Housing Price Prediction

Danfeng Li [dl66]

December 10, 2025

UIUC

AI Usage Statement

Generative AI tools (ChatGPT by OpenAI) were used only to assist with preparation, paper reading, editing, code debugging, and improving the clarity of written explanations. All data analysis, modeling decisions, interpretations, and final edits were performed and verified by the authors. No AI tools were used to generate or alter figures, research data, or results. The authors take full responsibility for the accuracy and integrity of the work.

Abstract

This study develops a complete statistical learning framework for housing price prediction using the Ames Housing dataset. After an extensive preprocessing pipeline that involved removing variables with high levels of missingness, engineering age related features, correcting implausible area values, performing tailored missing value imputation, and applying targeted outlier removal, we obtained a cleaned dataset suitable for both unsupervised and supervised modeling. Principal Component Analysis and K means clustering were then applied to uncover latent market structure, revealing three meaningful home tiers that differ systematically in size, quality, and neighborhood composition. These insights motivated the construction of an interpretable engineered feature, MarketSegment, designed to summarize this hidden structure.

Five predictive models, including LASSO, Random Forest, XGBoost, K Nearest Neighbors, and Support Vector Regression, were trained under a unified preprocessing pipeline. LASSO achieved the strongest performance with an RMSE of approximately 16,766 and an R squared of approximately 0.915, indicating that a sparse linear model effectively captures the primary drivers of housing value. Incorporating the engineered MarketSegment feature did not improve numerical accuracy. However, LASSO's exclusion of the new feature highlights redundancy with existing predictors and demonstrates the model's interpretability and regularization behavior. Overall, the results emphasize the importance of principled preprocessing, the value of combining unsupervised insights with feature engineering, and the role of model selection in balancing predictive accuracy with interpretability.

1. Introduction and Literature Review

Accurately predicting housing prices remains a central task in real estate analytics and statistical learning, requiring both effective feature engineering and appropriate model selection. De Cock's introduction of the Ames Housing dataset established a modern benchmark for housing price prediction and emphasized the value of well-structured, domain-informed variables in regression modeling (De Cock, 2011). Building on this foundation, subsequent research has compared a range of modeling strategies. Jha et al. (2023) showed that tree-based and ensemble models often outperform simple linear methods, particularly when capturing nonlinear interactions and heterogeneous market effects, while also highlighting that meaningful feature construction can improve accuracy more reliably than increasing model complexity.

Recent studies have additionally examined interpretability and computational efficiency. John et al. demonstrated that LightGBM achieves strong performance on mixed-type housing data but requires supplementary tools for interpretability, whereas Battocchi et al. (2019) argued that predictive feature importance alone may obscure deeper structural drivers of price variation, motivating approaches that model latent factors or subgroup-level effects.

Research on hybrid and ensemble learning further supports the use of diverse model classes. Truong et al. (2020) found that stacked regressors consistently outperform individual learners across housing datasets, reinforcing the value of rich feature representations and multi-model strategies in capturing market heterogeneity.

Together, prior work suggests three key themes: (1) engineered features that summarize underlying market structure are crucial, (2) non-linear and ensemble models are effective for complex housing data, and (3) interpretability remains important alongside predictive accuracy. These insights motivate our use of PCA-based clustering to uncover latent market patterns and the construction of an engineered categorical feature, *MarketSegment*, designed to represent meaningful housing tiers within the Ames dataset.

2. Methodology

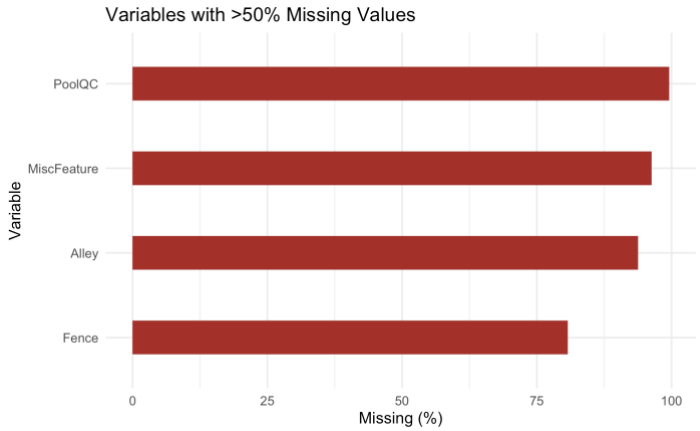
2.1 Data Preprocessing

A structured preprocessing pipeline was applied to enhance data quality and prepare the dataset for unsupervised and supervised learning. The original dataset contained 1,460 observations and 81 variables. Several systematic steps were implemented to clean the data, construct meaningful features, and address missingness and outliers.

2.1.1 Removing Variables with Excessive Missingness

The proportion of missing values was evaluated for every variable. Features with more than 50% missingness were removed because imputing them would introduce instability with limited benefit.

Using this threshold retains informative variables while avoiding noise from sparsely observed ones. In addition, the non-predictive identifier *Id* was dropped.



2.1.2 Feature Construction: Building Age and Renovation Age

To better capture structural characteristics related to a property's age, two engineered features were created:

$$age = 2025 - YearBuilt$$

$$renew\_age = 2025 - YearRemodAdd$$

These transformations provide more interpretable measures of property age and renovation recency. The original year variables were removed afterward to reduce redundancy and multicollinearity.

2.1.3 Minimum Value Constraints for Area Variables

To correct implausibly small or erroneous area measurements, lower bounds were imposed:

- GrLivArea was floored at 50 square feet,
- LotArea was floored at 100 square feet.

These constraints improve data realism while preserving the underlying distribution of reasonable values.

2.1.4 Handling Missing Values

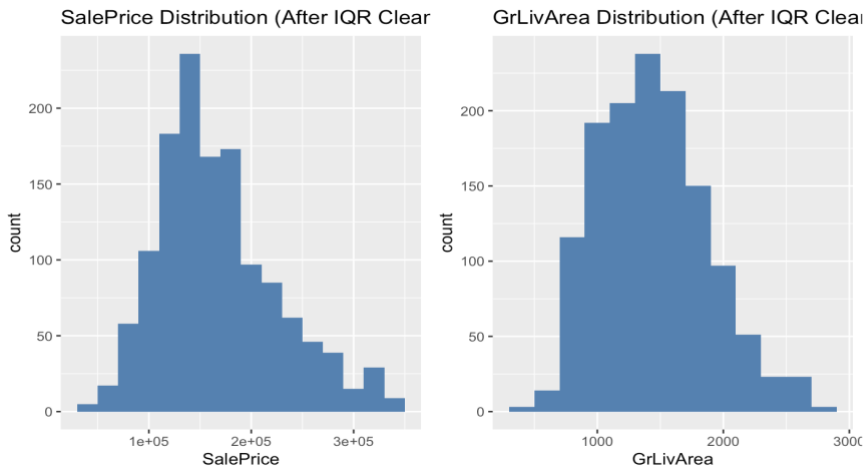
Missing data were treated separately by type. Numerical variables were median-imputed to reduce sensitivity to skewness and outliers. Categorical variables were imputed using the label "None," indicating the explicit absence of a category. The processed numerical and categorical fields were then recombined to form a complete dataset for analysis.

2.1.5 Outlier Removal Using the IQR Rule

Outlier detection was selectively applied to *SalePrice*, *GrLivArea*, and *LotArea*, which have long-tailed distributions that can disproportionately influence clustering and prediction. Observations outside the standard IQR range

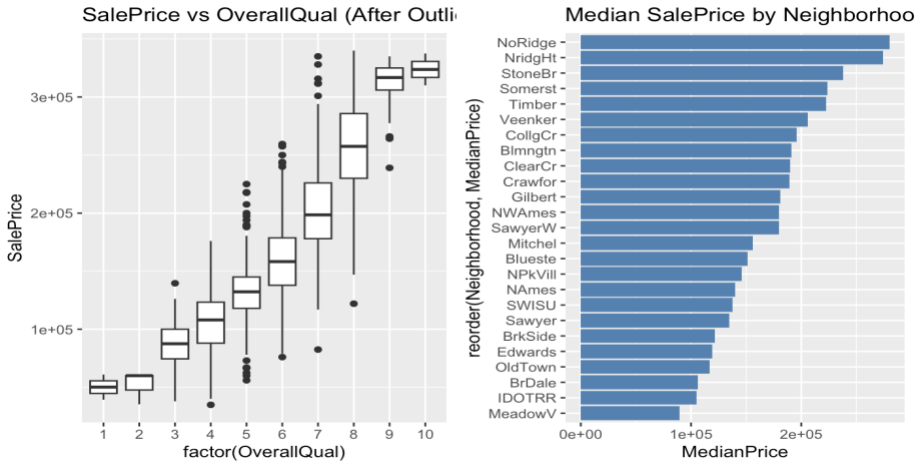
$$[Q1 - 1.5 \cdot IQR,\ Q3 + 1.5 \cdot IQR]$$

were removed. Limiting outlier removal to these key variables avoids excessive data loss while improving modeling stability.



2.1.6 Summary of the Cleaned Dataset

After preprocessing, the dataset contains 1,328 observations and 76 variables, including 37 numeric and 39 categorical features.



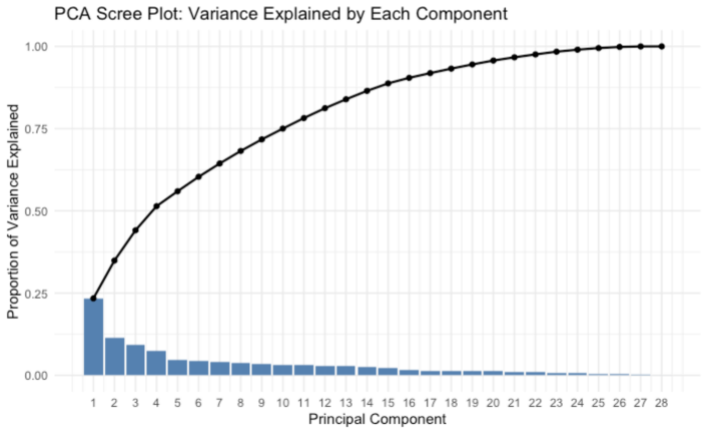
Exploratory visualizations reveal that *OverallQual* exhibits a strong positive association with *SalePrice*, and neighborhoods show substantial variation in median prices. These patterns confirm that structural quality and location are major determinants of housing value and should play a central role in predictive modeling.

2.2 Unsupervised Learning

To better understand the structure of the housing dataset before building supervised models, we applied several unsupervised learning methods to explore whether homes naturally group into submarkets and whether these groups relate to sale price or key categorical factors such as neighborhood.

2.2.1 PCA and Feature Preparation

Since clustering algorithms are sensitive to scale, we standardized all numeric predictors (excluding *SalePrice*) and applied principal component analysis (PCA).

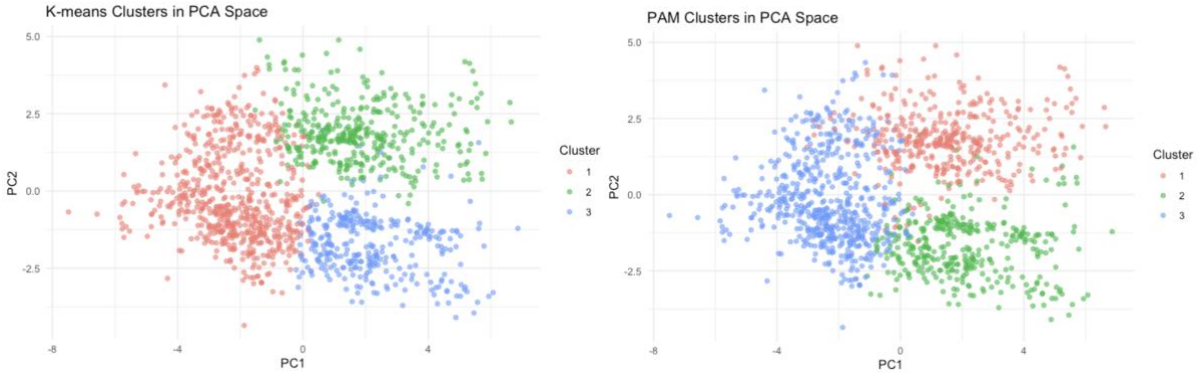


The scree plot shows that PC1 alone explains around 25% of the variance, and the first several components together capture more than 80%, indicating that PCA effectively compresses correlated housing features—such as size, quality, and lot characteristics—into a smaller set of meaningful dimensions. For visualization, we used PC1 and PC2, which already reveal clear spatial patterns that anticipate later cluster formation.

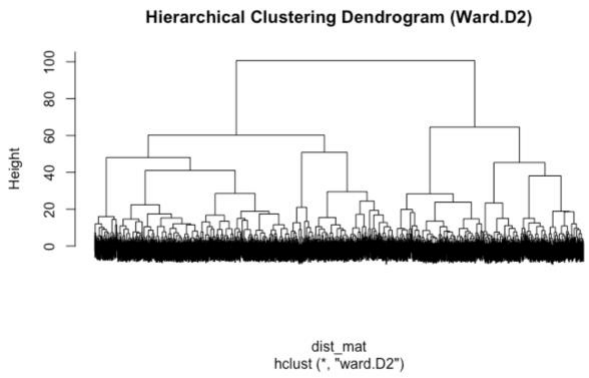
2.2.2 Clustering Results

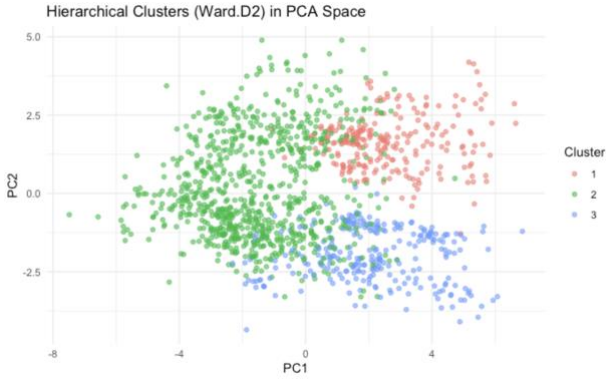
We applied three clustering algorithms—K-means, hierarchical clustering (Ward.D2), and PAM—on the PCA-transformed data.

- **K-means (k = 3)** produced three intuitive clusters. Homes in Cluster 3 (Blue) exhibit higher PC1 and PC2 values and correspond to larger, newer, higher-quality houses. Cluster 1 (Red) groups smaller or older homes, while Cluster 2 (Green) represents mid-range properties.



- **PAM (k-medoids)**, designed to be more robust to unusual points, also identified nearly identical segments.
- **Hierarchical clustering** revealed a similar three-branch structure in the dendrogram, and cutting the tree at $k = 3$ produced clusters consistent with K-means.

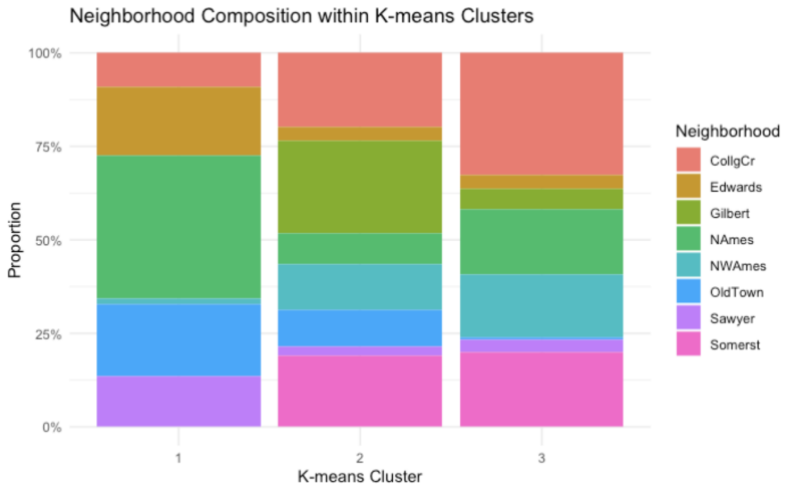




The agreement across all three algorithms suggests that the dataset contains stable, meaningful submarket divisions rather than noise-driven patterns.

2.2.3 Using Categorical Variables for Interpretation

We did not include categorical variables directly in clustering to avoid excessive dimensionality from one-hot encoding. Instead, after clustering, we examined how categories such as *Neighborhood* mapped onto the clusters. Higher-end neighborhoods like **CollgCr** and **Somerst** appear predominantly in Cluster 3, whereas Cluster 1 contains more homes from older or less expensive areas such as **Edwards**. This provides a real-world interpretation for the clusters and validates that they capture meaningful geographic and socioeconomic distinctions.



2.2.4 Relationship Between Clusters and SalePrice

Even though *SalePrice* was not used in clustering, the resulting groups differ significantly in value. Log-transformed sale prices show a clear ordering: Cluster 3 has the highest median price, Cluster 2 lies in the middle, and Cluster 1 has the lowest. This strong alignment with economic patterns reinforces the idea that unsupervised learning has uncovered genuine submarket structure within Ames.

2.2.5 Summary and Insights for Later Modeling

The unsupervised analysis reveals several useful insights:

- PCA identifies size, quality, and age as dominant drivers of variation.
- All three clustering algorithms consistently identify three major home groups.
- These clusters relate strongly to *SalePrice*, even though price was not used in the clustering process.
- Neighborhood composition also differs across clusters, suggesting important interactions between physical features and location.

These findings support the use of cluster labels as engineered features and motivate models capable of capturing non-linear patterns and market segmentation. They also justify our later construction of the MarketSegment variable as a structured representation of the underlying housing tiers in the dataset.

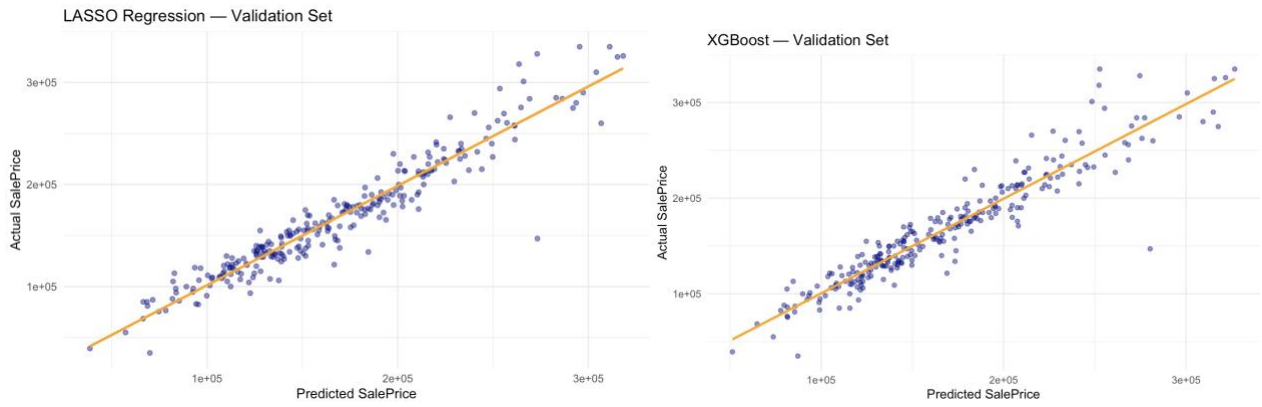
3. Model Development and Selection

3.1 Prediction Models

To compare different strategies for predicting housing prices, we evaluated five representative regression models—LASSO, Random Forest, XGBoost, k-Nearest Neighbors (KNN), and Support Vector Regression (SVM). These models span linear, tree-based ensemble, distance-based, and kernel-based learning paradigms. All models were trained using the same 80/20 train–validation split and a unified preprocessing pipeline including dummy encoding, normalization, median imputation, and handling of unseen categorical levels. Performance was assessed via RMSE and R^2 to allow fair comparison across modeling frameworks.

3.1.1 LASSO Regression

LASSO (Least Absolute Shrinkage and Selection Operator) is a regularized linear model that applies an L1 penalty to shrink coefficients and perform variable selection, improving generalization under multicollinearity and high dimensionality. The model was implemented using **glmnet** with 5-fold cross-validation to determine the optimal penalty parameter λ. Using λ that minimized cross-validated error, LASSO achieved an **RMSE of 16,766.06** and **R^2 of 0.9152**, the strongest performance among all approaches. These results suggest that a sparse linear structure is highly effective for capturing variability in Ames housing prices.



3.1.2 Random Forest

Random Forest is an ensemble learning method that builds multiple bootstrapped decision trees and aggregates their predictions to reduce variance and improve accuracy (Breiman, 2001). Following the framework outlined by Raschka and Mirjalili (2017), each tree is trained on a bootstrap sample while randomly selecting a subset of predictors at each split.

In this study, we implemented the Random Forest model using the randomForest package in R. The model was trained using all available predictors, with the following hyperparameters:

- ntree = 500: number of trees grown in the ensemble.
- mtry = floor(sqrt(p)): number of predictors randomly sampled at each split, following the standard rule for regression.
- nodesize = 5: minimum terminal node size, used to control tree depth and reduce overfitting.

After verifying that no missing values remained, the model was trained on the processed training set and evaluated on the validation set. Random Forest achieved an RMSE of 20,739.88 and R^2 of 0.8759, explaining about 88% of the variation in housing prices and demonstrating strong predictive accuracy relative to the other nonlinear models considered.

3.1.3 XGBoost

XGBoost is a gradient boosting framework optimized for speed and regularization (Chen & Guestrin, 2016). We implemented it using **xgboost** with DMatrix objects for computational efficiency (DMLC, 2019). After iteratively tuning the model, we selected the following set of hyperparameters:

- Set learning rate (eta) = 0.1
- Set number of boosting rounds (nrounds) = 200
- Tree-specific parameters: max_depth = 6, min_child_weight = 2, subsample = 1, colsample_bytree = 0.8
- Regularization parameters:lambda = 0.45, alpha = 0, gamma = 0.5
- Set the objective function to reg:squarederror

These settings balance model complexity and generalization performance by controlling tree size, limiting overfitting, and ensuring stable updates during training.

When evaluated on the validation dataset, the XGBoost model achieved strong predictive performance, with an RMSE of 18,827.92 and an R^2 of 0.8923, indicating high accuracy in predicting housing prices and explaining a substantial proportion of variance in the target variable.

3.1.4 K-Nearest Neighbors Regression

KNN predicts outcomes by averaging the target values of the k most similar observations. Because distance-based models degrade in high-dimensional spaces, standardized preprocessing was essential. Using caret with 5-fold cross-validation to tune k, the model achieved RMSE = 27,231.66 and R^2 = 0.7759, reflecting the limitations of KNN when applied to large, dummy-encoded feature sets.

3.1.5 Support Vector Regression (SVM)

SVR extends Support Vector Machines to continuous outcomes by fitting a function within an ε-insensitive margin. An RBF kernel was used to capture nonlinear relationships. The model was implemented using **e1071** with cost = 10, epsilon = 0.1, and a kernel width determined by gamma = 1/p.

On the validation dataset, SVR achieved a RMSE of 58,047.43 and an R^2 of 0.5478. This represents the weakest performance among all models evaluated, suggesting that SVR was unable to effectively capture the structure of the housing price data under the chosen hyperparameters.

3.2 Hyperparameter Tuning Strategy

Hyperparameter tuning was conducted using 5-fold cross-validation for all models to reduce overfitting and produce stable generalization estimates. Tree-based and nonlinear models (Random Forest, XGBoost, KNN, SVM) were tuned via grid search across key architectural parameters. XGBoost underwent the most extensive tuning due to its larger parameter space. LASSO's optimal penalty value was automatically selected by **cv.glmnet**. RMSE was used as the primary metric because of its interpretability for pricing data, with R^2 reported to quantify the proportion of variance explained.

4. Results and Interpretation

To compare predictive performance across models, five regression approaches—LASSO, Random Forest, XGBoost, KNN, and SVM—were trained and evaluated using a consistent preprocessing pipeline and an 80/20 train–validation split. RMSE and R^2 were used as primary performance metrics. Table below summarizes the results.

Model	RMSE	R2
Lasso Regression	16766.06	0.9152267
Random Forest	20739.88	0.8758659
XGBoost	18827.92	0.8922822
KNN	27231.66	0.7759455
SVM	58047.43	0.5477963

4.1 Comparative Model Performance

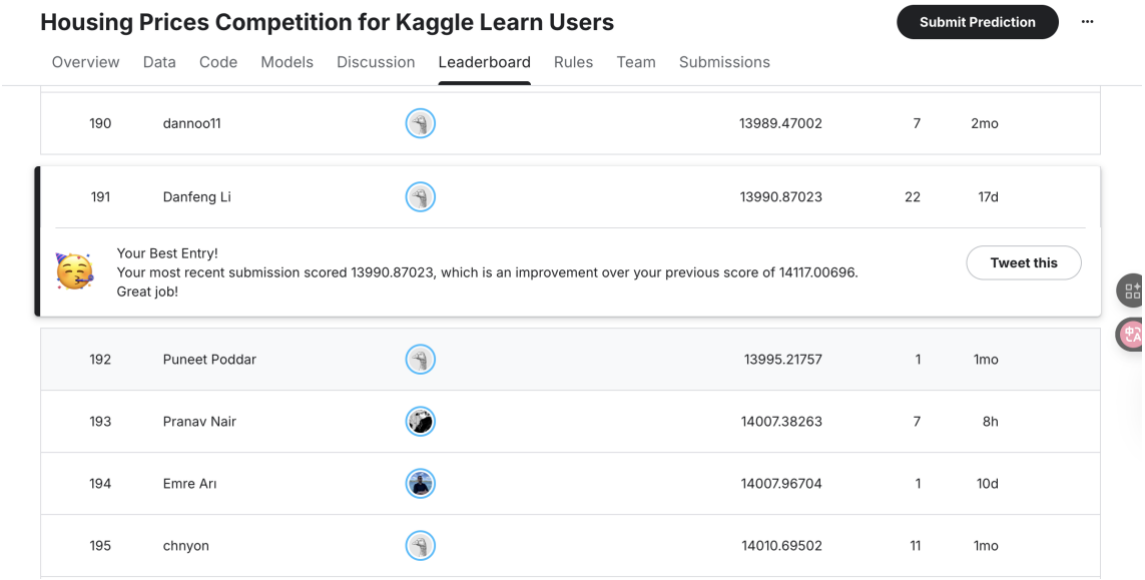
All models underwent hyperparameter tuning (e.g., cross-validated λ for LASSO, tree depth and feature sampling for Random Forest, boosting parameters and regularization for XGBoost). These procedures helped ensure fair comparisons by controlling model complexity.

LASSO achieved the lowest RMSE and highest R^2, indicating that a sparse linear structure generalizes best to this feature space. XGBoost and Random Forest performed strongly but less accurately,

suggesting that nonlinear patterns exist but do not outweigh the advantages of LASSO's regularization. KNN showed reduced accuracy, reflecting its limitations in high-dimensional, dummy-encoded data. SVM performed the worst, indicating difficulty capturing the underlying structure despite nonlinear kernels.

Given its accuracy, stability, and interpretability, **LASSO was selected as the final model**.

4.2 Performance on the Test Dataset



4.3 Interpretation and Practical Meaning

LASSO's feature-selection property highlights the most influential predictors of SalePrice. Key retained variables include:

- Overall Quality (OverallQual) – the strongest predictor, reflecting buyers' high sensitivity to construction quality and finishes.
- GrLivArea (Above-Ground Living Area) – larger homes command higher prices due to expanded usable space.
- Neighborhood indicators – location-based differences capture school district quality, accessibility, and local amenities.
- Lot characteristics – including lot size and frontage, affecting desirability and land value.
- Property age and renovation status – newer or recently renovated homes tend to be priced higher.

These findings align with established real-estate intuition and confirm that structural quality, size, and location remain the primary drivers of housing value. LASSO's interpretability ensures that these relationships are transparent and can directly inform valuation or investment decisions.

5. Feature Engineering Challenge

This section introduces an interpretable, data-driven feature—**MarketSegment**—designed to capture hidden structural differences in housing value. Building on the unsupervised learning results from Section 2.2, we derive this feature from K-means cluster assignments in PCA space. MarketSegment summarizes patterns in property size, quality, age, and neighborhood characteristics that are not explicitly encoded in the raw dataset.

5.1 Feature Design and Construction

Clustering analysis revealed that homes naturally group into three distinct clusters when projected onto the first two principal components. These clusters differ systematically in physical attributes, renovation status, and neighborhood distributions. To translate this structure into a modeling feature, we assign descriptive tier labels:

- Cluster 1 → Low Segment
- Cluster 2 → Medium Segment
- Cluster 3 → High Segment

This forms the new feature:

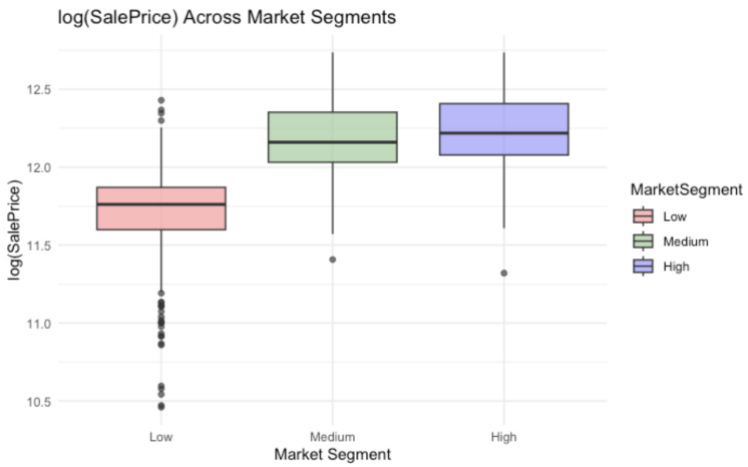
$$\texttt{MarketSegment} \in \{\texttt{Low}, \texttt{Medium}, \texttt{High}\}$$

The final distribution across segments is reasonably balanced (Low: 635, Medium: 365, High: 328), supporting the interpretability and stability of this new feature.

5.2 Interpretation in Housing Market Context

MarketSegment acts as a high-level abstraction of housing value by combining multiple correlated structural and neighborhood attributes into a single interpretable label. The three segments correspond to well-known economic tiers in real estate markets:

- Low Segment: Smaller or older homes in lower-priced neighborhoods
- Medium Segment:Typical suburban homes with moderate size and quality
- High Segment: Newer, larger homes in more desirable locations



The segments display clearly separated distributions of log sale prices, with the Low segment having the lowest median values and the High segment the highest. Since SalePrice was never used in constructing the clusters, this separation confirms that MarketSegment captures meaningful economic and structural differences among homes.

5.3 Effect of Including MarketSegment on Model Performance

To assess the predictive value of the engineered *MarketSegment* feature, we retrained the LASSO model from Section 4 using the same data split, preprocessing steps, and cross-validation procedure. Because LASSO performs automatic variable selection through its L1 penalty, it provides a straightforward way to determine whether the new feature contributes information beyond the existing predictors.

After adding *MarketSegment*, the model's performance remained unchanged (RMSE = 16,726.77; R^2 = 0.9156). All dummy variables for *MarketSegment* received coefficients of zero, indicating that LASSO considered the feature redundant with stronger predictors already in the dataset—such as *OverallQual*, *GrLivArea*, and detailed neighborhood indicators.

This result is consistent with the behavior of regularized linear models, which favor parsimonious representations and exclude features that do not add unique explanatory value. Although *MarketSegment* did not improve numerical accuracy, it remains conceptually useful: it summarizes latent market tiers uncovered through clustering and provides an interpretable lens for understanding variation in home values. This reflects the broader point that engineered features can enhance insight and interpretability even when they do not alter predictive performance.

5.4 Advantages, Limitations, and Extensions

The *MarketSegment* feature offers several **advantages**. Its tiered structure is highly interpretable and aligns with how analysts and buyers typically conceptualize housing markets. The feature is fully data-driven, derived from unsupervised clustering that summarizes complex relationships among size, quality, and neighborhood characteristics. It is also model-friendly, integrating naturally into both linear and non-linear predictive frameworks.

However, the approach has **limitations**. Cluster boundaries are specific to the Ames dataset and may not generalize without recalibration. The use of hard clustering forces each property into a single segment even when it lies between tiers. In addition, the segmentation is sensitive to PCA-related choices such as scaling and the number of retained components.

These limitations point to several **extensions**. Gaussian Mixture Models could replace K-means to yield probabilistic segment membership. Incorporating geographic coordinates would create spatially informed segmentation, and interacting *MarketSegment* with variables such as age or neighborhood could capture finer-grained market patterns. Overall, *MarketSegment* provides an interpretable, economically meaningful view of latent market structure, even if its contribution to predictive accuracy is limited.

References

Arshad, M. A., Kandanur, P., Sonawani, S., Batool, L., & Habib, M. U. (2024). *From predictive importance to causality: Which machine learning model reflects reality?* Department of Computer Science, Iowa State University; National University of Computer & Emerging Sciences; Lahore University of Management Sciences.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Distributed Machine Learning Community (DMLC). (2019). *xgboost* [GitHub repository]. https://github.com/dmlc/xgboost

Raschka, S., & Mirjalili, V. (2017). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow* (2nd ed.). Packt Publishing.

Schäfer, B., Purucker, L., Janowski, M., & Hutter, F. (2024). *How usable is automated feature engineering for tabular data?* University of Freiburg; University of Technology Nuremberg; ELLIS Institute Tübingen; Prior Labs.

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). *Housing price prediction via improved machine learning techniques.* In *Proceedings of the 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019).* Elsevier. https://creativecommons.org/licenses/by-nc-nd/4.0/

Appendix

This figure reports the distribution of the engineered MarketSegment feature (Low, Medium, High) and the corresponding performance of the extended LASSO model. After including MarketSegment as an additional predictor, the validation RMSE (16,726.77) and R^2 (0.9156) remain identical to the baseline model. The lack of improvement indicates that the LASSO penalty shrank all MarketSegment coefficients to zero, suggesting redundancy with existing high-signal predictors.

```
MarketSegment distribution:

   Low Medium   High
   329    370    629
Extended LASSO RMSE (with MarketSegment): 16726.77
Extended LASSO R2  (with MarketSegment): 0.9156384
Improvement in RMSE (baseline - extended): 0
```