AMS 325 Final Project

# Housing Price Predictions Using Statistical Modeling

Raymond Pepper, Sophia Sieli, Stephanie Dong, Dylan Maray, Nancy Huang, Madeline Groth

# Overview

Background & Importance:

In today's real estate market, understanding the key drivers behind housing prices is crucial for buyers, sellers, investors, and policymakers.

- Gain a deeper understanding of what drives housing prices and evaluate the performance of regularized regression techniques in predicting real estate values for the market.

Goal: Analyze how housing features affect housing prices and identify the most influential predictors, while assessing how well linear models can modify such relationships. More specifically, we estimated housing prices using multiple linear regression models, based on a dataset of residential properties located in the Delhi region.

Software and Library: Python (libraries including pandas, numpy, seaborn, matplotlib, and statsmodels)

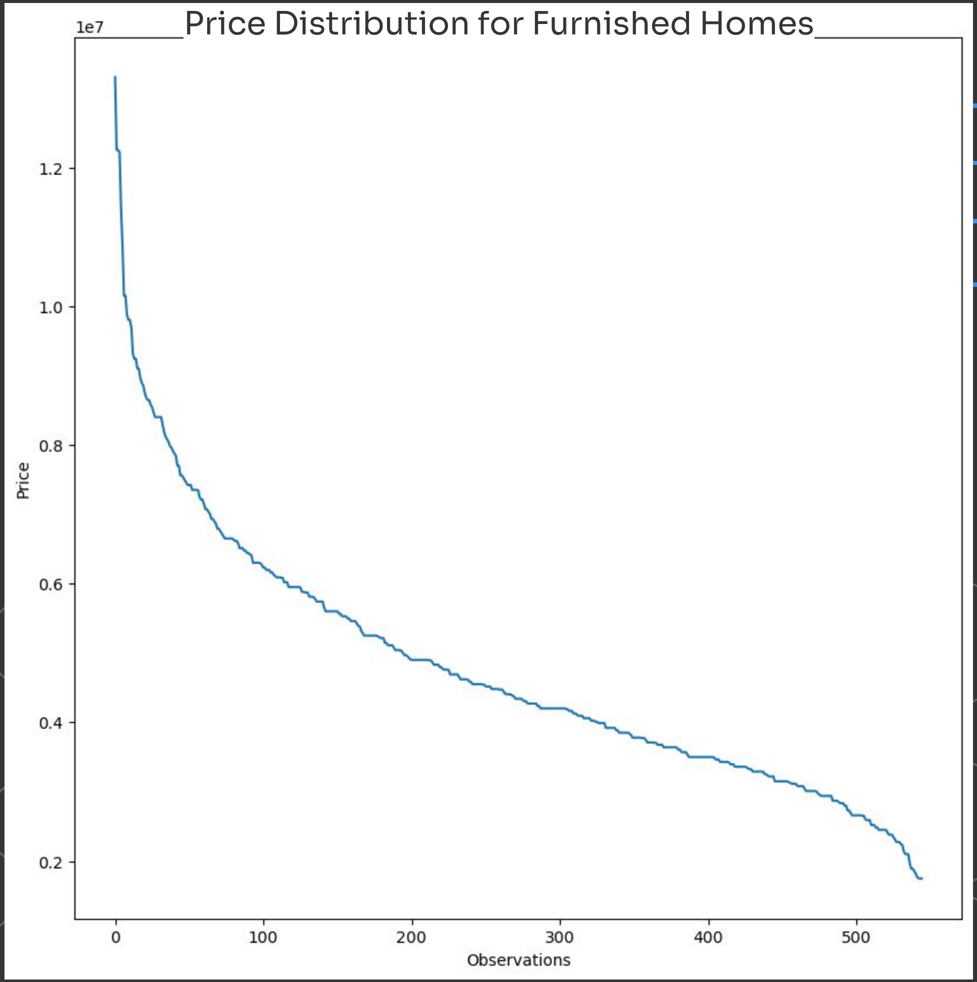
# Initial Price Visualization

Data set includes:
- 545 house listings
- Each described by 13 attributes

Our target “price” with predictors including:
- Numerical: area, bedroom, bathrooms, stories etc.
- Categorical: mainroad, guest room, air conditioning etc.

We narrow our focus to **furnished homes** (140 observations) to reduce variability caused by furnished status.

`Matplotlib.pyplot, seaborn were used`

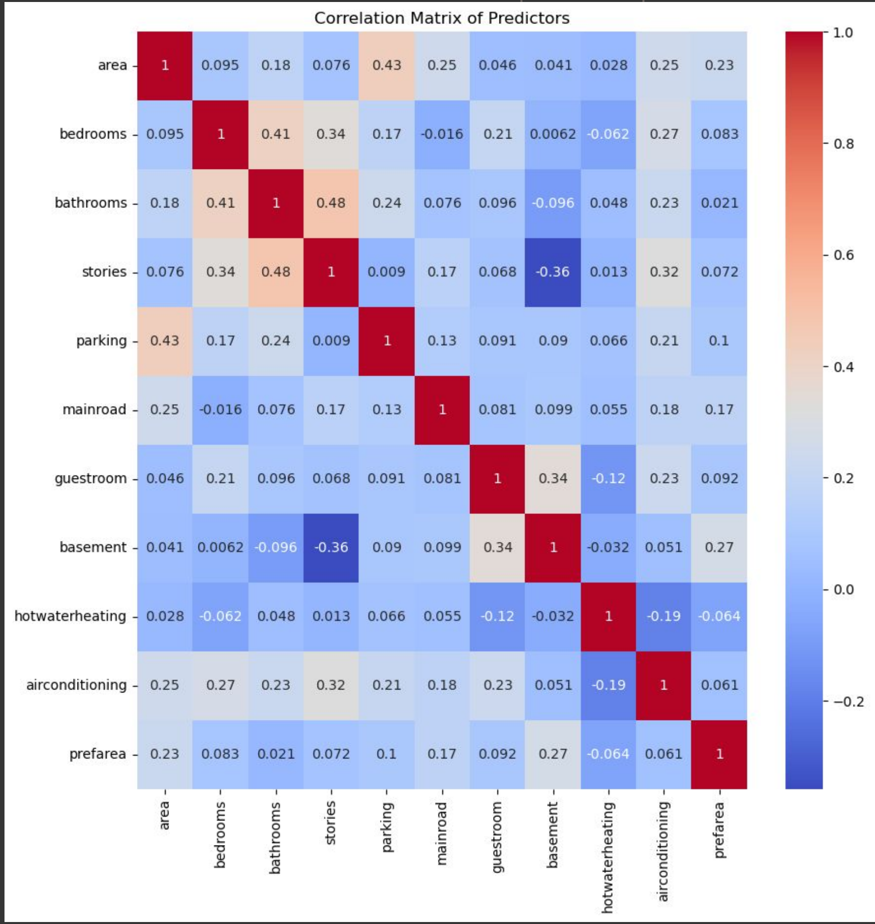


# Correlation Matrix & Multicollinearity

To ensure that our regression model won't suffer from multicollinearity, we computed a correlation heatmap of the predictors.

Most correlations are moderate and within acceptable ranges. Later, we validate this with a Variance Inflation Factor (VIF) analysis.

`Matplotlib.pyplot, seaborn were used`

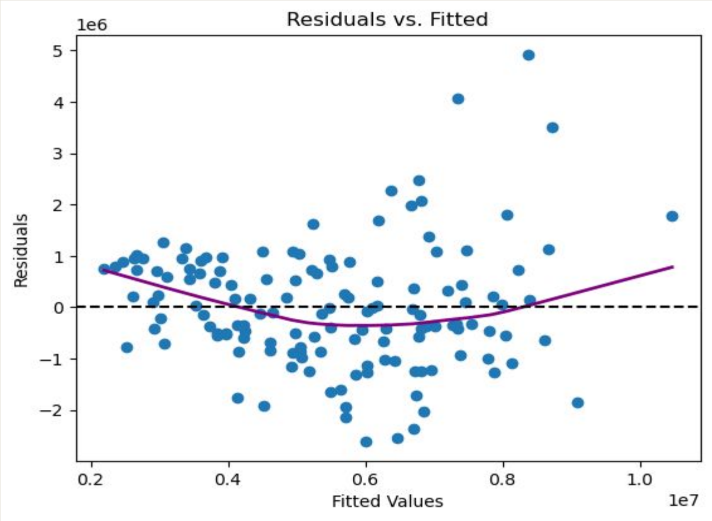


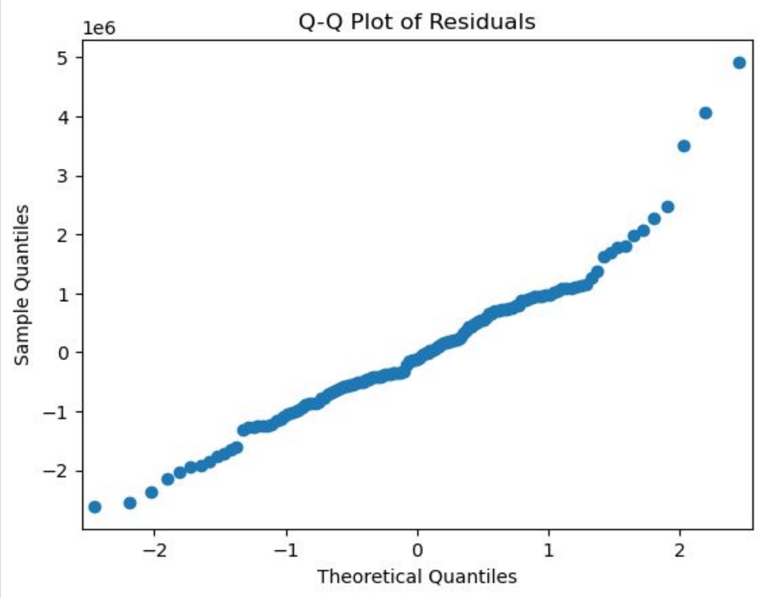

Red = Positive Correlation
Blue = Negative Correlation

# Initial OLS Regression

- Our OLS model is built using all predictors. The model yields an R-squared of 0.686, but assumptions tests show violations. The residuals vs fitted value plot shows curvature which indicates non-linearity and heteroscedasticity.
- The Q-Q plot and Shapiro-Wilk test reveals non-normal residuals.
- Since the data is cross-sectional, Independence is likely assumed.

`matplotlib.pyplot & statsmodels.api were used`

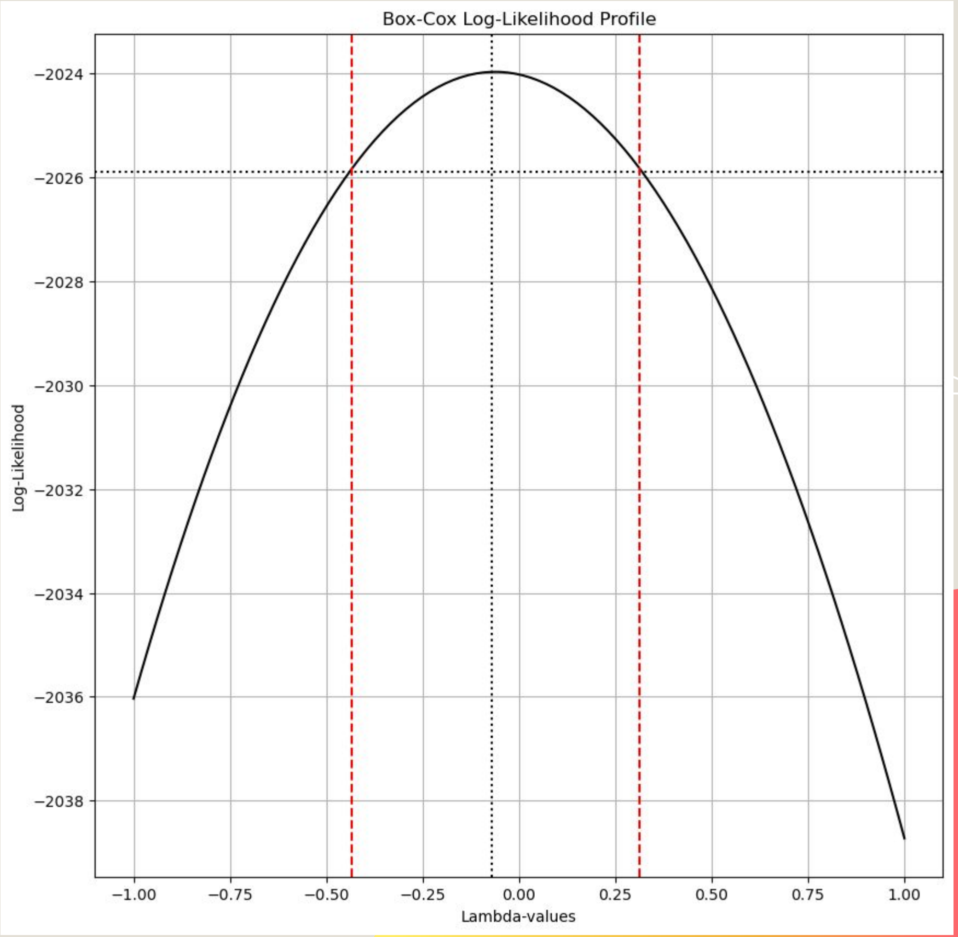


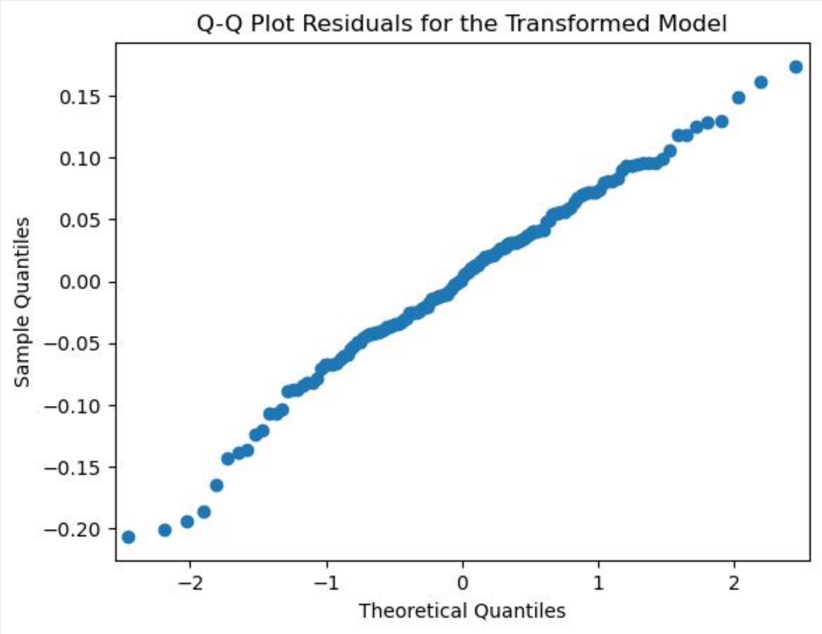
# Box-Cox Transformation

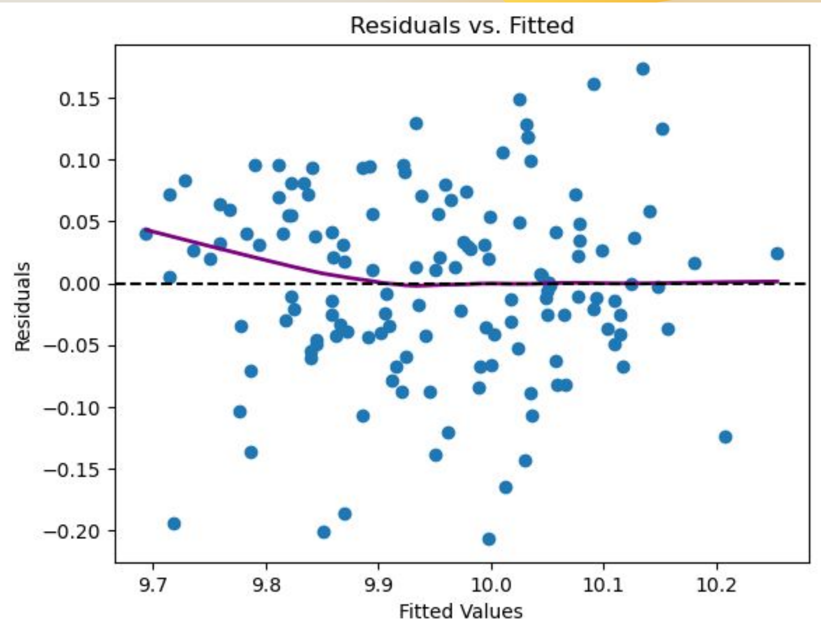
- To address the violations, we apply a Box-Cox transformation to the dependent variable (price).
- The optimal lambda is found to be approximately -0.06.
- This transformation improves model performance and helps meet regression assumptions without losing data.
  - Independence is likely to be assumed since the data is still cross-sectional.

```
numpy , matplotlib.pyplot, scipy were used
```



# Results



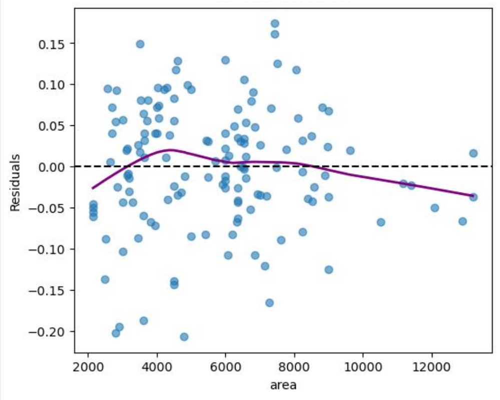




After transformation, the model's R-squared improves to 0.724. Residual diagnostics suggest that linearity, homoscedasticity, and normality have improved. The LOWESS line in the residual plot is now flatter, and the Shapiro-Wilk test no longer rejects normality ($p = 0.23$).

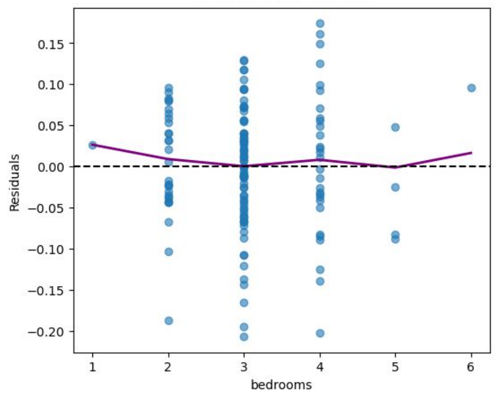
`statsmodels, matplotlib.pyplot were used`

# Results: Linearity & Homoscedasticity

Residuals vs Area



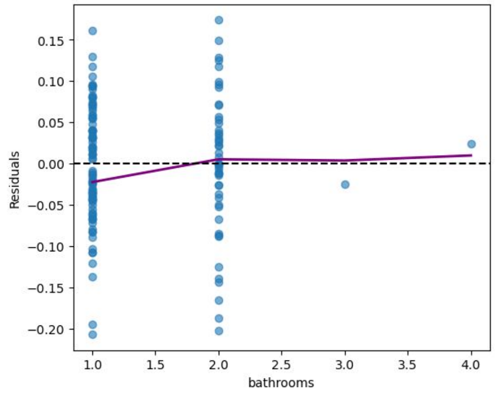

Residuals vs Bedrooms



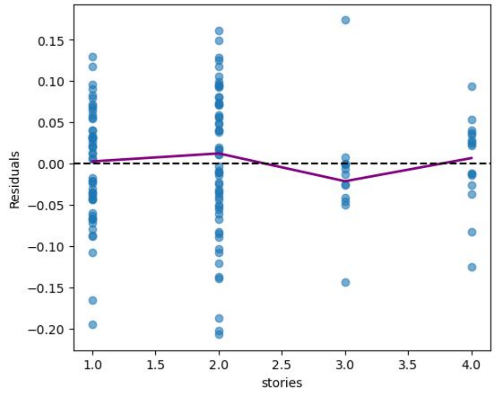

- This shows whether the residuals vary linearly and with constant spread against predictors like area, bedroom, bathrooms, stories, and parking.
- Visually demonstrates linearity and homoscedasticity are individually satisfied.

```
seaborn , numpy, matplotlib.pyplot were used
```

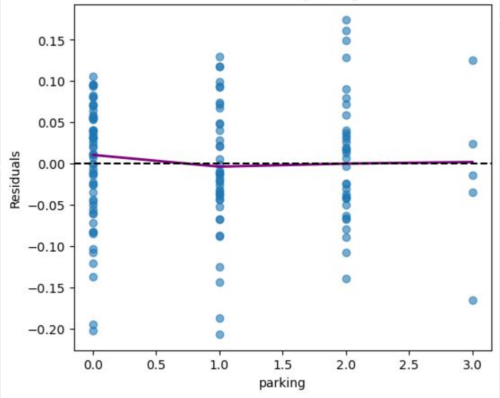
Residuals vs Bathrooms




Residuals vs Stories



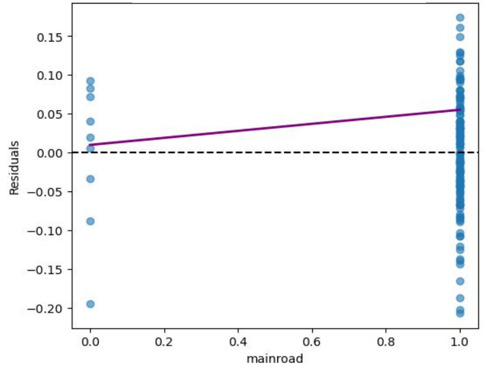

Residuals vs Parking



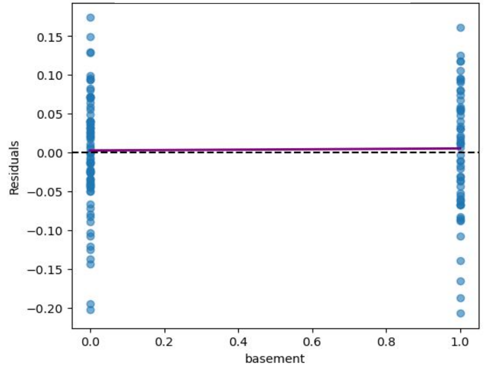
# Results: LOWESS

The comparison of residuals and categorical predictors such as mainroad, guess room, air conditioning, hot water heating, prefarea, and basement. This shows that linearity holds up.

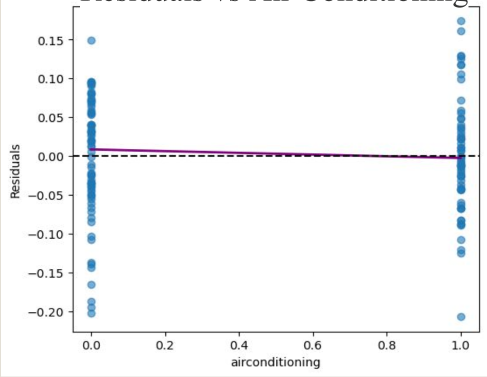
Residuals vs Mainroad



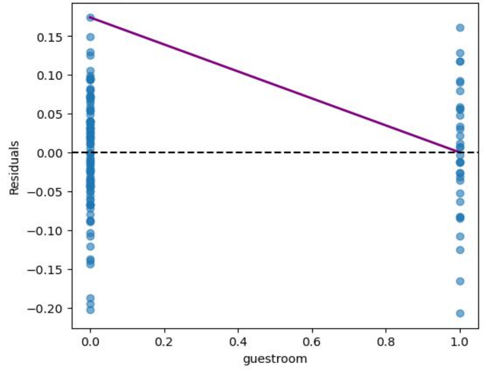

Residuals vs Basement



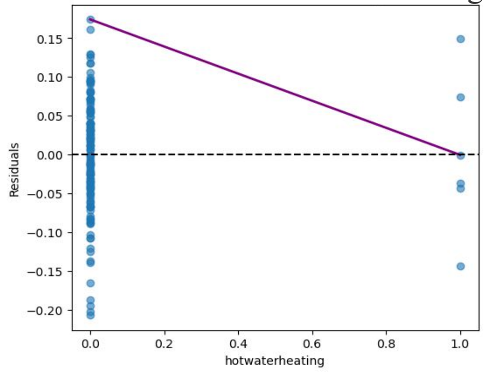

Residuals vs Air Conditioning



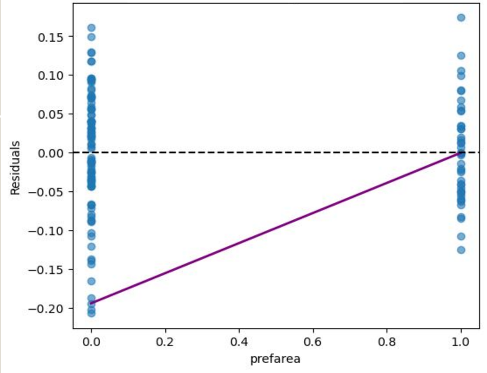

Residuals vs Guessroom




Residuals vs Hot Water Heating




Residuals vs Prefarea




Statsmodels.api,
Matplotlib.pyplot
Were used

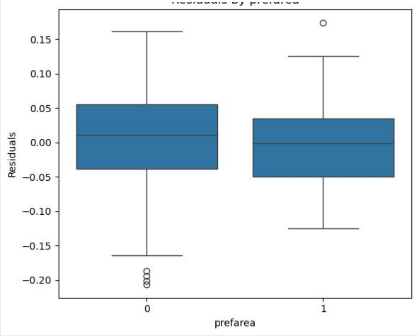
# Residual by Categorical Predictors

- Mainroad and air conditioning predictors: residuals shows similar spreads across groups which supports homoscedasticity.
- Guest room and basement shows mild vertical asymmetry, but not severe.
- Hot water heating shows a more spread and asymmetry which suggests that this predictor might be contributing to skew or instability.

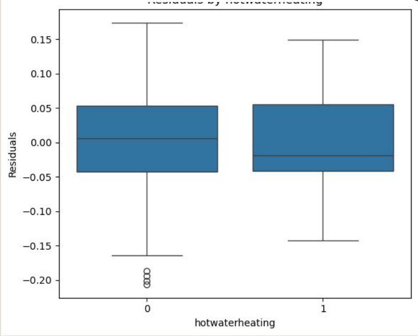
These results confirmed that our transformed model (Box-Cox) was well-behaved and did not systematically favor one group over another, strengthening the credibility of our conclusions.

`Matplotlib.pyplot, seaborn were used`

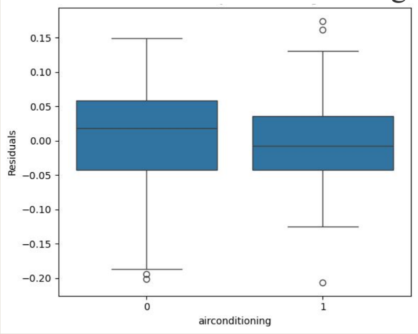
Residuals vs Prefarea



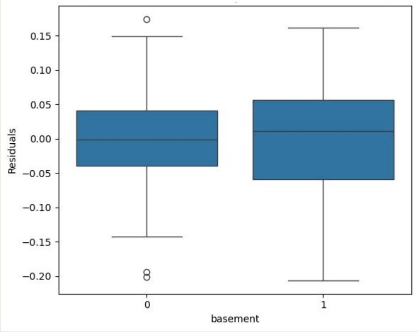

Residuals vs Hot Water Heating



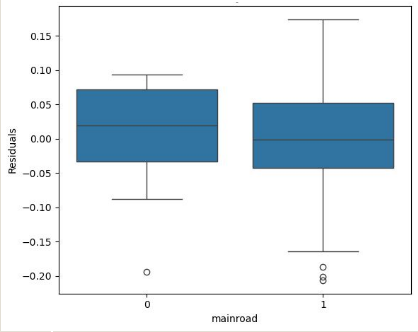

Residuals vs Air Conditioning



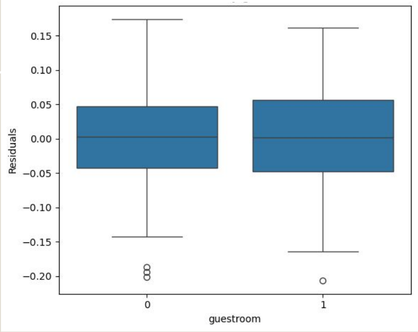

Residuals vs Basement




Residuals vs Mainroad




Residuals vs Guessroom

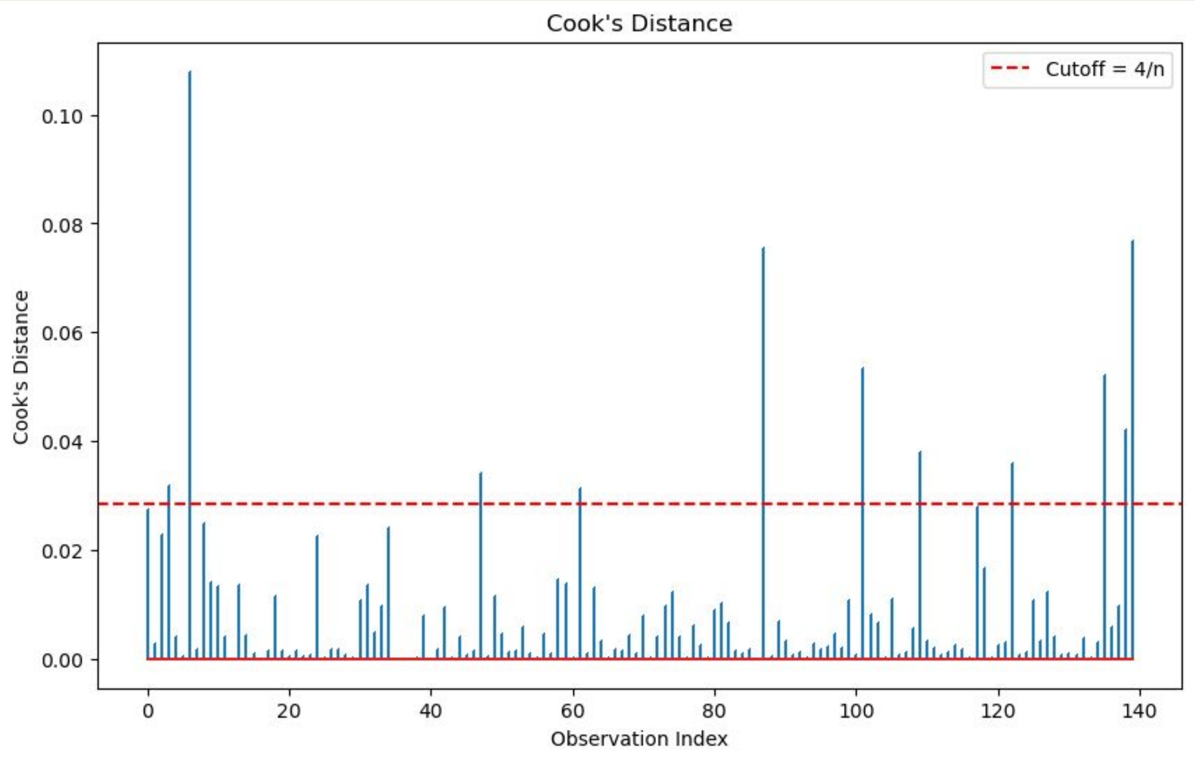


# Outliers & Influence

We used Cook's Distance to detect influential points:

- 11 points exceeded the threshold but are not numerous enough to distort the model
- These points are noted but retained in the model due to their minimal impact

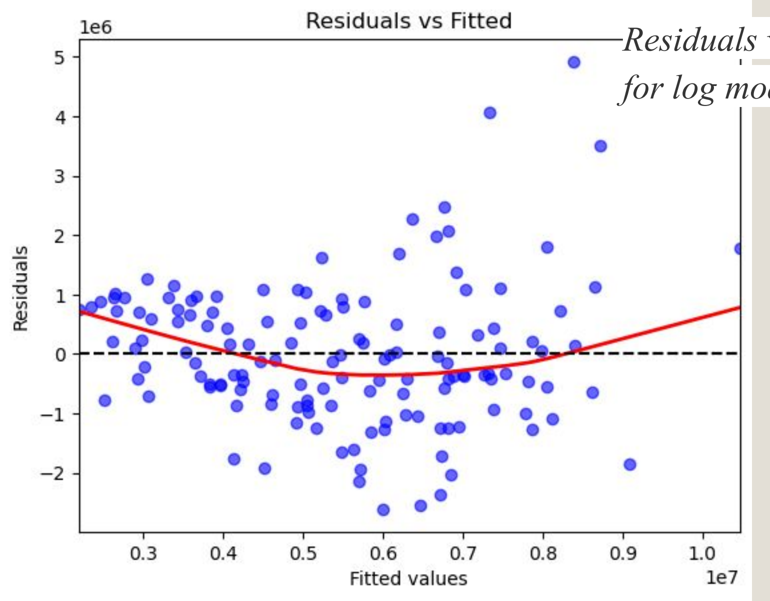
`Matplotlib.pyplot, numpy were used`



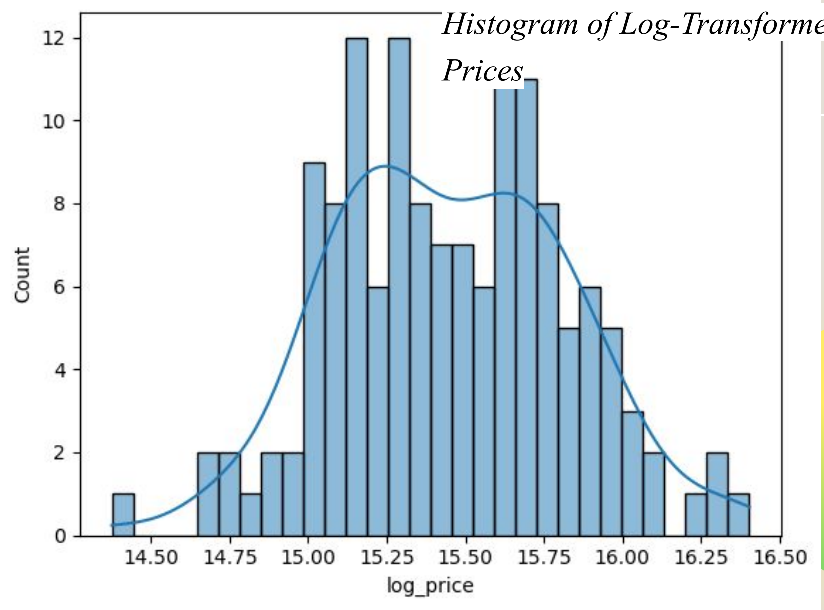
# Log Transformation

We also tested another transformation on the original data. Independence is still assumed since no data was lost (still cross-sectional). Given the noticeable skewness in the price data, we applied the log transformation to make the distribution more symmetric. This did improve normality but it did not satisfy linearity or homoscedasticity based on the residual plots and White's test.

`Seaborn were used`




*Residuals vs Fitted for log model*



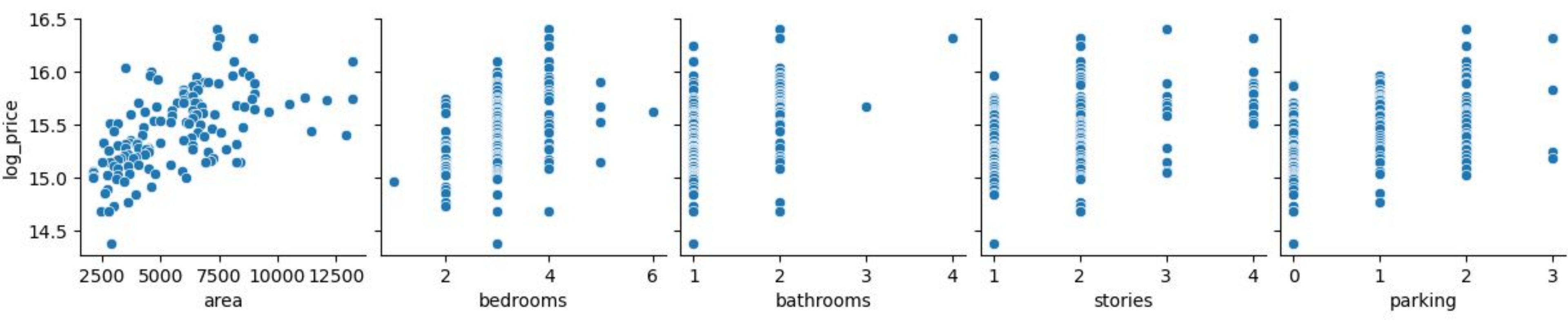

*Histogram of Log-Transformed Prices*

White's test: `'Test p-value': np.float64(0.2923232452609214)`

# Log Transformation Price vs. Quantitative Predictors

It shows a linear trend strength between log prices and the predictors. This helped us justify why log transformation improved normality (even though it did fail other assumptions).

- Log Price vs. Area: shows a clear positive linear trend which indicates that log transformation improved this relationship
- Log Price vs. Bedroom/Parking: tightened the vertical spread




`Seaborn, matplotlib.pyplot were used`

# Which is the Better Model?

| Box-Cox | Log |
| --- | --- |
| ✅ Linearity | ❌ Linearity |
| ✅ Normality | ✅ Normality |
| ✅ Homoscedasticity | ❌ Homoscedasticity |
| ✅ Independence | ✅ Independence |

Box-Cox transformation is the better model to use!

# Testing First Hypothesis

```
                            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:                        18:53:49   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
```

**Hypothesis 1:** Does the area of a house significantly affect its price?

$H_0$: Area has no effect on price.

$H_1$: Area significantly affects price.

**Results:** According to the summary of the OLS regression results, we reject the null hypothesis, $H_0$ ($p < 0.001$). Therefore, area is a significant predictor of housing prices.

`Numpy and statsmodels.api were used`

# Testing Second Hypothesis

statsmodels.api were used

**Hypothesis 2:** Does the effect of area depend on the number of bedrooms?

***$H_0$*:** No interaction between area and bedroom.

***$H_1$*:** There is an interaction .

**Results:** According to the OLS Regression Resul we fail to reject ***$H_0$*** (p = 0.828) which means the interaction between the number of bedrooms and the area is not significant.

```
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
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            trans_price   R-squared:                       0.724
Model:                            OLS   Adj. R-squared:                  0.698
Method:                 Least Squares   F-statistic:                     27.71
Date:                Thu, 08 May 2025   Prob (F-statistic):           7.58e-30
Time:                        18:53:49   Log-Likelihood:                 164.18
No. Observations:                 140   AIC:                            -302.4
Df Residuals:                     127   BIC:                            -264.1
Df Model:                          12
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const               9.5481      0.075    127.079      0.000       9.399       9.697
area             1.659e-05   1.21e-05      1.371      0.173   -7.36e-06    4.05e-05
bedrooms            0.0161      0.023      0.698      0.486      -0.029       0.062
bathrooms           0.0379      0.015      2.477      0.015       0.008       0.068
stories             0.0285      0.009      3.018      0.003       0.010       0.047
parking             0.0214      0.009      2.445      0.016       0.004       0.039
mainroad            0.0476      0.029      1.631      0.105      -0.010       0.105
guestroom           0.0369      0.017      2.121      0.036       0.002       0.071
basement            0.0233      0.017      1.369      0.173      -0.010       0.057
hotwaterheating     0.1182      0.035      3.392      0.001       0.049       0.187
airconditioning     0.0859      0.016      5.495      0.000       0.055       0.117
```

# Model Evaluation (OLS vs ElasticNet)

| | Model | RMSE |
|---|---|---|
| 0 | OLS (Interaction) | 0.069936 |
| 1 | ElasticNet | 0.076564 |

From the RMSE comparison table, we compared OLS and the ElasticNet using RMSE to test model generalization. Our findings were that Elastic Net did not improve predictions and OLS had a lower RMSE of ~0.07 vs. ~0.077 compared to Elastic Net. This indicates that multicollinearity was not a major issue in our dataset. It also validates our choice of choosing a simpler OLS model.

`Pandas were used`

# Conclusion

- This project built a statistically valid and interpretable model to predict housing prices using structural features of furnished homes. It demonstrated the importance of balancing interpretability with predictive performance.
- Hypotheses:
  - Area of a house is a statistically significant positive predictor of its market price.
  - Interaction effects between area and number of bedrooms are not statistically significant
- While more advanced methods like ElasticNet offer useful alternatives, our results show that a well specified linear model with appropriate transformation can offer both accuracy and clarity.
  - Box-Cox OLS had a lower RMSE of ~0.07 vs. ~0.077 compared to Elastic Net
- Future Studies: Expand by incorporating temporal or geographic variables, or applying nonlinear models to further enhance the predictive power of real estate valuation.

# THANK YOU!
# ANY QUESTIONS?

# Individual Contributions:

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