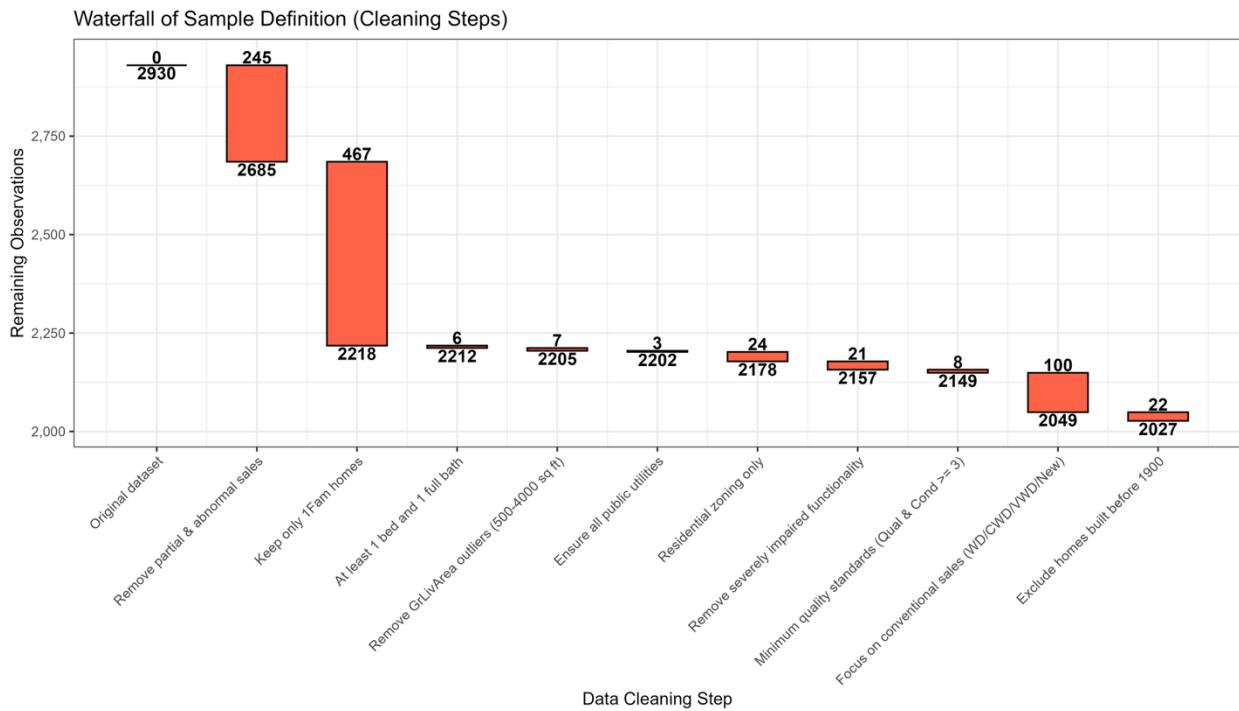


# Variable Selection in Ames Housing Data

## 1. Preparatory Work

Before the model was developed, several adjustments were made to the *Ames Housing* data to create a clean, representative sample of typical residential home sales in Ames. Starting with the original dataset, a series of filters were applied to focus the analysis on homes that represent normal transactions. Figure 1 below reports the waterfall drop-out conditions that were implemented to create the final sample population. The original dataset had the following conditions placed to determine final sample: 1) removing partial or abnormal sales; 2) keeping only single-family detached homes; 3) each home must have at one bedroom and one full bathroom; 4) removing homes with living areas outside the range of 500 to 4,000 square feet; 5) keeping homes with all public utilities; 6) keeping homes in residential zoning; 7) removing homes with severe functional impairment; 8) removing homes with very poor overall quality or overall condition ratings (i.e., < 3); 9) keeping homes with conventional sale types; 10) keeping homes built after 1900. These drop-out criteria help create a homogenous dataset, resulting in a final sample of 2,027 observations representing typical residential transactions in the Ames market.

*Figure 1. Waterfall Drop-Out Conditions to Determine Final Sample*



Categorical variables using first-principles reasoning and empirical validation determined top predictors of *SalePrice*. Real estate fundamentals suggest that location, quality, and condition drive property value. Simple linear regressions were fit (*SalePrice* ~ Variable) to calculate  $R^2$  values measuring each variable's independent predictive power. The top 5 variables based on their  $R^2$  value were then dummy coded for analysis. *OverallQual* had the highest individual  $R^2$

value with 0.632, closely followed by *Neighborhood* ( $R^2 = 0.609$ ), then *BsmtQual* ( $R^2 = 0.485$ ), then *ExterQual* ( $R^2 = 0.473$ ), and finally, *KitchenQual* ( $R^2 = 0.422$ ).

*OverallQual* exhibited the largest mean difference between its highest and lowest levels, with a spread of \$397,590. Specifically, homes at the highest level (*OverallQual* = 10) had an average of \$487,932, while homes at the lowest level (*OverallQual* = 3) in the filtered sample sell for an average of \$90,342. The large mean difference indicates strong discrimination power, that is the ability of a variable to effectively spread high-value properties from low-value properties. This suggests that the variable captures real, substantial differences in home values. Moreover, large mean differences improve prediction accuracy and model reliability, Figure A1 visualizes the mean differences among the top 10 categorical variables based on highest  $R^2$  value.

## 2. Predictive Modeling Framework

A modeling framework was developed with a 70/30 training/testing split in the final sample of observations with 2,027. The “train” set contains 70% of observations in the final sample that will be used to develop the predictive model, equating to 1,430 records; while the “test” set contains the remaining 30% of observations in the final sample that will be used to assess the accuracy of the predictions from the model, equating to 597 records. Table 1 below details the partition between the training and testing sets for the model.

Table 1. Training and Testing split partitions for Modeling

Dataset	N	Percentage (%)
<b>Total Sample</b>	2097	100
<b>Training Set</b>	1430	71.5
<b>Test Set</b>	597	49.5
<b>Verification</b>	$1430 + 597 = 2097$	$71.5 + 49.5 = 100$

## 3. Model Identification by Automatic Selection

### 3.1 Candidate Predictors

A pool of sixteen candidate predictor variable groups comprising of forty-eight individual predictors for modeling home sales. This pool strategically balances comprehensiveness of parsimony, while excluding the original categorical variables themselves to avoid multicollinearity.

Continuous variables capture properties of dimensions including *LotArea*, *GrLivArea* (above-grade living area), *TotalBsmtSF* (basement area), and *GarageArea*, measured in square feet. Temporal variables include *YearBuilt* and *YearRemodel*, capturing depreciation and renovation effects. Engineered features include *TotalSqftCalc* (total finished square footage: *BsmtFinSF1* + *BsmtFinSF2* + *GrLivArea*) and *QualityIndex* (*OverallQual* × *OverallCond* interaction), which aggregate multiple quality dimensions. Discrete variables count bedrooms, full bathrooms, and half bathrooms. Five categorical variables are represented through reference-coded dummy variables from section 1. Each categorical variable's first level alphabetically serves as the omitted reference category, with dummy variables capturing deviations from this baseline. This

reference coding approach avoids perfect multicollinearity while enabling flexible modeling of categorical relationships with sale price. Table 1A describes a list of the pool of candidate predictors.

### 3.2 Model Identification

To identify a parsimonious and predictive set of variables for modeling home prices, three automated variable selection procedures were applied: forward selection, backward elimination, and stepwise selection, using the training dataset. The three methods converged on a largely overlapping set of key predictors. These included measures of property size, overall and component quality indicators, property age, and amenities. In addition, multiple neighborhood indicators were consistently selected, reflecting substantial location-based price differentials. For example, properties located in neighborhoods such as *StoneBr*, *NoRidge*, *NridgHt*, *Crawfor*, and *Somerst* exhibited large positive sale prices relative to the baseline neighborhood, while others showed negative or smaller effects.

Forward selection produced the largest model, retaining a greater number of predictors, including some with weak statistical significance once other variables were controlled for. Backward elimination yielded a more parsimonious model by removing predictors that contributed little marginal explanatory power in the presence of stronger covariates. Stepwise selection produced an intermediate solution, retaining most of the influential predictors identified by forward selection while excluding several redundant variables.

Variance Inflation Factors (VIFs) were examined to assess multicollinearity in the final models obtained via forward, backward, and stepwise variable selection. While several indicator variables associated with quality categories exhibited moderate to high VIF values (approximately 10–16), this inflation is largely structural, reflecting the overlapping nature of categorical encodings and correlated quality constructs. More severe multicollinearity was observed among continuous size-related variables in the forward selection model, where multiple measures of living area (e.g., *GrLivArea*, *FirstFlrSF*, *SecondFlrSF*, and *TotalSqftCalc*) were simultaneously included, resulting in VIF values exceeding 100 for some predictors. Backward elimination and stepwise selection reduced the most extreme multicollinearity by excluding redundant size measures, while maintaining comparable explanatory power (Adjusted R<sup>2</sup> ≈ 0.93).

Overall, although the three selection procedures yielded slightly different final models, they consistently identified the same core predictors of housing prices, and multicollinearity concerns were primarily attributable to redundant representations of house size and correlated quality indicators rather than problematic dependence among neighborhood indicators.

### 3.3 Model Comparison

The four candidate models were compared using adjusted R<sup>2</sup>, AIC, BIC, mean squared error (MSE), and mean absolute error (MAE) computed on the training sample. The forward, backward, and stepwise selection models exhibited nearly identical in-sample performance, with adjusted R<sup>2</sup> values around 0.927 and similar prediction errors. In contrast, the benchmark “junk”

model performed substantially worse across all metrics, indicating poor explanatory power and predictive accuracy. Model rankings differed by metric: the stepwise model ranked first according to adjusted R<sup>2</sup> and AIC, the backward model ranked first according to BIC and MAE, and the forward model achieved the lowest MSE. These differences reflect the distinct objectives of each criterion, as AIC and BIC penalize model complexity differently, while MSE and MAE focus solely on predictive accuracy. Consequently, no single model dominates across all metrics, and the choice of a preferred model depends on whether the primary goal is predictive performance, parsimony, or interpretability.

#### 4. Predictive Accuracy

Out-of-sample predictive performance was evaluated using the test dataset and summarized using mean squared error (MSE) and mean absolute error (MAE). The MSE and MAE rankings across the four models are reported below in Table 2.

*Table 2. MSE and MAE Rankings Across Four Models*

Model	Test MSE	MSE Rank	Test MAE	MAE Rank
<b>Forward</b>	359,332,575	2	13,465.39	2
<b>Backward</b>	356,910,948	1	13,455.89	1
<b>Stepwise</b>	360,649,325	3	13,497.79	3
<b>Junk</b>	681,172,107	4	19,116.77	4

While the forward, backward, and stepwise models exhibited similar in-sample performance, their out-of-sample predictive accuracy differed slightly. The backward selection model achieved the lowest test-set MSE and MAE, making it the best-performing predictive model. Notably, the model that achieved the best in-sample fit (stepwise selection) did not achieve the best out-of-sample performance, reflecting the presence of overfitting in more complex specifications. This highlights the importance of evaluating predictive models on held-out data rather than relying solely on in-sample fit measures.

Both MSE and MAE provide useful but distinct perspectives on predictive accuracy: MSE places greater weight on large prediction errors, while MAE reflects typical absolute deviations. In this case, both metrics agree on the ranking of models, reinforcing the conclusion that the backward selection model generalizes best. Consequently, model evaluation should consider both metrics jointly to capture different aspects of prediction error.

#### 5. Operational Validation

Out-of-sample predictive performance was further evaluated using a business-oriented metric called *PredictionGrade*, which categorizes predictions based on their percentage error relative to actual sale prices: Grade 1 for predictions within 10%, Grade 2 within 10–15%, Grade 3 within 15–25%, and Grade 4 for errors greater than 25%. On the training dataset, the Forward, Backward, and Stepwise models correctly predicted sale prices within 10% of the actual value for approximately 73–74% of observations, while the Junk model achieved only 56%, with a

noticeably higher proportion of extreme errors (Grade 4, ~9%). These patterns are held on the test dataset, with the top three models achieving roughly 75% Grade 1 predictions and the Junk model remaining far less accurate.

According to GSE standards for Automated Valuation Models (AVMs), a model is considered “underwriting quality” if it predicts within 10% at least 50% of the time; thus, Forward, Backward, and Stepwise models meet this threshold, while the Junk model does not. These results illustrate that translating predictive accuracy into practical thresholds, such as *PredictionGrade*, provides actionable insights for business decision-making, offering a clearer picture of model reliability than traditional statistical metrics alone. Furthermore, while in-sample fit measures suggested similar performance among the top three models, their out-of-sample accuracy confirms that slight differences exist and underscores the importance of validating models on held-out data.

## 6. Final Model Development

The final model was developed in three phases, with the first phase outlined above. The second phase involved systematic refinement through a series of modifications including completing categorical variables by adding missing dummy levels, addressing multicollinearity by removing high-VIF variables, improving parsimony by removing low R-squared contributors, eliminating counter-intuitive coefficients, and addressing heteroskedasticity through log transformation of the response variable. The third phase focused on validation and selection, where all modifications were tested on the holdout test set and models were compared using mean squared error, mean absolute error, and mean absolute percentage error. Diagnostic assessment included VIFs, the Breusch-Pagan test for heteroskedasticity, residual plots, and evaluation of influential observations.

Performance was evaluated using multiple metrics to provide a comprehensive assessment of model quality. Mean squared error measures the average squared prediction error and is particularly sensitive to large errors. Mean absolute error provides the average absolute prediction error in dollar terms. Mean absolute percentage error (MAPE) expresses average error as a percentage of actual price, facilitating comparison across different price ranges. Additionally, the calculated prediction grades represent the percentage of predictions falling within 10%, 15%, and 25% error thresholds, providing a business-oriented view of prediction quality.

Six distinct refinement strategies to address the diagnostic issues identified in the baseline model. The first strategy involved completing categorical variables to ensure proper factor representation, with the expectation of better interpretability. The second strategy aimed to remove *QualityIndex* to reduce VIF in quality variables and enable valid inference. The third strategy focused on removing low contributors to improve parsimony and achieve a simpler model. The fourth strategy targeted removal of *BedroomAbvGr* due to its counter-intuitive negative sign, expecting more logical coefficients. The fifth strategy involved removing *TotalSqftCalc* to reduce VIF in size variables and decrease multicollinearity. The sixth strategy

employed log transformation to fix heteroskedasticity and obtain valid standard errors. Each modification was carefully evaluated on the test set to assess its impact on predictive performance.

A comprehensive performance summary reveals stark differences across the seven models tested. Model 2, complete categories, achieved the best test MSE of 352 million with test MAE of \$13,386 and test MAPE of 7.59%, using 38 predictors and exhibiting adjusted R-squared of 0.927 but Breusch-Pagan statistic of 435. Model 1, the backward selection original, achieved second-best test MSE of 357 million with MAPE of 7.64% using 34 predictors. Model 7, log-transformed, placed third with test MSE of 364 million and MAPE of 8.16% using 33 predictors but achieving Breusch-Pagan statistic of only 95. Models 3 through 6, representing various cleanup procedures, performed progressively worse, with Model 6 having the worst test MSE of 430 million and MAPE of 8.67%. Performance ranking by test MSE demonstrates that Model 2 complete categories perform best with no degradation from the optimal. Model 1 backward original performs 1.4% worse. Model 7 log-transformed performs 3.4% worse but has best diagnostics.

Several key observations emerge from this comparison. First, categorical completion improved performance, with Model 1 to Model 2 showing 1.4% improvement, teaching the lesson that categorical variable representation should always be complete. Second, cleanup steps degraded performance, with Model 2 to Model 3 losing 11.3% by removing QualityIndex and Model 5 to Model 6 losing 8.6% by removing TotalSqftCalc, for cumulative degradation from Model 2 to Model 6 of 22.2%, teaching the lesson that diagnostic fixes can substantially hurt prediction. Third, log transformation offered the best trade-off, with Model 2 to Model 7 showing only 3.4% prediction loss but 78% heteroskedasticity improvement, teaching the lesson that small performance costs can purchase major diagnostic benefits.

Diagnostic comparison shows clear differences between Model 2 and Model 7 across multiple dimensions. For heteroskedasticity, Model 2 has Breusch-Pagan statistic of 435 while Model 7 has 95, making Model 7 the clear winner. For complexity, Model 2 uses 38 predictors while Model 7 uses 33, favoring Model 7. For interpretability, Model 2 expresses effects in dollar changes while Model 7 uses percentage changes, generally favoring Model 7 for communication. For statistical validity, Model 2 cannot support valid inference while Model 7 can strongly favoring Model 7. Overall, Model 7 demonstrates clear superiority on diagnostics while Model 2 demonstrates superiority on prediction.

The decision framework leads to our primary recommendation of Model 7, the log-transformed specification. The rationale rests on several pillars. First, the model sacrifices only 3.4% worse prediction than the best model, an acceptable cost in most contexts. Second, it achieves 78% improvement in heteroskedasticity, a major benefit for statistical practice. Third, it enables valid statistical inference capability for hypothesis tests and confidence intervals. Fourth, it follows standard practice in housing economics, meeting publication expectations. Fifth, it offers better interpretability through percentage changes that communicate effectively to diverse

audiences. Sixth, it provides greater simplicity with 33 rather than 38 predictors, easing explanation and deployment.

## 7. Conclusions

After several weeks working intensively with the Ames housing data, attempting to predict sale prices from observable property characteristics, I have come to appreciate both the deceptive complexity of what initially appears to be a straightforward regression problem and the limitations of applying textbook diagnostic rules mechanically. The experience has fundamentally challenged my understanding of the relationship between model complexity, diagnostic cleanliness, and predictive performance.

The most instructive challenge throughout this analysis has been multicollinearity, though not in the straightforward way textbooks typically present it. I discovered that multicollinearity in this dataset manifests in three distinct forms, each requiring different treatment and challenging conventional wisdom about VIF thresholds. The first type is structural multicollinearity that exists by mathematical design. TotalSqftCalc represents total finished square footage and is derived by summing first floor square footage, second floor square footage, and total basement square footage. This creates VIF values in the range of four to five for these size variables. Standard textbook guidance suggests removing the redundant composite variable since the components should capture the same information. When we followed this guidance and removed TotalSqftCalc, test performance degraded by 8.6 percent. This puzzling result suggests that the aggregate measure captures something, perhaps non-linear relationships, scaling effects, or implicit interactions, that the individual components do not fully represent despite being mathematically equivalent on paper.

Severe heteroskedasticity represents the second major challenge inherent in housing price data. The Breusch-Pagan test yielded a statistic of 435 with p-value less than  $2.2e^{-16}$ , indicating that prediction errors vary systematically with the level of predicted prices. This is not surprising given the nature of housing markets. A fifty-thousand-dollar prediction error on a five-hundred-thousand-dollar home represents ten percent error and might be acceptable, while the same absolute error on a one-hundred-fifty-thousand-dollar home represents thirty-three percent error and is clearly unacceptable. The textbook solution is a log transformation of the response variable, which we applied successfully, reducing the Breusch-Pagan statistic from 435 to 95, an improvement of 78%. However, this solution creates its own complications. Log transformation introduces back-transformation bias because the expected value on the original scale does not equal the exponential of the expected value on the log scale due to Jensen's inequality.

The most direct path to improved predictive accuracy would be obtaining additional data on variables currently missing from the dataset. School district quality ratings are publicly available and could be merged using address identifiers or census tract codes. Crime statistics exist at neighborhood or precinct level and could be geocoded to properties. Walkability scores from services such as Walk Score could be matched to property locations. Age and condition of major systems might be extractable from property inspection records if such data are archived. Detailed

renovation history might be recoverable from building permit databases maintained by the city. I would expect incorporating school quality and crime data, which represent fundamental value drivers in most housing markets, to reduce mean absolute percentage error by one to three percentage points if these factors are particularly important in Ames.

Whether simpler models outperform complex models has a profoundly frustrating answer that emerged clearly from our systematic testing: it depends entirely on context. Our empirical results directly challenge the parsimony principle as it is often stated. Model 2 with 38 predictors achieved test set mean squared error of 352 million. Model 6 with 33 predictors, created by removing variables to achieve cleaner diagnostics and greater simplicity, achieved test set mean squared error of 430 million, representing performance degradation of 22 percent. Greater complexity performed substantially and unambiguously better. This contradicts the conventional wisdom that simpler models generalize better to new data. However, the contradiction is more apparent than real. The conventional wisdom about superior generalization of simple models applies specifically to situations where complex models overfit, showing training set performance that substantially exceeds test set performance. We observed the opposite pattern across all seven specifications, with training mean squared error consistently exceeding test mean squared error. No overfitting occurred in any model we tested. In the absence of overfitting, additional predictive variables that capture genuine signal improve performance rather than degrading it. The parsimony principle protects against a problem we did not have.

The question of whether we need maximum fit models or whether simpler but more interpretable models serve better presents what initially appears to be a fundamental trade-off but reveals itself upon examination to be a false dichotomy based on overly narrow definitions. The term "interpretability" has at least three distinct meanings that matter differently for different stakeholders. Coefficient interpretability asks whether we can explain what each individual coefficient means in practical terms. Model interpretability asks whether we can explain how the model generates predictions and trace the logic from inputs to outputs. Practical interpretability asks whether stakeholders can use the model insights to make decisions or take actions. These three types of interpretabilities are related but distinct, and a model might score well on one dimension while scoring poorly on another.

Our Model 2 represents the maximum fit specification with 38 predictors and linear scale exhibits mixed interpretability. Individual coefficients such as first floor square footage at \$35 per square foot have clear, intuitive meaning. However, the coefficient for bedrooms above grade at negative \$4,373 requires sophisticated statistical explanation about controlling for total space. The coefficients for quality level dummies require understanding of the reference category and can only be interpreted relative to baseline quality. The model uses dollar-based coefficients that are intuitive for some audiences but less natural than percentage changes for others. The severe heteroskedasticity with Breusch-Pagan statistic of 435 makes the model unusable for statistical inference requiring valid standard errors. For model-level interpretability, the linear additive structure is fully transparent, and we can trace every dollar of predicted price to specific features. For practical interpretability, the answer depends on audience. Real estate agents find 38

coefficients too complex for casual conversation. Automated pricing algorithms find it perfect for generating valuations. Academic researchers find it unusable because invalid inference prevents hypothesis testing.

Our Model 7 with 33 predictors and log scale might be called the more interpretable alternative, though this characterization requires nuance. It has five fewer predictors providing marginal simplification, expresses coefficients as elasticities where a 0.03 coefficient means three percent price increase per unit change in predictor which often communicates more clearly than dollar amounts, enables valid statistical inference through much-improved heteroskedasticity properties allowing hypothesis testing and confidence interval construction, but still requires explaining 33 coefficients which remains substantial complexity for non-technical audiences. Test set means that absolute percentage error of 8.16 percent is only 0.57 percentage points worse than Model 2, representing modest sacrifice. For coefficient interpretability, percentage changes are often clearer than dollar amounts especially when communicating with business stakeholders accustomed to thinking in relative rather than absolute terms. For model interpretability, the linear additive structure on log scale remains transparent. For practical interpretability, the model serves academic research requiring valid inference, supports mortgage underwriting requiring defensible methodology, but may not be meaningfully simpler than Model 2 for explaining to real estate agents since 33 coefficients still exceeds typical stakeholder patience.

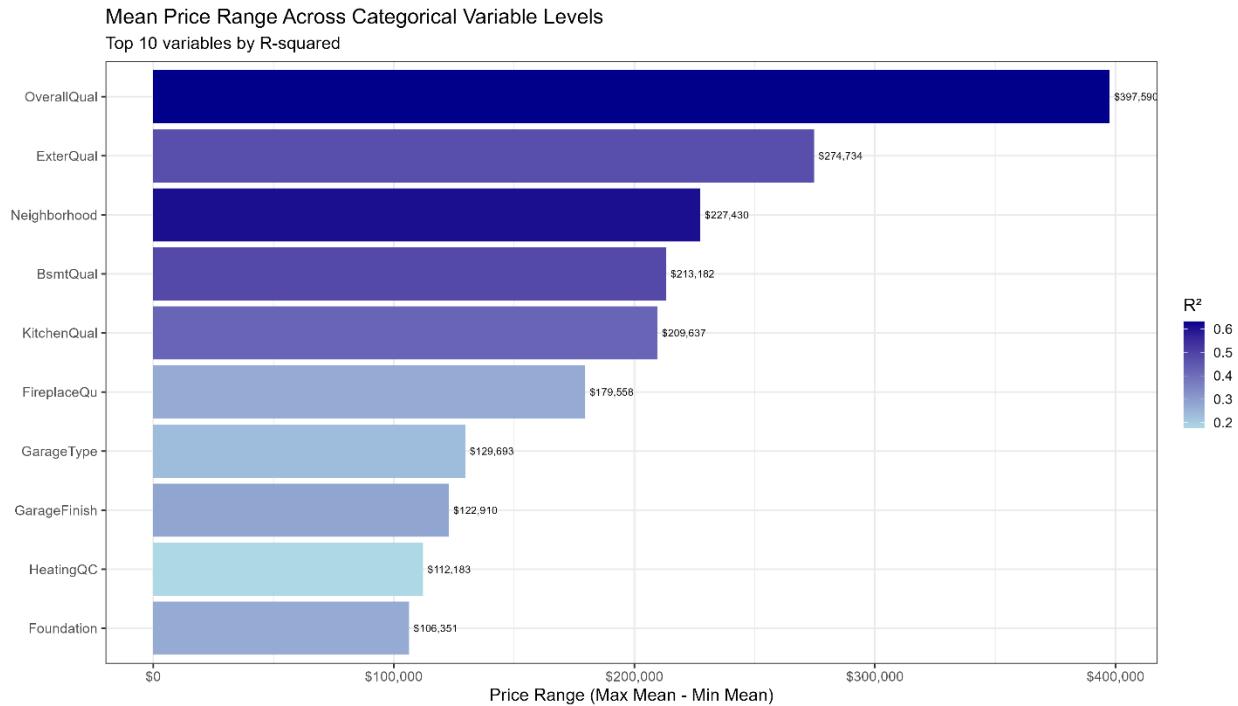
The systematic testing we conducted reveals that the supposed dichotomy between maximum fitness and interpretability is false for several reasons. First, interpretability has multiple distinct meanings and what counts as interpretable depends critically on audience and context. Model 2 is interpretable in its transparent linear structure but not interpretable for statistical inference. Second, objectives determine appropriate trade-offs and there is no universal right answer independent of how the model will be used in practice. Model 2 provides best predictions, Model 7 provides best inference capability, and a simplified variant would provide best communication. All three are correct for their respective purposes. Third, simplification can hurt both accuracy and interpretability simultaneously. Models 3 through 6 sacrificed predictive accuracy to achieve cleaner diagnostics and lower VIF, but they did not become more interpretable in any meaningful sense because 33 to 37 coefficients remain too complex for casual explanation. They achieved lose-lose outcomes. Fourth, small increases in complexity can sometimes improve interpretability rather than hurting it. Moving from 34 to 38 predictors by adding missing categorical dummy variables added modest complexity but improved interpretability by enabling proper comparison across all factor levels and making the model logically complete.

In the end, the Ames housing data taught me that statistical modeling is fundamentally a conversation between theory and evidence, where theory proposes relationships and evidence either confirms or rejects them, but neither dictates alone. The systematic testing of seven model specifications, with most 'improvements' actually degrading performance, proved more valuable than simply finding a perfect model because it revealed the dangers of mechanical rule-following and the necessity of empirical humility. The best model for predicting house prices is not the one

with the lowest VIF, the fewest predictors, or even the highest R-squared, but rather the one that honestly acknowledges its limitations, serves its intended purpose effectively, and earns the trust of the people who must use it to make real decisions about real properties in the real world.

## A. Appendix

*Figure 1. Mean Difference Among Top Categorical Variables*



*Table 1. List Containing Pool of Predictors for Modeling*

Candidate Predictor Pool				
Variables available for automatic selection				
No.	Variable Name	Type	Description	Range/Levels
1	LotArea	Continuous	Lot size in square feet	2,500 - 215,245
2	GrLivArea	Continuous	Above grade living area (sq ft)	520 - 3,820
3	TotalBsmtSF	Continuous	Total basement area (sq ft)	0 - 3,206
4	GarageArea	Continuous	Garage area (sq ft)	0 - 1,488
5	YearBuilt	Continuous	Original construction year	1900 - 2010
6	YearRemodel	Continuous	Remodel year (= YearBuilt if no remodel)	1950 - 2010
7	TotalSqftCalc	Continuous (Engineered)	Total finished square footage	520 - 5,185
8	QualityIndex	Continuous (Engineered)	OverallQual x OverallCond interaction	9 - 72
9	BedroomAbvGr	Discrete	Number of bedrooms above grade	1 - 5
10	FullBath	Discrete	Number of full bathrooms above grade	1 - 3
11	HalfBath	Discrete	Number of half bathrooms above grade	0 - 2
12	OverallQual_*	Categorical (Dummy Set)	Overall quality rating (7 dummies)	Reference: 3
13	Neighborhood_*	Categorical (Dummy Set)	Neighborhood location (20 dummies)	Reference: Blmgtn
14	BsmtQual_*	Categorical (Dummy Set)	Basement quality (4 dummies)	Reference: Ex
15	ExterQual_*	Categorical (Dummy Set)	Exterior quality (3 dummies)	Reference: Ex
16	KitchenQual_*	Categorical (Dummy Set)	Kitchen quality (3 dummies)	Reference: Ex