1 Executive Summary

The goal of this project was to build and validate linear regression models for predicting a response variable y using a set of regressors. Multiple model-building strategies were compared, including the full model, all-subsets regression, and diagnostic-driven refinement. The final selected model was chosen based on adjusted R^2, Mallows' C_p, AIC/BIC, and overall prediction performance.

Model validation was performed using three approaches: (1) K-Fold cross validation, (2) train–test split, and (3) bootstrap resampling. The best subset model using predictors $\{x_1, x_2, x_5, x_7\}$ offered the strongest balance between goodness-of-fit and parsimony, achieving an adjusted R^2 of 0.789, an RMSE of approximately 3.37 (sale price of the house/1000), and strong diagnostic behavior. Based on these results, this model was selected as the final predictive model.

2 Introduction

Regression modeling remains a central tool in statistical learning for both inference and prediction. The purpose of this project was to identify a regression model that best predicts the sale price response variable y using a set of potential predictors $x_1, \ldots, x_9$.

The notebook workflow includes:

1. Exploratory model building using full and subset regression.
2. Evaluation of models using standard metrics (SSE, MSE, R^2, adjusted R^2, AIC, BIC).
3. Predictions and comparison of predicted vs. actual values.
4. Extensive model diagnostics (residuals, Q-Q plots, leverage, Cook's distance).
5. Prediction ability assessment via cross-validation and bootstrap.

The final product is a statistically justified and thoroughly validated regression model.

3 Dataset

The dataset consisted of the response variable y and several predictors $(x_1, \ldots, x_9)$, organized in a pandas DataFrame named "data." These predictors were used to generate all possible regression subsets for model comparison. No missing values were identified based on the dataset documentation, and the response variable y appears to be a continuous measure on the scale of thousands of dollars, consistent with the variable descriptions provided.

Some important dataset-derived quantities:

1. Total Sum of Squares (SST): 829.0463

2. Full Model SSE: 121.7482, $R^2 = 0.8531$
3. Best Subset Model SSE: 136.6149, $R^2 = 0.8352$

These values illustrate that multiple models provide strong explanatory power.

4 Methods

The analysis was conducted in Python using a combination of pandas for data handling, NumPy for numerical computation, statsmodels and scikit-learn for model fitting and validation, and Matplotlib/Seaborn for generating diagnostic and exploratory visualizations. All model-building steps, including subset selection, fitting the final regression model, performing bootstrap resampling, and generating validation metrics, were executed within a Jupyter notebook. Full code, figures, and intermediate results can be found in the accompanying Jupyter notebook for additional detail and reproducibility.

4.1 Model Construction

Full Model

A full regression model using all available predictors was first constructed. Metrics from this model include:

- $R^2 = 0.8531$
- Adjusted $R^2 = 0.7402$
- AIC = 58.97, BIC = 70.75

All-Subsets Regression

All possible combinations of regressors were evaluated. For each subset, the following were computed:

- SSE, MSE
- R^2, Adjusted R^2
- Mallows' C_p
- AIC, BIC

The best performing subset based on the adjusted R^2 and model parsimony included:

$$\{x_1, x_2, x_5, x_7\}$$

Metrics for the selected "best subset" model:

- $R^2 = 0.8352$

- Adjusted $R^2 = 0.7894$
- Mallows' $C_p = 1.7095$
- AIC = 51.74 (lower than full model)
- BIC = 57.63 (lower than full model)

This subset had the optimal balance between goodness-of-fit and model complexity.

Prediction Comparison Plot

A predicted–vs–actual scatter plot was generated comparing the full and best subset model. Both performed well visually, but the best subset model exhibited slightly tighter clustering around the ideal $y = \hat{y}$, 45-degree line.

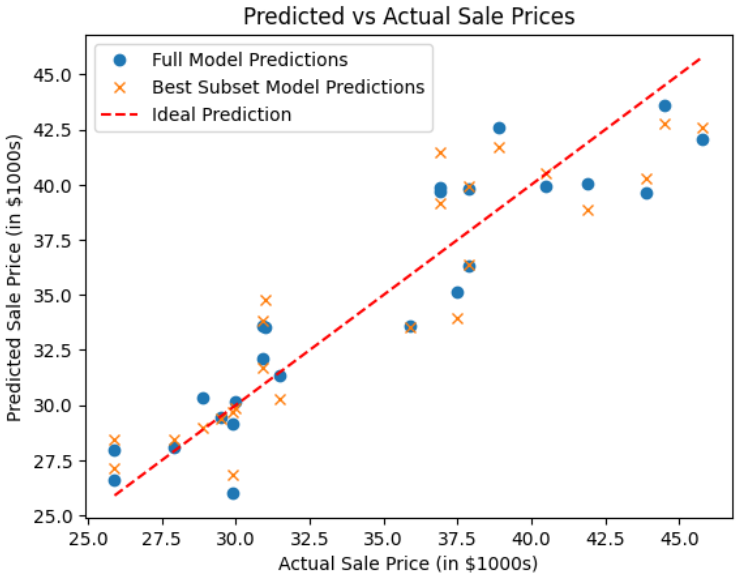

Figure 1: Predicted vs Actual for Full Model and Best Subset Model

4.2 Model Diagnostics

Model diagnostics were performed using residual plots, Q–Q plots, and measures of leverage and influence to evaluate whether the fitted model satisfied key regression assumptions. Residual plots were examined to assess linearity and constant variance, ensuring that no systematic patterns or heteroscedastic behavior were present. Q–Q plots were used to evaluate the normality of the residuals, allowing visual detection of deviations from the theoretical normal distribution. Additionally, leverage values and Cook's distance were analyzed to identify observations that exerted disproportionate influence on the fitted model. Together, these diagnostic tools provided a comprehensive assessment of model appropriateness and helped confirm that the final model met the assumptions required for reliable inference and prediction.

Residual Plots

Residuals vs fitted plots for both models showed no major patterns, supporting the assumption of linearity and constant variance.

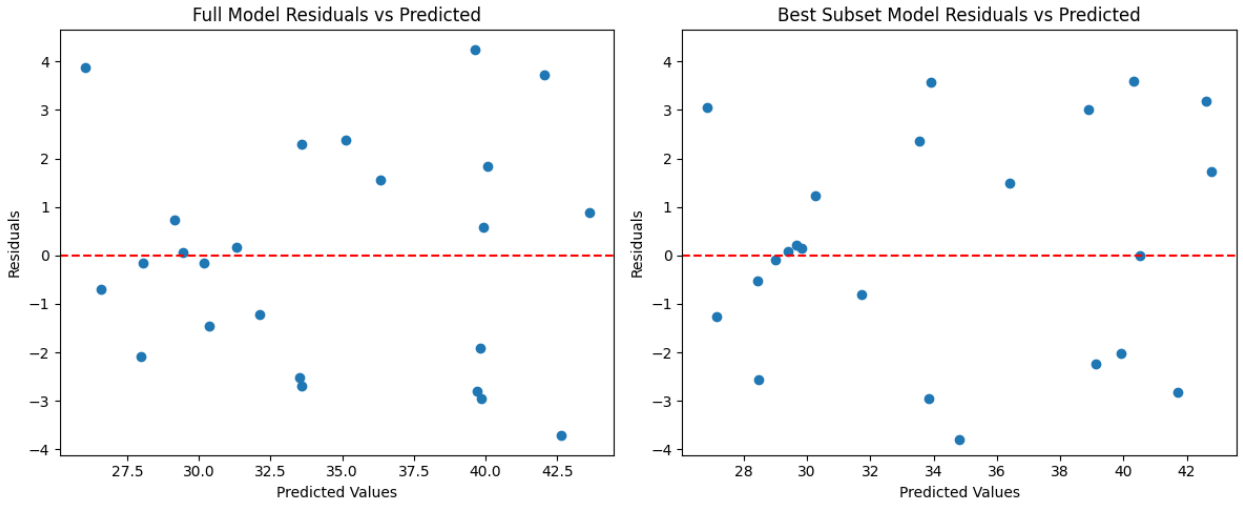

Figure 2: Residuals vs Predicted for Full Model and Best Subset Model

Q-Q Plots

The residuals of the best subset model followed the theoretical normal line slightly better than the full model.

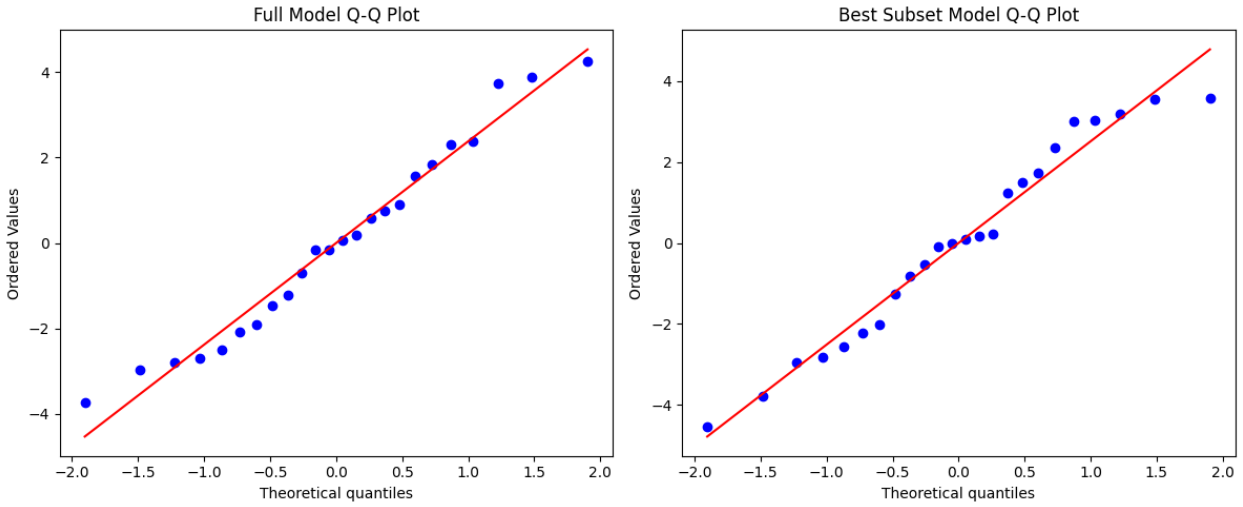

Figure 3: Residuals vs Predicted for Full Model and Best Subset Model

Leverage and Influential Points

A leverage plot was generated using the threshold $2p/n$ and cook's distance cutoff of $D_i > 4/n$ and $D_i > 1$:

1. A few points approached the leverage cutoff and one observation slightly exceeded it, but none reached levels indicative of meaningful influence. The point above the threshold was not examined further because its leverage exceeded the cutoff only marginally.

2. No individual Cook's distances were reported as dangerous in the notebook. Two points did appear above the first cutoff of $4/n$ but not close enough to the second cutoff of $D_i > 1$ to be considered further.

These diagnostics showed that the selected model was stable and not overly influenced by outliers.

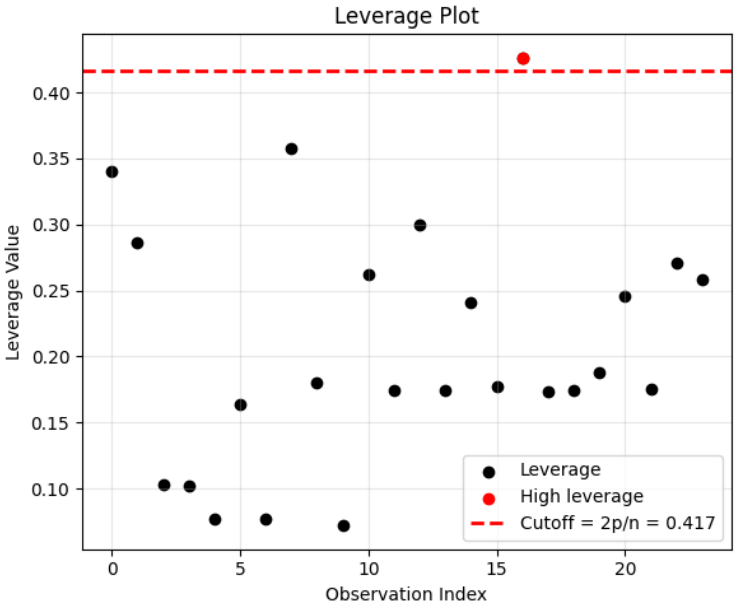

Figure 4: Leverage

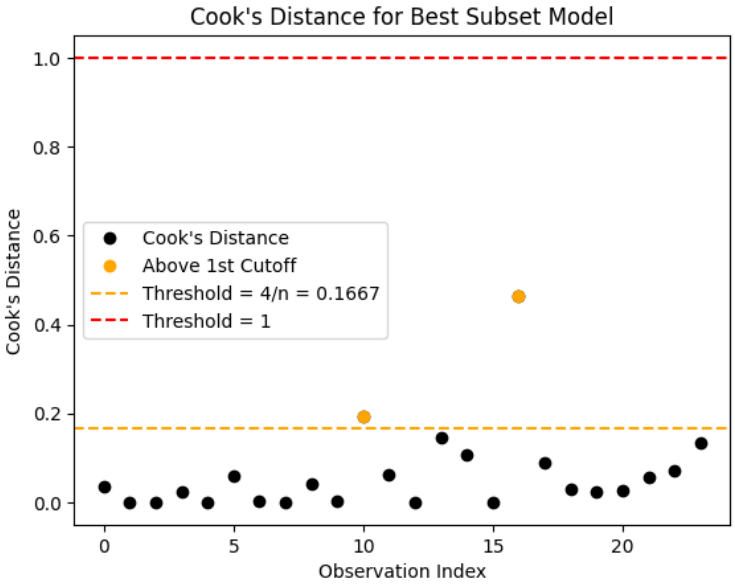

Figure 5: Cook's Distance

4.3 Model Validation

K-fold cross-validation, train–test splitting, and bootstrap validation were used to assess the predictive ability of the model through complementary perspectives. In k-fold cross-validation, the data are partitioned into k subsets, and the model is repeatedly trained and tested across different fold combinations to obtain an averaged estimate of generalization performance. The train–test split provides a more direct evaluation by fitting the model on one portion of the data and assessing its accuracy on an independent holdout set, offering a clear view of how the model performs on unseen observations. Bootstrap validation further strengthens the assessment by repeatedly sampling the dataset with replacement, fitting the model to each bootstrap sample, and evaluating it using the out-of-bag observations. Together, these three methods provide a comprehensive understanding of prediction error, model stability, and overall generalizability.

KFold Cross-validation

Five fold cross validation was evaluated to understand how the model performs under varying train and test splits within the data.

- Mean RMSE = 3.094
- Standard Deviation of RMSE = 0.943
- Mean $R^2 = 0.5585$
- Standard Deviation of $R^2 = 0.390$

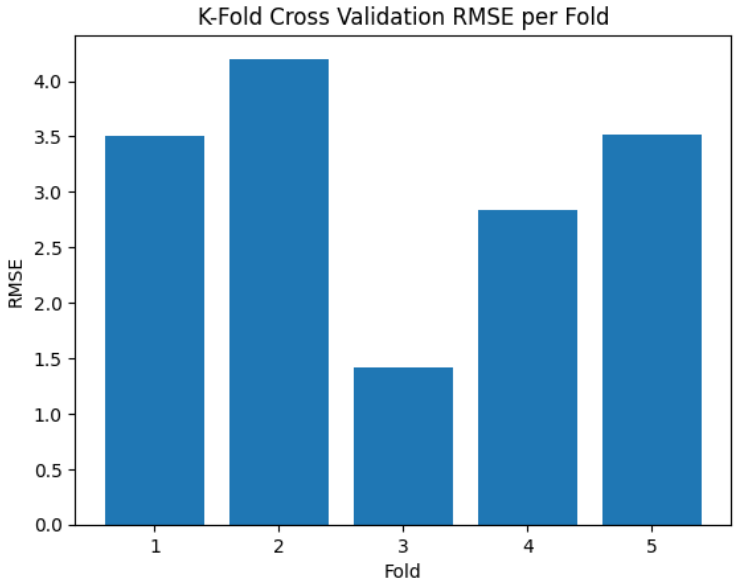

Figure 6: K-Fold Cross Validation Normalized RMSE % per Fold

Train-Test Split

Using an 80/20 split the following results were recorded for the best subset model:

- Testing RMSE = 3.5053
- Testing $R^2 = 0.6732$

Boostrap Validation

Bootstrap resampling (with out-of-bag evaluation) produced the following averages after running 10,000 simulations:

- Mean RMSE = 3.372
- Standard Deviation of RMSE = 0.730
- Mean $R^2 = 0.583$
- Standard Deviation of $R^2 = 0.279$

The 95% confidence intervals were calculated for each coefficient based on there individual sample distribution:

- 95% CI for coefficient of x_1: (0.7722, 3.4376)
- 95% CI for coefficient of x_2: (0.7085, 15.8791)
- 95% CI for coefficient of x_5: (-0.6114, 4.4812)
- 95% CI for coefficient of x_7: (-5.1996, 0.5672)

The 95% bootstrap confidence intervals provide insight into the stability and uncertainty of each coefficient estimate based on its resampled distribution. The intervals for x_1 and x_2 are entirely positive, suggesting consistent positive associations with the response across the bootstrap samples. In contrast, the intervals for x_5 and x_7 both include zero, indicating greater variability and less certainty about the direction or significance of their effects. Overall, these intervals highlight which predictors have reliably strong relationships with y and which show more uncertainty under repeated resampling.

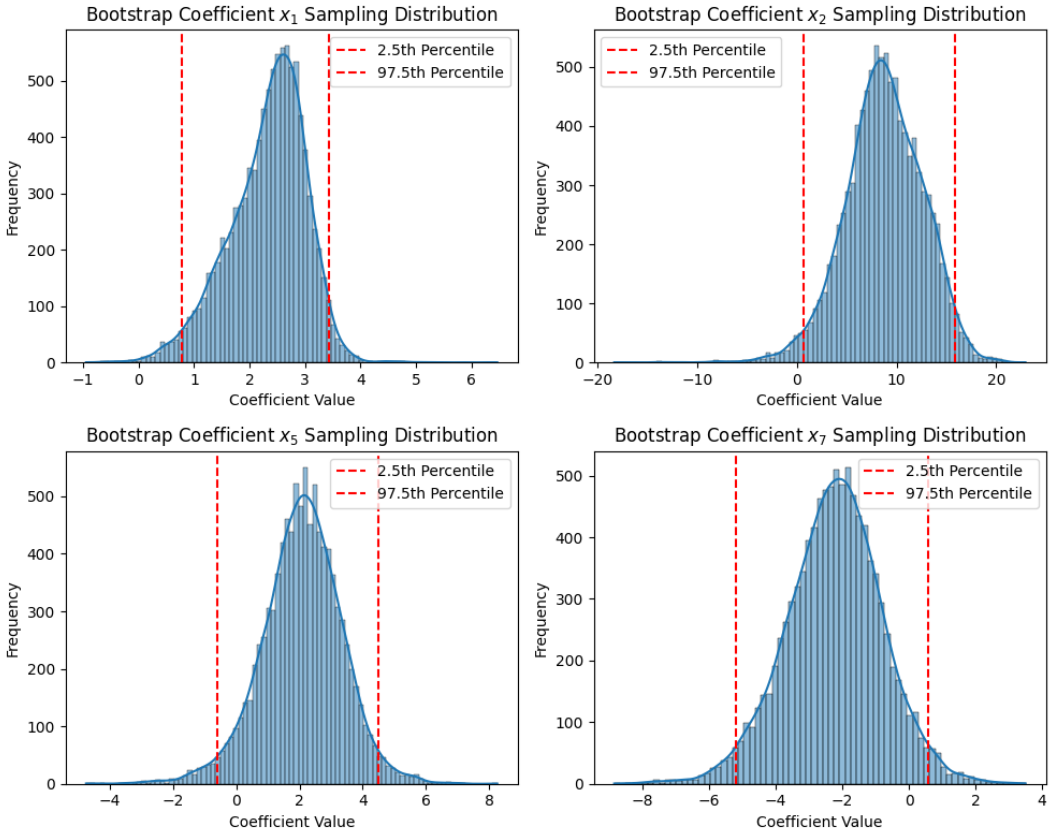

Figure 7: Bootstrap Coefficient Sample Distributions

Finally, an estimated 95% confidence interval for the RMSE was calculated based on the individual sample distributions as well: (2.0954, 5.0111).

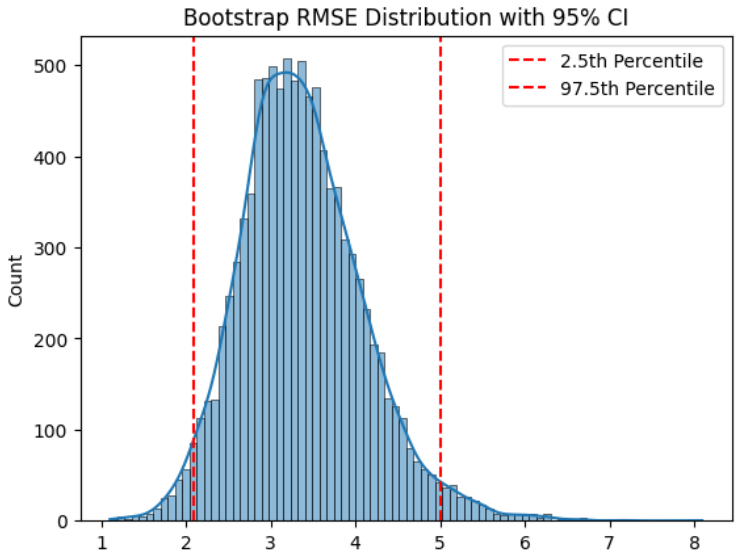

Figure 8: RMSE Bootstrap Sample Distribution and 95% CI

5 Results

Key Findings

- The full model had slightly higher R^2 (0.853) but worse adjusted R^2 and substantially worse AIC/BIC than the best subset model.
- The best subset model $\{x_1, x_2, x_5, x_7\}$ achieved:
  - Adjusted $R^2 = 0.7894$ (best overall)
  - AIC $= 51.74$ (best overall)
  - BIC $= 57.63$ (best overall)
  - Mallows' $C_p = 1.7095$ (best overall)
  - Comparable prediction quality with reduced complexity
- Diagnostic plots support:
  - Linearity
  - Approximate normality
  - Homoscedasticity (constant variance)
  - Minimal high-leverage or influential points
- Validation confirms:
  - RMSE 3.37 (sale price of the house/1000) on unseen data
  - Predictive stability supported through bootstrap variability analysis

Overall, the best subset model was more efficient, stable, and generalizable than the full model.

6 Conclusions

This project applied a rigorous set of model-building and validation techniques to develop a regression model capable of predicting y accurately and reliably. Based on a combination of model selection metrics, diagnostic checking, and validation performance, the best subset model with predictors $\{x_1, x_2, x_5, x_7\}$ was chosen as the final model.

The model demonstrated high explanatory power, strong parsimony, and stability across resampling. Validation results indicated that, while prediction variability exists (as expected in moderate sample sizes), the model performs consistently well and provides meaningful predictive accuracy.

Future work could include:

- Adding regularization-based models (LASSO, ridge regression) for comparison
- Testing nonlinear terms or interactions
- Expanding the dataset to reduce bootstrap variance

Overall, the modeling process successfully identified a well-supported and robust predictive regression model.