# Data Analysis on Housing Price Prediction

**AMS 325: Computing and Programming Fundamentals in Applied Mathematics and Statistics**

**Stony Brook University**

Dylan Maray: Group Coordinator, Code, Report/Presentation Reviewer
Nancy Huang: Presentation Creator
Raymond Pepper: Final Report and Presenter
Sophia Sieli: Proposal, Report/Presentation Format Editor, Presenter
Stephanie Dong: Proposal, Final Report, and Presenter
Madeline Groth: Final Report and Presentation Creator

## **I. Introduction:**

In today's real estate market, understanding the key drivers behind housing prices is crucial for buyers, sellers, investors, and policymakers. House prices are influenced by a variety of factors, including location, size, economic conditions, and property-specific features. Accurate price prediction models can aid in managing affordability, planning sustainable communities, and making smart house investment decisions, while also preventing speculative market bubbles.

This project focuses on building a statistical model to estimate housing prices using multiple linear regression models, based on a dataset of residential properties located in the Delhi region. The dataset that we will use in this analysis contains information on approximately 500 residential properties, including area (in square footage), number of bedrooms and bathrooms, number of stories, availability of air conditioning, basement, parking, and furnishing status. These are features that are commonly used in real estate to determine and estimate a housing market's value. To narrow our scope and reduce variability, we will limit our analysis to furnished houses only, as furnishing status may introduce uncontrolled differences in price that are not directly related to structural features of the house.

The primary goal of this project is to analyze how these housing features affect housing prices and identify the most influential predictors, while assessing how well a linear model can modify those relationships. We are particularly interested in examining the role of square footage (area) in predicting price and whether its effect changes depending on the number of bedrooms. To explore these questions, we form the following two hypotheses:

- Hypothesis 1: Does the area of a house significantly affect the sale price?
  - $H_0$: The area of a house does not significantly affect sale price (the coefficient of the area is zero).
  - $H_1$: The area of a house does significantly affect price (the coefficient of the area is not equal to zero).

- Hypothesis 2: Does the effect of area on price depend on the number of bedrooms?
  - $H_0$: There is no interaction between area and bedrooms (the interaction term between area and bedrooms has a coefficient of zero).
  - $H_1$: There is a significant interaction effect between area and bedrooms.

To ensure statistical validity, we will apply key linear regression assumption tests (including tests for normality, homoscedasticity, independence, and multicollinearity), explore model improvement strategies such as log and Box-Cox transformations, and use regularization techniques like ElasticNet to evaluate model performance under potential multicollinearity. Model performance is evaluated using metrics such as adjusted $R^2$, AIC, BIC, and test set RMSE.

Through this analysis, we aim to gain a deeper understanding of what drives housing prices and evaluate the performance of regularized regression techniques in predicting real estate values.

## II. Literature Review:

In examining existing approaches to housing price prediction, many analyses follow a structured methodology incorporating data cleaning, exploratory data analysis (EDA), feature engineering, model selection, and evaluation using performance metrics such as RMSE and $R^2$. These studies often utilize regression techniques such as regularized models, like Lasso and Ridge regression, due to their ability to handle multicollinearity and improve model generalization. Since our analysis aligns with this framework, we focus on understanding how housing characteristics,  location, and the number of bedrooms impact sale prices. After loading and preparing the dataset, we assess regression assumptions and apply regularized models to enhance robustness. This mirrors the sequence and modeling philosophy found in prior housing data analyses, where linear regression and its penalized variants are commonly employed.

In a 2005 study by Hui Zou and Trevor Hastie at Stanford University, the authors introduced Elastic Net, a regularization method that combines the strengths of Lasso and Ridge regression. Lasso regression is limited by the inability to select more than n variables when $p> n$

and its tendency to select only one variable from groups of highly correlated predictors. The Elastic Net applies both L1 and L2 penalties, enabling simultaneous feature selection and coefficient shrinking while encouraging a grouping effect. Therefore, strongly correlated predictors will tend to be selected together. This behavior is most useful in high-dimensional data, such as gene expression data, where grouped variables carry joint predictive power. Zou and Hastie found that ElasticNet often outperformed Lasso in prediction accuracy, especially in cases of multicollinearity, or when $p > n$. Based on these findings, our project includes Elastic Net regression to cross-validate predictive performance while leveraging collinearity and model generalization(Zou & Hastie, 2005). We also integrate residual analysis and hypothesis testing for a stronger assessment of regression assumptions, allowing us to explore interaction effects and diagnostic criteria in greater depth.

In conclusion, while previous studies have demonstrated the strengths of Lasso regression in predicting housing prices, our analysis adds value by using Elastic Net regression and integrating thorough model diagnostics and hypothesis testing. This supports the broader literature that shows regularized models provide reliable, interpretable, and often superior results in housing data applications.

## III. Methods:

### A. Data Preprocessing:

Our dataset contains 545 housing records with 13 variables; five quantitative (area, bedrooms, bathrooms, stories, parking) and seven categorical (mainroad, guestroom, basement, hot water heating, air conditioning, prefarea, furnishing status). We focus on furnished homes only to reduce the price variability caused by differences in furnishing, which leaves us with 140 observations. Categorical variables with binary outcomes were encoded using 0/1, and multicollinearity was checked using a correlation matrix. This process ensured our predictors were in a form suitable for linear regression.

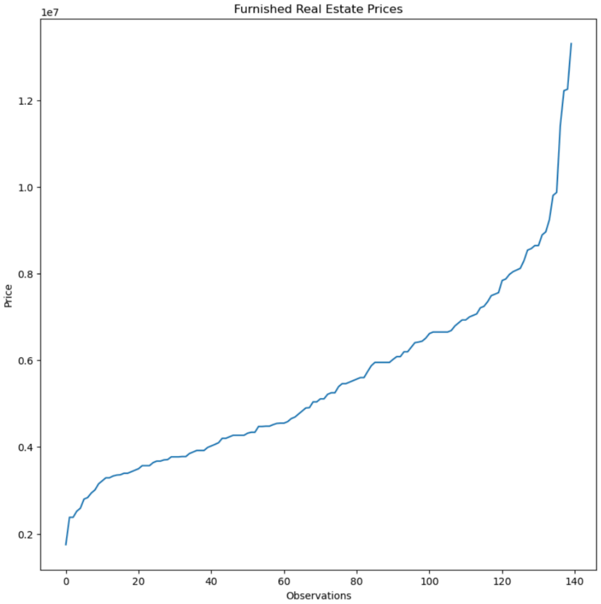


*Fig. 1: Housing Prices vs. Observation Index*

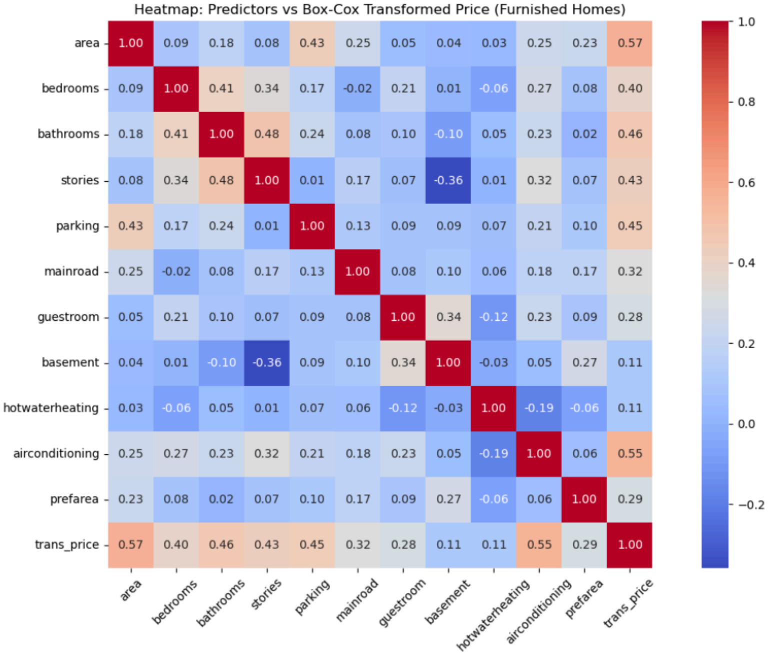


*Fig. 2: Correlation Heatmap of Predictors*

### B. Initial OLS Model and Assumption Testing:

The data is cross-sectional because observations were collected in a single point in time and without any explicit time order. In cross-sectional data, Independence is likely assumed. The initial OLS regression model included all predictors. It achieved an R-squared of 0.686, suggesting reasonable explanatory significance. However, diagnostic plots indicated violations of key assumptions.

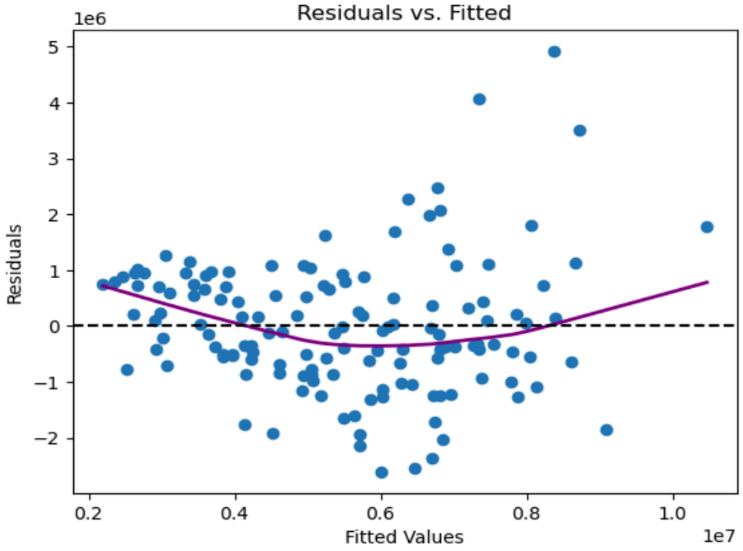


*Fig. 3: Residuals vs. Fitted (Pre-transformation)*

The residuals exhibited curvature, indicating non-linearity. Also, the Shapiro-Wilk test (p < 0.05) showed residuals were not normally distributed.

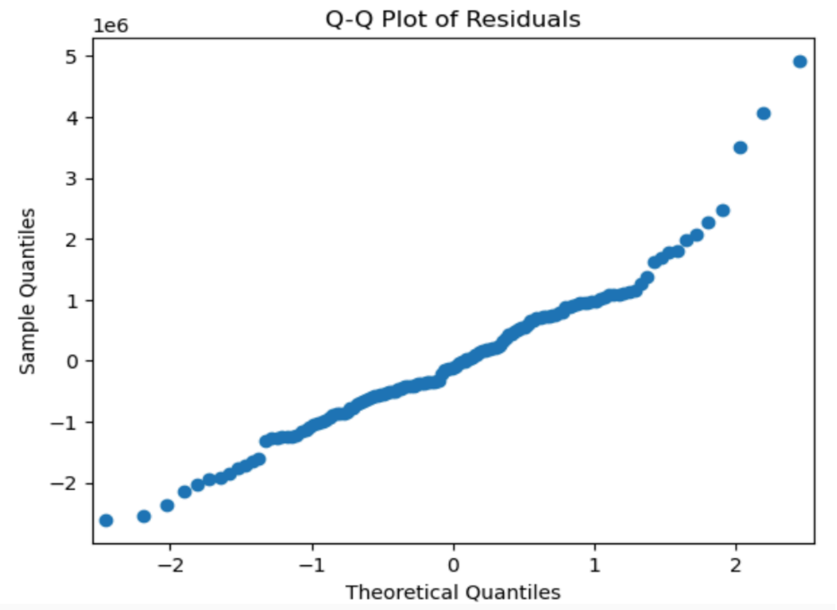


*Fig. 4: Q-Q Plot of Residuals (Pre-transformation)*

These findings necessitated transformation of the dependent variable to stabilize variance and improve normality.

C. Box-Cox Transformation:

Using the Box-Cox log-likelihood profile, we identified an optimal lambda of -0.0618. This transformation resulted in a better-fitting model (R-squared = 0.724).

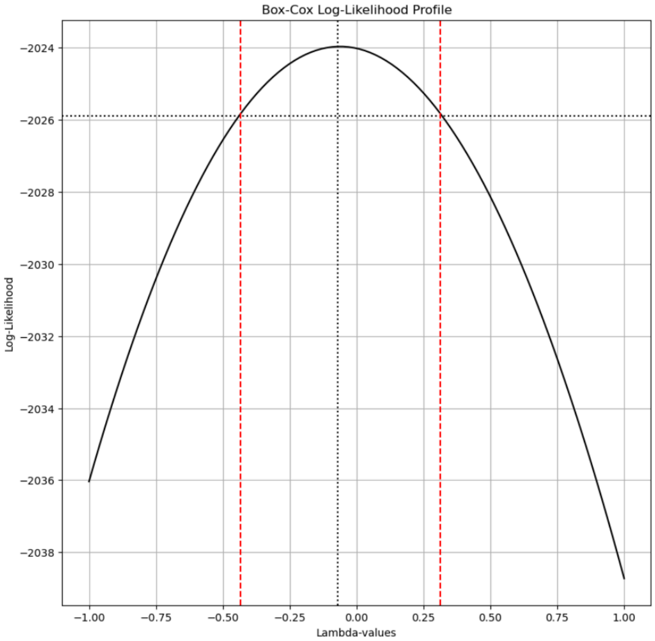


*Fig. 5: Box-Cox Log-Likelihood Profile*

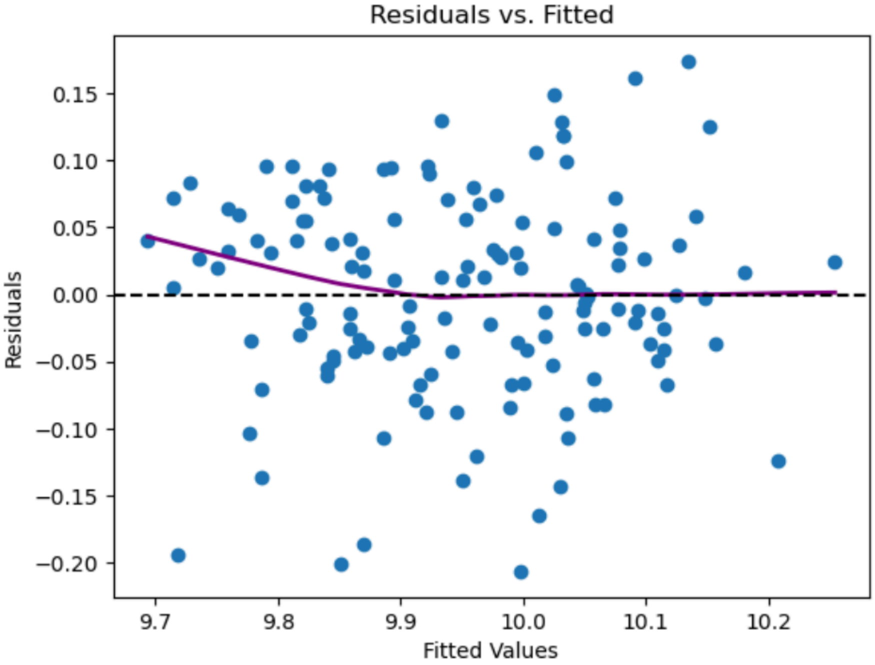
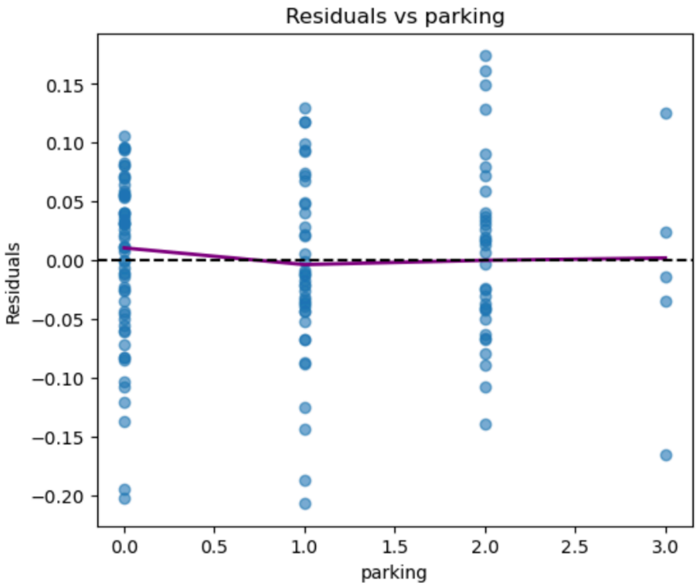
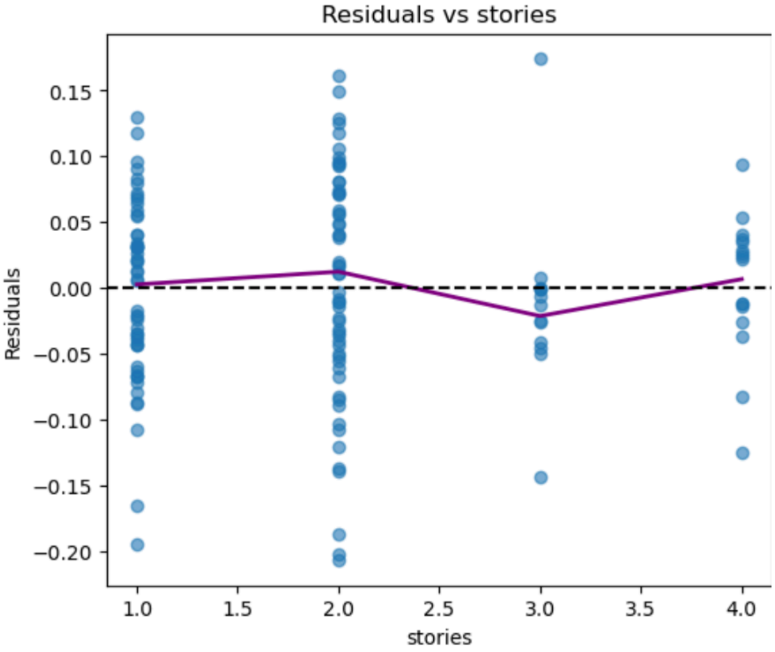
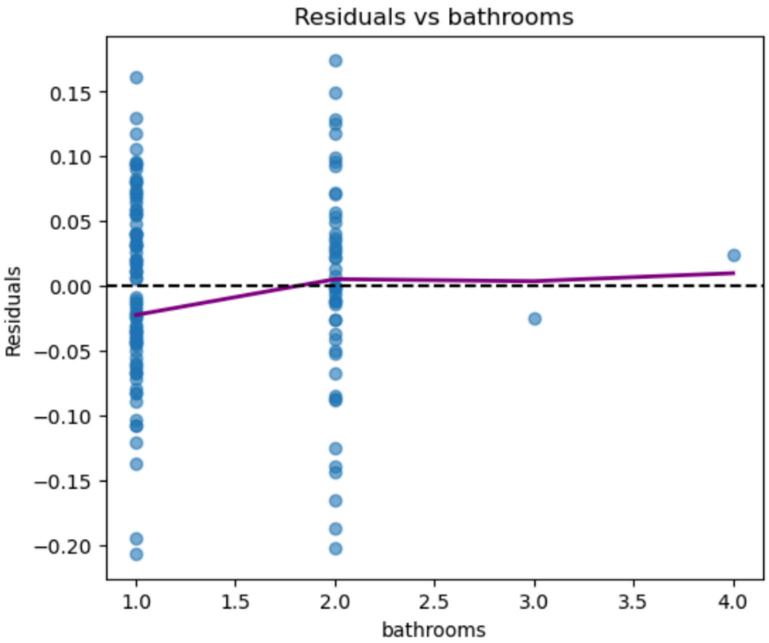


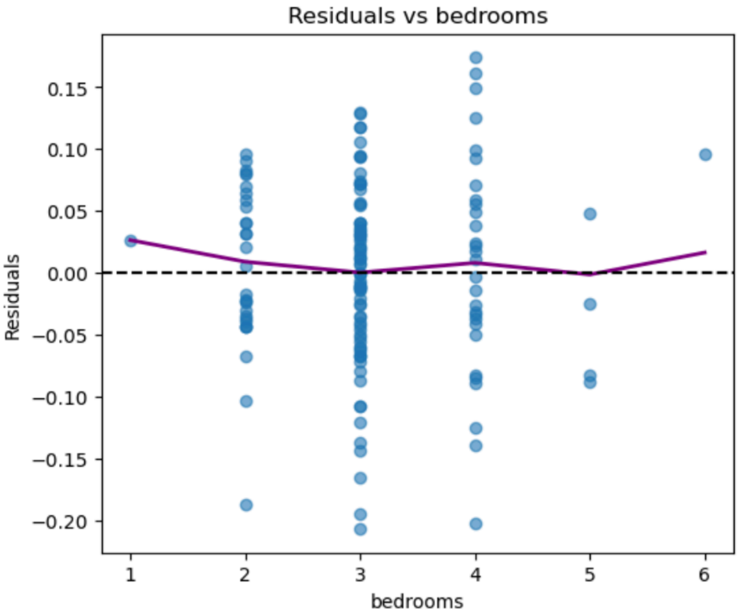
*Fig. 6: Residuals vs. Fitted (Box-Cox Transformed)*

Post-transformation residual plots showed no significant curvature and homoscedasticity appeared to be satisfied. Additional residual plots by predictor also confirmed no obvious non-linearities or variance issues.






Residuals vs bathrooms
Residuals
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
1.0
1.5
2.0
2.5
3.0
3.5
4.0
bathrooms


Residuals vs bedrooms
Residuals
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
1
2
3
4
5
6
bedrooms

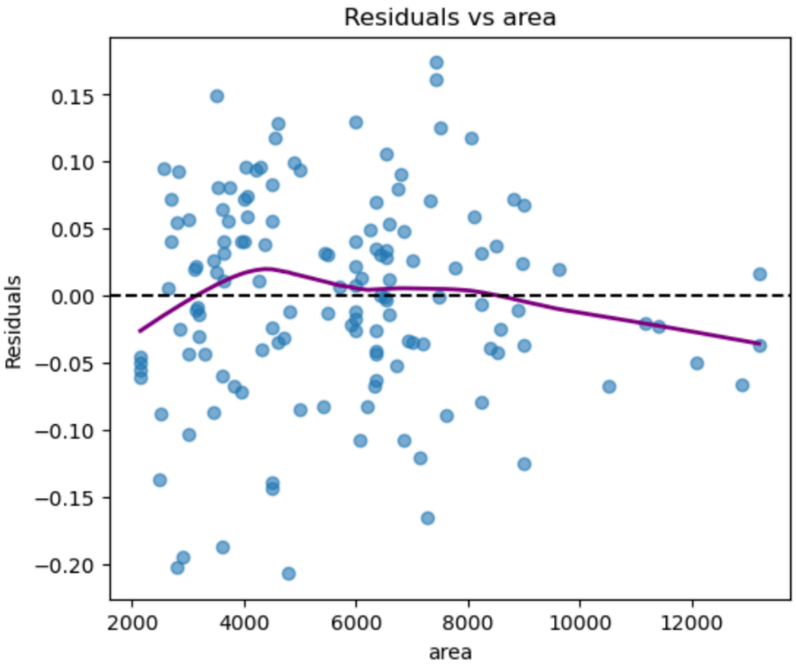
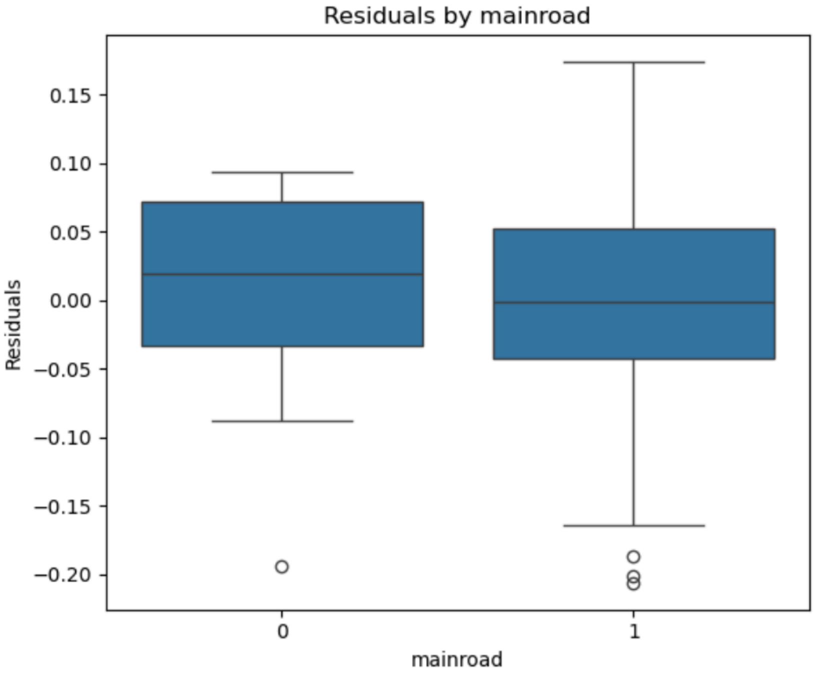
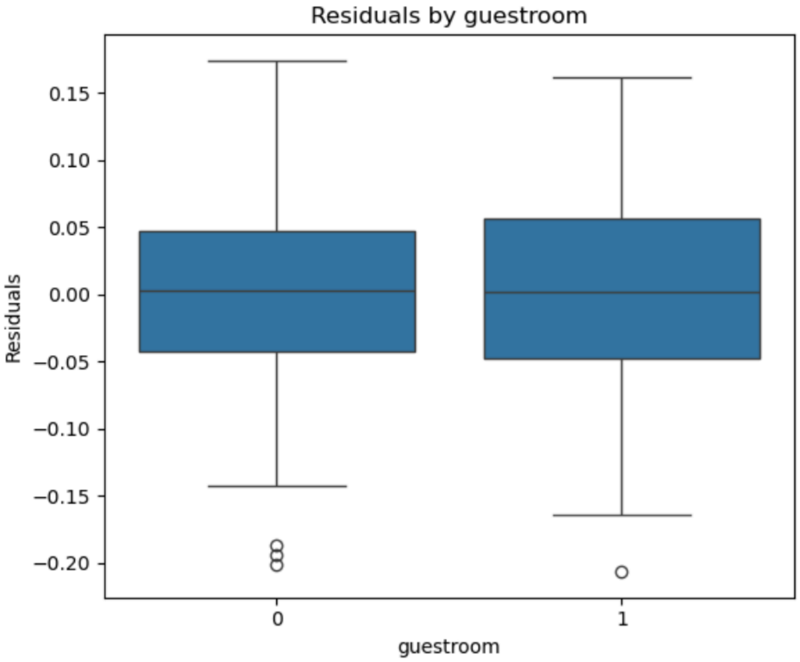
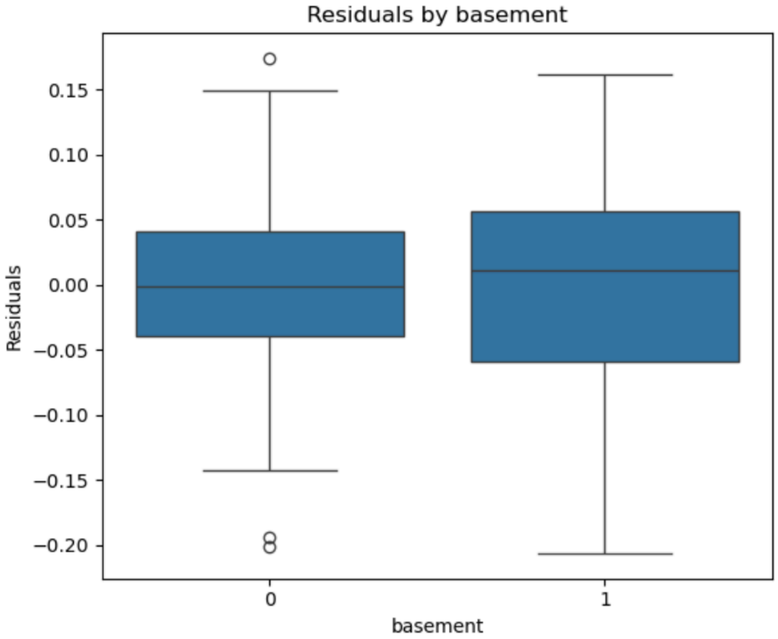
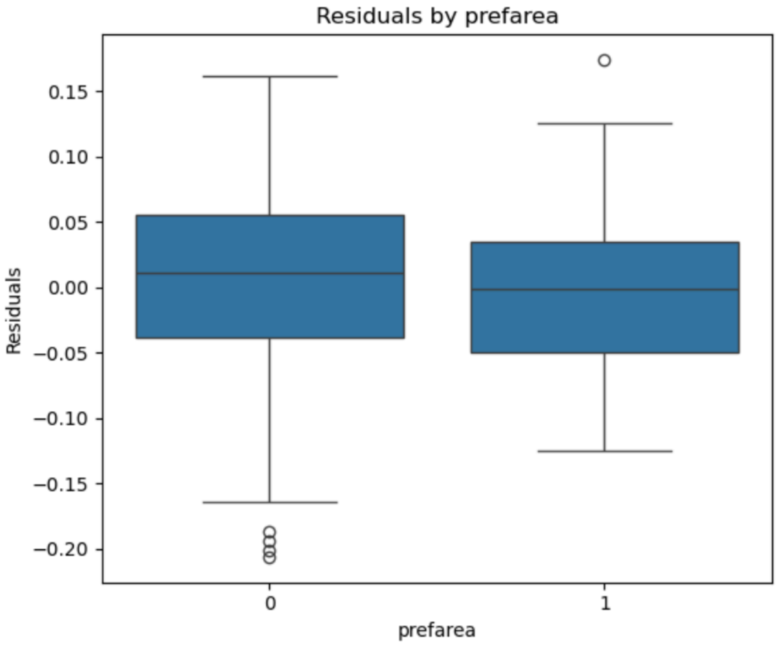


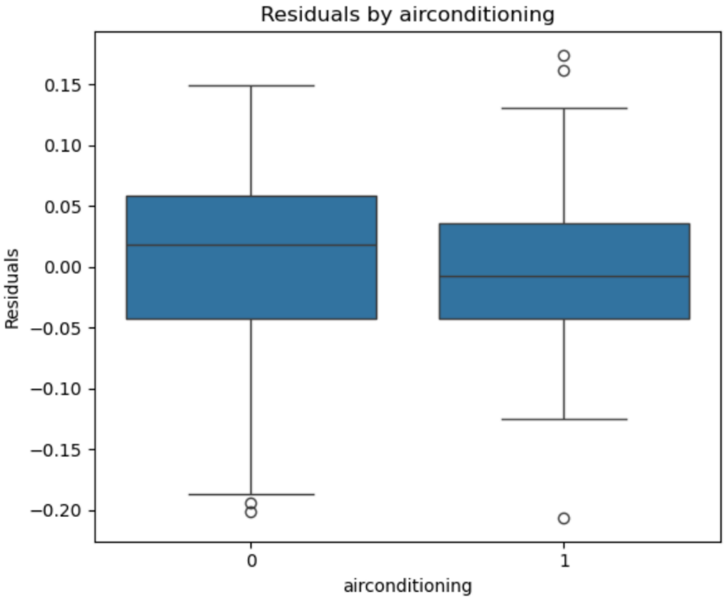
*Fig. 7: Residuals vs. Quantitative Predictors*




Residuals by guestroom
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
Residuals
0
1
guestroom


Residuals by basement
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
Residuals
0
1
basement


Residuals by prefarea
Residuals
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
0
1
prefarea


Residuals by airconditioning
Residuals
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
0
1
airconditioning

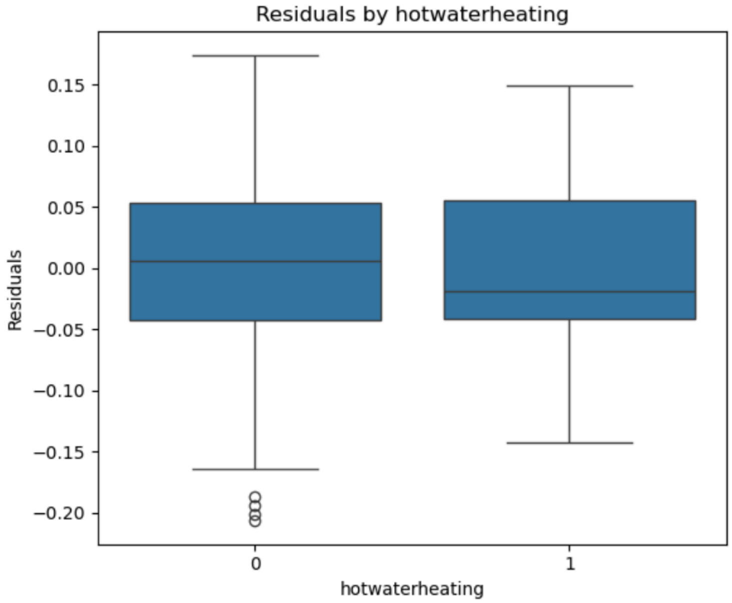


*Fig. 8: Residuals vs. Categorical Predictors (Box Plots)*

Breusch-Pagan test: p = 0.7244

Shapiro-Wilk test: p = 0.2305

Durbin-Watson: 1.308

Together, these suggest the assumptions of Linearity, Homoscedasticity, and Normality. Independence is likely assumed because the data is cross-sectional.

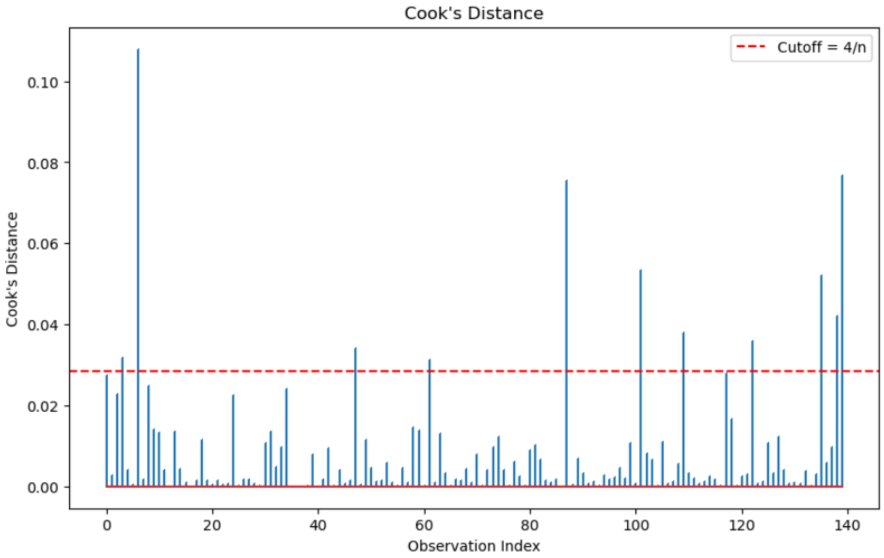


*Fig. 9: Cook's Distance Plot*

11 data points identified as influential, but none severely distorted the model.

D. Log Transformation (Alternative):

For robustness, we also tested a log transformation. This resulted in comparable R-squared (0.724), improved skewness (from 1.08 to 0.06), and better-behaved scatterplots.

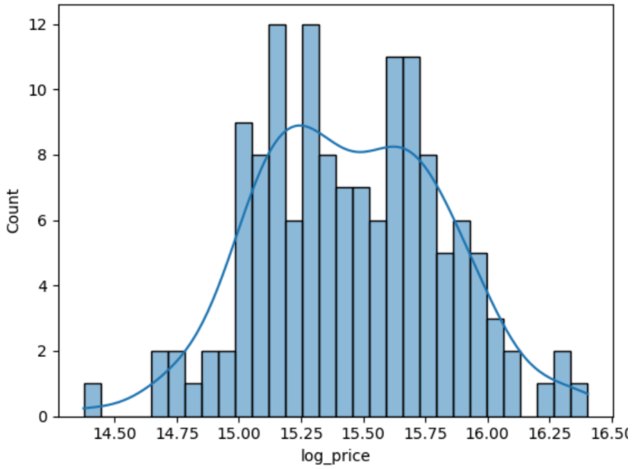


*Fig. 10: Histogram of Log-Transformed Prices*

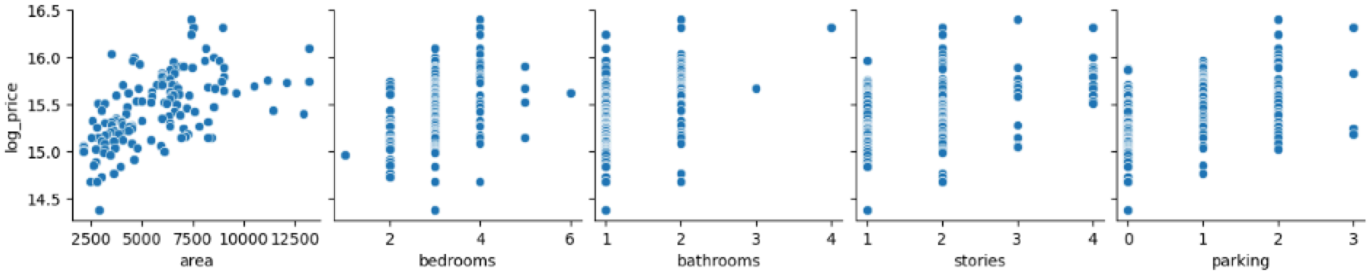


*Fig. 11: Pairplot: log_price vs. Predictors (Log-Transformed)*

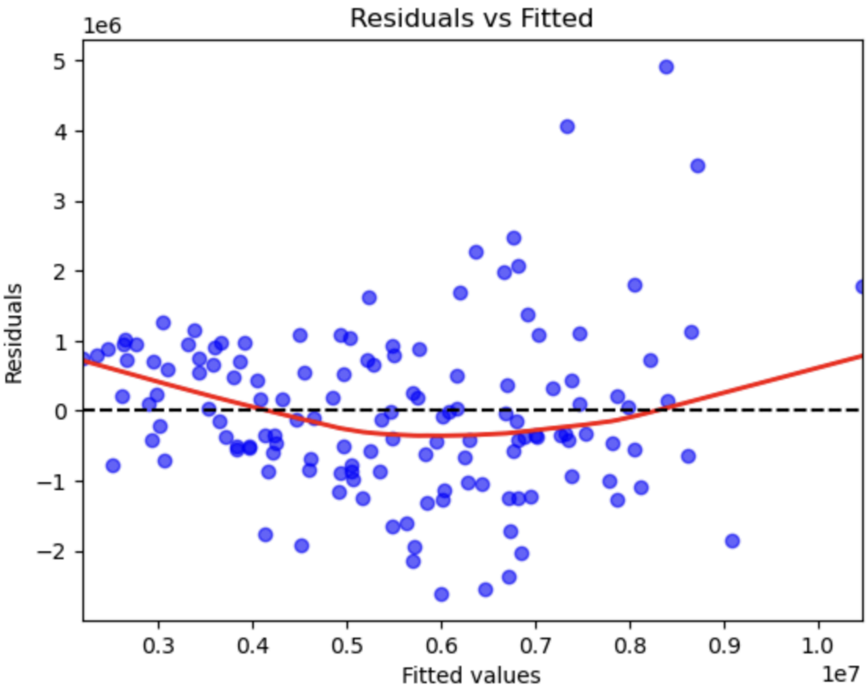


*Fig. 12: Residuals vs. Fitted (Log-Transformed)*

Despite satisfying the Normal assumption, the log transformation failed to satisfy the Linearity and Homoscedasticity assumption. From the Log-Transformed residual plot, there appears to be an upward parabola. On the other hand, the Box-Cox transformation satisfied the assumptions of Linearity and Homoscedasticity (see *Fig. 6*). Recall, that original data is already cross-sectional which means that Independence is likely assumed.

## E. Model Selection:

When deciding between the Log transformation and the Box-Cox Transformation, it is reasonable to conclude that Box-Cox is the better transformation to use since Linearity, Normality, Homoscedasticity, and Independence are satisfied.

Best Model: Box-Cox Transformation

## F. Model Comparison:

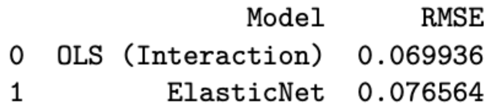
We evaluated three models:

Model 1: OLS without interaction (best model)

Model 2: OLS with interaction between area and bedrooms

Model 3: ElasticNetCV

ElasticNetCV had slightly higher test RMSE (0.077) than Box-Cox OLS (0.070), and VIF < 2 confirmed low multicollinearity. Thus, we favored the simpler OLS model.



```
                  Model      RMSE
0  OLS (Interaction)  0.069936
1         ElasticNet  0.076564
```

*Fig. 13: RMSE Comparison: OLS vs. ElasticNet*

## IV: Hypothesis

**A). Hypothesis 1:** Does the area of a house significantly affect sale price?

The first hypothesis is to evaluate whether the square footage of a house (area) has a statistically significant effect on its sale price. In real estate, area is often one of the most influential predictors of value, and we aim to confirm this through the regression analysis. We set up the following hypothesis and compared the model below:

$H_0$: The area of a house does not significantly affect sale price (the coefficient of the area is zero).

$M_1$: trans_price (all predictors except area) - a model that includes all predictors except area

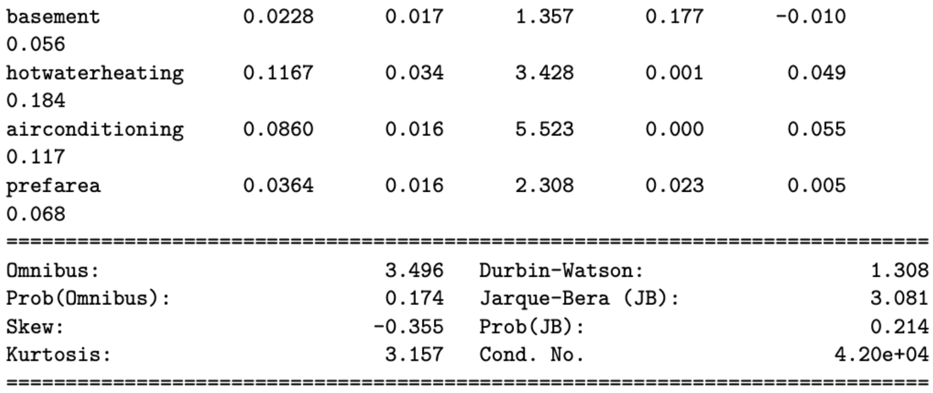
$H_1$: The area of a house does significantly affect price (the coefficient of the area is not equal to zero).

$M_2$: trans_price (all predictors including area) - a full model including area, along with all other predictors

Result and Conclusion:

The Box-Cox transformed OLS regression output (Figure 12) shows that the variable area has an estimated coefficient of $1.913 \times 10^{-5}$ with a standard error of $3.25 \times 10^{-6}$. The associated t-statistic is 5.893, and the p-value is less than 0.001, indicating there is a highly significant positive relationship between area and transformed price. In addition, the 95% confidence interval for the area coefficient range from $1.27 \times 10^{-5}$ $to$ $2.56 \times 10^{-5}$ does not contain zero, which further supports the statistical significance of the area. The overall model exhibits an adjusted R² value of 0.700, which means that approximately 70% of the variance in the transformed sale price is explained by the predictors in the model. The model's fit is supported by an AIC of -304.3 and BIC of -269.0, indicating that the model has a strong overall performance.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            trans_price   R-squared:                       0.724
Model:                            OLS   Adj. R-squared:                  0.700
Method:                 Least Squares   F-statistic:                     30.45
Date:                Thu, 08 May 2025   Prob (F-statistic):           1.37e-30
Time:                        18:53:47   Log-Likelihood:                 164.16
No. Observations:                 140   AIC:                            -304.3
Df Residuals:                     128   BIC:                            -269.0
Df Model:                          11
Covariance Type:            nonrobust
================================================================================
===
                  coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
---
const           9.5342      0.040    239.608      0.000       9.455
9.613
area         1.913e-05   3.25e-06      5.893      0.000    1.27e-05
2.56e-05
bedrooms        0.0206      0.010      2.107      0.037       0.001
0.040
bathrooms       0.0382      0.015      2.520      0.013       0.008
0.068
stories         0.0286      0.009      3.033      0.003       0.010
0.047
parking         0.0217      0.009      2.539      0.012       0.005
0.039
mainroad        0.0478      0.029      1.646      0.102      -0.010
0.105
guestroom       0.0369      0.017      2.128      0.035       0.003
0.071
```



```
basement              0.0228       0.017       1.357       0.177      -0.010
0.056
hotwaterheating       0.1167       0.034       3.428       0.001       0.049
0.184
airconditioning       0.0860       0.016       5.523       0.000       0.055
0.117
prefarea              0.0364       0.016       2.308       0.023       0.005
0.068
==============================================================================
Omnibus:                        3.496   Durbin-Watson:                   1.308
Prob(Omnibus):                  0.174   Jarque-Bera (JB):                3.081
Skew:                          -0.355   Prob(JB):                        0.214
Kurtosis:                       3.157   Cond. No.                     4.20e+04
==============================================================================
```

*Fig. 14: ANOVA Table: Inclusion of Area*

Based on the statistical evidence above, we reject the null hypothesis and conclude that the area is a statistically significant predictor of sales price. As the size of a house increases, the price tends to increase as well. This decision is supported by the small p-value (p-value <0.05), the 95% confidence interval excluding zero, and the positive coefficient. These results indicate a consistent and meaningful relationship between house size and price, aligning with both statistical evidence and real-world expectations.

**B). Hypothesis 2:** Does the effect of area on price depend on the number of bedrooms?

The second hypothesis is to evaluate whether the impact of the square footage of a house (area) on sale price varies depending on the number of bedrooms. In other words, we are testing for the presence of an interaction effect between area and bedrooms. While both features may individually influence price, this hypothesis investigates whether their combined influence is stronger or weaker than expected from their separate effects. We set up the following hypothesis and compared the model below:

$H_0$: There is no interaction between area and bedrooms (the interaction term between area and bedrooms has a coefficient of zero).

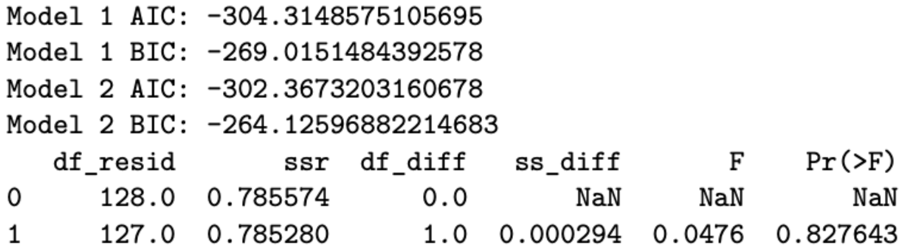
$M_1$: trans_price (area + bedrooms + other predictors) - a model that contains all predictors except the interaction term area × bedrooms

$H_1$: There is a significant interaction effect between area and bedrooms.

$M_2$: trans_price (area  bedrooms + other predictors) -  a full model including all predictors, including the interaction term area × bedrooms

Result and Conclusion:

The ANOVA output (Figure 13) shows that the interaction term does not significantly improve model performance. The F-statistic is 0.0476 with a p-value of 0.8276, which is significant above 0.05. This suggests that the interaction effect is not statistically meaningful. In addition, the model that includes the interaction term area × bedrooms (Model 2) has a higher AIC value of -302.37 and a BIC value of  -264.13 than the model that does not include the interaction term area × bedrooms (Model 1), which is AIC = -304.31 and BIC = -269.02. This indicates that the simpler model is preferred.



```
Model 1 AIC: -304.3148575105695
Model 1 BIC: -269.0151484392578
Model 2 AIC: -302.3673203160678
Model 2 BIC: -264.12596882214683
   df_resid       ssr  df_diff   ss_diff       F    Pr(>F)
0     128.0  0.785574      0.0       NaN     NaN       NaN
1     127.0  0.785280      1.0  0.000294  0.0476  0.827643
```

*Fig. 15: ANOVA Table: Interaction Effect*

Based on the statistical evidence, we fail to reject the null hypothesis and conclude that there is no significant interaction between area and the number of bedrooms in predicting house price. The relationship between area and price appears to be consistent across houses with varying bedroom numbers. Therefore, the interaction term does not add meaningful explanatory power to the model.

## **V: Conclusion**

In this project, we developed a regression model to predict housing price using various property features, including area, number of bedrooms and bathrooms, and several binary housing attributes for furnished houses. Through a series of modeling steps and assumption checks, we systematically refined our approach to ensure both statistical validity and practical interpretability. Our initial multiple linear regression model revealed that area was a highly significant predictor of sale price. This finding is also consistent with the expectations in real estate valuation. In order to address the normality and variance issues in the residuals, we applied both log and Box-Cox transformations, ultimately selecting the Box-Cox transformation due to its better diagnostic performance. The transformed model achieved a strong adjusted $R^2$ of 0.700 and passed all key regression assumptions, including linearity, normality, and homoscedasticity.

We also investigated two hypotheses. The first confirmed that the area of a house is a statistically significant positive predictor of its market price. This finding is consistent with intuition and supported by rigorous statistical testing. We also confirmed that interaction effects between area and number of bedrooms are not statistically significant, implying that the effect of square footage on price is relatively stable regardless of the bedroom count.

To further improve the model's reliability and access predictive performance, we explored regularization methods using ElasticNet regression. While ElasticNet provided a comparable RMSE on the test set, it did not improve the performance compared to the Box-Cox transformed OLS model due to the absence of strong multicollinearity. Thus, we selected the Box-Cox transformed OLS model without the interaction of area × bedrooms as our final model due to its stronger interpretability and superior performance.

Overall, this project produced a statistically reliable model for predicting house prices and also showed valuable insights in the modeling process. We demonstrated the importance of balancing interpretability with predictive performance by carefully addressing the assumptions, testing multiple transformations, and comparing both traditional and regularized regression approaches. While more advanced methods like ElasticNet offer useful alternatives, our results show that a well-specified linear model with an appropriate transformation can offer both accuracy and clarity. Future studies for this project may expand by incorporating temporal or geographic variables or applying nonlinear models to further enhance the predictive power of real estate valuation.

# Works Cited

Ashydv. (2019, March 12). *Housing Price Prediction ( linear regression )*. Kaggle. https://www.kaggle.com/code/ashydv/housing-price-prediction-linear-regression/notebook.

Manimala. (2017, August 3). *Boston House Prices*. Kaggle. https://www.kaggle.com/datasets/vikrishnan/boston-house-prices.

Seabold, S., & Perktold, J. (2010, May 1). *Statsmodels: Econometric and Statistical Modeling with Python - SciPy Proceedings*. scipy. https://proceedings.scipy.org/articles/Majora-92bf1922-011.

Zhang, T., & Zizheng, L. (2021). *(PDF) A Comparative Study of Regression Models for Housing Price Prediction*. Comparative Study on Regularized Regression for Housing Prediction. https://www.researchgate.net/publication/383112591_A_Comparative_Study_of_Regression_Models_for_Housing_Price_Prediction.

*Regularization and Variable Selection via the Elastic Net*, sites.stat.washington.edu/courses/stat527/s14/readings/zouhastie05.pdf. Accessed 20 May 2025.