Ames Housing Price Prediction

Advanced Apex Project - Real Estate Price Modeling

Team: The Outliers
Institution: BITS Pilani - Digital Campus

Course: Advanced Apex Project 1
Term: First Trimester 2025-26
Supervisor: Dr. Bharathi Dasari
Submission Date: November 2025

Table of Contents

1. Executive Summary

The Ames Housing Price Prediction project builds data-driven valuation models using 2,930 residential property records and 82 raw features from the Ames, Iowa market. Through rigorous preprocessing, feature engineering, and comparative modeling, the final Ridge Regression model achieves an R-squared of 0.85 on the held-out test set, with an average error (RMSE) of $34,713. The workflow supports buyers, sellers, investors, and lenders with objective pricing guidance and actionable insights.

• Complete handling of 15,749 missing values across 27 columns.

• Five engineered features capture holistic property characteristics (e.g., Total_SF).

• Comparative modeling (Simple LR, Multiple LR, Ridge) ensures robustness.

• Visual storytelling translates analytical results into stakeholder-friendly insights.

2. Dataset Overview & Quality Assessment

Dataset Summary

Metric	Value
Total Records	2,930
Original Features	82
Final Features Used	73
Price Range	$12,789 - $755,000
Mean Price	$180,796
Median Price	$160,000
Std Dev	$79,887
Total Missing Values	15,749
Columns with Missing Data	27

Quality checks confirmed zero duplicate records and a complete SalePrice target. Features with more than 50% missingness were removed, while remaining gaps were resolved via semantic imputations (e.g., 'None' for missing basement qualities, 0 for non-existent garages) followed by median/mode filling.

3. Preprocessing & Feature Engineering

Missing Value Strategy

A four-step pipeline addressed missing data: (1) drop high-missing features (>50%), (2) encode structural absence as 'None', (3) fill numerical amenities with 0 where applicable, and (4) median/mode imputation for residual gaps. This resulted in a 100% complete modeling dataset.

Engineered Features

- Total_Bathrooms = Full Bath + 0.5×Half + basement equivalents
- Total_Porch_SF = Sum of all porch square footage
- House_Age = Yr Sold - Year Built
- Years_Since_Remod = Yr Sold - Year Remod/Add
- Total_SF = Total Bsmt SF + Gr Liv Area

Total_SF (sum of basement and above-grade living area) achieved the second-highest correlation with SalePrice (r = 0.79), validating the domain-driven approach.

4. Modeling Strategy & Evaluation

Data was split 80/20 into training and testing sets with random_state=42 for reproducibility. Simple Linear Regression established a baseline using Overall Quality alone. Multiple Linear Regression leveraged all 73 engineered and cleaned features, while Ridge Regression introduced L2 regularization to counter multicollinearity (VIF > 10 across size/quality attributes).

Model	Train R^2	Test R^2	RMSE	MAE	Overfit Gap
Simple Linear Regression	0.6325	0.6512	$52,879	$36,141	0.0187
Multiple Linear Regression	0.8612	0.8492	$34,772	$21,615	0.0120
Ridge Regression (α=1.0)	0.8609	0.8497	$34,713	$21,551	0.0112

Ridge Regression (alpha = 1.0) provided the best balance of accuracy and generalization with R^2 = 0.8497 and RMSE = $34,713 on the test set, while maintaining the lowest overfitting gap (0.0112).

5. Key Insights & Recommendations

• Overall Quality and Total_SF dominate predictive power, aligning with real estate intuition.

• Quality upgrades (kitchen, bath, finishes) yield the highest ROI for sellers.

• Buyers should benchmark price-per-square-foot within the same neighborhood to capture location premiums.

• Investors can target properties where actual prices fall below model predictions for potential arbitrage.

• Ridge Regression's stability makes it suitable for deployment and underwriting support.

6. Visual Appendix

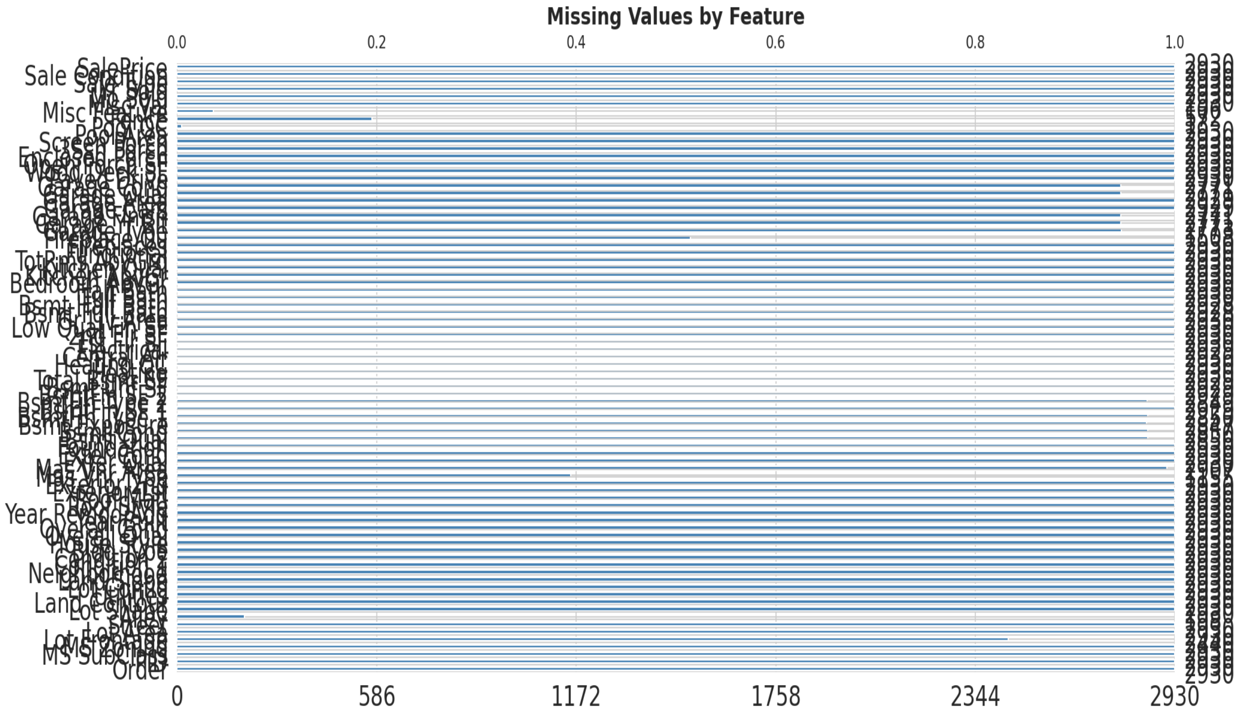

Figure 1: Missing Values by Feature

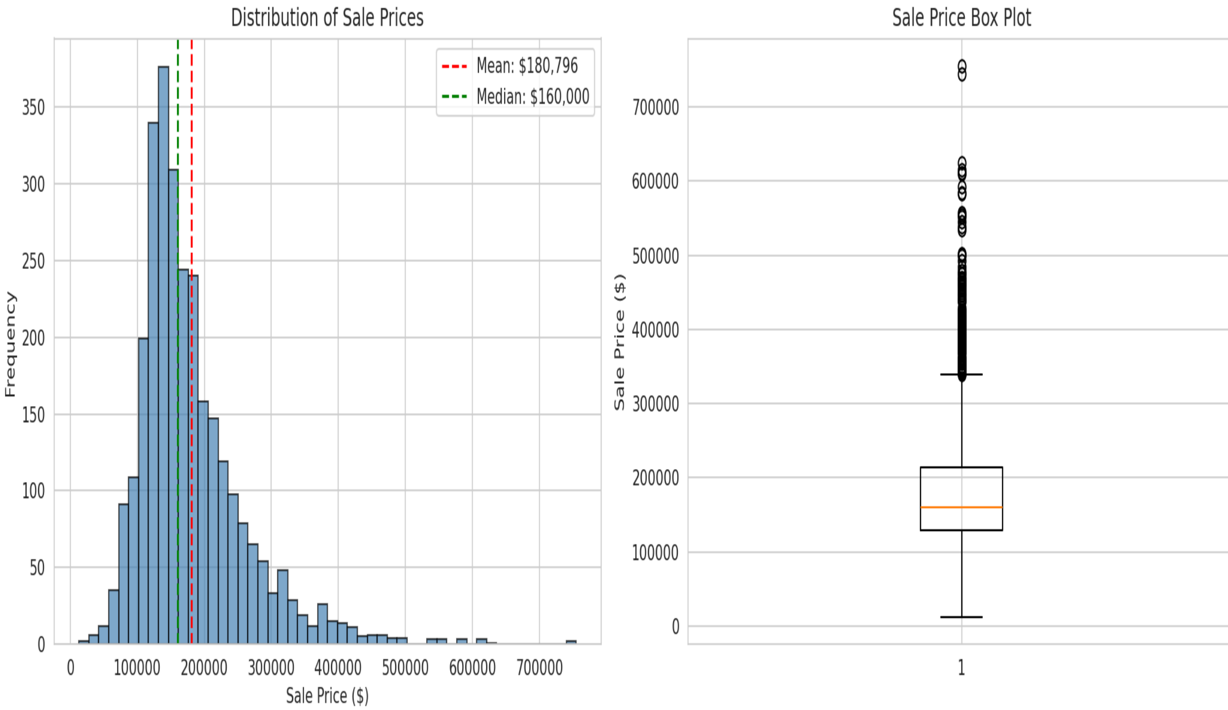

Figure 2: SalePrice Distribution

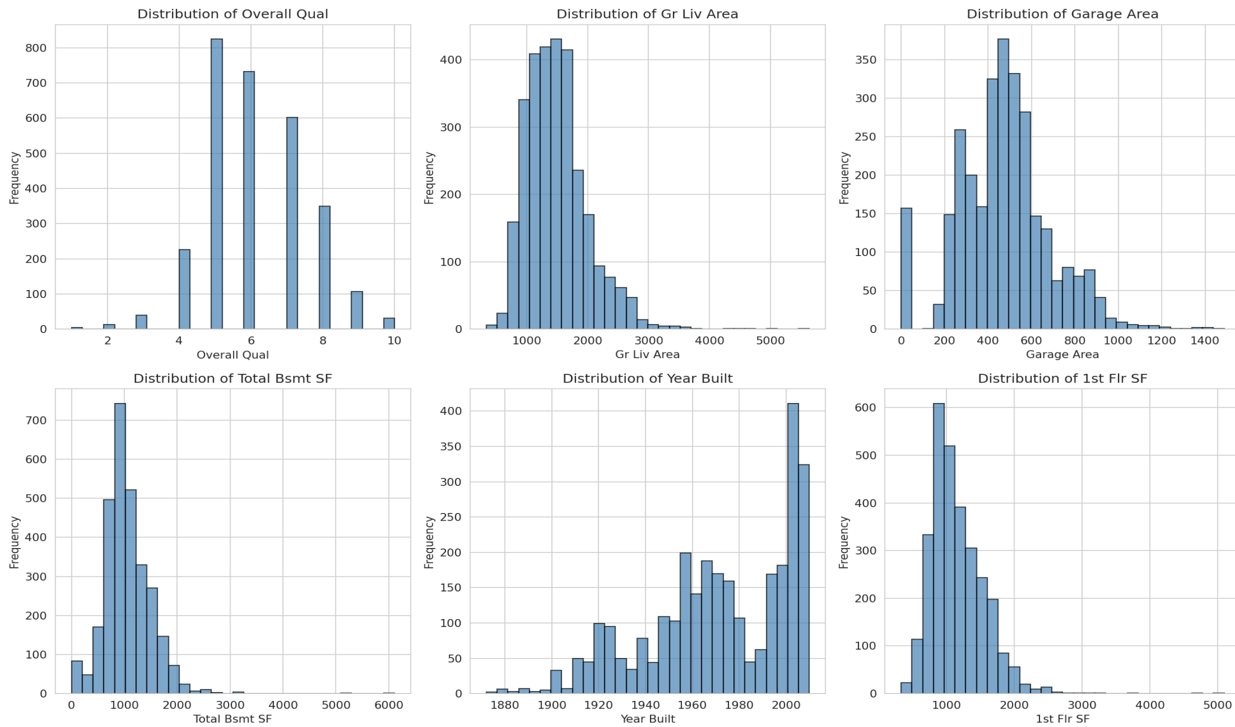

Figure 3: Key Feature Distributions

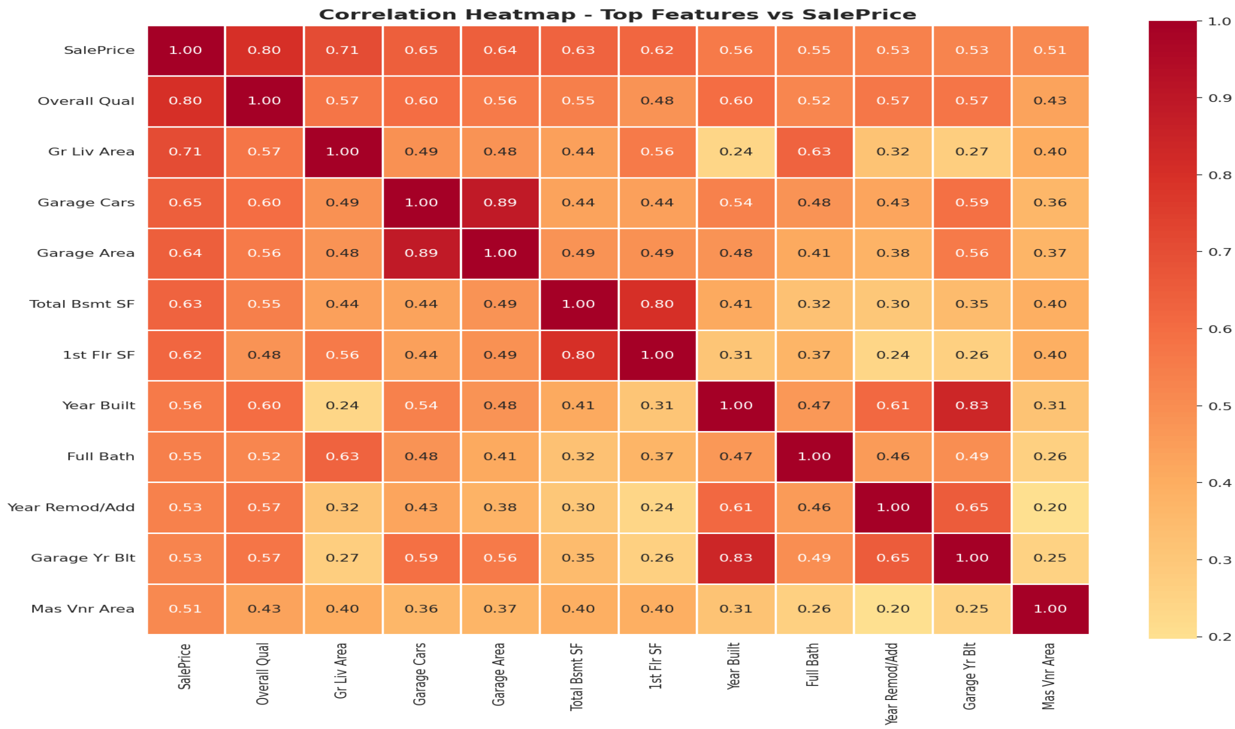

Figure 4: Top Feature Correlations

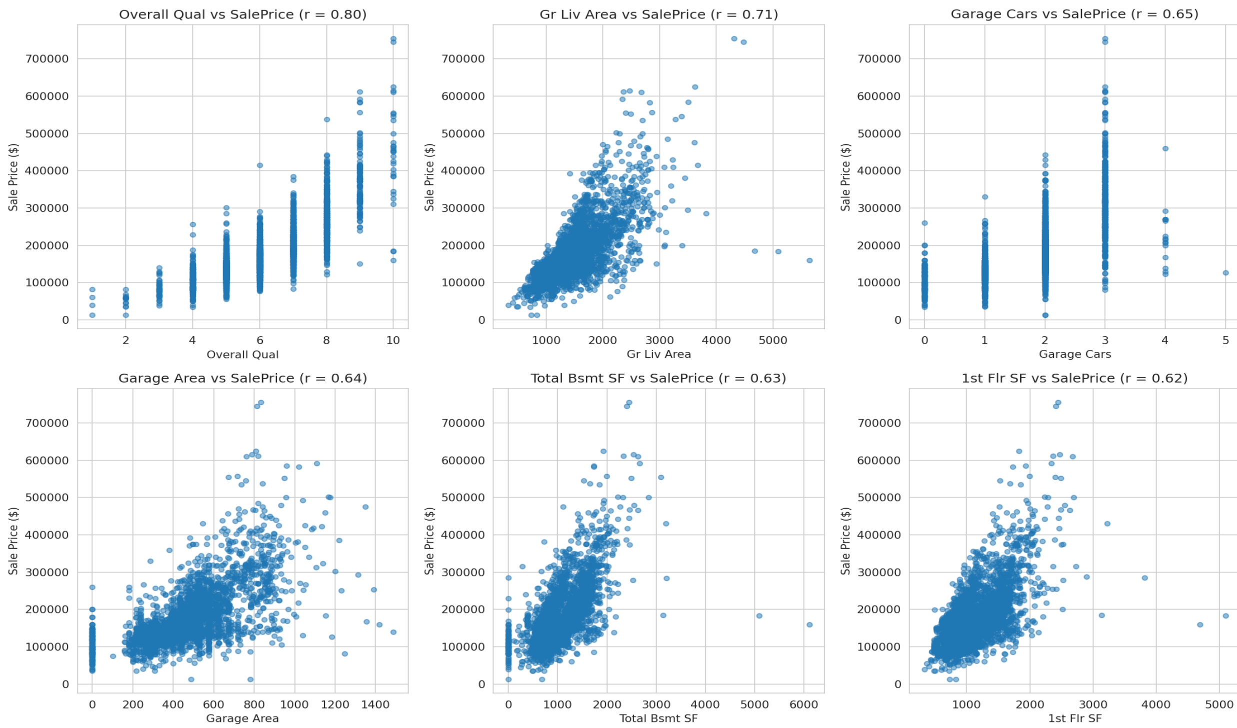

Figure 5: Top Predictors vs SalePrice

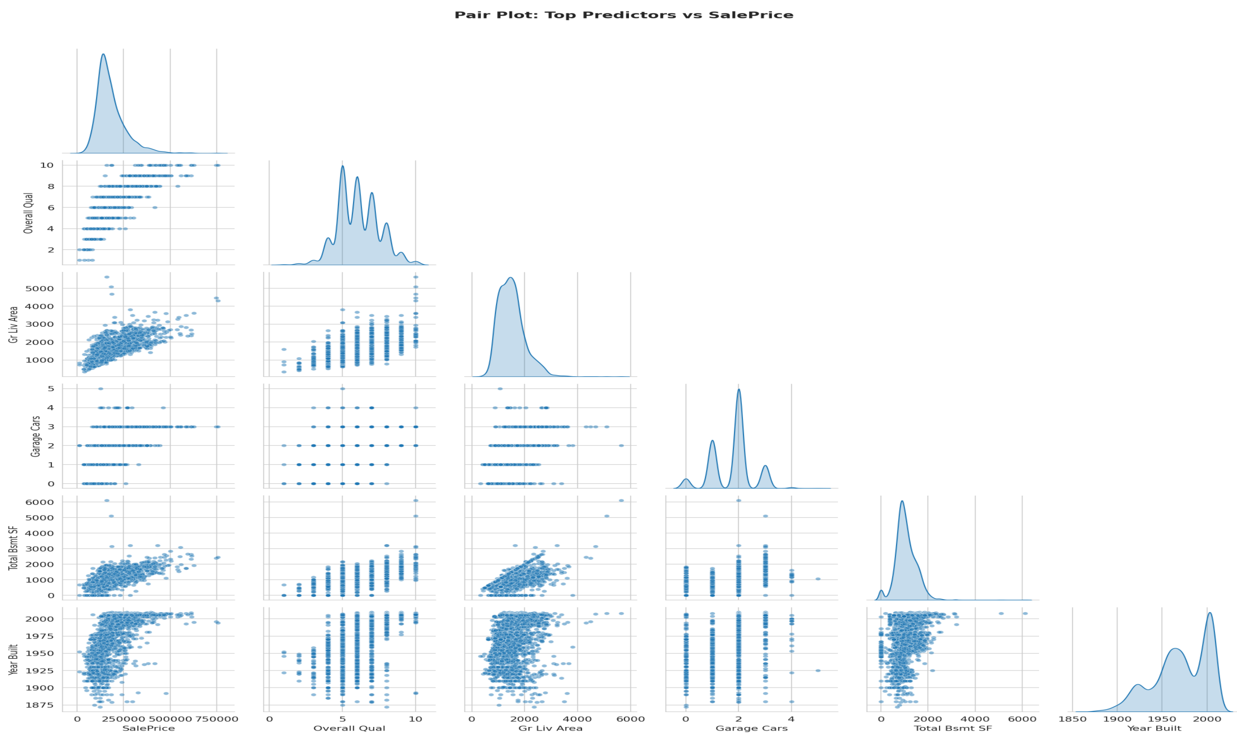

Figure 6: Pair Plot of Key Predictors

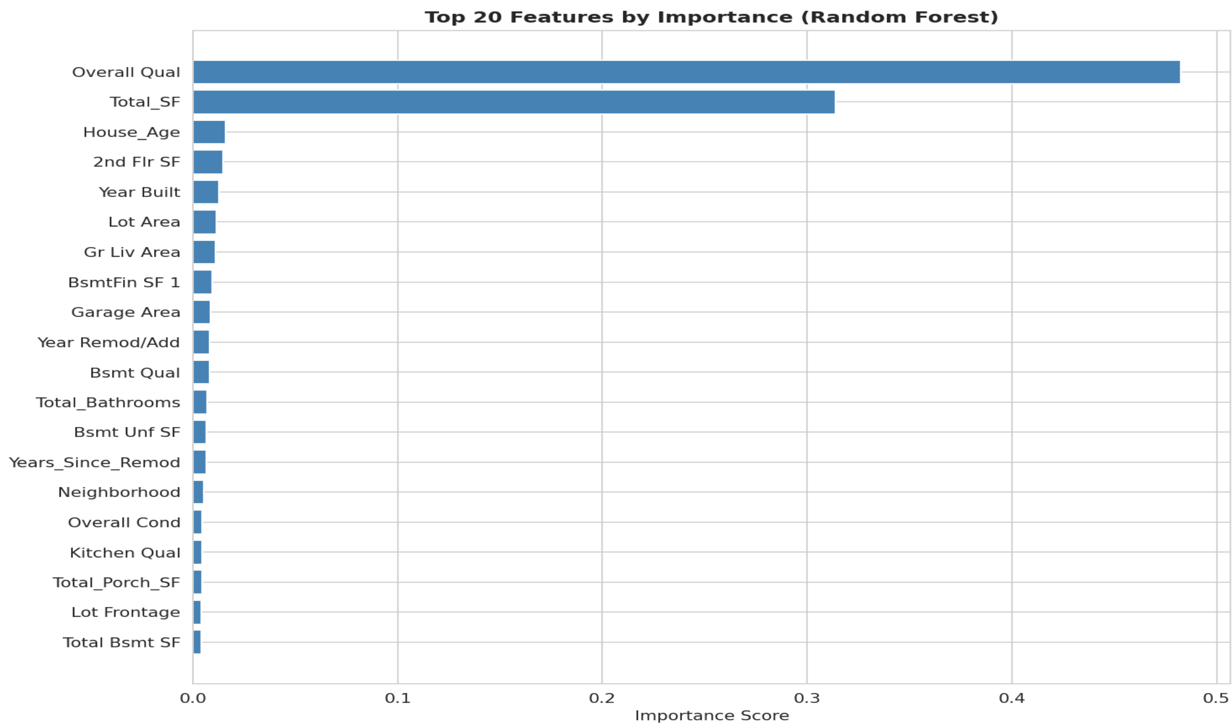

Figure 7: Top 20 Feature Importance

Figure 8: Model Comparison (R^2 & RMSE)

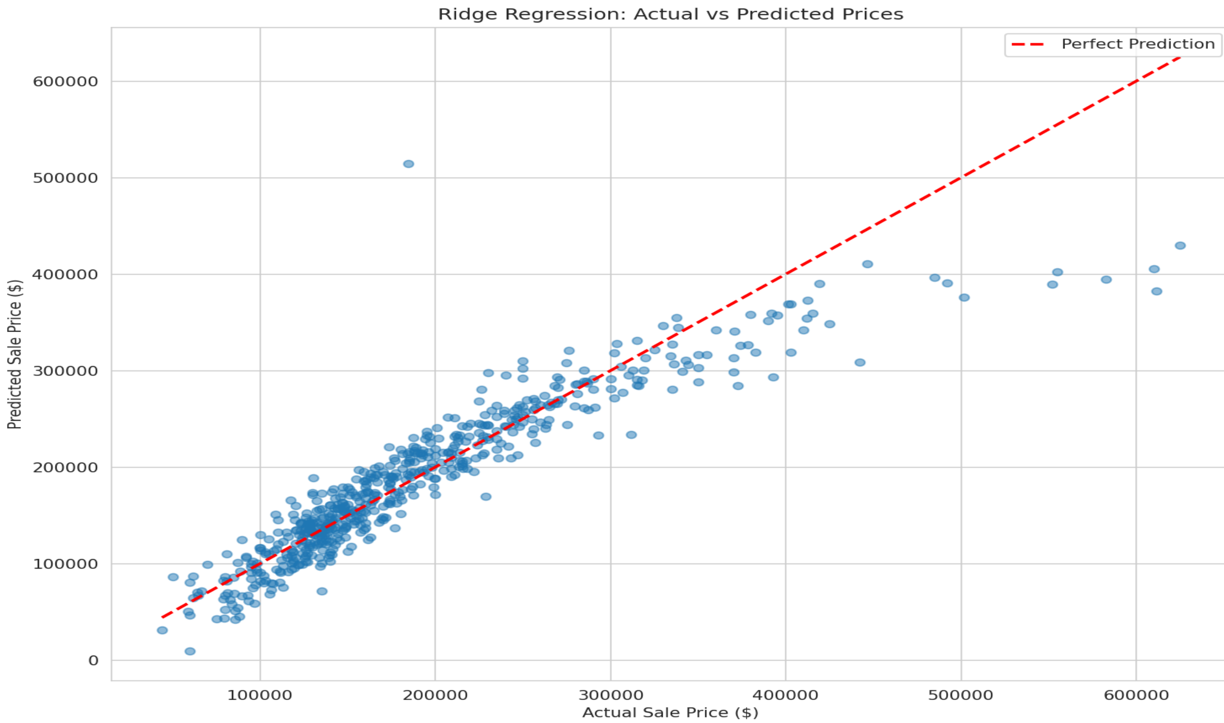

Figure 9: Ridge: Actual vs Predicted

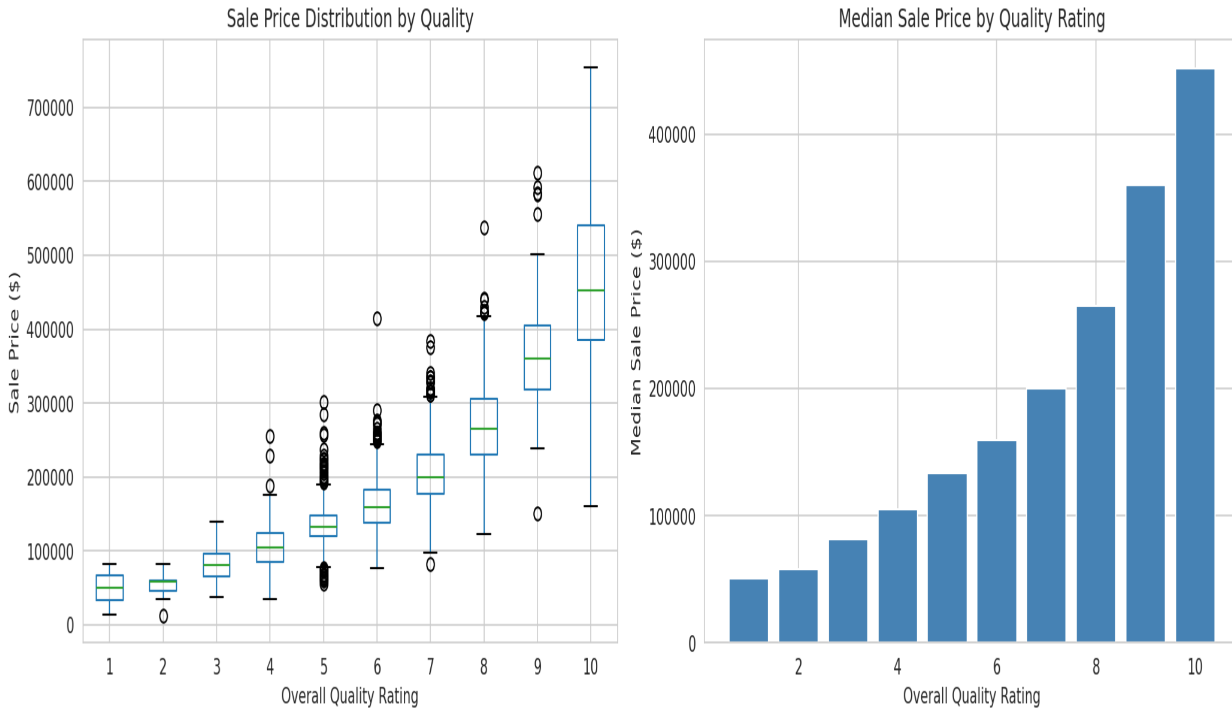

Figure 10: Price Distribution by Quality

Figure 11: Price vs Living Area by Neighborhood

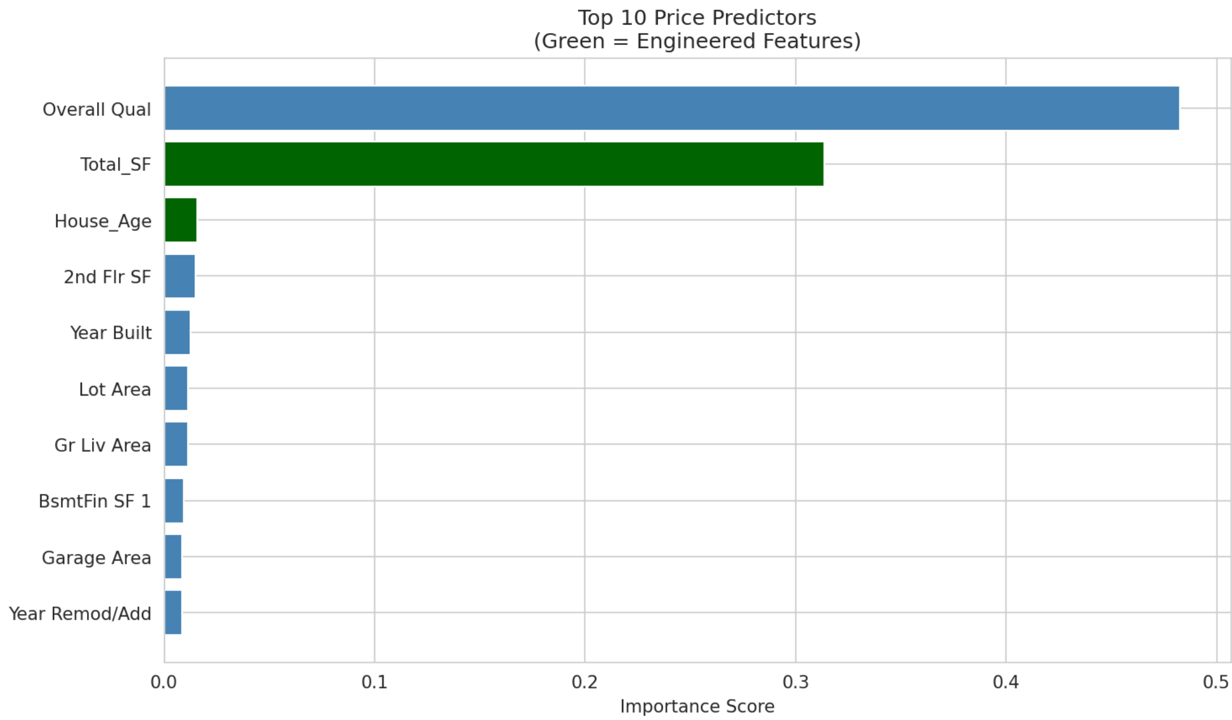

Figure 12: Top 10 Predictors (Engineered Highlighted)

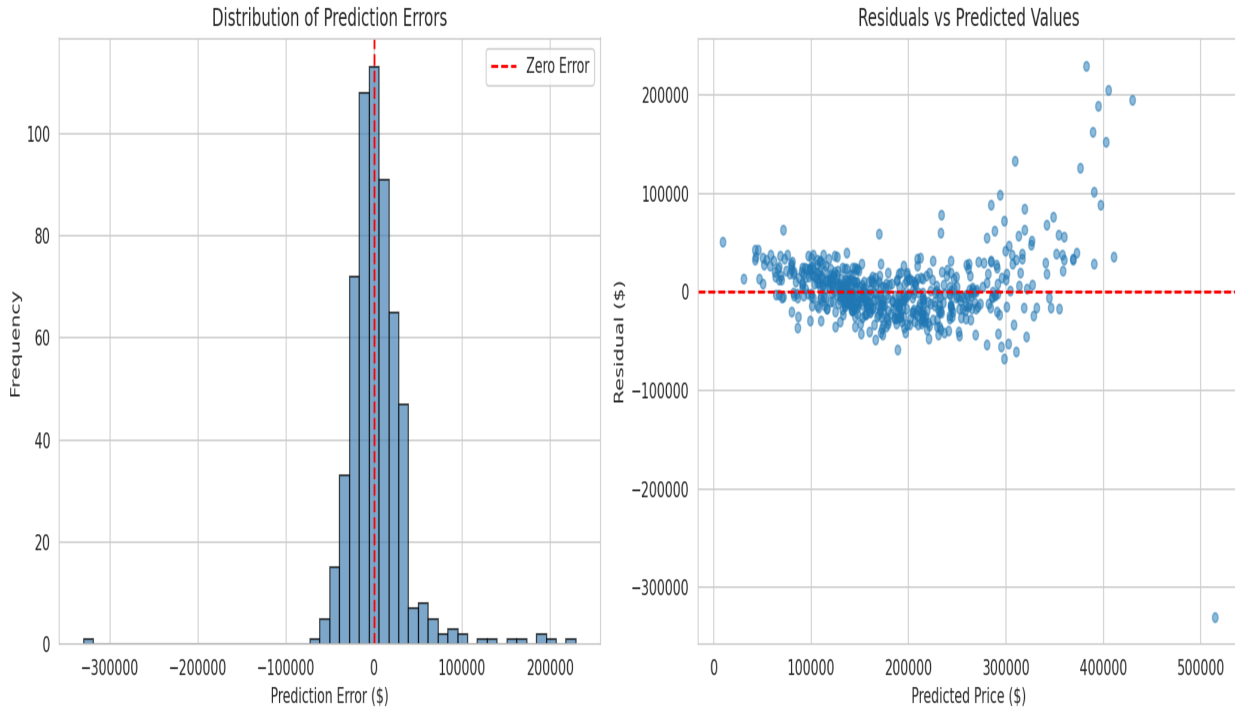

Figure 13: Residual Diagnostics

Project Summary Dashboard

```
AMES HOUSING PRICE PREDICTION - KEY METRICS

Dataset Overview
  - Properties Analyzed: 2,930
  - Features Used: 73
  - Target: Sale Price (12,789 – 755,000)

Model Performance (Ridge Regression)
  - R-squared: 84.97% of price variance explained
  - RMSE: $34,713 average error
  - MAE: $21,551 average absolute error

Top Predictors
  1. Overall Quality (r = 0.80)
  2. Total SF - Engineered (r = 0.79)
  3. Living Area (r = 0.71)
  4. Garage Cars (r = 0.65)
  5. Total Basement SF (r = 0.63)

Key Finding: Quality and size are the primary drivers of home value.
```

Figure 14: Executive Summary Dashboard