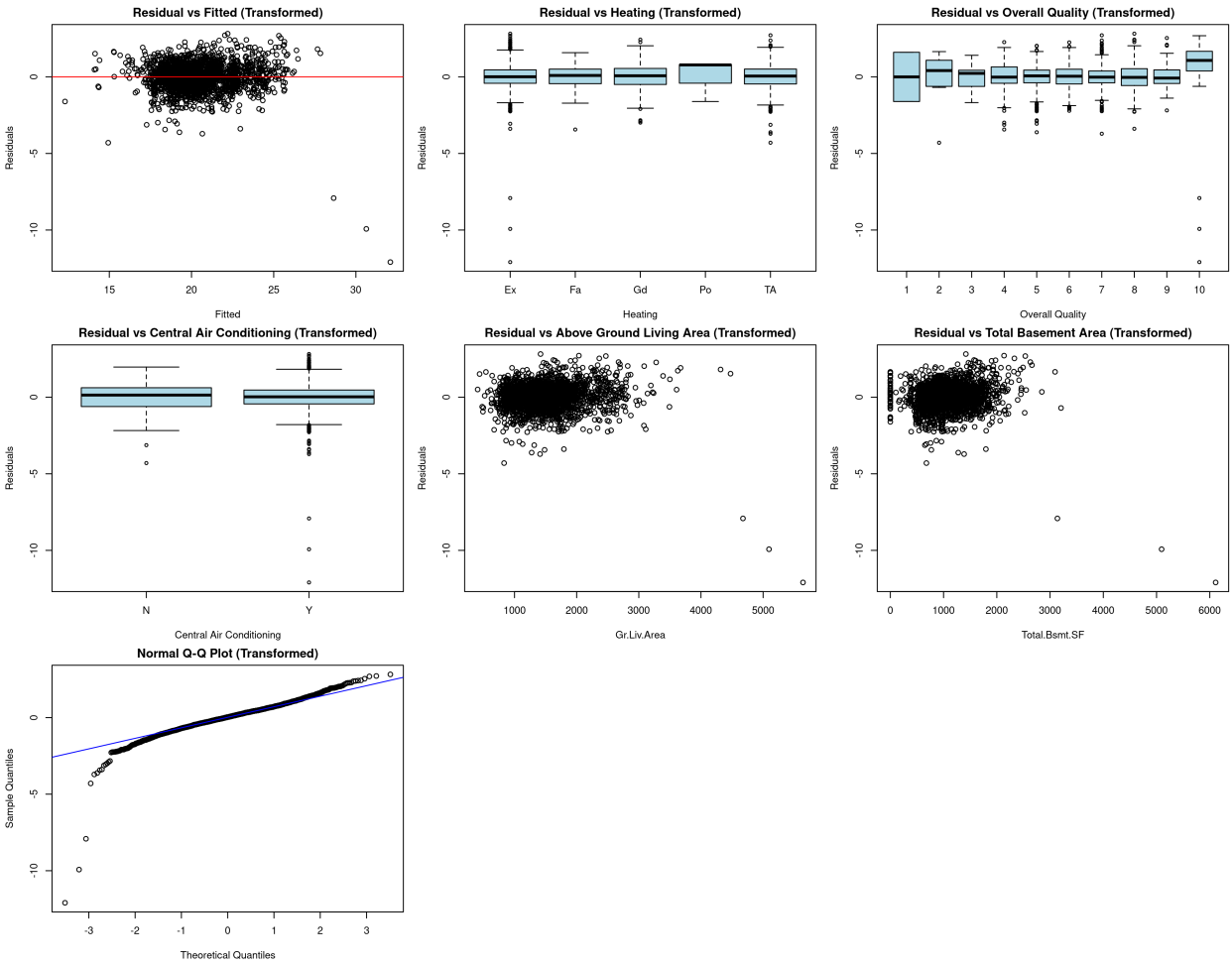

Figure 3: Diagnostic Plots for Transformed Model

Following the transformation, the **MLR conditions were reassessed**, and **VIF values were checked again**, confirming that the assumptions were now satisfactorily met. The VIF values have been included as **Table 1** below while the pairwise scatterplots of predictors for checking the conditional mean predictor condition has been included in the appendix section of this report as **Figure 5**:

Predictor	GVIF	Df	GVIF^(1/(2*Df))
Overall.Qual	3.287470	9	1.068352
Heating.QC	1.835478	4	1.078869
Central.Air	1.220194	1	1.104624
Gr.Liv.Area	1.648533	1	1.283952
Total.Bsmt.SF	1.680169	1	1.296213

Table 1: VIF Values Summary

Problematic Observations

Problematic points were identified using **leverage values, Cook's Distance, DFFITS, and DFBETAS**. **136 leverage points were identified**, likely representing houses with unusually large or small features such as square footage or basement area. These points reflect real-world variations in housing characteristics, aligning with our descriptive research goal to capture the full diversity of the dataset.

Similarly, **six regression outliers were found**, representing houses where actual sale prices deviated significantly from predicted values. These deviations may arise from unique factors like renovations or location-specific premiums and contribute to the tail behavior observed in the Q-Q plot.

Additionally, **one observation was flagged as influential by Cook's Distance**, while **113 points were flagged by DFFITS**, indicating that certain observations heavily influence the model. **DFBETAs revealed points with disproportionate impacts on specific coefficients**, likely due to unusual predictor combinations.

While these problematic points could have challenged model stability, they were **retained to ensure that the model remains comprehensive and reflective of real-world variability**, aligning with our descriptive research focus and to preserve its generalizability. **Table 2** below summarizes the types and number of problematic points:

Influence Diagnostic	Type of Problematic Point	Number of Problematic Points
Leverage Values	Leverage Point	136
Standardized Residuals	Outlier	6
Cook's Distance	Influential Point	1
DFFITS	Influential Point	113
DFBETAs (per coefficient)	Influential Point	[1], 7, 13, 17, 6, 6, 6, 7, 7, 7, 24, 57, 126, 6, 100, 96, 121, 121

Table 2: Problematic Points Summary

Statistical Testing

ANOVA tests verified the overall significance of the transformed model ($p < 0.05$), confirming that the predictors collectively explained a substantial portion of the variability in $SalePrice^{\frac{1}{4}}$.

Individual t-tests were also conducted for each predictor. Significant results ($p < 0.05$) confirmed that predictors such as Central Air Conditioning and Heating Quality levels had substantial impacts on the transformed response variable. These findings aligned with theoretical expectations and prior literature.

Curiously, **Overall Quality with a score of 2 had a p-value of 0.1669, indicating that it did not contribute significantly at the 0.05 threshold**. While non-significant predictors like this were considered as candidates for removal to improve model simplicity, they were **still included based on theoretical relevance to the research question and for maintaining continuity towards the Overall Quality Predictor**.

Table 3 in the appendix summarizes the results of the t-tests we conducted.

Final Model

The final model **retained all predictors and incorporated the quarter-power transformation of** $SalePrice$. The model is expressed as follows:

$$SalePrice^{\frac{1}{4}} = \beta_0 + \beta_1 \times Overall.Qual2 + \beta_2 \times Overall.Qual3 + \beta_3 \times Overall.Qual4 + \beta_4 \times Overall.Qual5 + \beta_5 \times Overall.Qual6 + \beta_6 \times Overall.Qual7 + \beta_7 \times Overall.Qual8 + \beta_8 \times Overall.Qual9 + \beta_9 \times Overall.Qual10 + \beta_{10} \times Heating.QCgd + \beta_{11} \times Heating.QCTA + \beta_{12} \times Heating.QCFa + \beta_{13} \times Heating.QCPo + \beta_{14} \times Central.AirY + \beta_{15} \times Gr.Liv.Area + \beta_{16} \times Total.Bsmt.SF$$

This model achieved an adjusted R^2 of **0.8263**, reflecting strong explanatory power while maintaining parsimony.

The **coefficients for the model including those for Heating Quality and Central Air Conditioning are highlighted in Table 4**, showcasing their significant contributions to house pricing. These results underline the importance of comfort and energy efficiency in determining property values.

Coefficient	Variable Name	Model Coefficient Estimate
β_0	Intercept	12.81
β_1	Overall Quality 2/10	1.018
β_2	Overall Quality 3/10	2.492
β_3	Overall Quality 4/10	2.741
β_4	Overall Quality 5/10	3.465
β_5	Overall Quality 6/10	3.898
β_6	Overall Quality 7/10	4.600
β_7	Overall Quality 8/10	5.624
β_8	Overall Quality 9/10	6.944
β_9	Overall Quality 10/10	6.542
β_{10}	Good Heating Quality	-0.2682
β_{11}	Typical/Average Heating Quality	-0.4820
β_{12}	Fair Heating Quality	-0.5351
β_{13}	Poor Heating Quality	-1.465

Coefficient	Variable Name	Model Coefficient Estimate
β_{14}	Central Air Conditioning	0.8647
β_{15}	Above Ground Living Area	0.001310
β_{16}	Total Basement Area	0.0007343

Table 4: Coefficients' Estimation Summary

Conclusions and Limitations

The primary research question of this study was: **How do heating quality and central air conditioning impact house sale prices?** The final multiple linear regression model indicates that **heating quality and central air conditioning significantly influence house prices**, even after controlling for other predictors like overall quality, above-ground living area, and basement area. Houses with better heating quality and central air conditioning were generally associated with higher transformed sale prices.

For example, the estimated coefficient for Central Air Conditioning is **0.865**, indicating that, on average, **houses with central air conditioning have an expected increase of 0.865 units** in $SalePrice^{\frac{1}{4}}$ compared to houses without central air, holding all other variables constant. This suggests central air conditioning has a strong positive association with transformed house prices. **For heating quality, houses rated as Po (poor) have a much lower expected transformed sale price than those rated as Ex (excellent), with an estimated coefficient of -1.465**, reflecting the substantial impact of heating quality on property value.

The findings are consistent with prior literature. **Hahn et al. (2018) found that energy-efficient systems raised property values**, aligning with our result that heating quality is a significant predictor. Similarly, **the importance of comfort amenities like central air conditioning has been noted in previous studies**, reinforcing our findings. These results are not surprising, given the increasing demand for energy-efficient and comfortable housing.

However, limitations exist in this analysis. First, **extreme observations (outliers) were identified in the data** and were addressed, but their potential influence on the model cannot be fully eliminated. Second, while diagnostic tests showed that most assumptions of linear regression were met, **minor violations such as residual non-normality and heteroscedasticity were observed**, which could affect the model's predictive accuracy. Finally, **the dataset was**

limited to Ames, Iowa, and may not generalize to other housing markets with different demographic or geographic factors.

In conclusion, the study provides a robust yet interpretable model that underscores the importance of heating quality and central air conditioning in determining house prices. **Addressing the identified limitations, such as extending the dataset to include diverse regions, could further enhance the applicability and accuracy of future analyses**.

Ethics Discussion

We opted for manual selection methods over automated ones for practical and ethical reasons. Manual selection allows for **deliberate consideration of the research context**, including theoretical grounding and prior studies, as demonstrated by the chosen predictors from the Ames Housing Dataset. Automated methods prioritize algorithmic efficiency and might **overlook nuances specific to the research question**, potentially leading to inappropriate or spurious predictors being included.

Ethically, **we believe automated methods carry a greater risk of negligence compared to manual selection**, particularly in the context of our research question on the impact of heating quality and central air conditioning on house sale-prices. Without careful oversight, these methods can **violate virtues like wisdom (zhi 智) and trustworthiness (xin 信)**, as described in the ethics module, by prioritizing algorithmic efficiency over thoughtful consideration of housing features and their relevance to pricing. For instance, wisdom may be compromised if an algorithm selects predictors based purely on statistical significance while overlooking theoretically relevant features like heating quality, which is supported by prior literature.

From a practical perspective, **manual selection aligns with our focus on interpretability and clarity, crucial for aiding stakeholders**. Automated methods are beneficial for larger, more complex datasets but **might compromise understanding in smaller, focused analyses like ours**. Using manual selection ensures predictors are meaningful, avoids overfitting, and reflects careful judgment.

Hence, **while both methods can be ethically acceptable, automated selection had a greater risk of negligence for our project**. We chose manual selection to **maintain relevance, ensure**

stakeholder clarity, and uphold virtues essential to responsible statistical practice. This approach avoids blameworthy practices and supports the project's goals.

Bibliography

Anselin, L., & Lozano-Gracia, N. (2008). Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics, 34*(1), 5–34. https://doi.org/10.1007/s00181-007-0152-3

Çağlayan, E., & Arıkan, E. (2011). Determinants of house prices in Istanbul: A quantile regression approach. *Quality & Quantity, 45*(2), 305–317. https://doi.org/10.1007/s11135-009-9296-x

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). https://jse.amstat.org/v19n3/decock.pdf

Hahn, J., Hirsch, J., & Bienert, S. (2018). Does "clean" pay off? Housing markets and their perception of heating technology. *Property Management, 36*(5), 575–596. https://doi.org/10.1108/PM-08-2017-0051

Prevek. (2024). Ames Housing Dataset. Retrieved from https://www.kaggle.com/datasets/prevek18/ames-housing-dataset

Appendix

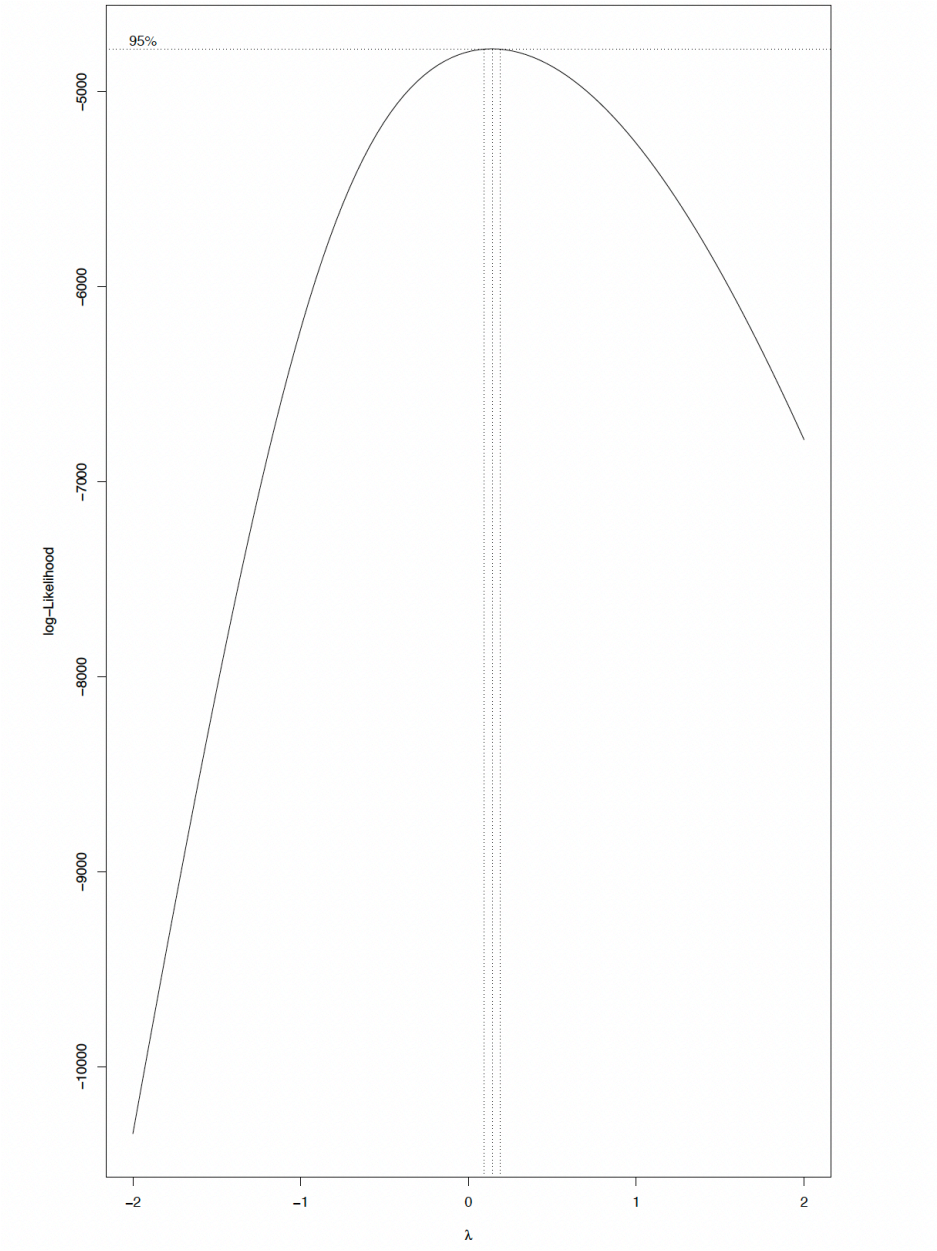

Figure 4: Box-Cox Suggestion on Initial Model

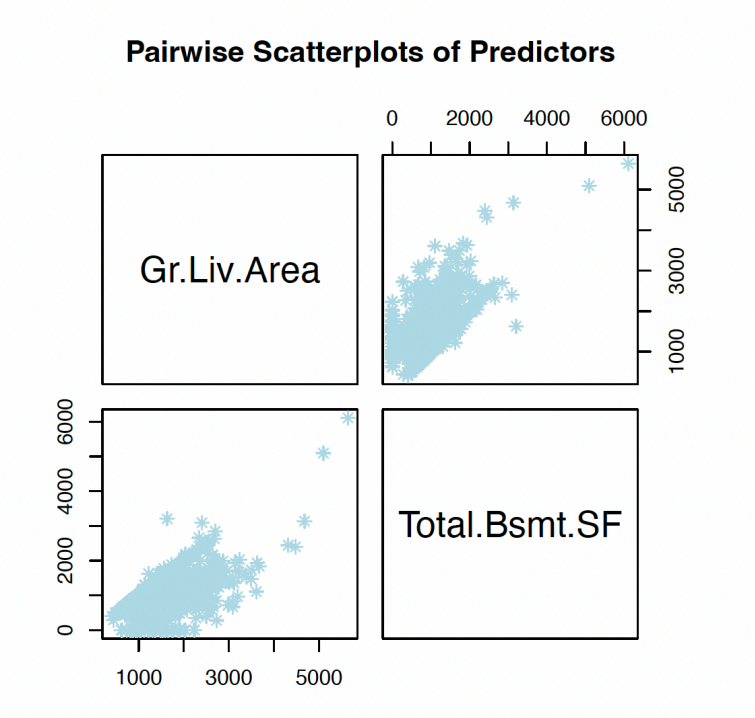

Figure 5: Pairwise Scatterplots of Predictors

Variable Name	T Value	Pr(>|t|)
Intercept	18.640	< 2E-16
Overall Quality 2/10	1.382	0.166965
Overall Quality 3/10	3.537	0.000413
Overall Quality 4/10	4.017	6.10E-05
Overall Quality 5/10	5.086	3.96E-07
Overall Quality 6/10	5.716	1.24E-08
Overall Quality 7/10	6.733	2.10E-11
Overall Quality 8/10	8.202	3.93E-16
Overall Quality 9/10	10.039	< 2E-16
Overall Quality 10/10	9.213	< 2E-16
Good Heating Quality	-4.354	1.40E-05
Typical/Average Heating Quality	-4.787	1.81E-06
Fair Heating Quality	-2.622	0.008805
Poor Heating Quality	-9.674	< 2E-16
Central Air Conditioning	9.598	< 2E-16
Above Ground Living Area	27.857	< 2E-16
Total Basement Area	13.798	< 2E-16

Table 3: T-tests Summary