Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

Predictive Modeling of House Prices: A Machine Learning Approach

1. Introduction

Accurate house price prediction is essential for buyers, sellers, and investors. This study uses the Kaggle "House Prices - Advanced Regression Techniques"[1] dataset to develop predictive models for estimating house prices. The research aims to analyze model performance across different regression techniques, identify key features affecting house prices, and examine non-linear relationships in the data.

2. Dataset Overview

The dataset consists of 1460 observations and 81 variables, including numerical and categorical attributes that describe property characteristics, quality, and market conditions. Below are key dataset statistics:

- Target Variable: SalePrice represents the selling price of houses in USD. The average house price is $180,921, with values ranging from $34,900 to $755,000
- Numerical variables: 38 (e.g., GrLivArea, TotalBsmtSF, OverallQual, GarageCars)
- Categorical variables: 43 (e.g., Neighborhood, HouseStyle, RoofStyle)

3. Data Pre-processing Pipeline

Given the complexity of the dataset, a structured data pre-processing pipeline was implemented to handle missing values, multicollinearity, and feature transformations. The key steps include:

1. Handling Missing Values:

   Features with more than 90% missing values were removed to avoid excessive imputation bias. These included:

   - PoolQC (~99% missing)
   - MiscFeature (~96% missing)
   - Alley (~94% missing)

2. Addressing Multicollinearity:

   Features exhibiting high multicollinearity (correlation > 0.8 with other predictors) were dropped to reduce redundancy and improve model interpretability. These included:

   - 1stFlrSF (high correlation with TotalBsmtSF)
   - TotRmsAbvGrd (high correlation with GrLivArea)
   - GarageYrBlt (high correlation with YearBuilt)
   - GarageArea (high correlation with GarageCars)

3. Imputing Missing Values:

   - Numerical features: Imputed using the `median` strategy to reduce the influence of outliers.
   - Categorical features: Imputed using the `most_frequent` strategy to maintain consistency.
   - Both strategies were implemented using Scikit-learn's `SimpleImputer`.

4. Feature Transformation:

   - Categorical features: Applied label encoding to convert categorical variables into numerical representations.
   - Numerical features: Standardized using z-score normalization to ensure uniform scaling across different magnitudes.

This pre-processing pipeline ensured that the dataset was well-prepared for machine learning models by improving data quality, reducing noise, and enhancing model performance.

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

4. Research Questions and Methodology Overview

This study aims to develop a robust machine learning model for predicting house prices while investigating key factors influencing the predictions. We address the following research questions:

RQ1: Can we develop a machine learning model to accurately predict house prices using the available features? How does model performance compare across different regression techniques?

RQ2: Which features have the highest predictive power? What feature engineering techniques enhance model accuracy?

RQ3: What is the statistical distribution of house prices, and what factors influence this distribution?

RQ4: Are there non-linear relationships between house prices and key features? Can advanced modeling techniques capture these relationships more effectively than linear models?

RQ5: How do temporal factors (year built and renovation dates) influence house prices?

Methodology Overview

To address these questions, we implemented a structured approach, which will be detailed in the following sections:

1. Asses statistical distribution of target variables:
    - Analyzed the statistical distribution of SalePrice to understand its properties and identify necessary transformations.
2. Feature Selection and Dimensionality Reduction:
    - Applied LASSO regression for feature selection to identify the most important predictors.
    - Used Principal Component Analysis (PCA) and Factor Analysis (FA) to reduce dimensionality while preserving variance.
3. Model Development and Comparison:
    - Evaluated multiple regression techniques, including:
        - Linear regression (Baseline model)
        - Tree-based models (Decision Tree, Random Forest, Gradient Boosting)
        - Generalized Linear Models (GLMs)
    - Compared models under different conditions:
        - Predicting SalePrice in its original form vs. log-transformed SalePrice.
        - Using all available features vs. a reduced feature subset derived from LASSO.
4. Non-linearity in Relationships:
    - Investigated non-linear relationships between house prices and key features using Polynomial Regression.
5. Impact of Temporal Factors:
    - Assessed how year built and renovation dates influence house prices.
    - Modeled these effects using a Mixed Effects Model, incorporating both fixed and random effects.
6. Model Evaluation and Interpretation:
    - Compared model performance using the following metrics:
        - Root Mean Squared Error (RMSE)
        - R-squared (R^2)
    - Assessed feature importance through:
        - Model-derived importance scores.
        - Interpretation of principal components (PCA) and latent variables (Factor Analysis).

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

5. Statistical Analysis

What is the statistical distribution of house prices?

Exploratory Data Analysis (EDA) revealed that **house prices exhibit skewness and contain outliers**, necessitating a closer examination of their distribution. To determine the most appropriate distribution, we fitted three candidate distributions: **Normal, Log-Normal, and Gamma**.

Visual Assessment: Histogram and Fitted Distributions

We plotted a histogram of **SalePrice** overlaid with the fitted distribution curves to assess the shape of the data (Figure 1).

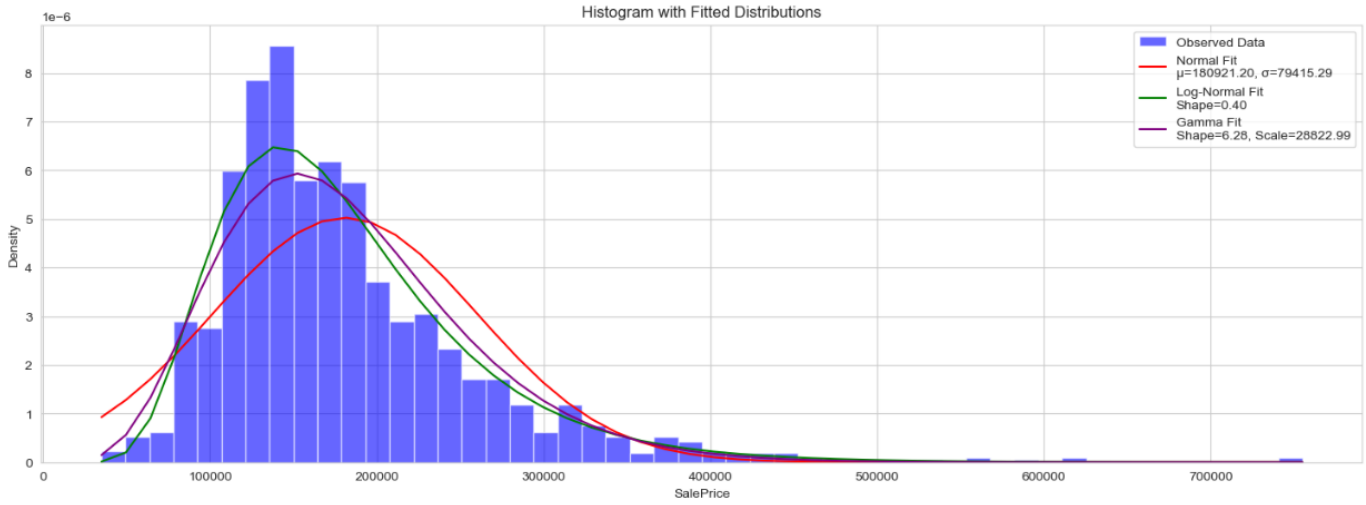

Figure 1. Histogram of SalePrice with Normal, Log-Normal, and Gamma distribution fits.

Q-Q Plot Analysis for Skewed Distributions

To further evaluate the suitability of these distributions, we used Q-Q plots, which compare observed quantiles against theoretical quantiles. If the data follows a specific distribution, the points should align closely with the reference line. The Log-Normal transformation of SalePrice demonstrated the best alignment, suggesting that house prices follow a Log-Normal rather than a Normal or Gamma distribution (Figure 2).

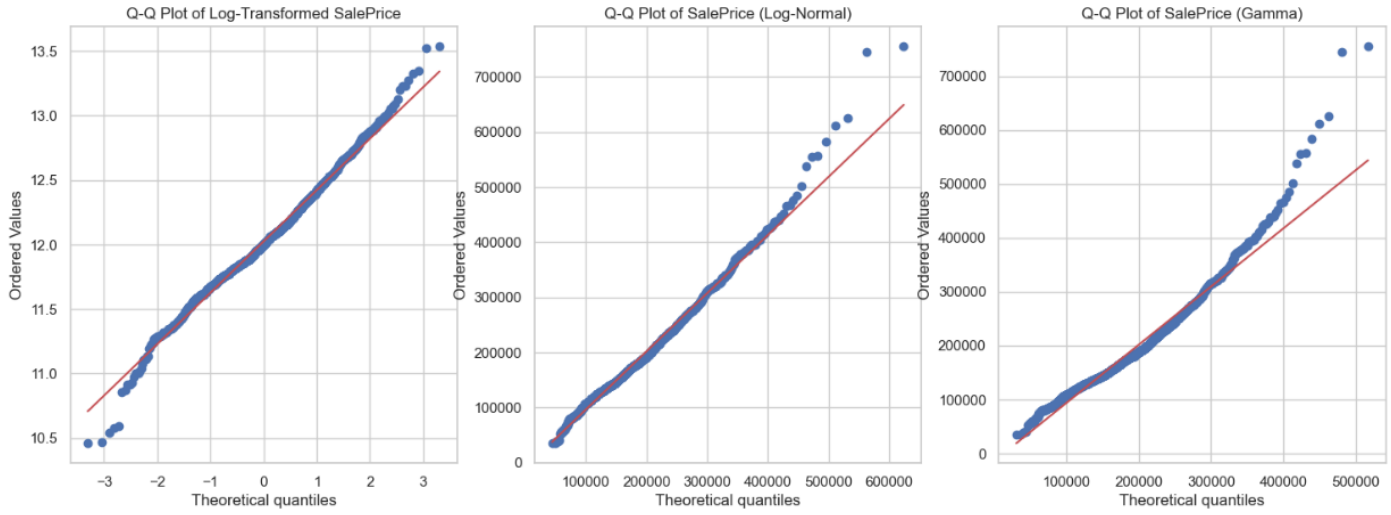

Figure 2. Q-Q Plots for Normal, Log-Normal, and Gamma distributions.

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

Best-Fit Distribution Selection Using AIC/BIC

To formally compare distributions, we computed the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values for each candidate model. Lower AIC/BIC values indicate a better fit (Refer to Table1).

Distribution	AIC	BIC
Log-Normal	**36,578.96**	**36,594.81**
Normal	37,092.04	37,102.62
Gamma	42,812.92	42,828.78

Table 1. AIC and BIC values for Normal, Log-Normal, and Gamma distributions fits

Both **AIC and BIC confirm that the Log-Normal distribution provides the best fit** for house prices, reinforcing our findings from the histogram and Q-Q plot analysis.

Feature selection using LASSO Regularization

LASSO (Least Absolute Shrinkage and Selection Operator) performs automatic feature selection by applying an L1 penalty, which forces some regression coefficients to zero. This helps eliminate features that contribute little to predicting house prices, improving model interpretability and reducing overfitting. In our dataset, which originally contained 72 numeric and categorical features, LASSO **reduced the number of relevant predictors to 40** by shrinking less important coefficients to zero.

Larger absolute coefficient values indicate higher predictive importance. A positive coefficient suggests that an increase in the feature's value leads to a higher predicted house price, while a negative coefficient implies the opposite. GrLivArea and OverallQual have the highest positive coefficients, indicating that larger living areas and higher overall quality significantly increase house prices. GarageCars also has a positive impact, suggesting that homes with more garage spaces tend to be priced higher (Refer to Figure 3).

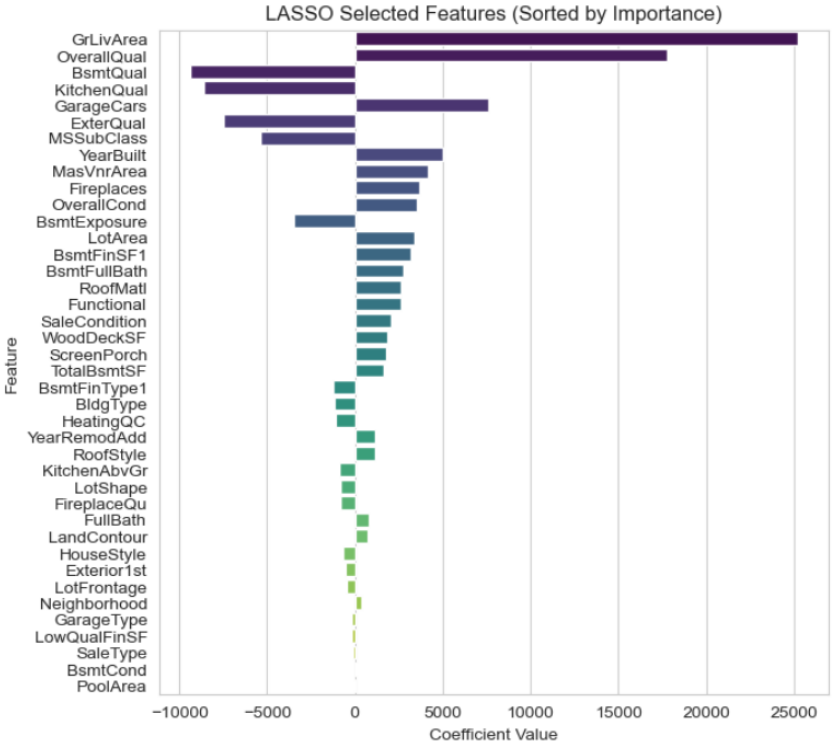

Figure 3: Bar plot of the 40 LASSO-selected features ranked by importance

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

By removing irrelevant variables and retaining the most impactful ones, LASSO enhances model efficiency while preserving predictive power. The next section will explore how these selected features contribute to different regression models.

Dimensionality reduction techniques

a. PCA

We applied Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, ensuring that at least 95% of the variance was retained. After standardizing the numerical variables, we identified that the first 24 principal components explain 95% of the data variability. The first 5 components explain 42.2% of the variance, while the first 10 explain 72.5%, demonstrating that a significant portion of the dataset's information can be represented with a smaller number of variables. This will allow us to optimize the predictive models without sacrificing accuracy. Figure 4 below shows the distribution of the variance explained by each principal component.

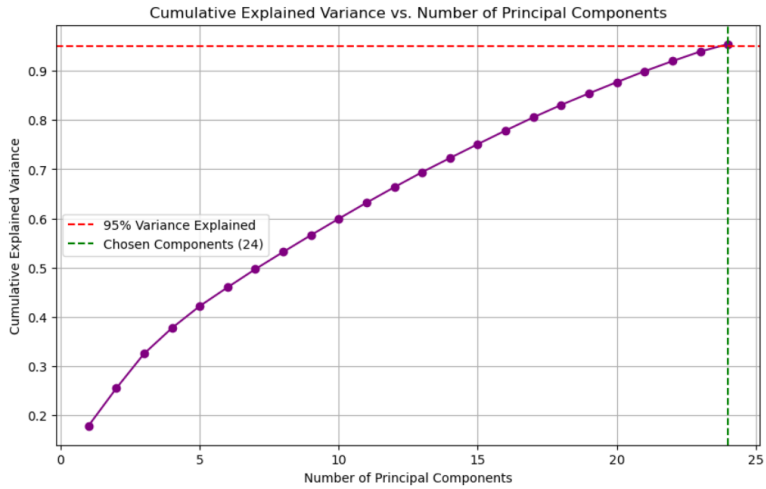

Figure 4. Scree plot of Explained variance for each Principal Component

	OverallQual	GarageCars	GrLivArea	YearBuilt	1stFlrSF
PC1	0.345528	0.325843	0.316679	0.279872	0.279150
PC2	-0.024822	-0.071846	0.347648	-0.370916	0.062873
PC3	0.097411	0.026601	0.121862	0.126042	-0.285726
PC4	-0.001560	0.002028	-0.023024	-0.050304	0.198331
PC5	-0.154979	0.033622	0.010817	0.096420	0.208370

Table 2. Top 5 Principal Components (PCs) and Top 5 variables (largest PC loadings) in PC1

b. Factor Analysis

Since PCA is a dimensionality reduction technique suited only for numerical features, we also implemented Factor analysis to handle the dataset's categorical features. We initially set the number of factors to 10 and generated a scree plot (Refer Figure 5). The largest factor explains 23.54 units of variance, suggesting that it may represent a dominant structure in the data. Several factors (Factors #1, 2, 4, 6, 7) explain a moderate amount of variance, while other factors explain much less. We then analyzed the factor loading for the factor with the largest explained variance (Refer Table 3). The feature 'Neighbourhood' exhibits a notably high loading of 5.2, indicating a strong correlation with the factor that explains the highest variance, and therefore, it is a significant predictor of SalePrice.

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

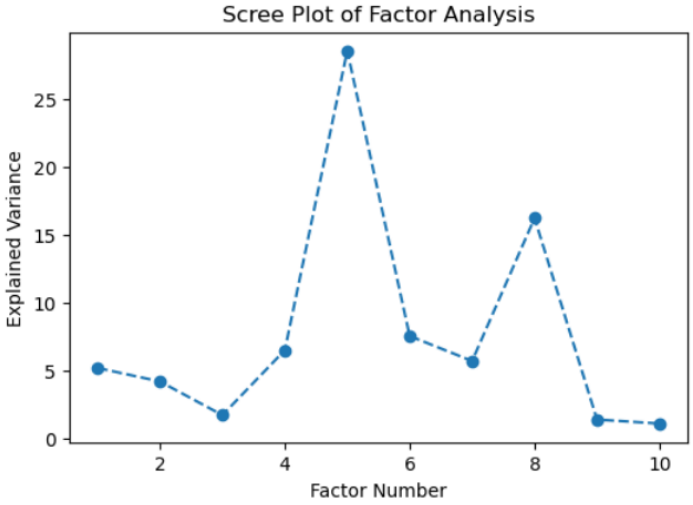

Figure 5. Scree plot of Explained variance for each Factor

Feature	Factor Loading
Neighborhood	5.210752
GrLivArea	0.298910
2ndFlrSF	0.282978
MSSubClass	0.263665
HouseStyle	0.243677
KitchenAbvGr	0.217977
GarageType	0.195528
BedroomAbvGr	0.191528
EnclosedPorch	0.162754
BsmtExposure	0.156954

Table 3. Factor loadings for the factor with largest Explained variance (Factor # 5)

Develop machine learning models to accurately predict house prices using the available features and compare performance across different regression techniques

The primary objective of this study is to build a robust predictive model for house prices and compare the performance of different regression techniques. We explored linear regression models and tree-based models to capture both linear and non-linear relationships.

Since house prices exhibit a right-skewed distribution, this can introduce biases in standard regression models. Because ordinary least squares (OLS) regression assumes normally distributed residuals, we also experimented with Generalized Linear Models (GLMs) to improve accuracy and account for distributional differences.

Modeling Approach

We trained and evaluated models using the following methodologies:

- Target Variables: Models were fitted to both the original SalePrice and its log-transformed version, log(SalePrice) to address skewness.
- Feature Selection: We trained models on both all available features and the LASSO-selected subset to assess the impact of feature reduction.
- Regression Models Used:
  - Linear Regression (using sklearn)
  - Tree-Based Models (using sklearn)
  - GLM (using statsmodels)
- Training Process:
  - 80-20 Train-Test Split was used for model evaluation.
  - 5-Fold Cross-Validation (CV) was applied to ensure model robustness.

This comparative approach provides insights into the trade-offs between simplicity, interpretability, and predictive power across different modeling techniques.

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

Original SalePrice

Model	Hyper-parameter tuning	Test RMSE	R2
Gradient Boosting	{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200}	26717.077857	0.906940
Random Forest	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}	28235.911987	0.896058
Linear Regression	NA	36224.992500	0.828919
Decision Tree	{'max_depth': None, 'min_samples_split': 10}	39210.883754	0.799553

Table 4. Performance of regression and tree-based models using all features, ranked by increasing Test RMSE

Model	Hyper-parameter tuning	Test RMSE	R2
Gradient Boosting	{'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 200}	28868.011298	0.891352
Random Forest	{'max_depth': None, 'min_samples_split': 10, 'n_estimators': 100}	30454.435118	0.879083
Linear Regression	NA	33947.764924	0.849752
Decision Tree	{'max_depth': 15, 'min_samples_split': 2}	43781.772277	0.750096

Table 5. Performance of regression and tree-based models using LASSO-selected features, ranked by increasing Test RMSE

Log-transformed SalePrice

Model	Hyper-parameter tuning	Test RMSE	R2
Gradient Boosting	{'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 100}	28235.420500	0.896062
Random Forest	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}	28604.066226	0.893330
Linear Regression	NA	28970.878133	0.890577
Decision Tree	{'max_depth': 5, 'min_samples_split': 2}	40398.015153	0.787232

Table 6. Performance of regression and tree-based models using all features, ranked by increasing Test RMSE

Model	Hyper-parameter tuning	Test RMSE	R2
Gradient Boosting	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200}	28102.709497	0.897037
Random Forest	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}	28942.482930	0.890791
Linear Regression	NA	29937.016943	0.883157
Decision Tree	{'max_depth': 10, 'min_samples_split': 10}	35505.837874	0.835644

Table 7. Performance of regression and tree-based models LASSO-selected features, ranked by increasing Test RMSE

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

Effect of Log Transformation (Comparison Between Original and Log-Transformed SalePrice)

- The Test RMSE values for log-transformed models are marginally higher than those for the original SalePrice.
- This indicates that the log transformation may not be as beneficial for improving model accuracy and predictive performance in this case.
- Among all models, Gradient Boosting consistently has the lowest RMSE after log transformation.

Effect of Feature Selection (LASSO)

- Using LASSO-selected features generally increases Test RMSE slightly compared to using all features.
- This implies that while feature selection simplifies the model, it may result in a small loss of predictive accuracy.
- However, feature selection can still be beneficial if interpretability or computational efficiency is a priority.

Model Performance Insights

- Gradient Boosting consistently outperforms other models across all scenarios.
- Random Forest is the second-best model, though its performance slightly lags Gradient Boosting.
- Linear Regression has higher error than tree-based models, indicating that it may not capture the complexities in the data as well.
- Decision Tree performs the worst, suggesting that a single tree is not sufficient to model the data effectively.

GLM Results for Log(SalePrice)

Generalized Linear Model Regression Results

Dep. Variable:	log(SalePrice)	No. Observations:	1168
Model:	GLM	Df Residuals:	937
Model Family:	Gaussian	Df Model:	230
Link Function:	log	Scale:	0.011015
Method:	IRLS	Log-Likelihood:	1091.2
Date:	Fri, 21 Mar 2025	Deviance:	10.321
Time:	17:16:37	Pearson chi2:	10.3
No. Iterations:	100	Pseudo R-squ. (CS):	1.000
Covariance Type:	nonrobust		

Figure 7. GLM output, log-transformed ‘SalePrice’ as target variable

We implemented a Generalized Linear Model (GLM) with a log link function, training it on 80% of the dataset (1,168 observations) and evaluating it on the remaining 20% (292 observations). The model's Root Mean Squared Error (RMSE) was 25,631.10, indicating that, on average, predictions deviated from actual house prices by this amount. Given the wide price range, this suggests some predictions were significantly off. Deviance (10.321) and Pearson Chi-Square (10.3) were low and closely aligned, suggesting that the variance structure was well captured. However, a Pseudo R-squared of 1.000 suggests an almost perfect fit, which may indicate overfitting.

Despite incorporating 230 predictors, the model’s relatively high RMSE (25.6K) suggests potential limitations in generalization. Some variables may lack predictive power, and multicollinearity or noise in the features could impact performance. The model's complexity, reflected in 937 residual degrees of freedom, suggests a risk of overfitting, which could reduce its effectiveness on unseen data. Further improvements could involve feature selection, regularization techniques, or alternative modeling approaches to enhance generalizability while maintaining interpretability.

The GLM model fitted on log(SalePrice) has a lower test RMSE compared to the Gradient Boosting model. However, as we will discuss later, the dataset exhibits strong non-linear relationships and contains many categorical variables, which require the creation of dummy variables. This results in a substantial increase in the number of predictors (230 in total).

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

Despite the slightly higher Test RMSE for Gradient Boosting, it remains the better option due to its ability to handle non-linear relationships and efficiently manage the large number of categorical variables.

Which features have the highest predictive power?

- We analyzed the strength and direction of the correlation values between numerical features and SalePrice (Table 8). Strong positive correlations exist between SalePrice and features such as OverallQual (material and finish quality), GrLivArea (above-ground living area), and GarageCars (garage size), indicating that better quality, larger homes, and bigger garages increase property values.

  Conversely, KitchenAbvGr (number of kitchens above ground), EnclosedPorch (square footage of enclosed porches), show negative correlations, suggesting that multiple kitchens and enclosed porches do not necessarily drive higher prices. Unlike OverallQual(material and finish quality), OverallCond (general upkeep 1-Poor to 10-Excellent) is negatively correlated with sale price, possibly because condition reflects maintenance rather than premium quality features. Older homes that are well-maintained might still be valued lower compared to newer homes with higher-end materials.

Top 3 positive correlations		Top 3 negative correlations	
OverallQual	0.790982	OverallCond	-0.077856
GrLivArea	0.708624	EnclosedPorch	-0.128578
GarageCars	0.640409	KitchenAbvGr	-0.135907

Table 8. Correlation of numerical variables with SalePrice

- PCA loadings revealed that OverallQual, GarageCars, GrLivArea, YearBuilt, and 1stFlrSF have strong positive correlations with the first five principal components, which together explain 42.2% of the variance in the data.

- Factor Analysis identified Neighbourhood as the most influential factor, showing the highest positive loading not only in the component explaining the largest variance (23%) but also across all 10 predictors, highlighting its strong impact on sale price.

- LASSO regularization ranked GrLivArea, OverallQual, and GarageCars as the top three most important features, reinforcing their significance as key predictors of sale price.

- The best model, **Gradient Boosting with original SalePrice using all features**, reveals the relative contribution of each variable in predicting house prices. Higher feature importance values signify a stronger influence, while lower values suggest a lesser impact. OverallQual (0.52) emerges as the most influential feature, indicating that the overall material and finish quality of a house has the greatest effect on its price, with higher quality ratings significantly boosting property value. GrLivArea (0.14), representing the above-ground living area in square feet, is the second most important factor, as larger homes tend to command higher prices. GarageCars (0.06), which denotes the number of garage spaces, plays a key role in house valuation, suggesting that buyers highly value sufficient parking and storage space (Refer Figure 8).

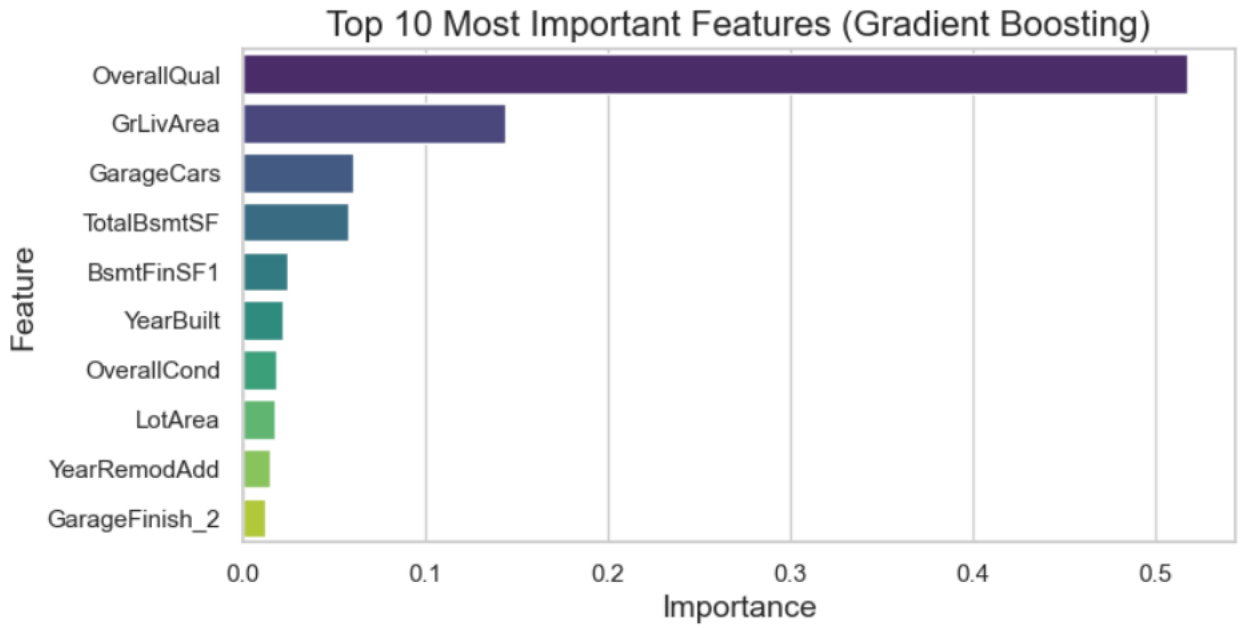

Figure 8. Feature Importance of Best Model (Gradient Boosting trained on original SalePrice using all featues)

Are there non-linear relationships between house prices and key features? Can advanced modeling techniques capture these relationships more effectively than linear models?

An analysis of the relationship between SalePrice and key features suggests the presence of both linear and non-linear patterns (Refer to Figure 9). Some variables, such as OverallQual, GrLivArea, GarageCars, and TotalBsmtSF, exhibit a positive, near-linear relationship with SalePrice, indicating that these features can be effectively modeled using simple or multiple linear regression. However, several features display non-linear relationships that cannot be adequately captured by a linear model. These include YearBuilt, LotArea, MasVnrArea, and BsmtFinSF1, which show curved trends and variations in their impact on SalePrice, suggesting the need for more flexible models. In such cases, relying solely on a linear approach could lead to inaccurate predictions. Additionally, variables like Fireplaces and 2ndFlrSF exhibit dispersed distributions and non-uniform trends, reinforcing the necessity for more advanced modeling techniques.Conversely, features such as KitchenAbvGr, BedroomAbvGr, MoSold, and YrSold do not show a strong or clear relationship with SalePrice, indicating that they might be removed from the model without significantly affecting predictive performance.

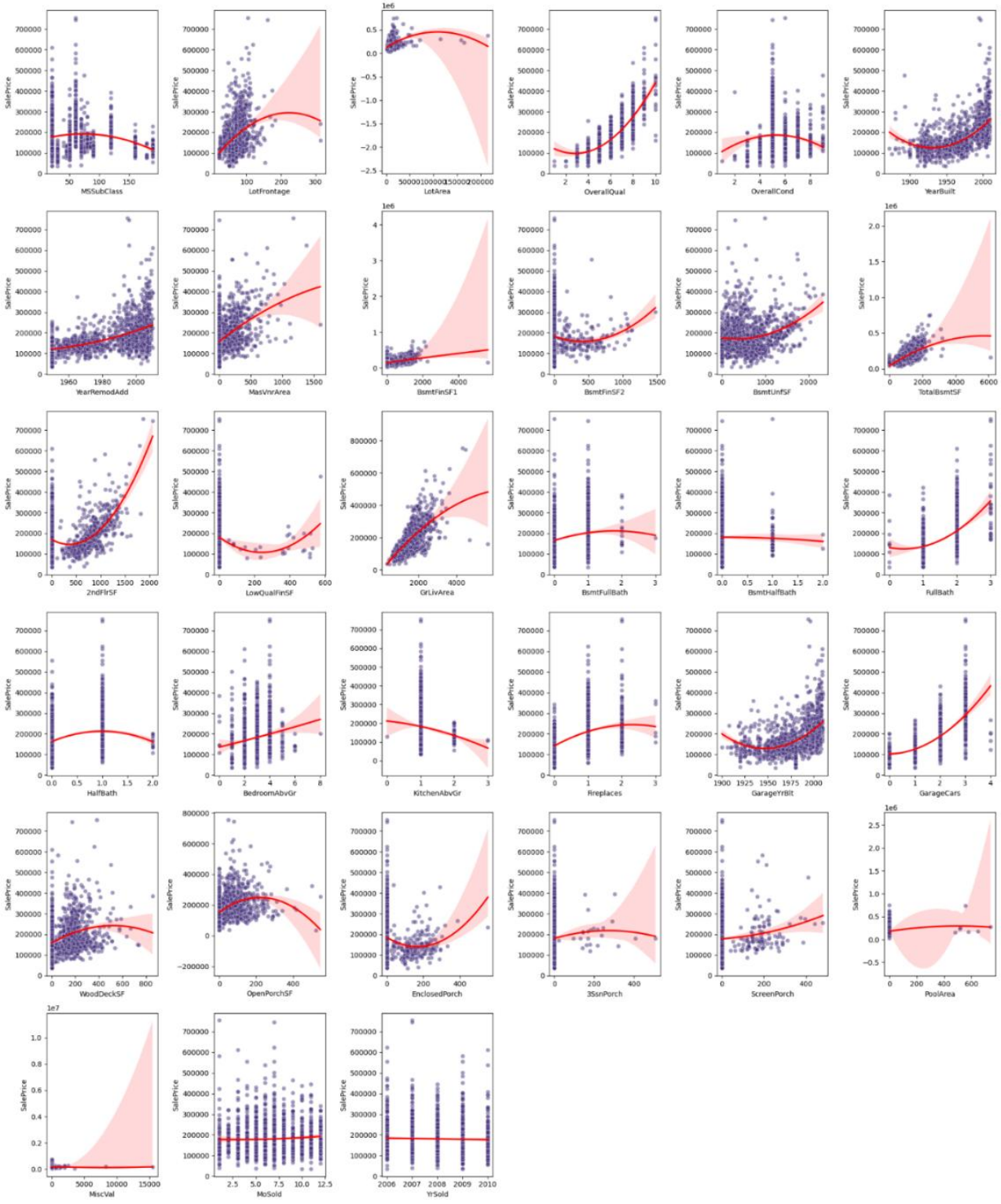

Figure 9. Variable relationships with SalePrice with Non-Linear Fit

Given the prevalence of non-linear relationships, advanced modeling techniques are likely to better capture these patterns compared to a simple linear regression model. To explore these non-linear relationships, we have already implemented Random Forests, Gradient Boosting, we will further evaluate models including Polynomial Regression. In this section.

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

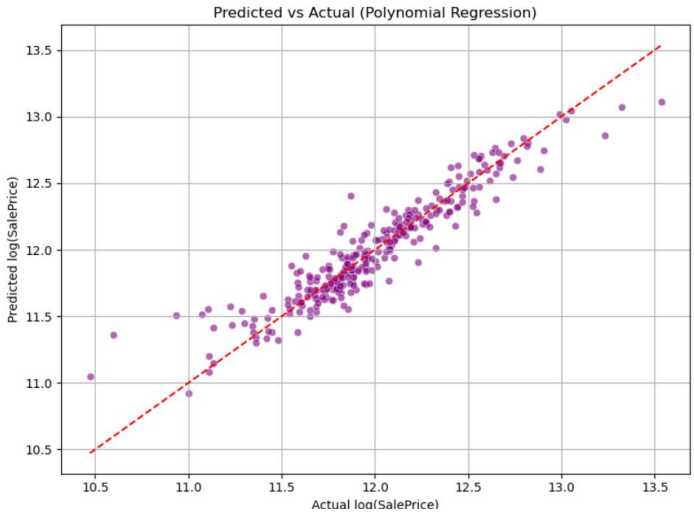

Figure 10: Predicted vs Actual (Polynomial Regression)

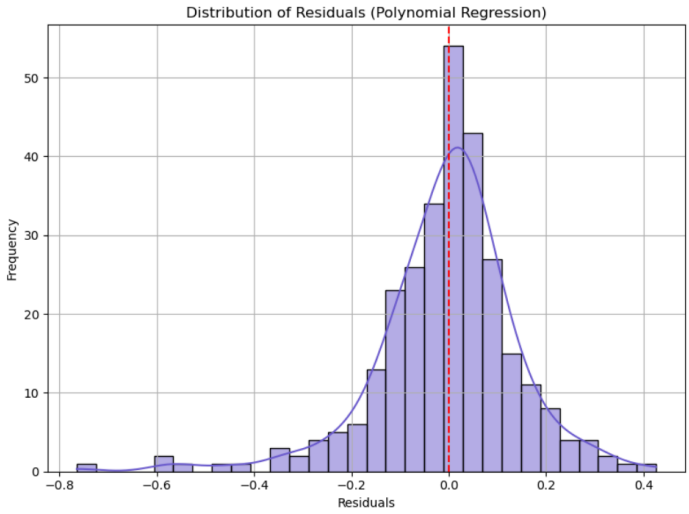

Figure 11: Distribution of Residuals (Polynomial Regression)

A second-degree polynomial regression model was implemented using the 10 most relevant variables previously selected using Lasso regression. This combination allows for capturing nonlinear relationships and interactions between characteristics without relying on highly complex models. The model achieved remarkable performance, with an R2 = 0.8860 and an RMSE of 0.1459 in predicting log(SalePrice), which represents a significant improvement over the traditional linear models. This demonstrates that by incorporating nonlinear and interaction terms, the model can capture more complex relationships between variables such as GrLivArea, OverAllQual, GarageCars, and YearBuilt.

However, the Test RMSE for Polynomial Regression using 10 top features from LASSO, is still larger than the Test RMSE for Gradient Boosting. (Refer to Tables 6 and 7). This indicates that Gradient Boosting has higher predictive accuracy as is more suited for modelling non-linear relationships. Taken together, these results confirm that substantive nonlinear relationships exist between house prices and key variables, and that advanced modeling techniques such as polynomial regression can capture these dynamics more effectively than simple linear models. Furthermore, the use of variable selection (such as Lasso) further boosts performance by focusing on the most informative features. These findings support the use of nonlinear and regularization techniques as recommended strategies for improving predictive accuracy in house price modeling.

RQ5: How do temporal factors influence house prices?

This study examines how construction year and last renovation date influence housing prices. A Mixed Effects Model is particularly well-suited for this analysis, as it accounts for both fixed effects (overall market trends) and random effects (variations across neighborhoods). This approach will help determine whether older houses systematically sell for lower prices and whether renovations mitigate depreciation effects. Additionally, the study explores whether appreciation rates vary by neighborhood, potentially revealing localized real estate booms or declines.

Figure 12 illustrates the trend in average sale prices over time based on Year Built and Year Remodeled. A general upward trend in sale prices is observed, particularly after 1960, with a sharp increase in recent years. While the Year Remodeled trend follows a similar trajectory, it remains lower than the Year Built trend, suggesting that remodeling increases house prices, but newly built homes tend to command higher prices. The fluctuations in sale prices before 1950 may reflect market instability or limited data availability.

Figure 13 highlights price variation across neighborhoods. Older houses (pre-1940) generally have lower sale prices, with some exceptions in specific neighborhoods. Certain neighborhoods (e.g., NridgHt, StoneBr, and NoRidge) exhibit higher sale prices, particularly for homes built after 2000. The wide spread of sale prices within each construction year suggests that location (neighborhood) plays a more significant role in pricing than construction year alone.

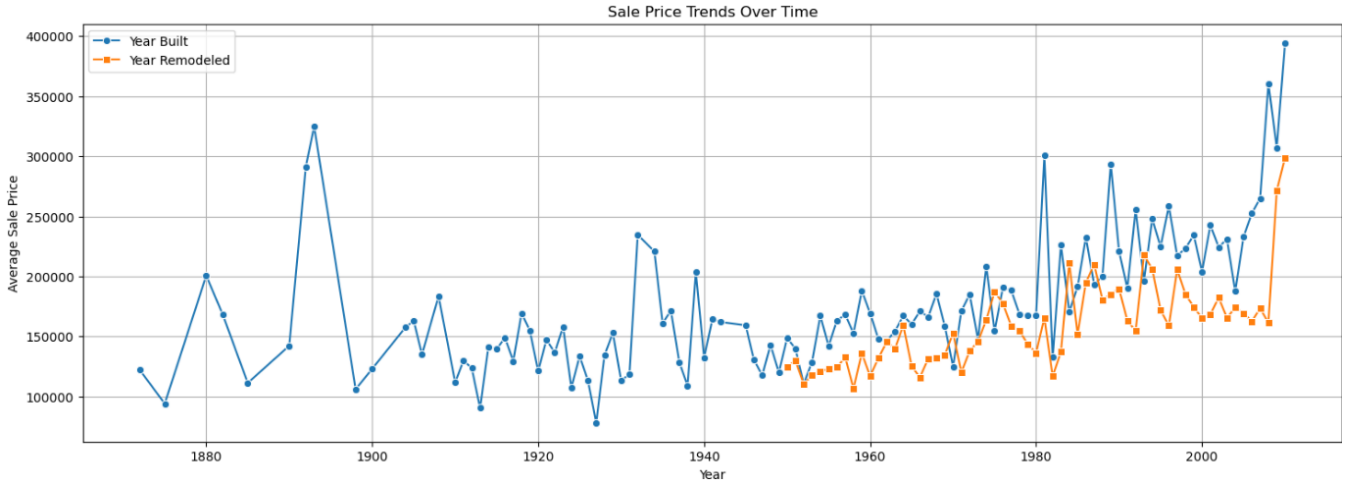

Figure 12. Sale Price Trends for Original construction date and Remodel date

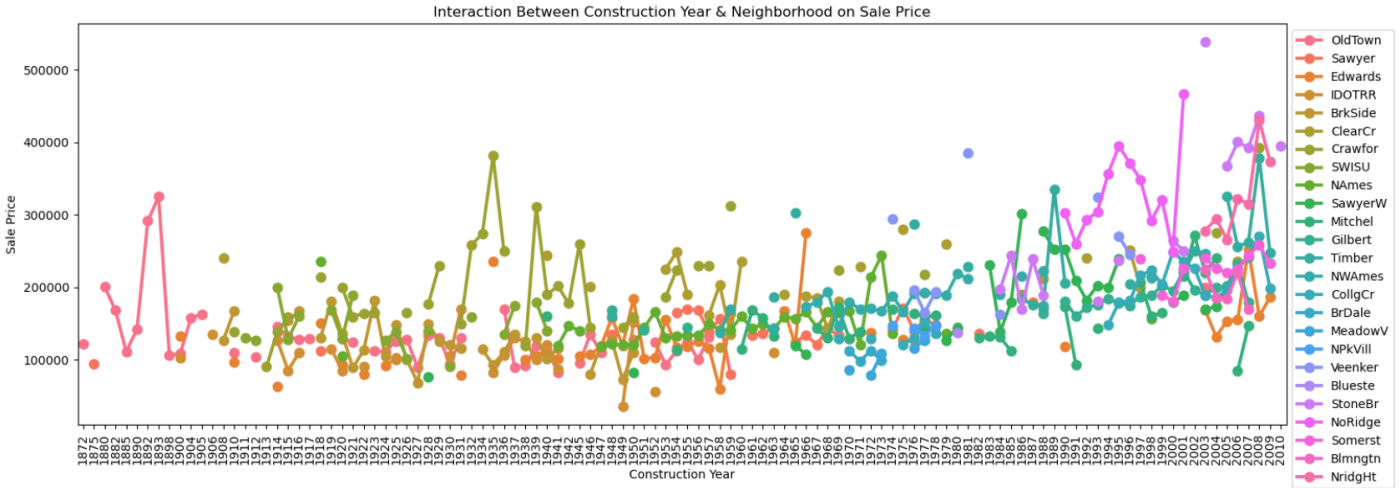

Figure 13. Interaction Between Construction Year & Neighborhood on Sale Price

```
            Mixed Linear Model Regression Results
============================================================
Model:            MixedLM Dependent Variable: Q('log(SalePrice)')
No. Observations: 1460    Method:             REML
No. Groups:       25      Scale:              0.0657
Min. group size:  2       Log-Likelihood:     -140.3464
Max. group size:  225     Converged:          Yes
Mean group size:  58.4
------------------------------------------------------------
             Coef.   Std.Err.      z      P>|z|    [0.025    0.975]
------------------------------------------------------------
Intercept    2.920      0.924    3.160    0.002     1.109     4.731
YearBuilt    0.005      0.000    9.880    0.000     0.004     0.006
Group Var    0.076      0.091
============================================================
```

Figure 14. Mixed Linear Model Regression Results

A mixed-effects regression model was fitted(refer to Figure 14), modeling log(SalePrice) as a function of year built (YearBuilt) as a fixed effect, and including Neighborhood as a random effect. This model captures both the overall trend in the time effect and differences in base price across neighborhoods.

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

The results indicate that the coefficient on YearBuilt is positive and highly significant (coef = 0.005, p < 0.001), implying that for each additional year since construction, the logarithmic value of the sale price increases by an average of 0.5%. This suggests that newer homes tend to sell for higher prices, which is consistent with the expectation that more recent constructions are in better condition or incorporate modern materials and standards.

Additionally, the model estimates a between-neighborhood variance of 0.076, indicating that there are substantial differences in base prices across neighborhoods, even after controlling for the year of construction. That is, the positive effect of building more recently remains, but the base price varies by area, justifying the use of a mixed-effects model instead of a simple linear regression.

Taken together, these results support the hypothesis in RQ5: temporal factors significantly influence housing prices, and this relationship is modulated by differences across neighborhoods. This finding is key to understanding how a property's value varies not only based on its age, but also on the spatial context in which it is located.

OLS Regression Results

Dep. Variable:	Q('log(SalePrice)')	R-squared:	0.417
Model:	OLS	Adj. R-squared:	0.416
Method:	Least Squares	F-statistic:	521.6
Date:	Fri, 21 Mar 2025	Prob (F-statistic):	1.43e-171
Time:	02:01:40	Log-Likelihood:	-337.16
No. Observations:	1460	AIC:	680.3
Df Residuals:	1457	BIC:	696.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-10.9770	0.772	-14.211	0.000	-12.492	-9.462
YearBuilt	0.0051	0.000	15.600	0.000	0.004	0.006
YearRemodAdd	0.0065	0.000	13.527	0.000	0.006	0.007

Omnibus:	70.480	Durbin-Watson:	1.905
Prob(Omnibus):	0.000	Jarque-Bera (JB):	142.128
Skew:	0.324	Prob(JB):	1.37e-31
Kurtosis:	4.384	Cond. No.	2.71e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.71e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 15. OLS Regression Results

A linear regression model with fixed effects was estimated(refer to Figure 15), where the dependent variable was log(SalePrice) and the independent variables were YearBuilt and YearRemodAdd. This model seeks to isolate the effect of home age and renovations on sales price, ignoring geographic variations (such as neighborhood).

Both predictors were found to be positive and highly significant:

- The coefficient on YearBuilt is 0.0051 (p < 0.001), indicating that for each more recent year of construction, the logarithmic price of the home increases by 0.51% on average.
- The coefficient on YearRemodAdd is 0.0065 (p < 0.001), suggesting that more recent renovations have an even larger effect: each additional year is associated with a 0.65% increase in log(SalePrice).

The model achieves an R2 = 0.417, implying that approximately 41.7% of the variability in home prices can be explained solely by these two temporal factors. Although this value is not extremely high, it does demonstrate that temporal factors have a substantial and systematic influence on home values.

These results clearly support the hypothesis posed in RQ5: temporal factors such as year of construction and year of remodeling have positive and statistically significant effects on home prices. Furthermore, the fact that year of remodeling has an even higher coefficient than year of construction suggests that improvements made over time may have a more direct impact on perceived property value than its original age.

6. Summary Results

1. Data Preprocessing Needs

Handling missing values, addressing multicollinearity, and applying transformations such as log transformation for SalePrice are essential steps before modeling. These preprocessing techniques improve model stability and predictive accuracy.

2. Feature Importance

The key features identified as the most significant predictors of house prices include OverallQual(Overall material and finish quality), GrLivArea(Above ground living area square feet), GarageCars(Size of garage in car capacity), and Neighborhood. These results were consistently corroborated by PCA, Factor Analysis, and LASSO regularization, underscoring their strong impact on housing prices. In the best model (Gradient Boosting with original SalePrice using all features), OverallQual, GrLivArea, GarageCars, TotalBsmtSF(Total square feet of basement area), and YearBuilt emerged as dominant factors. This suggests that overall quality and livable space are the primary determinants of house prices, while basement and garage space are more influential than lot size. Although renovations add value, newly built homes continue to command higher prices. Aesthetic features, such as fireplaces, contribute value but remain secondary to size and quality.

3. Feature Selection

LASSO regularization reduced the number of explanatory variables from 72 to 40, ensuring that only the most relevant features contributed to the model.

4. Statistical Distribution of House Prices

Analysis of the histogram, Q-Q plot, AIC, and BIC values confirms that the log-normal distribution provides the best fit for house prices.

5. Modeling Non-Linear Relationships

While certain variables, such as OverallQual and GrLivArea, exhibit near-linear relationships with SalePrice, others, including YearBuilt, LotArea, and BsmtFinSF1, show non-linear patterns. Gradient Boosting outperformed Polynomial Regression in accuracy.

6. Temporal Trends

Housing prices have shown an upward trend since 1960, with remodeling positively influencing prices, though newly built homes still command higher values. Neighborhood effects are significant, with a between-neighborhood variance of 0.076 in the mixed-effects model, confirming that base prices vary by location even when controlling for construction year. These results support the hypothesis that temporal and spatial factors significantly influence housing prices.

7. Best Model

The best overall model was **Gradient Boosting with original SalePrice and all features**. If interpretability is a priority, Linear Regression is an option despite lower accuracy. For computational efficiency, LASSO feature selection helps reduce model complexity with minimal loss in performance.

Group 11
Amali Jayatileke
Vidal Mendoza Tinoco

8. Statistical Assumptions and Their Violations

Despite confirming the log-normal fit of SalePrice and improved model performance with log transformation (reported for the GLM model), Shapiro-Wilk and Anderson-Darling tests indicate violations of lognormality assumptions. This suggests that while the transformation enhances model behavior, some statistical assumptions remain partially unmet.

9. Modeling Challenges and Shortcomings

One major challenge is interpretability, as many coefficients remain even after LASSO regularization. Non-linear relationships further complicate interpretation, with Gradient Boosting providing the best fit but being inherently difficult to explain. More time should be spent understanding latent variables and the underlying factors they represent.

10. Interesting Findings

Approximately 41.7% of the variability in home prices can be explained solely by YearBuilt and YearRemodAdd. While this is not an extremely high proportion, it demonstrates that temporal factors have a substantial and systematic impact on housing values. This insight underscores the importance of accounting for property age and renovation history in predictive modeling.

7. Conclusions and Recommendations

This study successfully developed a predictive model for house prices by integrating data preprocessing, feature selection, and advanced modeling techniques. Key predictors such as OverallQual, GrLivArea, GarageCars, and Neighborhood played a significant role in price estimation, with Gradient Boosting emerging as the most accurate model. Additionally, temporal trends and neighborhood effects were found to significantly influence pricing, highlighting the importance of location and property age in real estate valuation.

To enhance model performance, future work should explore additional property-related features such as crime rates, school proximity, and economic factors. Addressing interpretability challenges in non-linear models through SHAP values or Generalized Additive Models (GAMs) could provide deeper insights. SHAP values use a game-theoretic approach to explain individual predictions by distributing the model’s output among input features, making it possible to quantify each feature's positive or negative contribution. While GAMs may not always outperform Gradient Boosting in predictive accuracy, they offer greater interpretability by modeling relationships as smooth, additive functions rather than complex trees. Additionally, incorporating spatial autocorrelation techniques and time-series forecasting could further enhance the understanding of regional price variations and future market trends.

Overall, this study underscores the importance of robust preprocessing, flexible modeling techniques, and domain-specific insights in house price prediction. Future improvements in data quality, statistical refinements, and hybrid modeling approaches could further enhance predictive power and practical applicability in real estate analytics.

References

1. Dataset : https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Appendix

Data fields

- **SalePrice** - the property's sale price in dollars. This is the target variable
- **MSSubClass**: The building class
- **MSZoning**: The general zoning classification
- **LotFrontage**: Linear feet of street connected to property
- **LotArea**: Lot size in square feet
- **Street**: Type of road access
- **Alley**: Type of alley access
- **LotShape**: General shape of property
- **LandContour**: Flatness of the property
- **Utilities**: Type of utilities available
- **LotConfig**: Lot configuration
- **LandSlope**: Slope of property
- **Neighborhood**: Physical locations within Ames city limits
- **Condition1**: Proximity to main road or railroad
- **Condition2**: Proximity to main road or railroad (if a second is present)
- **BldgType**: Type of dwelling
- **HouseStyle**: Style of dwelling
- **OverallQual**: Overall material and finish quality
- **OverallCond**: Overall condition rating
- **YearBuilt**: Original construction date
- **YearRemodAdd**: Remodel date
- **RoofStyle**: Type of roof
- **RoofMatl**: Roof material
- **BsmtFinSF1**: Type 1 finished square feet
- **BsmtFinType2**: Quality of second finished area (if present)
- **BsmtFinSF2**: Type 2 finished square feet
- **BsmtUnfSF**: Unfinished square feet of basement area
- **TotalBsmtSF**: Total square feet of basement area
- **Heating**: Type of heating
- **HeatingQC**: Heating quality and condition
- **CentralAir**: Central air conditioning
- **Electrical**: Electrical system
- **1stFlrSF**: First Floor square feet
- **2ndFlrSF**: Second floor square feet
- **LowQualFinSF**: Low quality finished square feet (all floors)
- **GrLivArea**: Above grade (ground) living area square feet
- **BsmtFullBath**: Basement full bathrooms
- **BsmtHalfBath**: Basement half bathrooms
- **FullBath**: Full bathrooms above grade
- **HalfBath**: Half baths above grade
- **Bedroom**: Number of bedrooms above basement level
- **GarageType**: Garage location
- **GarageYrBlt**: Year garage was built
- **GarageFinish**: Interior finish of the garage
- **GarageCars**: Size of garage in car capacity

- **Exterior1st**: Exterior covering on house
- **Exterior2nd**: Exterior covering on house (if more than one material)
- **MasVnrType**: Masonry veneer type
- **MasVnrArea**: Masonry veneer area in square feet
- **ExterQual**: Exterior material quality
- **ExterCond**: Present condition of the material on the exterior
- **Foundation**: Type of foundation
- **Kitchen**: Number of kitchens
- **KitchenQual**: Kitchen quality
- **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
- **Functional**: Home functionality rating
- **Fireplaces**: Number of fireplaces
- **FireplaceQu**: Fireplace quality
- **BsmtQual**: Height of the basement
- **BsmtCond**: General condition of the basement
- **BsmtExposure**: Walkout or garden level basement walls
- **BsmtFinType1**: Quality of basement finished area
- **GarageArea**: Size of garage in square feet
- **GarageQual**: Garage quality
- **GarageCond**: Garage condition
- **PavedDrive**: Paved driveway
- **WoodDeckSF**: Wood deck area in square feet
- **OpenPorchSF**: Open porch area in square feet
- **EnclosedPorch**: Enclosed porch area in square feet
- **3SsnPorch**: Three season porch area in square feet
- **ScreenPorch**: Screen porch area in square feet
- **PoolArea**: Pool area in square feet
- **PoolQC**: Pool quality
- **Fence**: Fence quality
- **MiscFeature**: Miscellaneous feature not covered in other categories
- **MiscVal**: $Value of miscellaneous feature
- **MoSold**: Month Sold
- **YrSold**: Year Sold
- **SaleType**: Type of sale
- **SaleCondition**: Condition of sale

Data field information (.info())

Range Index: 2919 entries, 0 to 2918

Data columns (total 81 columns):

#	Column	Non-Null Count	Dtype
0	Id	2919 non-null	int64
1	MSSubClass	2919 non-null	int64
2	MSZoning	2915 non-null	object
3	LotFrontage	2433 non-null	float64

4 LotArea 2919 non-null int64

5 Street 2919 non-null object

6 Alley 198 non-null object

7 LotShape 2919 non-null object

8 LandContour 2919 non-null object

9 Utilities 2917 non-null object

10 LotConfig 2919 non-null object

11 LandSlope 2919 non-null object

12 Neighborhood 2919 non-null object

13 Condition1 2919 non-null object

14 Condition2 2919 non-null object

15 BldgType 2919 non-null object

16 HouseStyle 2919 non-null object

17 OverallQual 2919 non-null int64

18 OverallCond 2919 non-null int64

19 YearBuilt 2919 non-null int64

20 YearRemodAdd 2919 non-null int64

21 RoofStyle 2919 non-null object

22 RoofMatl 2919 non-null object

23 Exterior1st 2918 non-null object

24 Exterior2nd 2918 non-null object

25 MasVnrType 1153 non-null object

26 MasVnrArea 2896 non-null float64

27 ExterQual 2919 non-null object

28 ExterCond 2919 non-null object

29 Foundation 2919 non-null object

30 BsmtQual 2838 non-null object

31 BsmtCond 2837 non-null object

32 BsmtExposure 2837 non-null object

33 BsmtFinType1 2840 non-null object

34 BsmtFinSF1 2918 non-null float64

35 BsmtFinType2 2839 non-null object

36 BsmtFinSF2 2918 non-null float64

37 BsmtUnfSF 2918 non-null float64

38 TotalBsmtSF 2918 non-null float64

39 Heating 2919 non-null object

40 HeatingQC 2919 non-null object

41 CentralAir 2919 non-null object

42 Electrical 2918 non-null object

43 1stFlrSF 2919 non-null int64

44 2ndFlrSF 2919 non-null int64

45 LowQualFinSF 2919 non-null int64

46 GrLivArea 2919 non-null int64

47 BsmtFullBath 2917 non-null float64

48 BsmtHalfBath 2917 non-null float64

49 FullBath 2919 non-null int64

50 HalfBath 2919 non-null int64

51 BedroomAbvGr 2919 non-null int64

52 KitchenAbvGr 2919 non-null int64

53 KitchenQual 2918 non-null object

54 TotRmsAbvGrd 2919 non-null int64

55 Functional 2917 non-null object

56 Fireplaces 2919 non-null int64

57 FireplaceQu 1499 non-null object

58 GarageType 2762 non-null object

59 GarageYrBlt 2760 non-null float64

60 GarageFinish 2760 non-null object

61 GarageCars 2918 non-null float64

62 GarageArea 2918 non-null float64

63 GarageQual 2760 non-null object

64 GarageCond 2760 non-null object

65 PavedDrive 2919 non-null object

```
 66  WoodDeckSF     2919 non-null   int64
 67  OpenPorchSF    2919 non-null   int64
 68  EnclosedPorch  2919 non-null   int64
 69  3SsnPorch      2919 non-null   int64
 70  ScreenPorch    2919 non-null   int64
 71  PoolArea       2919 non-null   int64
 72  PoolQC         10 non-null     object
 73  Fence          571 non-null    object
 74  MiscFeature    105 non-null    object
 75  MiscVal        2919 non-null   int64
 76  MoSold         2919 non-null   int64
 77  YrSold         2919 non-null   int64
 78  SaleType       2918 non-null   object
 79  SaleCondition  2919 non-null   object
 80  SalePrice      1460 non-null   float64
dtypes: float64(12), int64(26), object(43)
```