Fall 25: CSci 4521—Applied Machine Learning

Homework 3
(Prepared by Brock Bye)

1. Data Visualization and Analysis

The number of apartments with different amounts of bedrooms and bathrooms are not even close to being normally distributed.

Most common amount of each:

- Bedrooms: 1 br
- Bathrooms: 1 bath

These distributions are skewed toward single bed/bath units because apartments typically don't house families. The small living space is more suited to tenants in lower numbers.

The histoplot distributions appear to be non-normal and right-skewed, indicating outliers on the high end of each histogram. The majority of apartments are clustered at the lower end, suggesting a prevalence of smaller, more affordable units. This aligns with earlier histograms showing a dominance of single-person apartments.

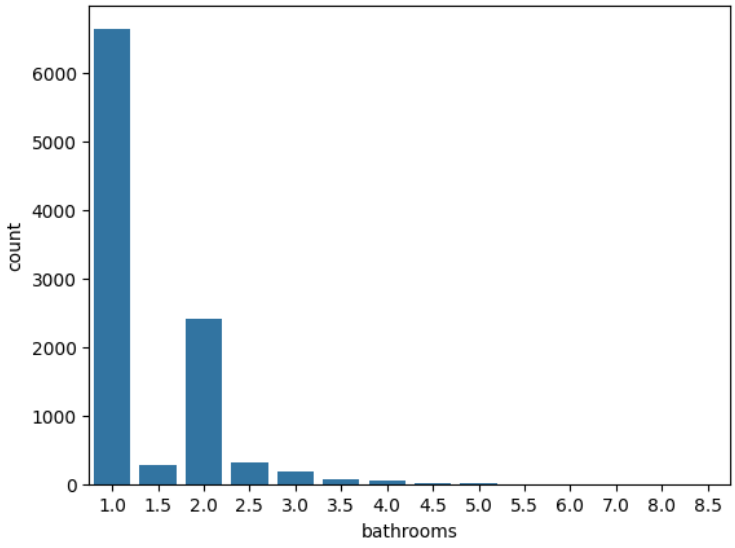

Figure 1: Bathrooms Histogram

Feature Selection and Correlation Analysis

I chose to include the following features: `bathrooms`, `bedrooms`, `price`, `square_feet`, `latitude`, and `longitude`. Each of these metrics was selected for inclusion in the correlation heat map because

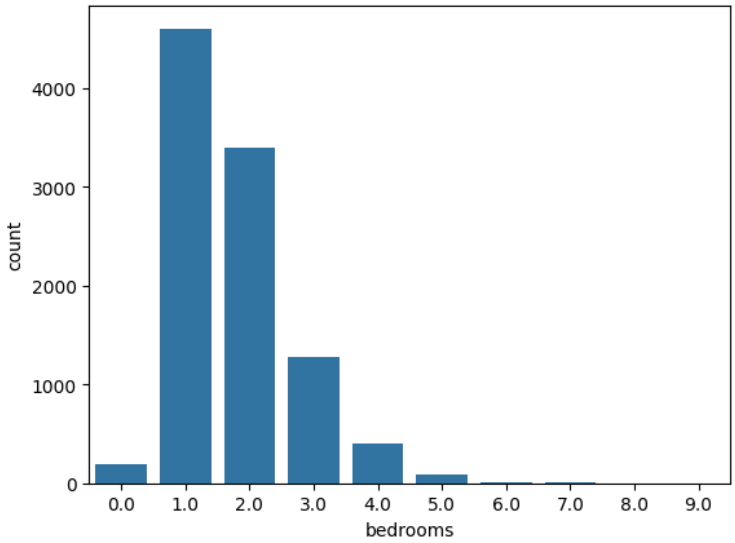

Figure 2: Bedrooms Histogram

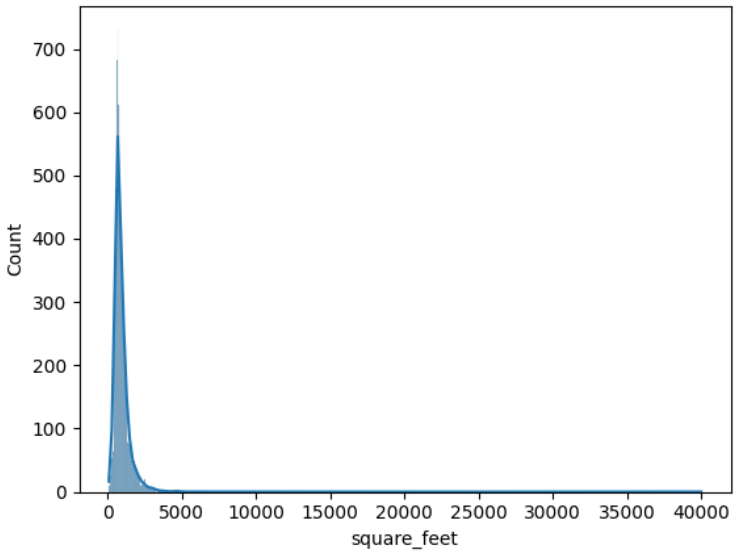

Figure 3: Square Feet Histoplot

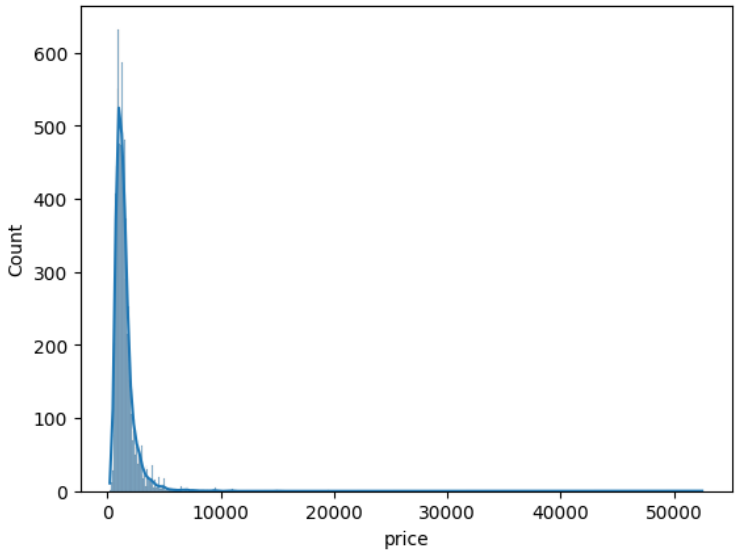

Figure 4: Price Histoplot

they are quantitative in nature.

`Bedrooms`, `bathrooms`, and `square_feet` are strongly correlated in a positive direction. These correlations exist because both `bathrooms` and `bedrooms` are directly related to `square_feet`. The more bathrooms and bedrooms an apartment has, the greater its square footage tends to be.

This observation suggests that for future analysis, only one of these features should be used. Including all of them may introduce multicollinearity, which could mislead the analysis and distort model performance.

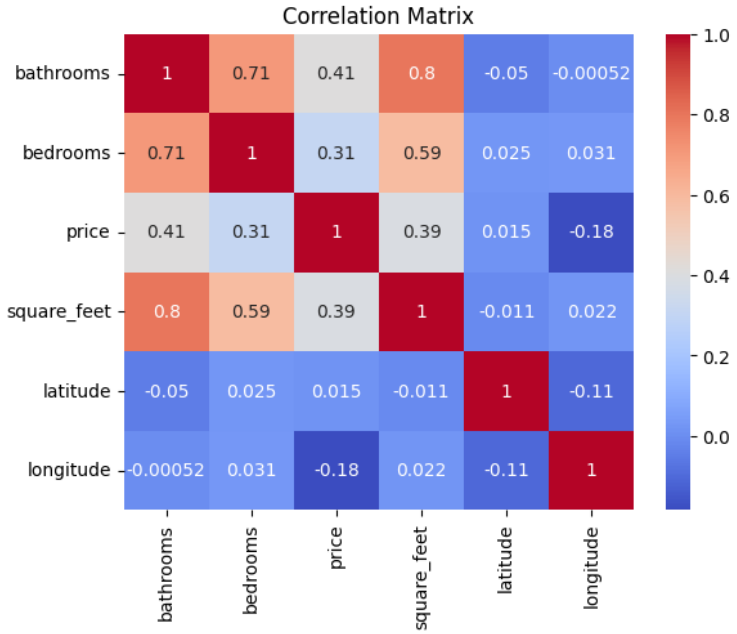

Figure 5: Correlation Heatmap

2. Parametric Modeling

Linear, single feature

I chose `square_feet` as my input because it is quantitative, and I felt that it was a solid linear regression to explore as a simple baseline.

For an un-normalized equation:

$$\text{price} = 1.43x + 1.21$$

with a learning rate of 8×10^{-7} that converges at loss = 1,082,125. The loss looks very large because the scales are orders of magnitude different between price and square feet.

Evaluation (Un-normalized)

- MSE: 2,016,290.46
- MAE: 563.45
- R^2: -1.89

Mean squared error shows predictions are off by \$1,100 to \$1,200 squared—large but not terrible for un-normalized data. Mean absolute error shows predictions are off by \$565 on average. The

negative and large R^2 indicates the model performs worse than simply predicting the mean price for every apartment.

I normalized both price and square feet by computing their z-scores.

Evaluation (Normalized)

- MSE: 0.983
- MAE: 0.482
- R^2: -0.631

With normalized values, evaluation scores are much better:

- MSE: squared prediction error is less than 1 standard unit.
- MAE: absolute prediction error is less than 0.5 standard unit.
- R^2: model explains -63% of the variance in price—still poor.

The boxplot is initially skewed in Q5, shrinking other quartiles due to outliers. After filtering Q5, distributions for Q2–Q4 show MAE around 0.3 with symmetrical upper/lower standard deviation of about 0.17. The model performs consistently across the lower 80% of prices.

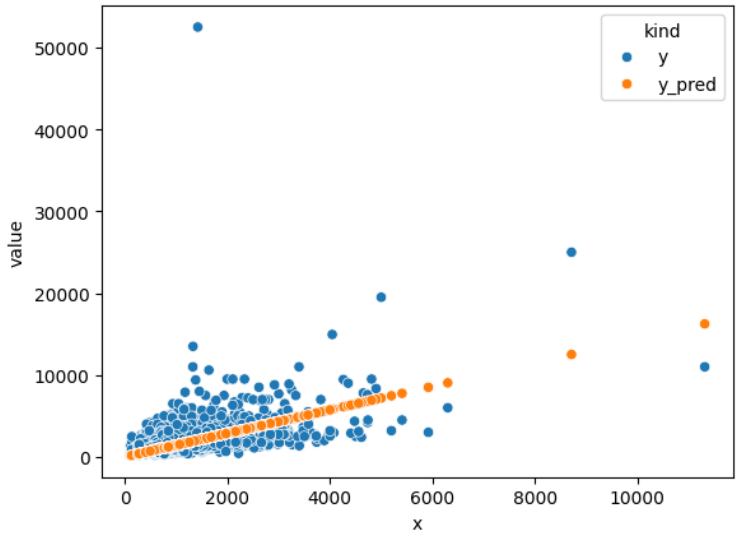

Figure 6: Single Feature Scatterplot

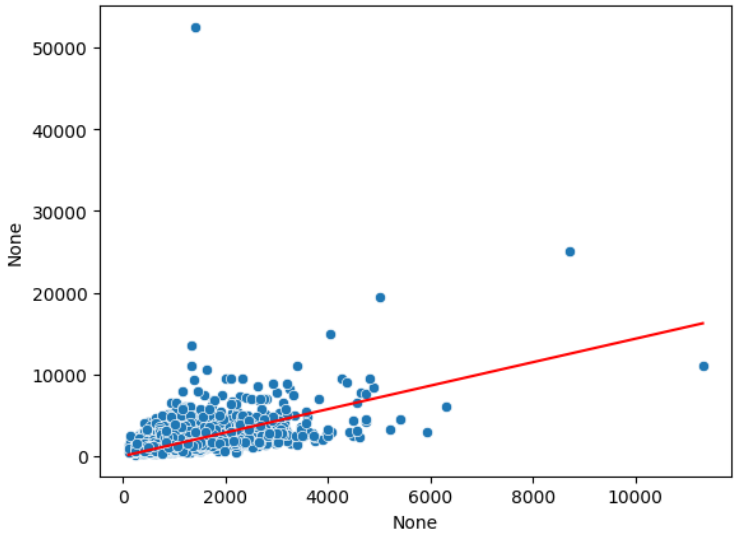

Figure 7: Single Feature with Line of Best Fit

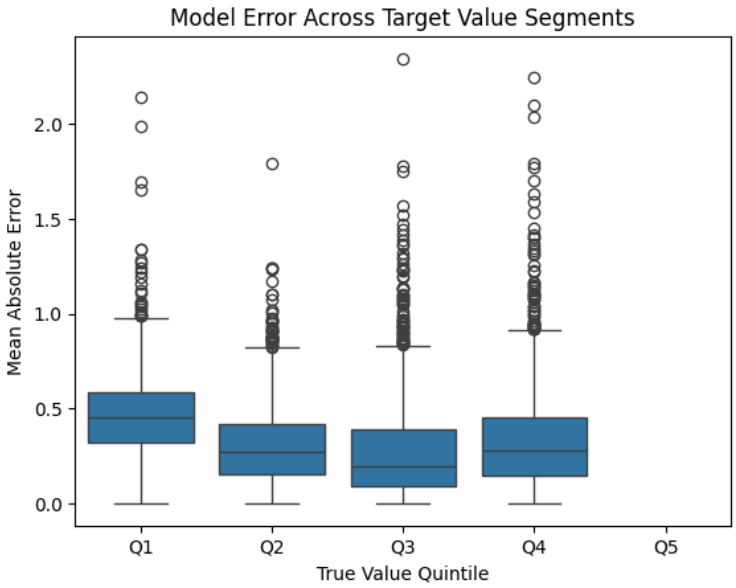

Figure 8: Single Feature Boxplot

Linear, Three Features

I chose `square_feet`, number of amenities, and number of pets to predict price. I selected three uncorrelated features and avoided using more than one of the correlated features: bathrooms, bedrooms, and square feet.

Normalized regression equation:

$$\text{price} = 0.603x_1 - 0.048x_2 - 0.021x_3 + 0.0189, \quad \text{loss} = 0.9057$$

Un-normalized regression equation:

$$\text{price} = 1.346x_1 + 11.536x_2 + 26.748x_3 + 23.9051, \quad \text{loss} = 1,062,380.90$$

Evaluation (Normalized)

- MSE: 0.974
- MAE: 0.479
- R^2: -0.616

The boxplot is initially skewed in Q5 due to outliers. After filtering Q5, distributions for Q2–Q4 show MAE around 0.3 with symmetrical upper/lower standard deviation of about 0.17. The model performs consistently across the lower 80% of prices.

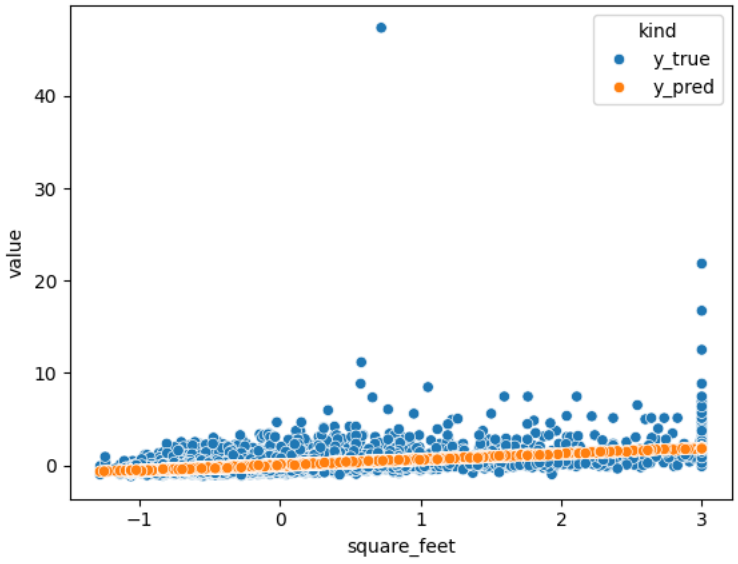

Figure 9: Multiple Feature: Square Feet Scatterplot

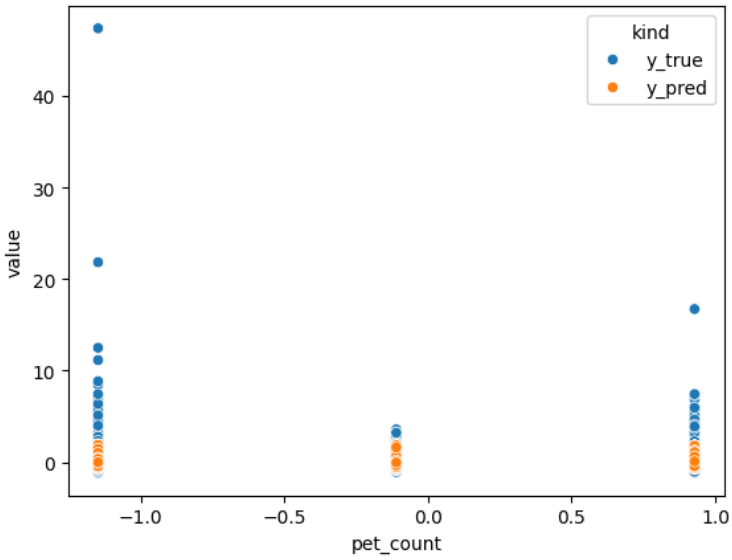

Figure 10: Multiple Feature: Pet Scatterplot

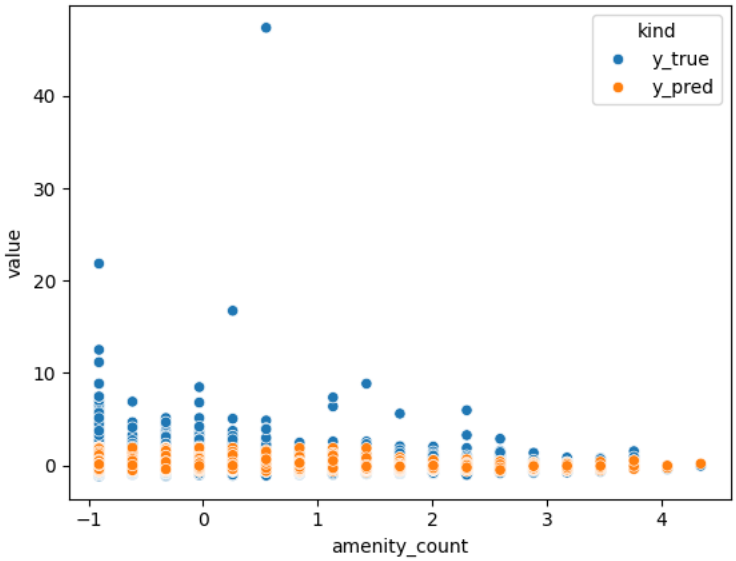

Figure 11: Multiple Feature: Amenity Scatterplot

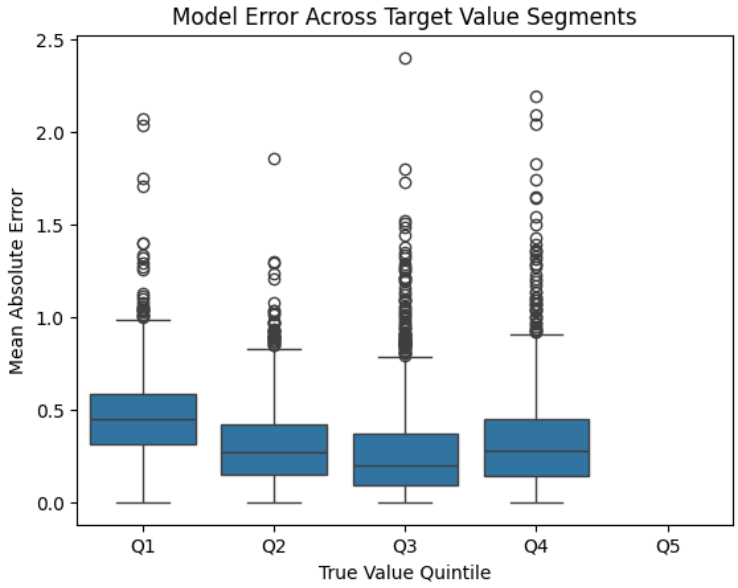

Figure 12: Multiple Feature Boxplot

Non-linear, Five Features

I chose `square_feet` and `bedrooms`, along with interaction terms: `square_feet`^{2}, `bedrooms`^{2}, and `square_feet` $\times$ `bedrooms`. These features had a correlation of 0.59, suggesting moderate correlation.

Un-normalized regression equation:

$$\text{price} = 0.616x_1 - 0.149x_2 - 0.258x_3 + 0.191x_4 + 0.160x_5 - 0.1029, \quad \text{loss} = 0.8728$$

Evaluation (Normalized)

- MSE: 0.514
- MAE: 0.475
- R^2: 0.219

These outputs are better than the single and multiple linear regression models in terms of variance explained. The MSE is also much lower, which is a good sign.

The boxplot is initially skewed in Q5 due to outliers. After filtering Q5, distributions for Q2–Q4 show MAE around 0.3 with symmetrical upper/lower standard deviation of about 0.17. The model performs consistently across the lower 80% of prices.

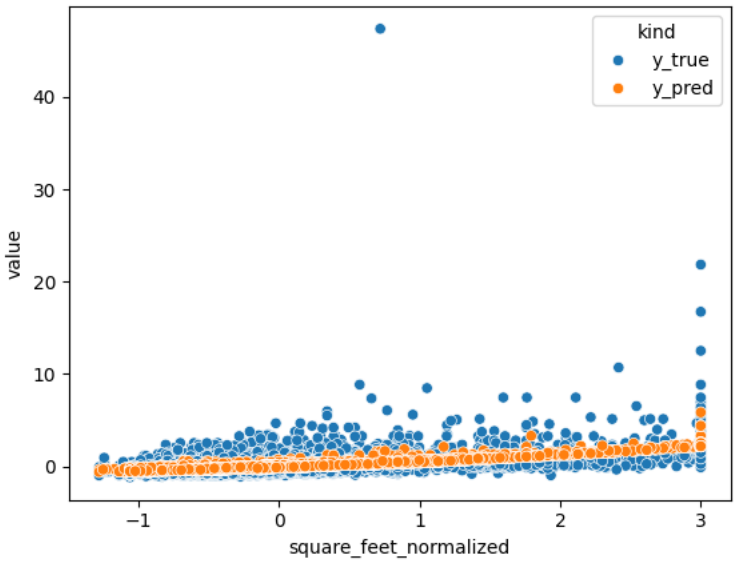

Figure 13: Nonlinear Feature: Square Feet

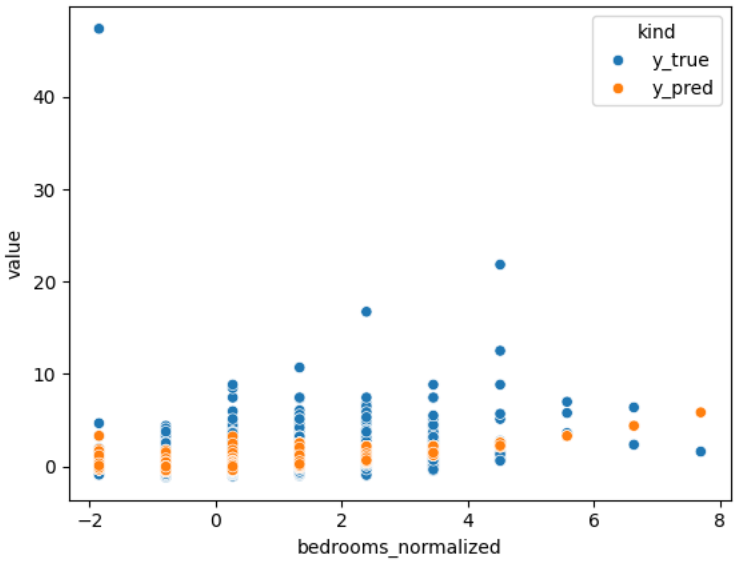

Figure 14: Nonlinear Feature: Bedrooms

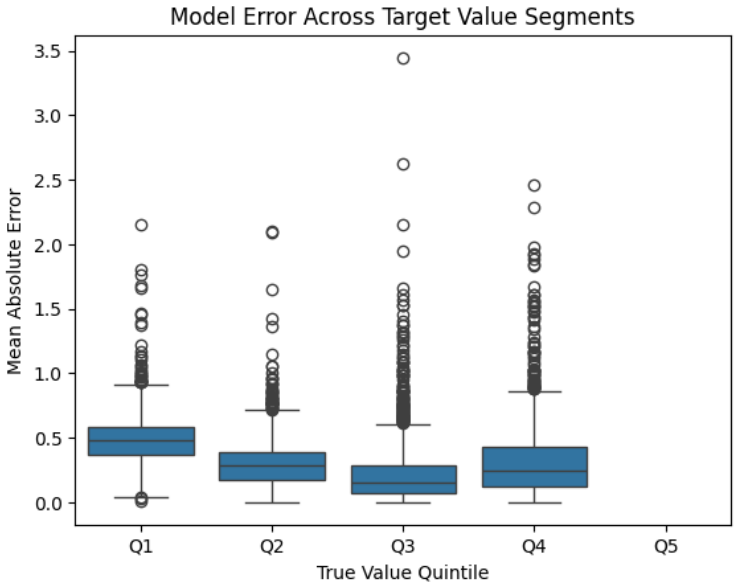

Figure 15: Nonlinear Feature Boxplot

3. Testing Accuracy

Model	**T = 0.25**	**T = 0.5**	**T = 1.0**
Single	0.29	0.57	0.87
Multiple	0.0005	0.001	0.008
Nonlinear	0.28	0.56	0.87

Table 1: Thresholded accuracy for each model at varying tolerance levels (T).

The nonlinear and single models perform similarly, whereas the multiple model lags behind significantly. This underperformance may be due to poor feature choices such as pet counts and amenities counts. In contrast, bedrooms and square feet appear to be strong, predictive features.

4. Cross Validation

The results revealed a notable shift in performance. Both the single and nonlinear models produced similar accuracies across all thresholds ($T = 0.25$, 0.5, 1.0), with slightly higher values than in the original single-split evaluation. However, the multiple features model, which previously showed extremely low accuracy, now performed comparably to the other two models. This suggests that the original test set may have contained outliers that affected the multiple model's predictions. AI tools were used to guide the implementation of cross-validation. I think that there was a disconnection on my behalf in regards to cross-validation and its implementation, because I don't think that the multiple features output could've shifted that much.

Model	**T = 0.25**	**T = 0.5**	**T = 1.0**
Single	0.34	0.638	0.89
Multiple	0.33	0.64	0.89
Nonlinear	0.34	0.64	0.90

Table 2: Summary of cross-validated thresholded accuracy across three models and three tolerance levels.