STAT 306 2024W2 - Project: Report

Group Project Report

Table of Contents

Introduction

Housing is one of the most fundamental elements of our lives, and the volatility in housing prices directly affects many people[1] (UN, n.d.). As a result, housing prices have become a topic of significant interest, not only as an economic indicator but also in social and personal contexts. The goal of this project was to build a linear regression model to find out the relationship [SB1] house sale prices using the Ames Housing dataset. The dataset includes 2,930 observations and 82 variables, covering aspects such as house size, location, and various other factors in relation to **SalePrice**. Given the large number of variables, we narrowed down the covariates through data cleaning and feature selection to make the dataset manageable. Additionally, we addressed potential violations of regression assumptions by applying transformations to ensure the model's validity.

During the model-building process, we examined linear relationships and applied variable transformations to address issues like non-constant variance and heavy-tailed residual distribution. We also used forward selection to choose the most appropriate covariates. After selecting the final model, we performed diagnostic checks and considered the impact of outliers and high leverage points. Finally, we explored the effects of interaction terms between categorical and numerical variables.

//TODO [SB1]Choose at most 3 explanatory variables & their relationship with response variable (salesprice)

Dataset Overview & Data Cleaning

For this project, we used the Ames Housing dataset, which includes 2,930 observations and 82 variables describing residential properties in Ames, Iowa. The variables cover a wide range of features—like the house's size, year built, number of rooms, and some ratings on quality and condition. There are both numeric variables (e.g., **Gr.Liv.Area, Garage.Area, Lot.Area, Year.Built**) and categorical ones (e.g., **Lot.Shape, Electrical, Central.Air**).To make the dataset more manageable and suitable for modeling, we first selected 18 candidate variables based on interpretability and data completeness. Then we applied a few more filtering steps:

Since 82 variables are too many for a linear regression model, we narrowed it down in two steps. First, we picked 18 variables that seemed interpretable and didn't have too many missing values. Then we filtered them further using the following criteria:

- Missing data: Variables like **Pool.Area, Fence**, and **Misc.Feature** had too many *NAs* or were rarely used, so we removed them early.
- Low correlation: Variables that didn't show a meaningful relationship with **SalePrice** (correlation less than 0.3 in absolute value) were excluded.
- Redundancy: When two variables meant basically the same thing (like **Year.Built** and **Year.Remod.Add**), we kept only one.
- Too many levels: Categorical variables like **Neighborhood** or **Bldg.Type** had too many categories, which could lead to overfitting, so we dropped them.
- Subjectivity: Some rating-based variables were too subjective to be reliable.

After these steps, we finalized a list of 14 variables: 13 covariates and **SalePrice** as the response. Here's a summary of the main ones used:

- **SalePrice**: House sale price (divided by 1,000 for scale)
- **Overall.Qual**: Overall quality rating (1–10)
- **Gr.Liv.Area**: Above-ground living area (sqft)
- **Garage.Area**: Garage area (sqft)
- **Year.Built**: Construction year
- **Lot.Area**: Lot size (sqft)
- **Bedroom.AbvGr**: Number of bedrooms above ground
- **Kitchen.AbvGr**: Number of kitchens above ground
- **Lot.Shape**: Shape of the lot (Reg, IR1, IR2, IR3)[SB1]
- **Utilities, Electrical, Central.Air, Yr.Sold:** Other relevant categorical/numeric info

//TODO [SB1]More description for IR1…

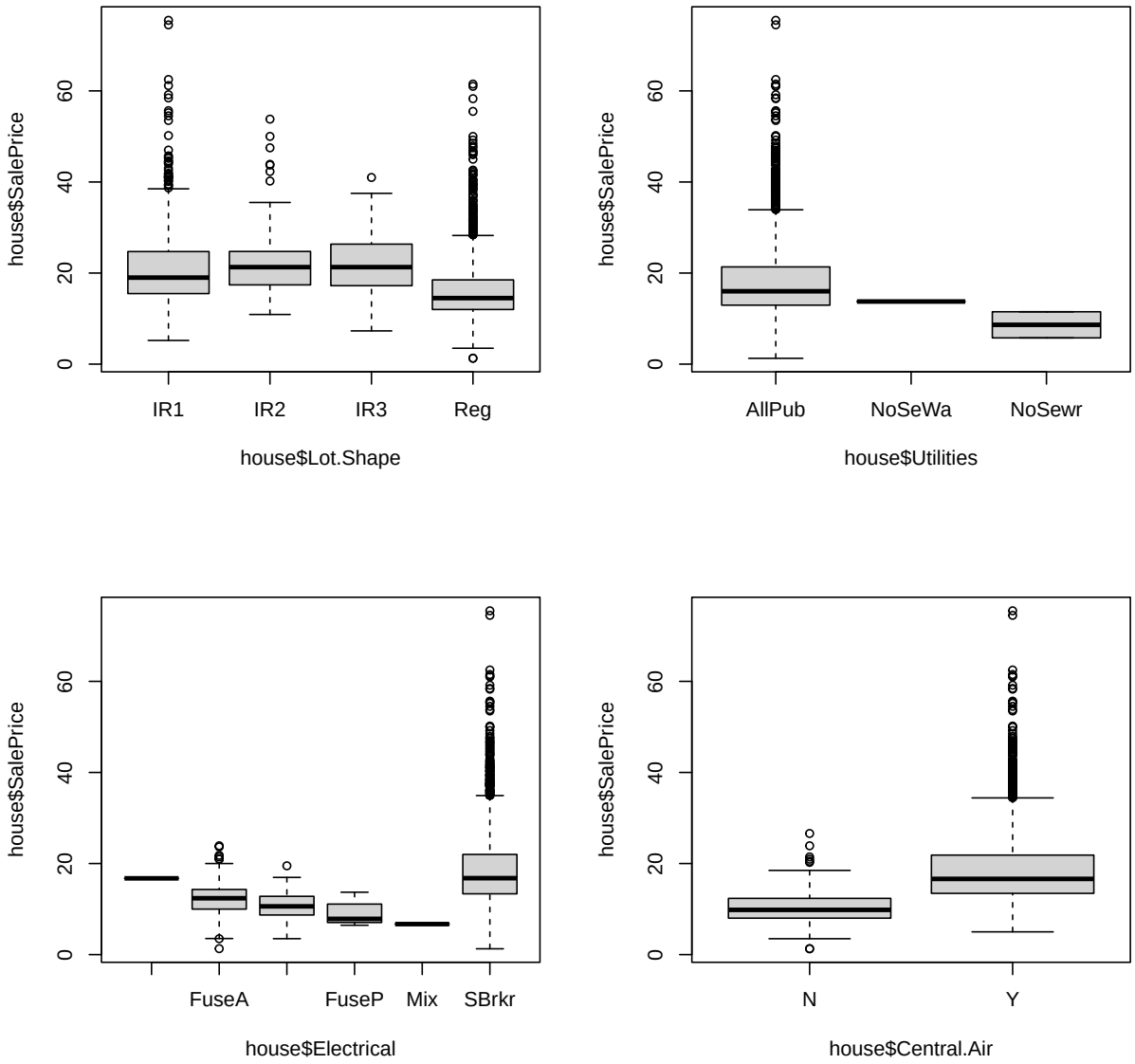
As for data cleaning, we didn't need to do any restructuring since the dataset came in clean CSV format. But we did some basic prep work: for instance, we converted character-based categorical variables like **Lot.Shape** into factor type using as.factor(). We also checked for missing values using `colSums(is.na())`, but since we already removed problematic variables, our final dataset was clean.

Finally, we adjusted a few column types—like turning **Year.Built** and **Bedroom.AbvGr** into integers—to make the modeling process easier. We also ran `summary()` to explore the variable distributions and noticed some skewed distributions and outliers, especially in **SalePrice**, which we handled later with log transformation.

Explorative Data Analysis

Before modeling[SB1] , we visualized the relationship between **SalePrice** and some categorical variables using boxplots. For example, houses with **Central Air** tended to have higher sale prices than those without it, and homes with **SBrkr** electrical systems also showed higher average prices compared to other types.

//TODO [SB1]Discuss later whether we mention we need log trans



Right skew show on the plot is a signal a do log(Y) transformation + assumption check also

Model Diagnostic

To check if our linear regression model satisfied the required assumptions, we examined two key diagnostic plots: the residuals vs fitted values plot and the Q-Q plot.

In the residuals vs fitted values plot, we observed a funnel shape, where the spread of residuals increased as the fitted values grew, indicating a violation of the homoscedasticity assumption (the assumption that the variance of errors is constant). Additionally, in the Q-Q plot, the residuals deviated from the theoretical reference line, especially at the tails, suggesting that the normality assumption might not hold.

These issues imply that our model could have non-constant variance and heavy-tailed residual distribution[SB1] , which can affect the validity of statistical inferences. To address this, we decided to apply a log transformation to the **SalePrice** variable.

//TODO [SB1]Too difficult term for reader -> easy way to describe -> expect more data points..

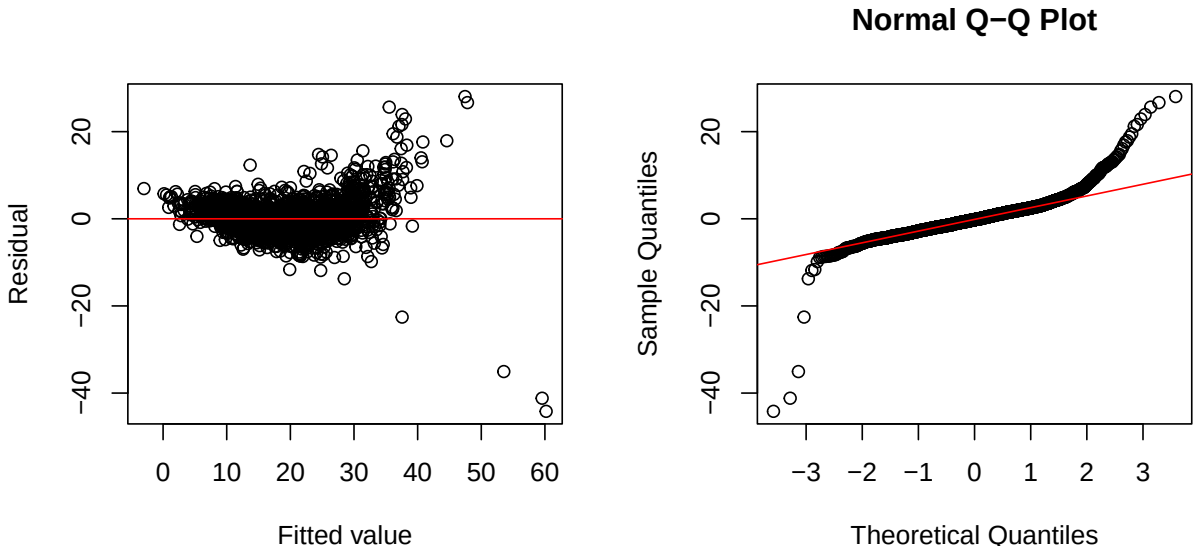
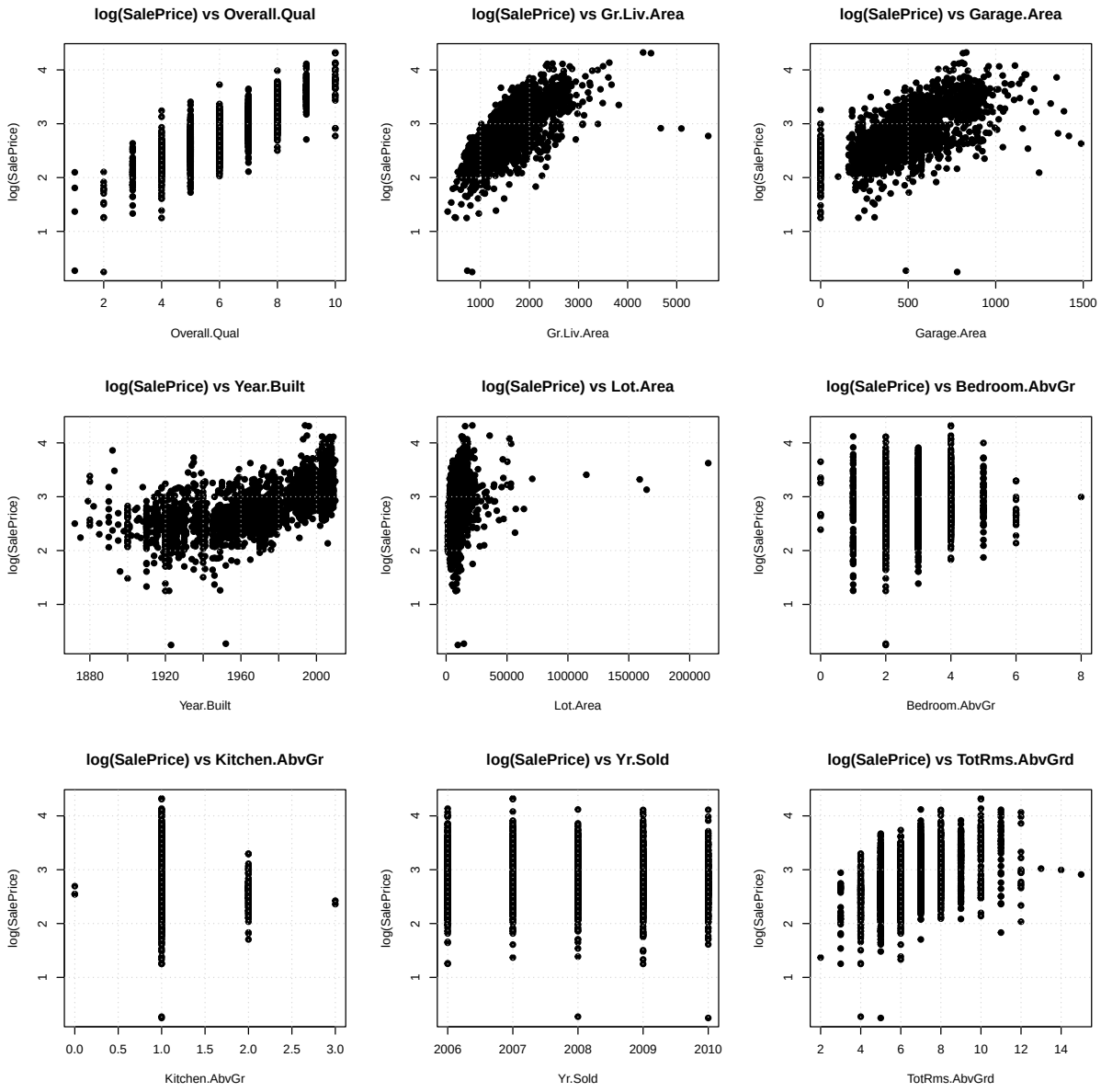


Fig.2: Residuals vs Fitted plot (left) and Normal Q-Q plot (right) from the initial model

Log Transformation & Model Rebuilding

To address the issues we found in the diagnostic plots—like the funnel-shaped residuals and the non-normal distribution—we applied a log transformation to the response variable, **SalePrice**. This step was taken to stabilize variance and make the error distribution closer to normal, both of which are important for linear regression.

After applying the transformation, we examined the scatterplots between **log(SalePrice)** and each continuous parameter. As shown in the figure below, the relationships with **Overall.Qual**, and **Gr.Liv.Area** appear more linear after the log transformation, which suggests the model is now better suited for linear regression.



However, the variable **Year.Built** still showed some curvature, even after the transformation. To capture this non-linear pattern[SB1] [SB2] [SB3] [SB4] , we included a quadratic term for `Year.Built` in the model:

// TODO [SB1]+ avoid multicollien.

// TODO [SB2]+ formula : why we follow this formula as mathematic way?

-> many explain..? -> just mention muliti colli….

// TODO [SB3]Correct formula

// TODO[SB4]Attach latex

$$(X_{Year.Built} - \bar{X_{Year.Built}})^2$$

After refitting the model using `log(SalePrice)` as the response variable, the residuals appeared more randomly scattered, and the Q-Q plot aligned more closely with the theoretical line. These improvements suggest that the key regression assumptions are now better satisfied, making the model more suitable for inference.

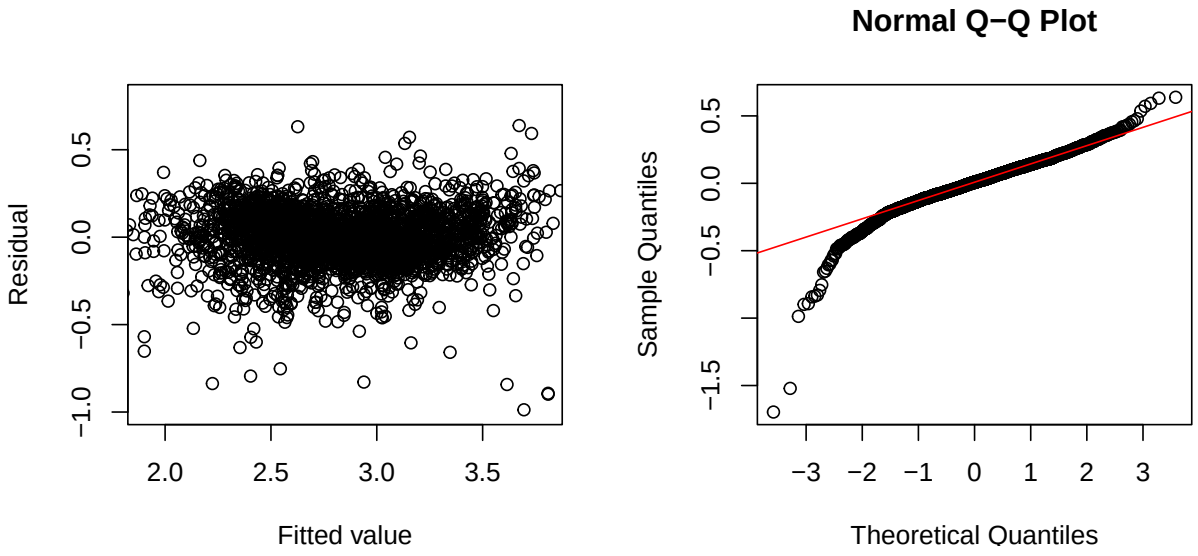


Fig.4: Model Diagnostics After log(SalePrice) Transformation: Residuals vs Fitted and Q-Q Plot

Multicollinearity check

Before finalizing the model, we checked for multicollinearity among the predictors using the Variance Inflation Factor (VIF). This helps to identify if any of the explanatory variables are highly correlated with each other, which can distort the interpretation of the regression coefficients. We used the commonly accepted threshold of 10, and all VIF values were well below that level. The highest was around 5.4 for `Gr.Liv.Area`, which is still considered acceptable. This suggests that multicollinearity is not a serious concern in our model, and the predictors are sufficiently independent to proceed with regression analysis.

```
          Overall.Qual            Gr.Liv.Area            Garage.Area
              2.889419               5.419218               1.758177
            Year.Built               Lot.Area           Bedroom.AbvGr
              2.768125               1.149620               2.160246
          Kitchen.AbvGr               Yr.Sold           TotRms.AbvGrd
              1.285590               1.004751               4.695560
 Quatratic_Gr.Liv.Area  Quatratic_Year.Built  Quatratic_Garage.Area
              2.218409               1.778837               1.154445
Quatratic_TotRms.AbvGrd
              1.831163
```

Fig 5: VIF Values for Explanatory Variables

Model Selection

To select the final set of explanatory variables, we used forward selection. Starting with an empty model, we added variables one by one based on how much they improved the model fit. Although the original dataset contained over 80 variables, we narrowed it down to 14 after filtering and diagnostic checks[SB1] [SB2] .

To determine the best stopping point, we considered both Mallows' Cp and AIC values. The Cp plot showed that the score stabilized around 14 variables, with little improvement beyond that. AIC also reached its lowest point with 14 predictors.

Following the principle of parsimony, we decided to go with the simplest model, selecting 14 variables as the final choice. [SB3]

//TODO [SB1]We include the quad term

//TODO [SB2](includes quadratic terms

//TODO [SB3]Change the plot's x axis title Covariate? Predictor? what else

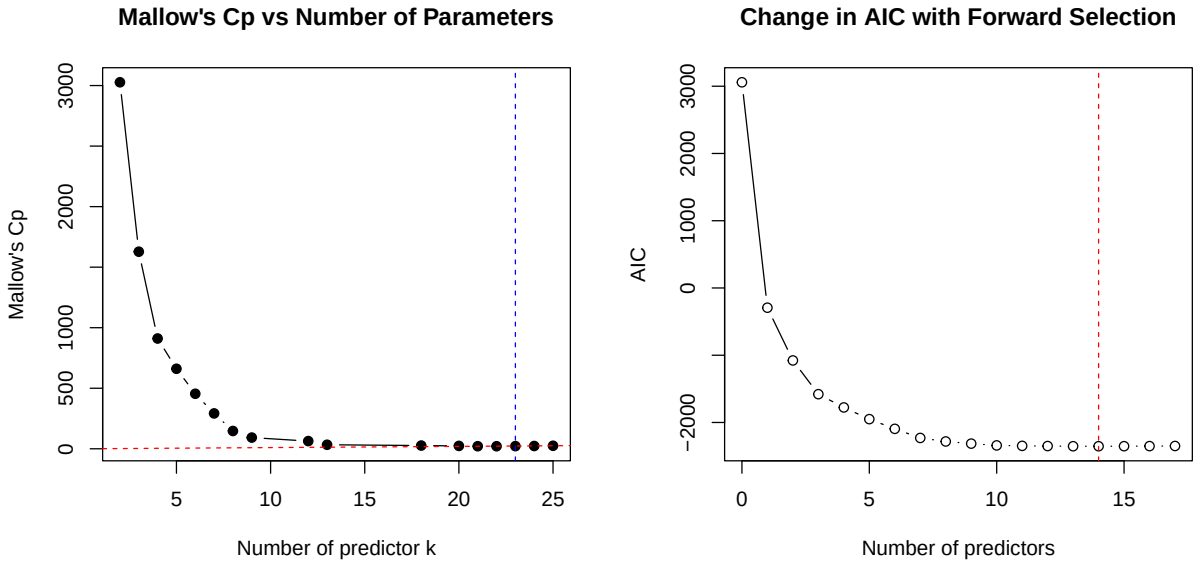


Fig 6: Model Selection using Mallows' Cp and AIC

Final Model Diagnostics

After fitting the final model, we revisited the regression assumptions by checking the diagnostic plots. The residuals vs fitted values plot showed a relatively random spread, which is a good indication that the homoscedasticity assumption (constant variance of errors) holds. The Q-Q plot showed that the residuals mostly followed a normal distribution, with only minor deviations at the tails.

The scale-location plot did not show any clear trends, suggesting that the variance of the residuals is consistent across fitted values. The residuals vs leverage plot did not reveal any influential data points that could excessively impact the model, and there were no values of Cook's distance greater than 1. Cook's distance is a measure of how much a data point influences the model, and values greater than

1 indicate a significant impact. Since no values exceeded 1, this indicates that there are no outliers significantly affecting the model.

These diagnostic results suggest that the final model meets the assumptions required for linear regression, making it suitable for interpretation and further analysis.

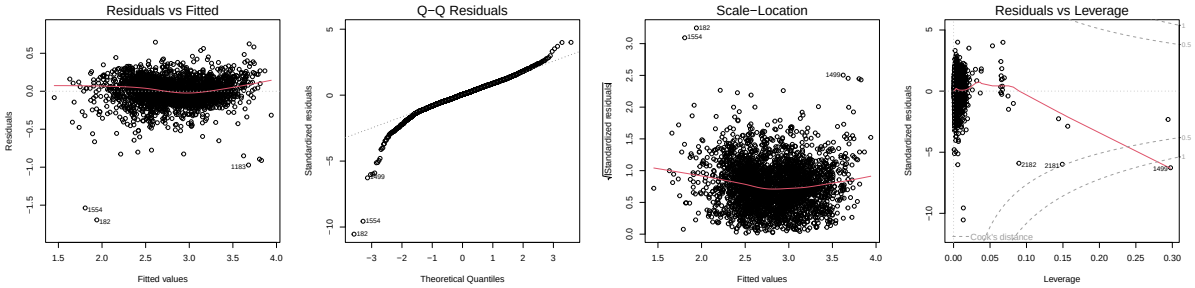


Fig. 7: Final Model Diagnostics, Assumption Check, Scale-Location and Residuals vs Leverage Plots for Final Model

Interaction Term

We explored all possible interactions between categorical variables (such as `Central.Air`, `Lot.Shape`, `Electrical`, `Utilities`) and numeric variables. Although the model's fitness increased slightly when these interactions were included, we decided to select only the significant interaction terms for the final model, named `simpler_interactive_model`.

Here is a summary of the model comparison:

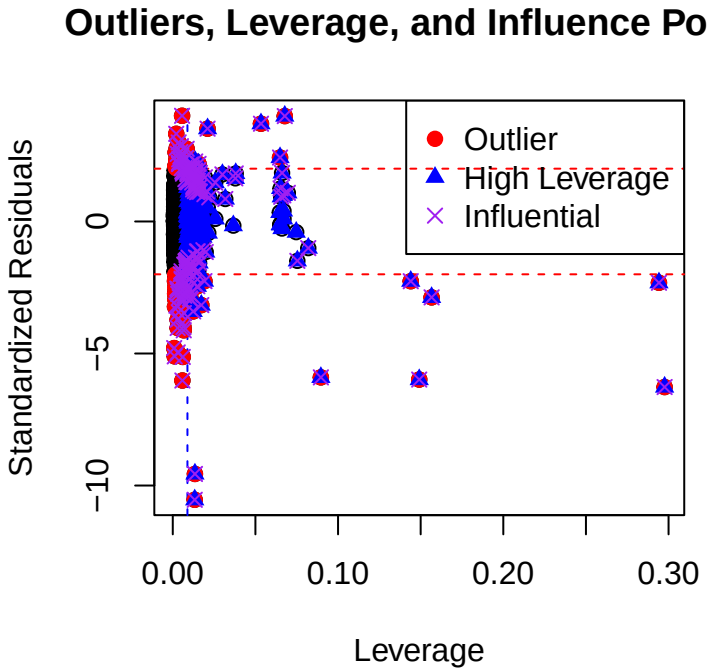
	number of parameter (p)	adj R^2
additive_model	25	0.8434
simpler_additive_model	22	0.8434
final_model	13	0.8423
interactive_model	39	0.8531
simpler_interactive_model	15	0.8443

The simpler interactive model showed an adjusted R^2 close to the interactive model but with fewer parameters, making it a more efficient choice.

The fit of the model with interactions improved slightly, but for the sake of model simplicity and clarity, we decided to include only the most significant interaction terms in the final model. A complete list of these interaction terms can be found in the appendix (page 13).

Outliers, Leverage, Influence (Limitation)

After building our model, we examined the potential outliers, leverage points, and influential data points. This step helps in assessing the reliability of the model and identifying any data points that might distort the results. The plot above highlights the outliers, points with high leverage, and influential data points in our dataset.



- **Outliers** (marked as red circles) represent data points that lie far from the rest of the data. These can significantly affect the model's coefficients and predictions.
- **High leverage points** (indicated by blue triangles) are points that have extreme values for the independent variables. While not necessarily outliers in terms of the response variable, they can still influence the model strongly, especially if they are far from the center of the data.
- **Influential points** (shown as purple crosses) are points that have both high leverage and residuals that significantly deviate from the model's predictions. These points can have a major effect on the regression coefficients and, therefore, on the overall model results.

In our case, we observed that there are a significant number of outliers and high leverage points in the data, which may pose limitations to the model. The presence of these points suggests that the model may be sensitive to these extreme values, which could lead to less reliable predictions and affect the generalizability of the results. These outliers and influential points need to be carefully addressed, either by transforming the data or removing problematic points, to ensure the robustness of the model.
+ confounding variables (+refer)[SB1]

//TODO [SB1]Nega -> confouding. And add some reference to descirbe

Discussion and Conclusion [SB1]

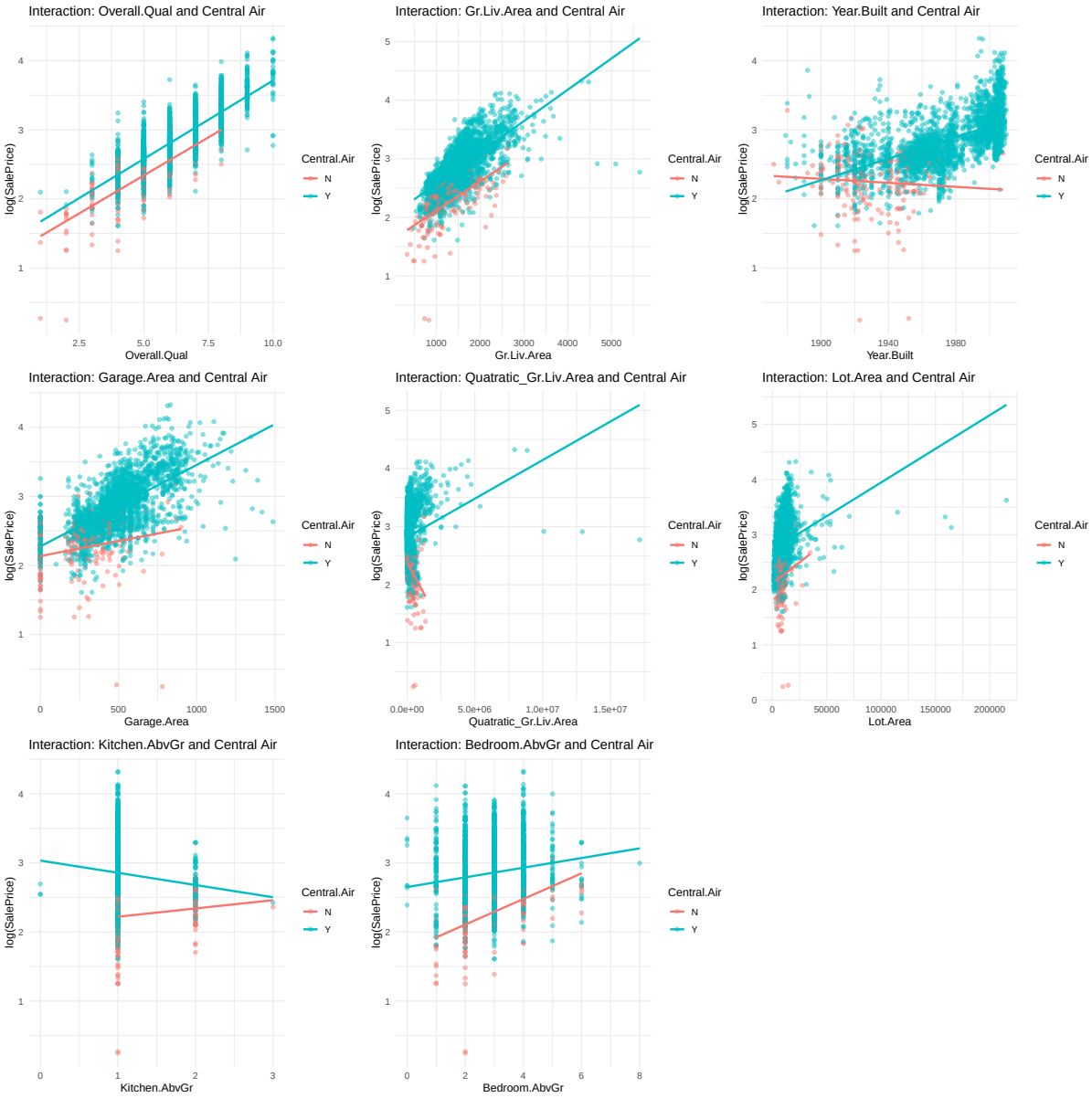
In conclusion, the final linear regression model based on 14 predictor variables satisfies the essential assumptions for regression analysis and provides meaningful insights into the key factors affecting house prices. By applying a log transformation to `SalePrice`, we stabilized the variance and made the error distribution more normal, addressing the issues found in the initial model's diagnostic plots. We also used forward selection to refine the model by selecting only the most relevant variables.

However, a few limitations were identified in the analysis. Outliers, high leverage points, and influential data points suggest that the model is sensitive to extreme values. These issues were addressed during the diagnostic phase, but further exploration and refinement may be needed to ensure the robustness of the model. The model with interaction terms showed a slight improvement in fit, but for simplicity, only the most significant interaction terms were included in the final model.

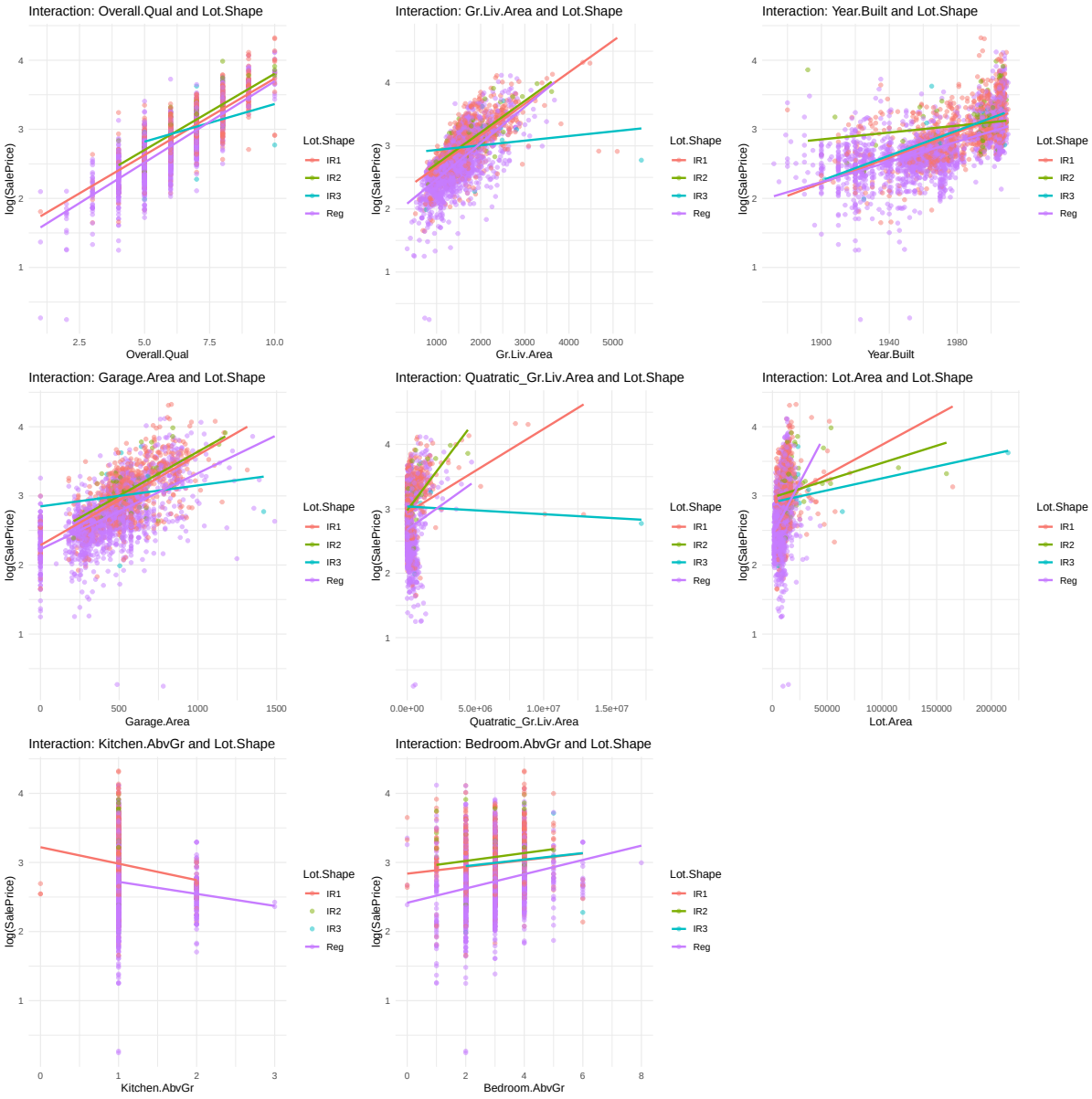
Overall, the model shows promising performance, but there is still room for improvement by addressing outliers and considering more complex interactions. As a result, this model provides a solid foundation for understanding the key factors influencing house prices, but further refinements are needed to enhance its predictive power and generalizability.

//TODO [SB1]Which top 3 we are describing ?

Appendix Interaction Term




```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```


Interaction: Overall.Qual and Lot.Shape
log(SalePrice)
Overall.Qual
Interaction: Gr.Liv.Area and Lot.Shape
log(SalePrice)
Gr.Liv.Area
Interaction: Year.Built and Lot.Shape
log(SalePrice)
Year.Built
Interaction: Garage.Area and Lot.Shape
log(SalePrice)
Garage.Area
Interaction: Quatratic_Gr.Liv.Area and Lot.Shape
log(SalePrice)
Quatratic_Gr.Liv.Area
Interaction: Lot.Area and Lot.Shape
log(SalePrice)
Lot.Area
Interaction: Kitchen.AbvGr and Lot.Shape
log(SalePrice)
Kitchen.AbvGr
Interaction: Bedroom.AbvGr and Lot.Shape
log(SalePrice)
Bedroom.AbvGr
Lot.Shape
IR1
IR2
IR3
Reg