UCLouvain

Linear Model Report
LSTAT2120

Giovanni Cavallo,
NOMA: 00172510
giovanni.cavallo@student.uclouvain.be

January 2, 2026

This report is part of the mandatory project for the class in Linear Models. The purpose of this is to report and explain the approach and results of the linear model project built on R on the dataset approved by the teacher's assistant, following the material taught during the class. The Rmd files coming with this report are supposed to work if the structure of the README.md file has been followed, but for an optimal visualization of the project, I suggest opening the .html files coming from the knitted markdowns.

Contents

1 Introduction

The project required finding a dataset that could provide meaningful insights on a specific topic through the analysis of a response variable (Y) and explanatory variables (X). Additionally, the dataset had to satisfy the following requirements: include at least 7 quantitative variables and 2 qualitative variables.

1.1 Dataset

The dataset I chose is the **AmesHousing** dataset, which focuses on housing prices in Ames, Iowa (USA).
I retrieved the dataset from **Kaggle** at this link.
The original dataset is composed of 2930 rows and 82 columns; however, for the purpose of this project, I selected a subset of the variables:

QUANTITATIVE VARIABLES:

- **GrLivArea**: Above-grade (ground) living area square footage.
- **TotalBsmtSF**: Total square footage of the basement area.
- **GarageArea**: Size of the garage in square feet.
- **LotArea**: Lot size in square feet.
- **OverallQual**: Rates the overall material and finish of the house (1 = Very Poor, 10 = Very Excellent).
- **YearBuilt**: Original construction date.
- **YearRemodAdd**: Remodel date (same as construction date if no remodeling or additions).

QUALITATIVE VARIABLES:

- **Neighborhood**: Physical locations within Ames city limits. In the US market, this is a strong proxy for school districts, distance to the university, and neighborhood prestige, which drastically affects value independent of the house itself.
- **HouseStyle**: Describes the vertical layout of the house. Common types include:
  - 1-Story: "Ranch" style (all ground level).
  - 2-Story: Bedrooms typically upstairs.
  - 1.5-Story: Older homes with converted attics (slanted ceilings).
  - Split-Level: Staggered floors with short flights of stairs.

The response variable chosen is **SalePrice**, which represents the price of the house in dollars.

2 Exploratory Data Analysis

The complete exploratory data analysis code and results can be found in the `exploratoryDataAnalysis.Rmd/html` file.

2.1 Missing Values

Only two features contained missing values, with exactly one missing observation each, as shown in Table 1. The missing values were located in rows 1342 and 2237. The response variable did not contain any missing values.

The approach adopted for these rows was to remove them from the dataset entirely.

Table 1: Missing Values Count by Variable

Variable	Missing Values
Gr.Liv.Area	0
Total.Bsmt.SF	1
Garage.Area	1
Lot.Area	0
Overall.Qual	0
Year.Built	0
Year.Remod.Add	0
Neighborhood	0
House.Style	0

2.2 Quantitative Variables

Table 2 presents the summary statistics for the quantitative variables. We can observe that some maximum values exceed the 3rd quartile by a significant margin, which is a potential sign of outliers. The same pattern is observed in the response variable (Table 3).

Table 2: Descriptive Statistics of Key Variables

Statistic	GrLivArea	TotalBsmtSF	GarageArea	LotArea	OverallQual	YearBuilt	YearRemodAdd
n_obs	2930	2929	2929	2930	2930	2930	2930
n_miss	0	1	1	0	0	0	0
mean	1499.69	1051.62	472.82	1 0147.92	6.09	1971.36	1984.27
sd	505.51	440.62	215.05	7880.02	1.41	30.25	20.86
min	334	0	0	1300	1	1872	1950
q25	1126.00	793.00	320.00	7440.25	5.00	1954.00	1965.00
median	1442.00	990.00	480.00	9436.50	6.00	1973.00	1993.00
q75	1742.75	1302.00	576.00	11555.25	7.00	2001.00	2004.00
max	5642	6110	1488	215245	10	2010	2010
skewness	1.27	1.16	0.24	12.81	0.19	-0.60	-0.45
kurtosis	7.13	12.12	3.95	267.57	3.05	2.50	1.66

2.2.1 Distributions

From Figure 1, it is possible to identify a few outliers in the extreme ranges. The distributions for the remaining quantitative variables can be found in the Appendix.

2.2.2 Linearity and Correlation

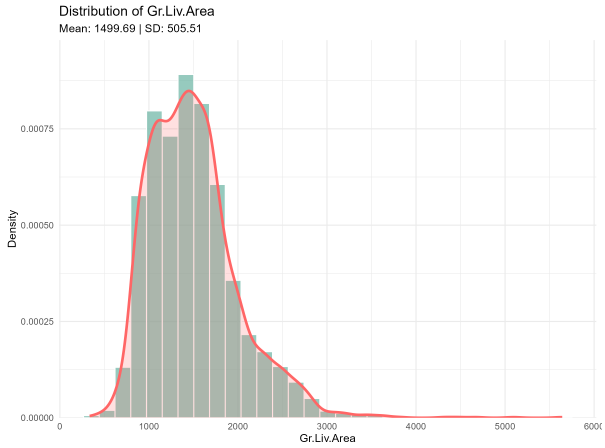
I checked for linearity by plotting the response variable against individual features and adding a smoothed regression line (using `ggplot`).
As Figure 2 illustrates, the response variable exhibits a strong linear relationship with some quantitative variables and a milder relationship with others.
The remaining linearity plots are included in the Appendix.

I also examined the correlation between quantitative variables, as shown in Figure 3.
"Overall Quality" and "Ground Living Area" appear to be the features most strongly correlated with the response variable.
Subsequent analysis regarding multicollinearity and autocorrelation will be presented later in the report.

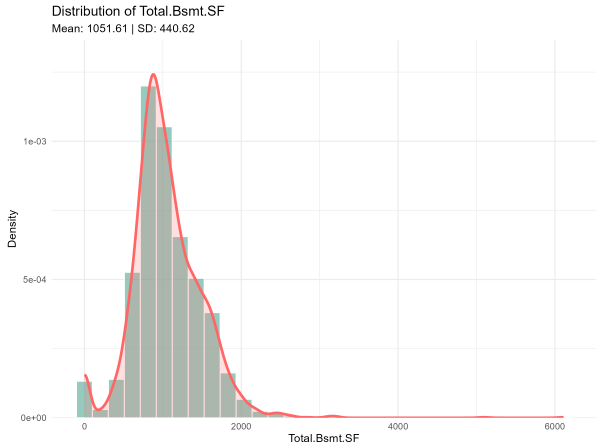
Table 3: Descriptive Statistics for SalePrice

Statistic	Value
n_obs	2,930
n_miss	0
mean	180,796.06
sd	79,886.69
min	12,789
q25	129,500
median	160,000
q75	213,500
max	755,000
skewness	1.74
kurtosis	8.11

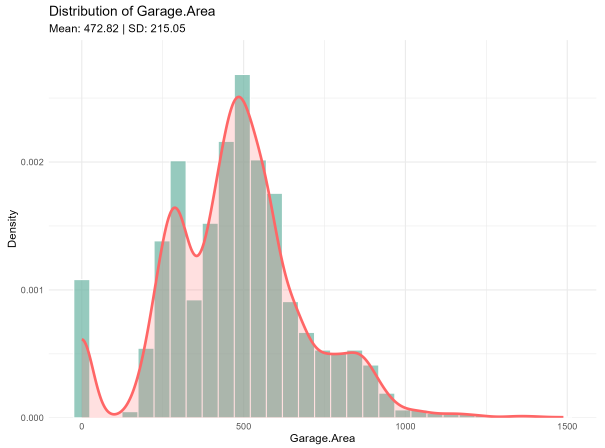
Figure 1: Histograms of the first 4 explanatory variables



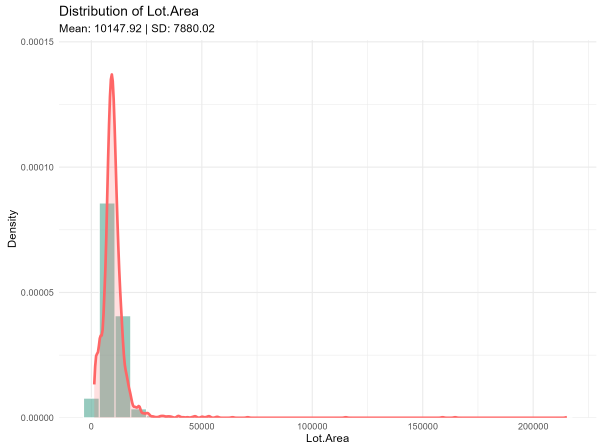
(a) Above Grade Living Area



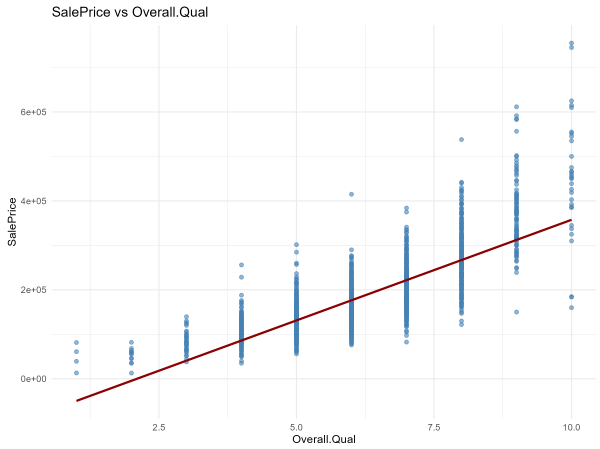
(b) Total Basement SF



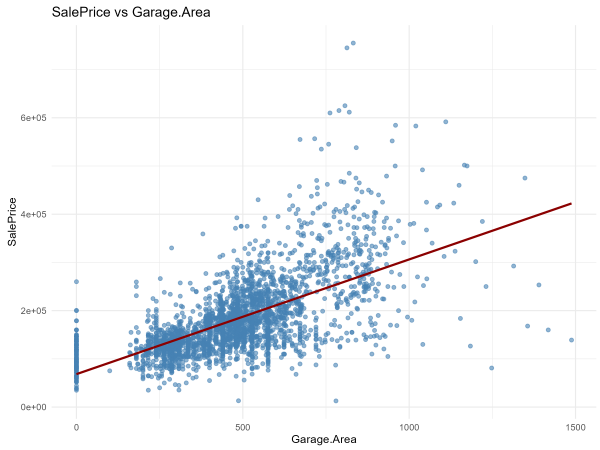
(c) Garage Area



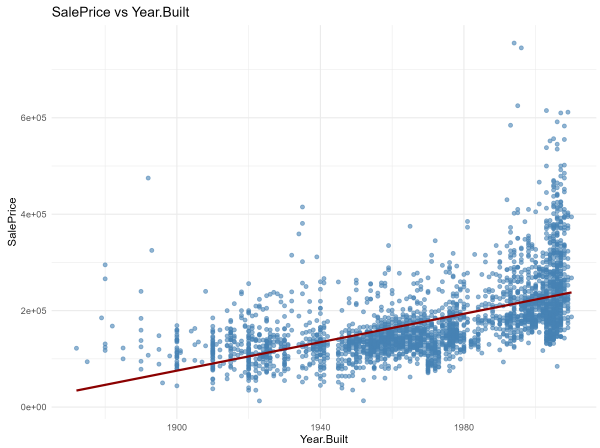
(d) Lot Area



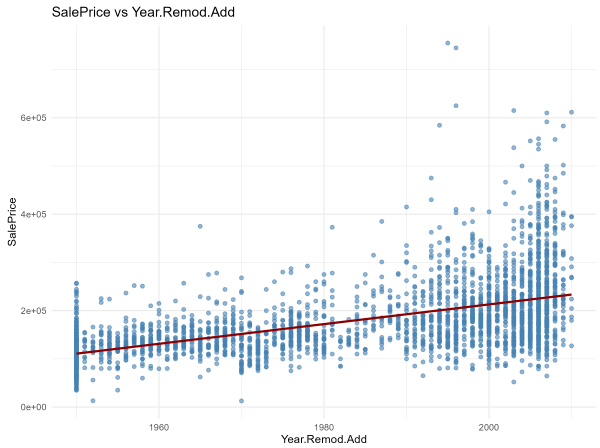
(a) Sale Price vs. Overall Quality



(b) Sale Price vs. Garage Area



(c) Sale Price vs. Year Built



(d) Sale Price vs. Year Remodeled

Figure 2: Scatter plots illustrating the linear relationship between Sale Price and key continuous predictors. The red line represents a linear regression fit.

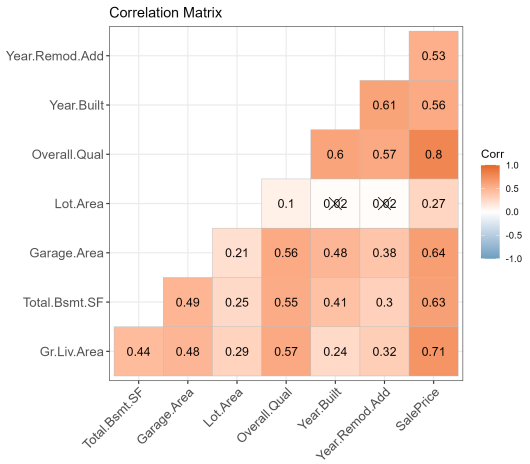
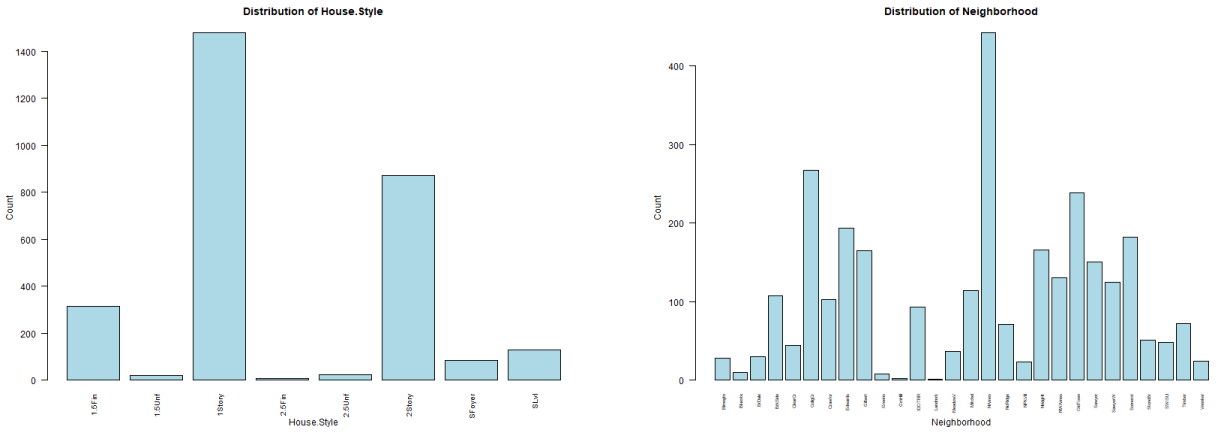


Figure 3: Correlation Matrix

2.3 Qualitative Variables

2.3.1 Distributions

I used bar plots (Figure 4) to visualize the distribution of levels for the two selected qualitative variables.

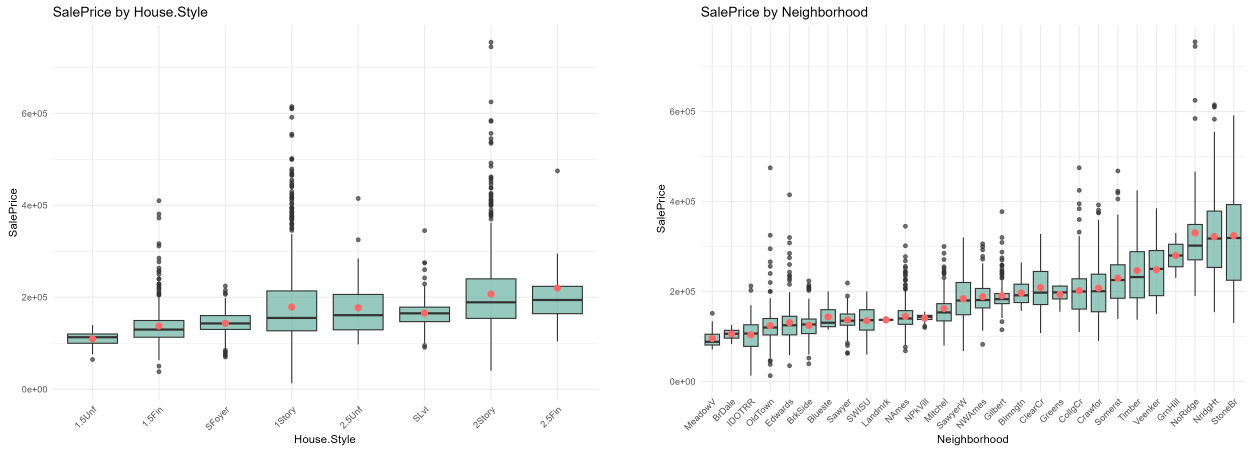


(a) House Style (b) Neighborhood

Figure 4: Bar Plots of Qualitative Variables

2.3.2 Linearity

To check for potential linear relationships with the categorical variables, I used box plots. I highlighted the mean (red circle) and ordered the levels based on their median values (Figure 5).



(a) House Style (b) Neighborhood

Figure 5: Box plots of Qualitative Variables

3 Model Evaluation

This section outlines the methodology used to prepare the data for modeling, including the train-test split strategy, multicollinearity checks, and feature selection techniques.

3.1 Split

Following the general machine learning approach and the project requirements, I split the dataset into distinct subsets.
The dataset was initially split into two portions: 80% was sampled randomly to form the **train** dataset, and the remaining 20% was subsequently split equally into 50% **validation** and 50% **test** sets.
The rationale behind these splits is to use the training dataset for fitting the models, the validation dataset to select the best performing models, and the test dataset to determine the final test RMSE (providing an unbiased estimate of performance).
All splits were performed on the original dataset after removing rows with missing values.

Table 4: The Dataset Split

Dataset	Rows	Columns
original_data	2928	10
test	293	10
train	2343	10
validation	292	10

3.2 Multicollinearity

The first step I took to assess potential issues with the linear model was to check for multicollinearity.
Based on the correlation matrix in Figure 3, the following variables showed potential multicollinearity due to their higher correlation (> 0.5) with other features:

- Year.Built × Year.Remod.Add
- Overall.Qual × Year.Built
- Garage.Area × Overall.Qual
- Gr.Liv.Area × Overall.Qual
- Overall.Qual × Year.Remod.Add
- Total.Bsmt.SF × Overall.Qual

Multicollinearity occurs when two or more predictors are highly correlated, making it difficult to isolate their individual effects. This can lead to unreliable coefficient estimates. The standard approach is often to remove one of the correlated variables.
To investigate this possibility, I calculated the Variance Inflation Factors (VIF).
I trained a linear model using only those 6 variables and calculated their VIF values.
The results are shown in Table 5.
The highest VIF observed was 2.67. Since no variable exceeded the standard thresholds (usually 5 or 10), I ruled out significant multicollinearity among these variables.

```{r}
vif_reg <- lm(SalePrice ~ Year.Built +
  Year.Remod.Add + Overall.Qual +
  Garage.Area + Gr.Liv.Area +
  Total.Bsmt.SF , train)
vif(vif_reg)
```

Table 5: VIF of the analyzed 6 variables

	Year.Built	Year.Remod.Add	Overall.Qual	Garage.Area	Gr.Liv.Area	Total.Bsmt.SF
VIF	2.08	1.82	2.67	1.77	1.73	1.59

3.3 One-Hot Encoding

Throughout the project, I employed two different approaches for handling categorical variables, resulting in two distinct models that I evaluated and trained in parallel.
In both cases, I used one-hot encoding, which involves creating binary variables (0 or 1) indicating the absence or presence of a specific category level for each observation.
To avoid perfect multicollinearity (the dummy variable trap), the encoding was performed for all levels except one, which served as the reference level for the intercept.
While many R packages automatically handle one-hot encoding, I performed this transformation manually to maintain better control over the variables and the encoding process.
The two approaches consisted of:

- **Full:** Encoding all levels (minus the reference) of each categorical variable, resulting in a high number of features.
- **Alternative:** Reducing the cardinality of categorical variables by grouping levels, then encoding these new, synthesized levels.

For the **Alternative** approach, the `Neighborhood` variable was simplified by grouping neighborhoods into price ranges based on their median values.
A similar strategy was applied to the `House.Style` variable, where styles with similar characteristics and median `SalePrice` values were merged into broader categories.
Table 6 illustrates the differences in variables between the two models.

Table 6: Number of Features in the 2 Different Models

Model	Obs	Variables
Full	2343	41
Alternative	2343	14

3.4 Full Model

The rationale for retaining all categorical levels (and thus increasing the number of features) is to preserve potentially useful information and improve model precision.
However, the primary drawbacks are the increased risk of overfitting (modeling noise rather than the signal) and the higher computational cost required for training and evaluation.

3.4.1 Coefficients and Confidence Interval

The full model contains 41 variables. Testing the significance of this many coefficients individually increases the risk of false positives (finding significance purely by chance).
To mitigate this multiple comparisons problem, I adjusted the significance threshold using the Bonferroni correction (further details are available in the `modelAnalysis.Rmd/html` file).
If the confidence interval of a coefficient contains 0, the coefficient is considered not significant.
Under these strict criteria, 26 variables were deemed not significant; notably, all of them were levels of the two categorical variables.

3.4.2 Hypothesis Testing

To verify whether the coefficients identified as insignificant were truly zero, I performed a hypothesis test.
I tested the null hypothesis that this subset of coefficients was jointly equal to zero. As shown in Table 7, this hypothesis was rejected because the RSS (Residual Sum of Squares) was significantly lower in the full model compared to the reduced model.
Therefore, we cannot assume that these coefficients are effectively zero, as they collectively contribute to the model's predictive power.

Table 7: Hypothesis Testing for Full Model

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2327	2.578e+12				
2301	2.510e+12	26	6.8022e+10	2.3984	9.415e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3.4.3 Feature Selection

Although individual coefficients appeared insignificant, the hypothesis test proved they are important for reducing the model's error (RSS).
Standard model selection criteria (e.g., AIC) or stepwise selection methods were computationally expensive given the high dimensionality of this model.
I reserved those techniques for the **Alternative** model (which has fewer variables). For the Full Model, I opted for regularization (Lasso) to perform feature selection.
The Lasso regression indicated that the optimal model retained 35 features, as shown in Figure 6, excluding 6 features out of the original 41.
The rejected features were all specific levels of the categorical variables.

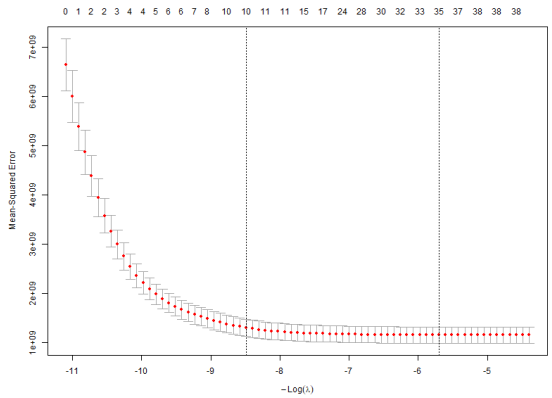


Figure 6: Lasso for the Full Model

Table 8: Coefficients Reduced to 0 in the Full Lasso Model

Feature	Coefficient
Neighborhood.Sawyer	.
Neighborhood.Greens	.
Neighborhood.CollgCr	.
House.Style.SLvl	.
House.Style.1.5Unf	.
House.Style.2.5Fin	.

3.5 Alternative Model

The Alternative Model contains fewer features than the Full Model due to the reduced number of levels in the categorical variables. The rationale behind this approach is threefold:

1. To reduce the number of variables analyzed during feature selection.
2. To prevent the model from using meaningless features (noise), thereby simplifying inference regarding variable impact.
3. To address the issue of low-frequency levels in the data (as observed in Figure 4).

3.5.1 Coefficients and Confidence Interval

The Alternative Model consists of 14 variables. Testing the significance of each coefficient individually still carries the risk of false positives (Type I errors).
Therefore, I applied the Bonferroni correction to lower the significance threshold (detailed results are available in the `modelAnalysis.Rmd/html` file and Supplementary Table S1).
Under this correction, `House.Style.SFoyer`, `House.Style.Split`, and `Neighborhood_Price_Range.Low` were deemed not significant. Notably, all three are levels derived from categorical variables.

3.5.2 Hypothesis Testing

I formulated two different hypotheses to test these specific coefficients.
The first hypothesis tested whether the coefficient for `Neighborhood_Price_Range.Low` was equal to zero.
As shown in Table 9, this hypothesis was rejected because the RSS (Residual Sum of Squares) was significantly lower in the model with the non-zero coefficient.
The second hypothesis tested whether the coefficients for `House.Style.SFoyer` and `House.Style.Split` were equal, given that they belong to the same parent categorical variable.
This hypothesis was also rejected. Consequently, neither simplifying assumption was supported by the data.

Table 9: Linear Hypothesis Test Results

Hypothesis	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
$Neighborhood\_Price\_Range.Low = 0$	2329	2.6332e+12				
	2328	2.6286e+12	1	4.5202e+09	4.0032	0.04553 *
$House.Style.SFoyer = House.Style.Split$	2329	2.6361e+12				
	2328	2.6286e+12	1	7.421e+09	6.5723	0.01042 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3.5.3 Feature Selection

I employed multiple approaches for feature selection on this model, all of which led to the same conclusion: retain all coefficients.
The first approach used model selection criteria (e.g., AIC, BIC), which apply a penalty for every added feature to the RSS. This tests whether the addition of a new feature leads to a substantial reduction in error.
The results are displayed in Figure 7.

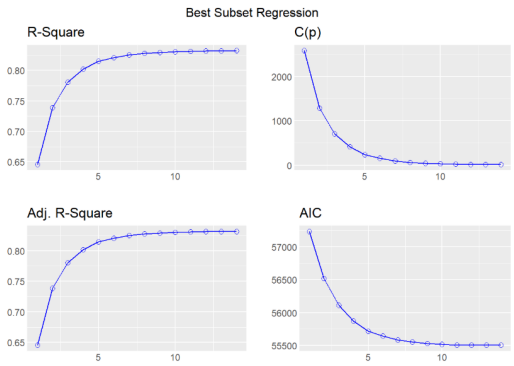


Figure 7: Plots of the different model selection criteria

The second type of feature selection involved stepwise selection methods.
Both Forward and Backward selection converged on the same result: retaining all features yields a significantly better model.
Table 10 presents the results from the forward selection process.

3.6 Normality

One of the primary objectives of this analysis was to verify if the normality assumption required for linear models was satisfied.
A common method to check this is by examining the distribution of residuals using a QQ-plot.
Both models yielded similar results; for simplicity, I present only the alternative model here (results for both models are available in the `modelAnalysis.Rmd/html` file).

Table 10: Forward Stepwise Summary

Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Base Model	59662.675	59674.194	53010.485	0.00000	0.00000
1	Overall.Qual	57233.719	57250.997	50581.928	0.64567	0.64552
2	Gr.Liv.Area	56520.508	56543.544	49868.986	0.73889	0.73866
3	Neighborhood_Price_Range.Very_High	56114.507	56143.302	49463.404	0.78062	0.78033
4	Total.Bsmt.SF	55874.262	55908.817	49223.597	0.80216	0.80183
5	Year.Built	55719.620	55759.934	49069.395	0.81496	0.81456
6	Lot.Area	55642.791	55688.865	48992.835	0.82108	0.82062
7	Garage.Area	55584.895	55636.728	48935.219	0.82560	0.82507
8	Year.Remod.Add	55552.242	55609.834	48902.761	0.82816	0.82757
9	Neighborhood_Price_Range.High	55533.811	55597.162	48884.465	0.82965	0.82899
10	House.Style.2Story	55518.839	55587.949	48869.632	0.83088	0.83015
11	House.Style.1.5	55510.102	55584.971	48860.999	0.83165	0.83086
12	Neighborhood_Price_Range.Low	55508.029	55588.658	48858.975	0.83194	0.83108
13	House.Style.Split	55506.258	55592.645	48857.256	0.83221	0.83128
14	House.Style.SFoyer	55505.279	55597.426	48856.326	0.83243	0.83142

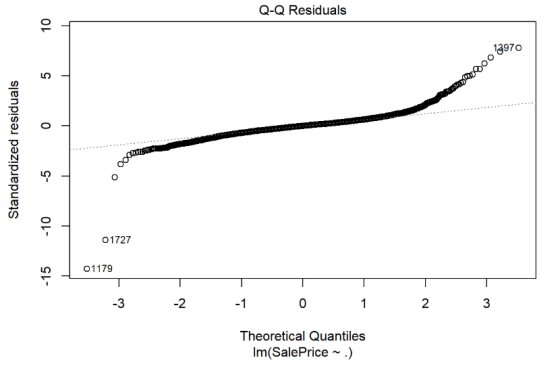


Figure 8: QQ-plot of Residuals for the Alternative Model

The QQ-plot in Figure 8 does not follow the theoretical line expected if the normality assumption were satisfied.
To verify this finding more rigorously, I applied the Shapiro-Wilk test (results in Table 11). The test rejected the null hypothesis, thereby confirming that the residuals do not follow a normal distribution.

Table 11: Shapiro-Wilk Normality Test Results

Test	**W**	**p-value**
Shapiro-Wilk normality test	0.82938	$< 2.2e-16$

A possible reason for this deviation is the presence of outliers, particularly visible as extreme points in Figure 8. Another common method to address extreme values and skewness is to apply a logarithmic transformation to the response variable.
Although a detailed outlier analysis is presented later, here I illustrate the improvement in residual distribution when using a logarithmic scale, and when using a logarithmic scale after removing outliers (Figure 9). Despite these adjustments, the normality test was still rejected. However, given the large sample size, we can rely on the Central Limit Theorem (CLT), which suggests that the linear model remains a valid approximation even if residual normality is not perfectly met.

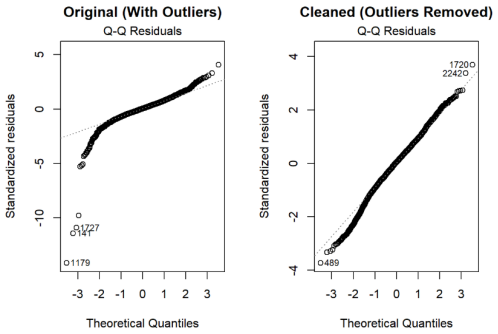


Figure 9: QQ-plot of Residuals After Log Transformation

3.7 Heteroskedasticity and Autocorrelation

Checking for heteroskedasticity, like normality, can be performed both visually and through formal statistical tests.
Plotting standardized residuals against fitted values allows us to assess if the variance of the residuals changes across different predicted values.
Figure 10 illustrates this behavior for the alternative model, which is further confirmed by White's Test (Table **??**).
Additionally, the Breusch-Godfrey Test confirmed the presence of autocorrelation.

Table 12: Residual Diagnostic Tests

statistic	**p.value**	**parameter**	**method**
1949	0	119	White's Test
51.608	6.776e-13	1	Breusch-Godfrey test

Due to the presence of heteroskedasticity and autocorrelation, the standard confidence intervals for the coefficients are likely unreliable (often presenting a too-optimistic scenario).
To address this, I calculated robust standard errors to verify if the significance of the coefficients changed.
The results for the alternative model are presented in Table 13. Using a p-value threshold of 0.05, no changes in significance were observed after applying robust standard errors.

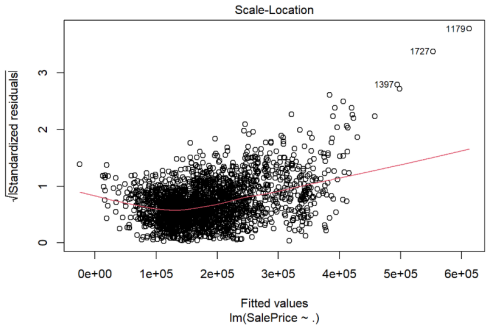


Figure 10: Plot for Heteroskedasticity

Table 13: Significance of Regression Coefficients

Variable	**Significance**
(Intercept)	***
Gr.Liv.Area	***
Total.Bsmt.SF	.
Garage.Area	***
Lot.Area	***
Overall.Qual	***
Year.Built	**
Year.Remod.Add	***
House.Style.2Story	**
House.Style.1.5	**
House.Style.SFoyer	**
House.Style.Split	.
Neighborhood_Price_Range.High	***
Neighborhood_Price_Range.Very_High	***
Neighborhood_Price_Range.Low	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

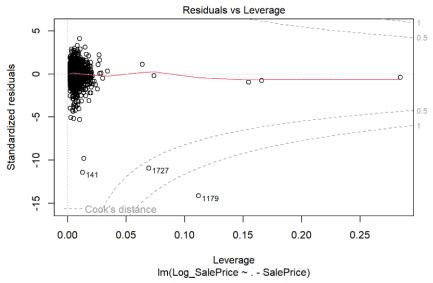


Figure 11: Plot Showing Cook's Distance for Alternative Model

3.8 Outliers and Influential Observations

Cook's distance was used to detect potential outliers and influential observations within the training dataset. Figure 11 displays the data points and their leverage, highlighting influential observations based on their Cook's distance values.
Points identified as influential based on Cook's distance were subsequently removed from the training dataset to prevent them from disproportionately affecting the model.

4 Predictions

Finally, I evaluated three different models with varying feature sets to determine which would perform best on the validation dataset.
The RMSE (Root Mean Squared Error) for the training predictions is shown in Table 14. Based solely on these training results, the Full Model appeared to be the best predictive model. However, as shown in Table 15, the best predictive model for the validation dataset proved to be the **Alternative Model**.
This discrepancy was expected. The Full Model contains many non-significant features that artificially lower the error on the training set (overfitting) but fail to generalize to new data.
Conversely, the Alternative Model contains fewer features, but a higher proportion of them are statistically significant.
This parsimony reduces the risk of overfitting, resulting in better performance on data outside the training set.
The Lasso model followed a similar pattern: it performed slightly better than the Full Model on the validation set, but still worse than the Alternative Model.

Table 14: RMSE - Training Dataset

Model	**RMSE**	**Features**
Full	32730	41
Alternative	33495	14
Lasso	36055	35

Table 15: RMSE - Validation Dataset

Model	**RMSE**	**Features**
Full	24243	41
Alternative	23836	14
Lasso	24154	35

Prior to generating predictions, the validation dataset was cleaned of outliers using boxplot statistics. Based on these results, the Alternative Model (with its reduced feature set) was selected as the optimal model.

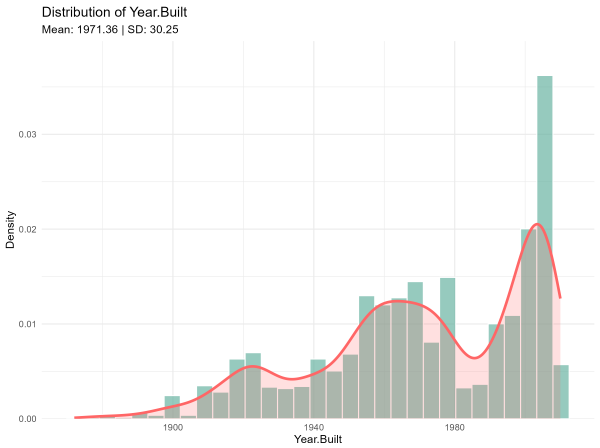
Therefore, I applied this model to the final Test dataset. The resulting Test RMSE serves as the most unbiased estimate of the model's performance on unseen data.
The final Test RMSE of the alternative model is **20190.78**.

5 Conclusions

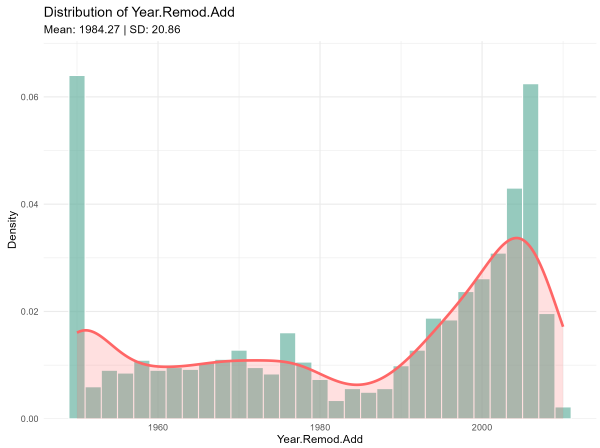
In this project, I assessed the assumptions of the linear model, specifically testing for the **Normal Distribution** of residuals, **Heteroskedasticity**, **Autocorrelation**, and **Multicollinearity** among predictors. Additionally, I conducted a comprehensive exploratory data analysis and identified influential **Outliers**.
Various linear models were tested and evaluated using multiple **feature selection** strategies.
The analysis demonstrated the distinct advantages of splitting the dataset into **Train**, **Validation**, and **Test** sets.
The observation that the Validation and Test RMSE were lower than the Training RMSE warrants further discussion. Possible causes include data leakage or simply the smaller size of the evaluation sets.
Although strict measures were taken to avoid data leakage, the smaller sample sizes of the validation and test datasets (compared to the training set) could naturally lead to a lower overall RMSE.
Given that the range and absolute values of the response variable are quite broad, even a few discrepancies between fitted and actual values can lead to a significant increase in RMSE, due to the way it is calculated.

6 Appendix

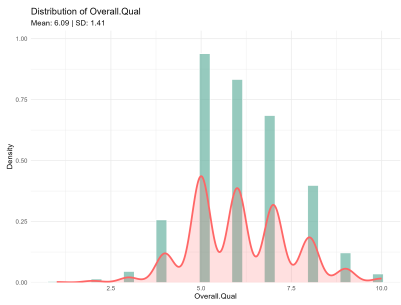
Supplementary Figure S1: Remaining Histograms for the Quantitative Variables



(a) Year Built

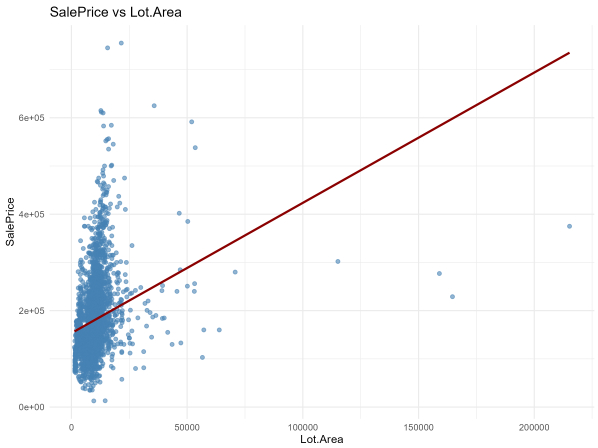


(b) Year Remodeled



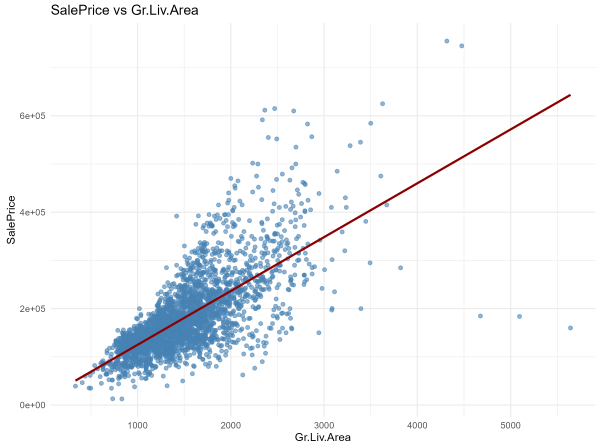
(c) Overall Quality

Supplementary Figure S2: Remaining Scatter Plots illustrating the linear relationships



(a) SalePrice vs Lot Area

(b) SalePrice vs Total Basement Surface



(c) SalePrice vs Ground Living Area

Supplementary Table S1: Confidence Interval Bonferroni Corrected for the Coefficients of the Alternative Model

Variable	**0.333 %**	**99.667 %**
(Intercept)	-1.085342e+06	-5.446712e+05
Gr.Liv.Area	4.674499e+01	5.983261e+01
Total.Bsmt.SF	1.039342e+01	2.431627e+01
Garage.Area	2.046920e+01	4.408744e+01
Lot.Area	3.962470e-01	8.857735e-01
Overall.Qual	1.479565e+04	1.939575e+04
Year.Built	3.834558e+01	2.559933e+02
Year.Remod.Add	1.154546e+02	3.662296e+02
House.Style.2Story	-1.853517e+04	-5.331765e+03
House.Style.1.5	-1.595192e+04	-1.036676e+03
House.Style.SFoyer	-4.231537e+03	1.887411e+04
House.Style.Split	-1.504056e+04	3.106891e+03
Neighborhood_Price_Range.High	4.058815e+03	1.576063e+04
Neighborhood_Price_Range.Very_High	5.472280e+04	7.279134e+04
Neighborhood_Price_Range.Low	-1.025181e+04	1.553512e+03