Module 1 Assignment - Regression Diagnostics with R

Jose De Leon

Northeastern University

College of Professional Studies

ALY6015: Intermediate analytics

Professor: Jean-Sebastien Provost

April 16, 2025

Contents

Introduction

This report presents a detailed analysis of housing prices in Ames, Iowa using a multi-step regression diagnostics and modeling workflow. The dataset, widely adopted in pedagogical and professional contexts, offers a high-quality alternative to the Boston Housing dataset, comprising 2,930 observations and 82 variables collected from the Ames Assessor's Office. The objective of the analysis is to build a statistically sound regression model to predict SalePrice, the sale value of homes, using meaningful and well-validated features.

The analytical process is structured in 14 systematic steps, including exploratory data analysis, handling missing values, model fitting, diagnostics, and model selection via all-subsets regression. The result is a validated regression model with high predictive performance and interpretability, useful for educational, analytical, or real estate decision-making applications.

Analysis

Exploratory data analysis and descriptive statistics

The analysis began with the import and inspection of the dataset. Initial exploration revealed that the target variable, SalePrice, was right-skewed, indicating potential for outliers and heteroscedasticity. Histograms (Figure 1 in the appendix) and boxplots (Figure 2 in the appendix) confirmed the presence of extreme values, and summary statistics (Table 1 in the appendix) helped identify variables with missing data. Early visualization using scatterplots (Figures 3, 4 and 5 in the appendix) and scatterplot matrices (Figure 6 in the appendix) reveal linear relationships between SalePrice and predictors like Gr.Liv.Area, Garage.Area, and Total.Bsmt.SF, guiding our selection of initial model inputs.

Data cleaning and preparation

Further, the dataset was cleaned and prepared for modeling. Missing data was addressed using the median for most numeric variables due to the presence of skewed distributions or outliers that would distort the mean. For Mas.Vnr.Area, zero was chosen as imputation since most homes do not have a masonry veneer and the missing rate was under 1%. Based on dataset documentation and best practices for model accuracy, we removed atypical properties such as those with extremely large Gr.Liv.Area values (greater than 4000 square feet) and transactions classified as Abnorml, Family, or Partial in SaleCondition. These steps enhance the consistency and robustness of the modeling process by ensuring the dataset reflects typical residential home sales.

Correlation matrix

A correlation matrix (Figure 7 in the appendix) was computed and visualized to reveal the relationships between the variables. Strong relationships were found between SalePrice and features such as Overall.Qual (0.80), Gr.Liv.Area (0.71), and Garage.Area (0.64). This informed the selection of variables for regression modeling.

Correlations

Scatterplots visualize how the strength and shape of relationships differ across predictors. The scatterplot between SalePrice and Overall Quality(Figure 8 in the appendix) reveals a strong positive relationship. The trend is fairly linear, indicating that higher overall quality ratings are associated with higher sale prices. This pattern supports the use of Overall.Qual variable as a predictor in the regression model. On the other hand, Lot.Area (Figure 9 in the appendix) showed a weak correlation, a non-linear pattern, and a high variance against SalePrice especially for larger lot sizes. Finally, we also examined the correlation between SalePrice and FullBathrooms(Figure 10 in the appendix) which exhibits a moderate positive trend, with sale prices increasing as the number of full bathrooms increases. These visualizations confirm the linear suitability of Overall.Qual, while highlighting potential non-linearities and variability in predictors like Lot.Area and Full.Bath.

Fit the regression model and interpretation

Model created: SalePrice = -28,826.999 + 70.697 × Gr.Liv.Area + 59.077 × Total.Bsmt.SF + 85.079 × Garage.Area

Based on the coefficients from table 2 in the appendix, we got the formula for the linear model using the 3 variables. Each coefficient is statistically significant at the 0.001 level. For every additional square foot of above-ground living area (Gr.Liv.Area), the model predicts an increase of approximately $70.70 in sale price, holding other variables constant. Similarly, an additional square foot in basement area (Total.Bsmt.SF) increases the predicted sale price by about $59.08. Finally, for every extra square foot in the garage area (Garage.Area), the sale price is expected to increase by roughly $85.08. The intercept is negative, which is common in models with no meaningful zero-value for the predictors. In table 3(from the appendix), we can find the Adjusted R-squared of 0.7266 and a Residual Standard Error (RSE) of 36,610.

Plot the regression model

In figure 11 from the appendix, standard regression diagnostic plots were analyzed to assess assumptions like linearity, normality, and homoscedasticity. The residuals vs. fitted plot revealed some curvature, suggesting minor non-linearity. The Q-Q plot showed some deviation from normality in the tails, indicating mild skewness or heavy-tailed residuals. The scale-location plot presents a mild heteroscedasticity, where the variance of residuals increases with fitted values, while Residuals vs Leverage indicated that most observations fall within acceptable bounds, though a few points with high leverage and large standardized residuals are noted.

Overall, while the model assumptions are mostly met, there is evidence of mild heteroscedasticity and non-linearity that could be improved in future model iterations.

Multicollinearity

Table 4 from the appendix, assessed multicollinearity using Variance Inflation Factors (VIF), which all VIF values are well below the commonly used threshold of 5, indicating that multicollinearity is not a concern in this model. These predictors do not exhibit strong linear relationships with one another, allowing each to independently contribute to explaining variance in SalePrice.

Outliers, High Leverage, and Influential Observation Detection

In table 5, figure 12 and figure 13 from the appendix we can find the observations 433, 1064, and 2446 that consistently appeared across all three diagnostic checks: as outliers (with large studentized residuals), as high leverage points (with large hat values), and as influential observations (with large Cook’s distances). Because these data points can skew the model's coefficients and diagnostic accuracy, we decided to remove them from the dataset for the improved model analysis in the next step.

This cleanup ensures our regression model better represents the general patterns in the data without being overly affected by extreme or unusual cases.

Model Improvement and Best Subset Selection

After identifying the outliers and influential points, we attempted to remove them. After doing so, we will fit a new model using the same three predictors: Gr.Liv.Area, Total.Bsmt.SF, and Garage.Area. The cleaned model (table 6 in the appendix) had a Residual Standard Error of 36630 and an Adjusted R-squared of 0.7267. Since the previous model had an Adjusted R-Squared of 0.7266, we can say that the initial model was already robust.

From table 7 in the appendix, we can get the formula of the cleaned model.

Cleaned model

SalePrice = -28904.1 + 70.695 × Gr.Liv.Area + 59.051 × Total.Bsmt.SF + 85.29 × Garage.Area

To further improve model performance, we applied all subsets regression using the regsubsets() function from the leaps package. We considered eight potential predictors: Gr.Liv.Area, Total.Bsmt.SF, Garage.Area, Overall.Qual, Year.Built, X1st.Flr.SF, Full.Bath, and Garage.Cars. The best-performing model (based on Adjusted R-squared), which is shown in figure 14 in the appendix, includes the following 7 predictors:

- Gr.Liv.Area
- Total.Bsmt.SF
- Garage.Area
- Overall.Qual
- Year.Built
- X1st.Flr.SF
- Full.Bath

The Garage.Cars variable was not selected, indicating that the garage area is a more informative variable than car capacity.

Comparison and Model Preference

The 7-variable model demonstrated a significant improvement in model performance. In table 8 from the appendix we can see that it achieved an Adjusted R-squared of 0.8344 and a Residual Standard Error of 28510, compared to the 3-variable model's Adjusted R-squared of 0.7267 and Residual Standard Error of 36630. This suggests that the additional predictors captured more variance in house prices and improved the overall accuracy of the model.

Therefore, the preferred model is the 7-variable model:

SalePrice = -828900 + 58.78(Gr.Liv.Area) + 26.88(Total.Bsmt.SF) + 35.61(Garage.Area) + 18360(Overall.Qual) + 385.6(Year.Built) + 17.43(X1st.Flr.SF) - 10420(Full.Bath)

We can find the coefficients in table 9 from the appendix.

This model balances predictive power with interpretability and includes features that are both statistically significant and meaningful from a real estate valuation perspective.

We also compared the two models using AIC and BIC, which evaluate model quality while penalizing for complexity:

- AIC (3-variable model): 58331.41
- BIC (3-variable model): 58360.42

- AIC (7-variable model): 57110.21

- BIC (7-variable model): 57162.43

Both AIC and BIC are lower for the 7-variable model, further confirming that it provides a better balance between fit and complexity.

Conclusion

This analysis demonstrates a complete regression modeling pipeline applied to real-world housing data. The process of filtering, cleaning, modeling, and validating results produced not only a high-performing predictive model but also valuable insights into the underlying drivers of home prices in Ames, Iowa. Key findings include the critical impact of Overall.Qual, the added value of newer construction (Year.Built), and the superiority of square footage (Garage.Area) over car count (Garage.Cars) in explaining price variation.

The final model with 7 variables offers a powerful, interpretable, and statistically sound approach to real estate valuation. Its superior Adjusted R-squared, lower residual error, and favorable AIC/BIC values confirm its usefulness for predictive and explanatory purposes.

References

De Cock, D. (2011). *Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project*. Journal of Statistics Education, 19(3). https://jse.amstat.org/v19n3/decock.pdf

Provost, J.-S. (2025). *Regression diagnostics with R* [Video]. Canvas. https://northeastern.instructure.com/courses/221013/assignments/2683591

OpenAI. (2025, April 16). *When to impute missing values with mean, median, or zeros* [Large language model]. ChatGPT. https://chat.openai.com/

Provost, J.-S. (2025). *Module 1 – Recap: Correlation and regression* [PDF file]. Department of College of Professional Studies, Northeastern University.

Appendix

Visualizations and tables

Distribution of Sale Price

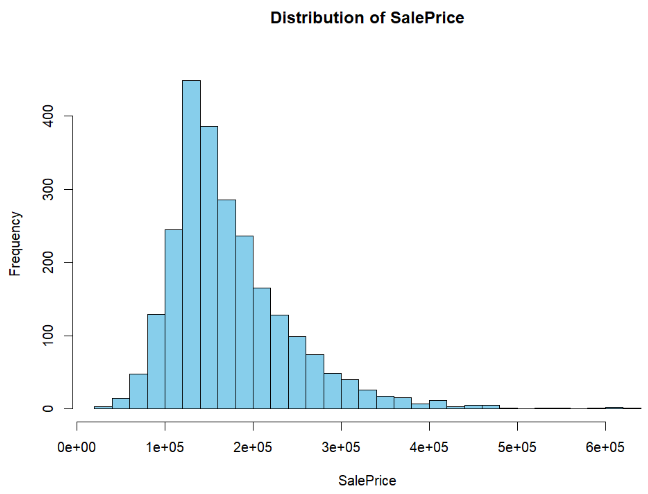


Figure 1: Histogram of Distribution of Sale Price

The histogram presents how house prices are spread out in the Ames Housing dataset, and it’s clear that most homes sell between $100,000 and $200,000, with a lot of them clustered around $150,000. The shape leans to the right, with a long tail that points to a smaller number of high-priced homes.

Boxplot of Sale Price

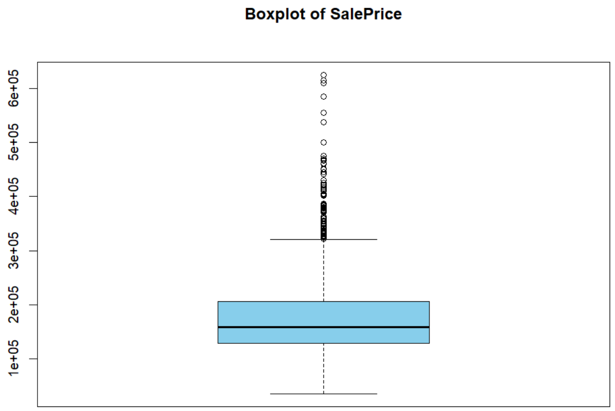


Figure 2: Boxplot of Sale Price

The boxplot of SalePrice presents that the median home price is around $175,000, with most sales falling between about $130,000 and $220,000. A lot of the data sits within this range, but there are some clear outliers on the higher end. These represent the more expensive homes in the market.

Summary statistics of Sale Price

Statistic	SalePrice
Min.	35000
1st Qu.	129000
Median	158000
Mean	174870
3rd Qu.	205963
Max.	625000

Table 1: Summary statistics of Sale Price

The summary statistics for SalePrice show that home prices in the Ames dataset range from $35,000 to $625,000, which is quite a spread. The median price is $158,000, and the average is slightly higher at $174,870, which makes sense given the data is skewed to the right. Most homes fall between $129,000 and $205,963, covering the middle 50 percent of all sales.

Scatterplot: Sale Price vs GrLivArea

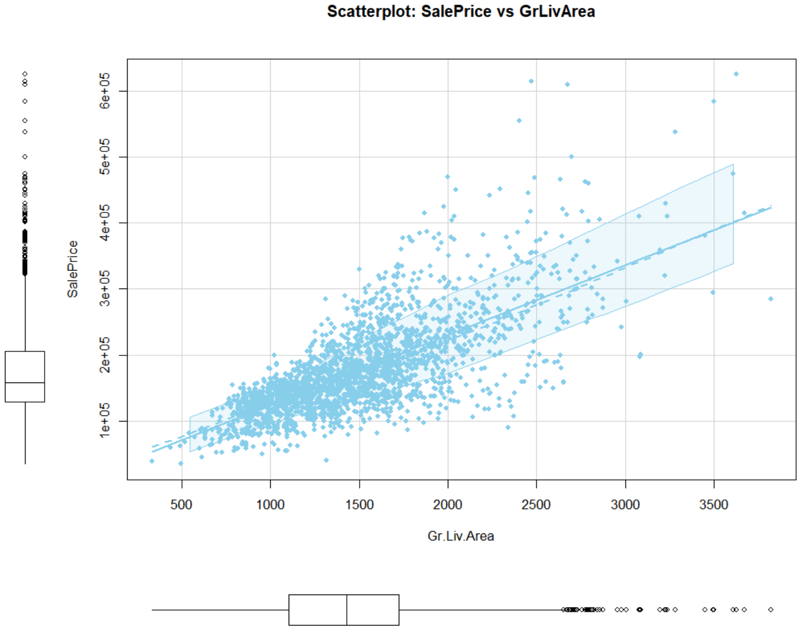


Figure 3: Scatterplot: Sale Price vs GrLivArea

In the scatterplot of SalePrice against Gr.Liv.Area we can see a strong upward trend; bigger homes tend to sell for more. Most of the points follow the line pretty closely, which suggests a consistent relationship between living area and price. There are a few outliers with really high values, but overall, the pattern is clear.

Scatterplot: Sale Price vs TotalBsmtSF

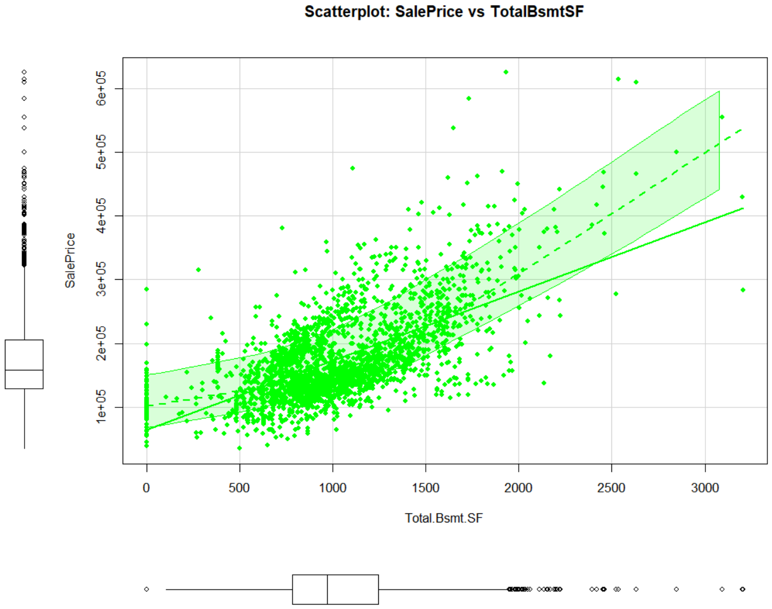


Figure 4: Scatterplot: Sale Price vs TotalBsmtSF

The scatterplot of SalePrice versus TotalBsmtSF indicates a clear upward trend, as homes with larger basements generally sell for more. However, the points are more spread out compared to living area, meaning basement size does impact price, but not as strongly or consistently. There's more variation in how much value a bigger basement adds.

Scatterplot: Sale Price vs Garage Area

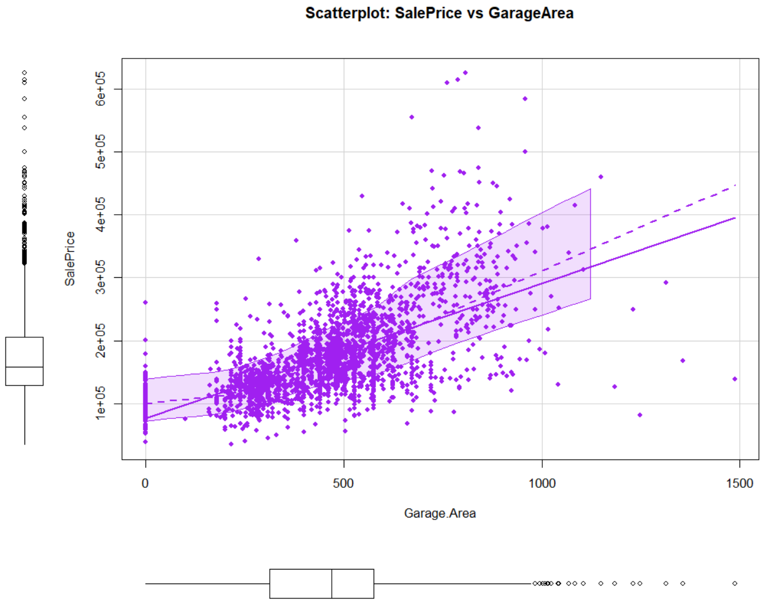


Figure 5: Scatterplot: Sale Price vs GarageArea

This scatterplot indicates the relationship between sale price and garage area. Overall, homes with bigger garages tend to sell for more, and you can see a clear upward trend in the data. There's a bit more spread in the points as garage size increases, but the pattern still suggests that garage space adds value to a home.

Scatterplot Matrix: Ames Housing Variables

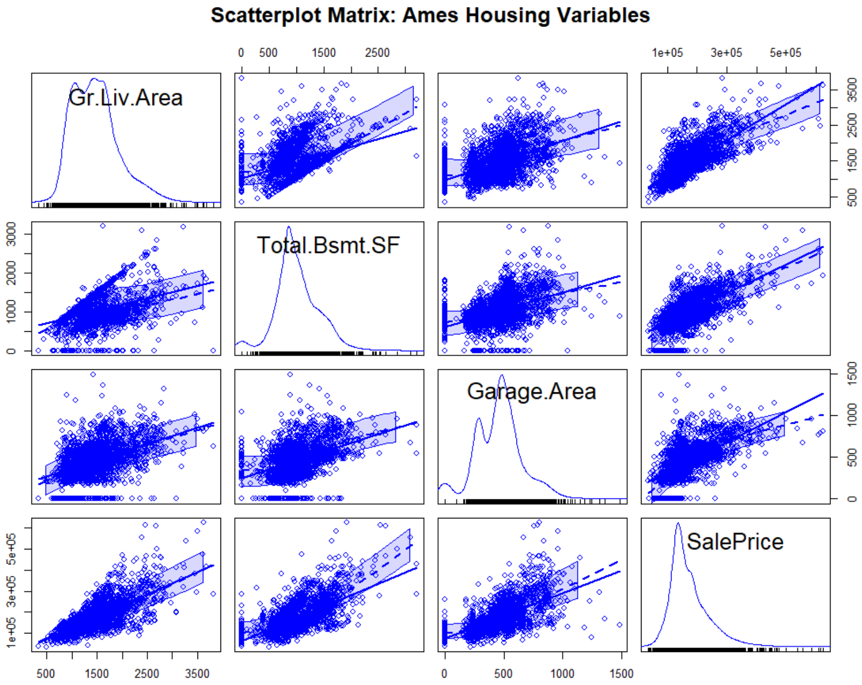
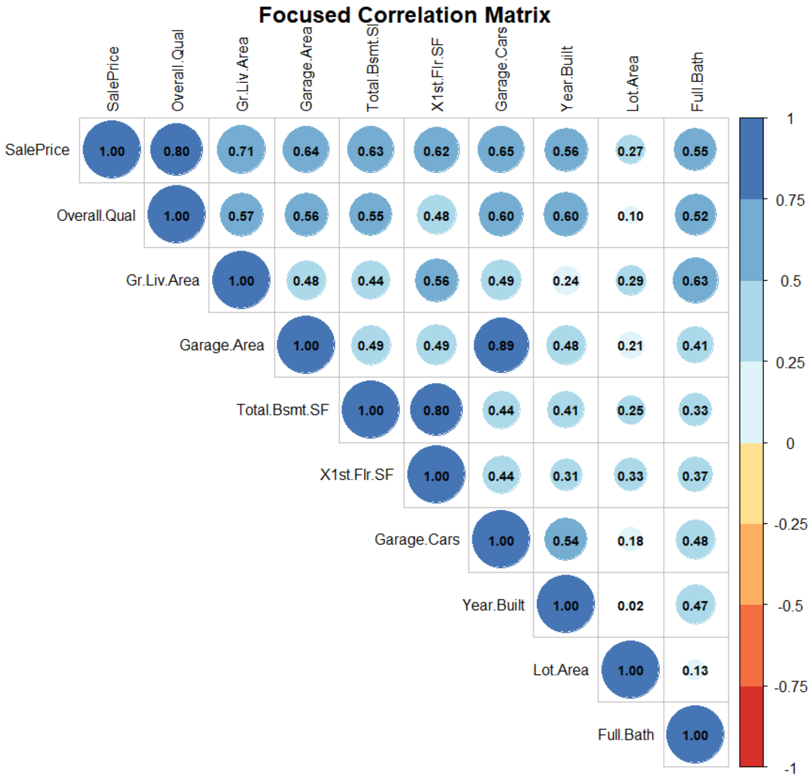


Figure 6: Scatterplot Matrix - Ames Housing Variables

This scatterplot matrix gives a quick view of how some key features like living area, basement size, and garage area relate to each other and to sale price. In general, as those features increase, the sale price tends to go up too. The plots along the diagonal show the shape of each variable's distribution, and you can clearly see that sale price has a long tail on the higher end.

Correlation Matrix



	SalePrice	Overall.Qual	Gr.Liv.Area	Garage.Area	Total.Bsmt.S	X1st.Flr.SF	Garage.Cars	Year.Built	Lot.Area	Full.Bath
SalePrice	1.00	0.80	0.71	0.64	0.63	0.62	0.65	0.56	0.27	0.55
Overall.Qual		1.00	0.57	0.56	0.55	0.48	0.60	0.60	0.10	0.52
Gr.Liv.Area			1.00	0.48	0.44	0.56	0.49	0.24	0.29	0.63
Garage.Area				1.00	0.49	0.49	0.89	0.48	0.21	0.41
Total.Bsmt.SF					1.00	0.80	0.44	0.41	0.25	0.33
X1st.Flr.SF						1.00	0.44	0.31	0.33	0.37
Garage.Cars							1.00	0.54	0.18	0.48
Year.Built								1.00	0.02	0.47
Lot.Area									1.00	0.13
Full.Bath										1.00

Figure 7 : Correlation Matrix

This correlation matrix shows how strongly each variable is related to sale price and to one another. Overall quality has the highest correlation with sale price, followed by living area, garage area, and basement size. Some features like lot area and garage car capacity have weaker relationships. The darker the circle, the stronger the connection, so it's easy to spot which variables might be more useful in predicting home prices.

Highest correlation: Sale Price vs Overall Quality

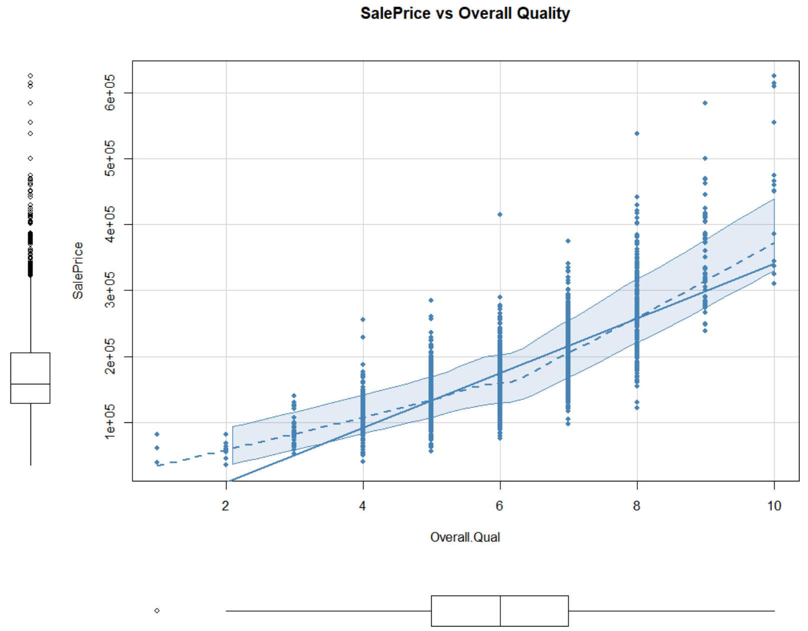


Figure 8: Highest correlation - Sale Price vs Overall Quality

This plot shows how sale price changes with overall quality, which is a rating of the home's materials and finish. There's a clear trend here, as the quality rating goes up, sale price tends to go up too. Houses rated closer to 9 or 10 sell for much more than those rated around 4 or 5, and the pattern looks pretty steady across the scale.

Lowest correlation: Sale Price vs Lot Area

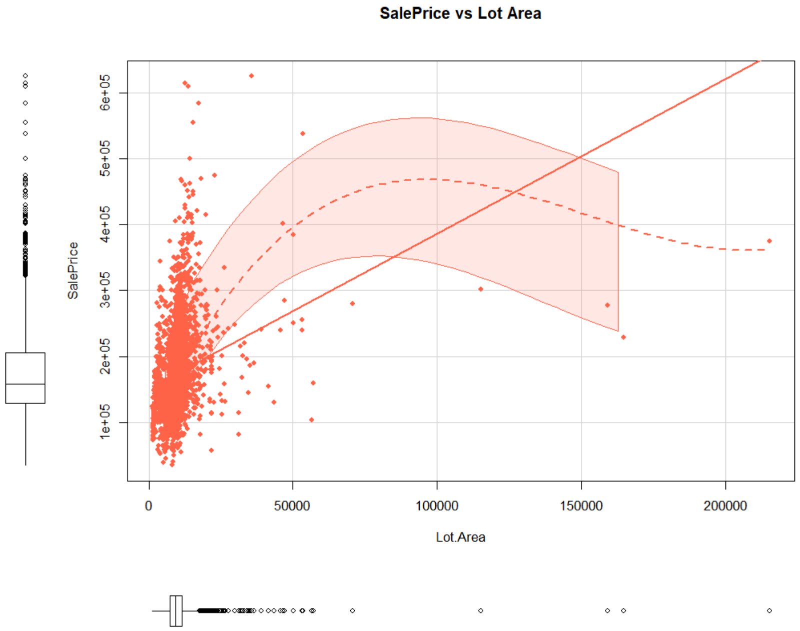


Figure 9: Lowest correlation - Sale Price vs Lot Area

This plot indicates the relationship between sale price and lot area. In general, homes with larger lots can sell for more, but the connection isn't very clear. Most of the data is packed in the lower range, and beyond a certain point, the pattern gets messy and less consistent. There are some really large lots that don't necessarily come with higher prices, which adds a lot of variation.

Correlation: Sale Price vs Full Bathrooms

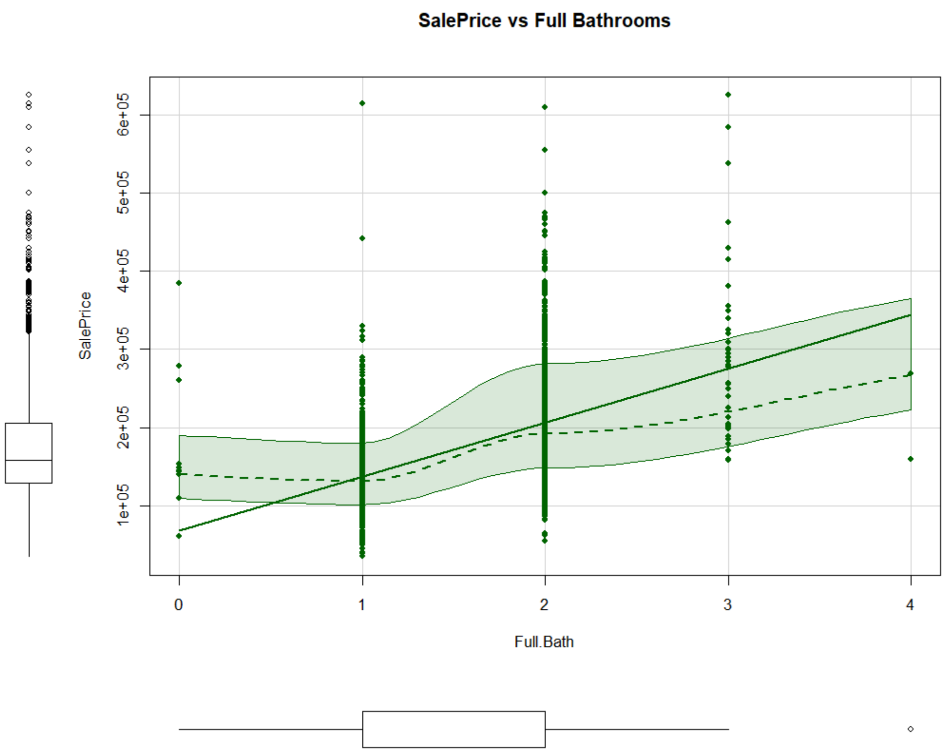


Figure 10: correlation - Sale Price vs Full Bathrooms

This plot exhibits how sale prices change with the number of full bathrooms in a house. In general, homes with more bathrooms tend to sell for more, especially when going from one to two. After that, the increase in price is less consistent. There's still a positive trend, but the price difference isn't as strong once you pass two or three bathrooms.

Coefficients of the model

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-28,826.999	2645.02	-10.9	<2e-16
Gr.Liv.Area	70.697	1.78	39.71	<2e-16
Total.Bsmt.SF	59.077	2.053	28.78	<2e-16
Garage.Area	85.079	4.341	19.6	<2e-16

Table 2: Coefficients of the model

This table shows the results from a regression model that predicts house prices based on living area, basement size, and garage area. All three variables have a strong effect on price and are statistically significant. On average, for every extra square foot of living space, the price goes up by about 71 dollars. A square foot of basement adds around 59 dollars, and a square foot of garage space adds about 85 dollars. The intercept is negative, which just helps the math work in the model, it doesn't have a real-world meaning in this case.

Summary of the model

Metric	Value
Residual standard error	36610
Degrees of freedom	2444
Multiple R-squared	0.7269
Adjusted R-squared	0.7266
F-statistic	2169
F-statistic DF	3 and 2444
F-statistic p-value	< 2.2e-16

Table 3: Summary of the model

This summary gives us an idea of how well the model fits the data. The R squared is about 0.73, which means the model explains around 73 percent of the variation in house prices using just three features, living area, basement size, and garage area. The adjusted R squared is almost the same, which shows the model is consistent. The residual standard error is about 36,610 dollars, so that's roughly how far off the predictions are, on average.

Plotting the regression model

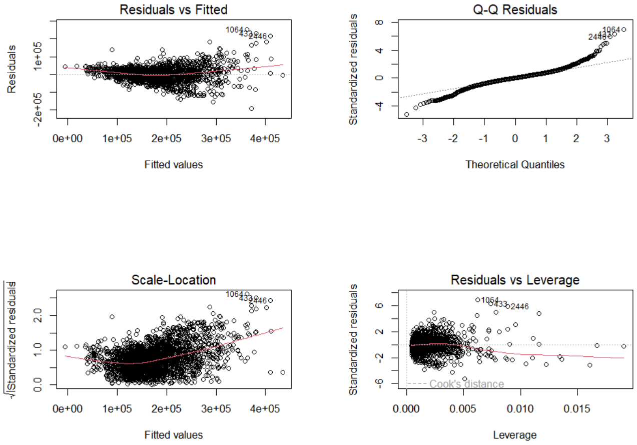


Figure 11: Plotting the regression model.

These plots help us see how well the model fits the data. The residuals show a slight curve, which suggests the relationship isn't fully linear. The spread of the points increases a bit, hinting at some uneven variance. A few points have high influence and could be affecting the results more than others.

Multicollinearity

Variable	Standardized Coefficient
Gr.Liv.Area	1.334709
Total.Bsmt.SF	1.29735
Garage.Area	1.41197

Table 4: Multicollinearity

This table manifests the multicollinearity check using VIF values for the three variables in the model. All the values are close to 1, which means there's no serious overlap between the predictors. Each variable brings in its own information without being too closely related to the others.

Outliers

Index	Rstudent	unadjusted p-value	Bonferroni p
1064	6.979924	3.7904E-12	9.2788E-09
433	6.343799	2.6624E-10	6.5177E-07
2446	5.931987	3.4172E-09	8.3653E-06
2593	-5.28135	1.3956E-07	0.00034164
2342	5.033558	5.1653E-07	0.0012645
2451	5.012524	5.757E-07	0.0014093
424	4.803154	1.6564E-06	0.0040548
16	4.565796	5.2228E-06	0.012785

Table 5: Outliers

This table presents the rows of the dataset that are outliers. These are data points where the predicted price was way off from the actual value.

High leverage observations

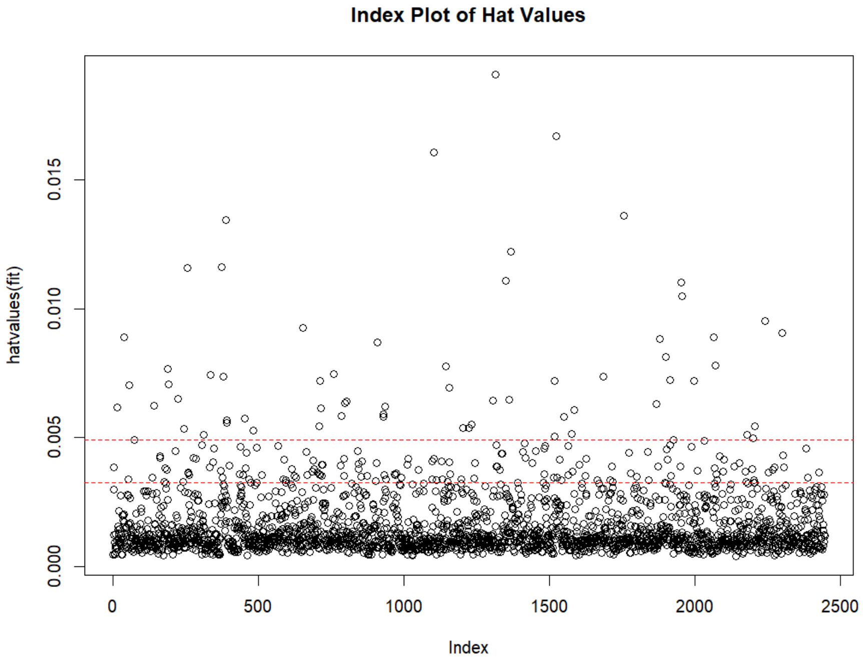


Figure 12: High leverage observations

This plot reveals which observations have high leverage, meaning they have unusual combinations of predictor values. Most of the points stay low, but a few stand out well above the red line. These higher points could have a stronger influence on the model, so they should be double checked.

Influential observations

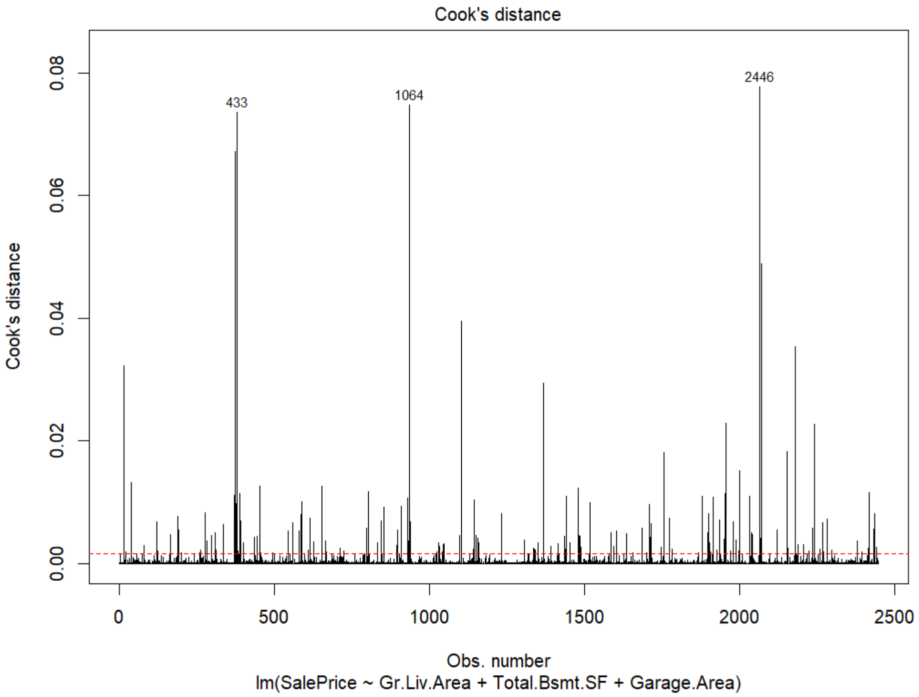


Figure 13: Influential observations

This plot manifests which data points have the biggest impact on the model. A few stand out, especially observations 433, 1064, and 2446, they’re way higher than the rest. These points have a lot of influence on the results, so it’s worth checking them out.

Summary of the cleaned model

Metric	Value
Residual standard error	36630
Degrees of freedom	2441
Multiple R-squared	0.727
Adjusted R-squared	0.7267
F-statistic	2167
F-statistic DF	3 and 2441
F-statistic p-value	< 2.2e-16

Table 6: Summary of the cleaned model

This summary is for the model after removing outliers and influential points. The R squared is still around 0.73, so the model explains about 73 percent of the variation in sale prices. The adjusted R squared barely changed, which means the model stayed consistent even after the cleanup. The residual error is about 36,630 dollars, so on average, predictions are off by that much.

Coefficients of the cleaned model

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-28904.1	2647.117	-10.92	<2e-16
Gr.Liv.Area	70.695	1.781	39.69	<2e-16
Total.Bsmt.SF	59.051	2.054	28.75	<2e-16
Garage.Area	85.29	4.347	19.62	<2e-16

Table 7: Coefficients of the cleaned model

These are the results from the updated model after cleaning the data. The coefficients stayed almost the same, which means the relationship between house features and price didn't really change. For each extra square foot of living space, the price goes up by about 71 dollars. Basement and garage space also add value, with each adding around 59 and 85 dollars per square foot, respectively. All of these effects are statistically strong.

Performance of subsets using Adjusted R^2

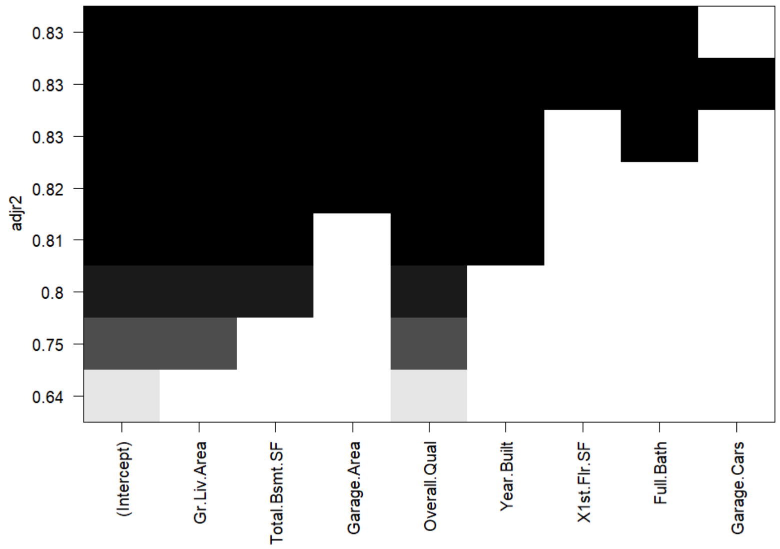


Figure 14: Performance of subsets using Adjusted R^2

This chart exhibits how the model improves as more variables are added. The adjusted R squared goes up the most with the first few features, like living area and overall quality. After that, each new variable helps a bit less. It's a good way to see when adding more stuff stops making a real difference.

Summary of the model with 7 variables

Metric	Value
Residual standard error	28510
Degrees of freedom	2437
Multiple R-squared	0.8349
Adjusted R-squared	0.8344
F-statistic	1760
F-statistic DF	7 and 2437
F-statistic p-value	< 2.2e-16

Table 8: Summary of the model with 7 variables

This summary shows that the model with seven variables does a great job explaining house prices. The adjusted R squared is about 0.83, which means it captures over 83 percent of the variation in prices. That's a big improvement compared to the simpler model. The average prediction error dropped to around 28,510 dollars, so it's not only more accurate but also more reliable.

Coefficients of the model with 7 variables

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-828900	50090	-16.547	< 2e-16
Gr.Liv.Area	58.78	1.907	30.826	< 2e-16
Total.Bsmt.SF	26.88	2.3	11.687	< 2e-16
Garage.Area	35.61	3.651	9.752	< 2e-16
Overall.Qual	18360	644.5	28.495	< 2e-16
Year.Built	385.6	26.36	14.625	< 2e-16
X1st.Flr.SF	17.43	2.638	6.61	4.71e-11
Full.Bath	-10420	1500	-6.946	4.79e-12

Table 9: Coefficients of the model with 7 variables

This table communicates how each of the seven variables affects house price in the model. Bigger homes, newer construction, better quality, and more garage or basement space all lead to higher prices. Overall quality and living area have the strongest impact. The negative value for full bathrooms is a bit surprising, but it probably means their effect is already covered by other features in the model. All of these numbers are statistically strong, so they each help explain the price in a meaningful way.

R code

```
library(dplyr)

library(ggplot2)

library(corrplot)

library(RColorBrewer)

library(car)

library(leaps)




#import the data

ames<-read.csv("AmesHousing.csv")



#step 2

#Exploratory data analysis for the dataset

#summary information for the data

summary(ames)

head(ames)

str(ames)

dim(ames)


#Descriptive statistics for sale price

summary(ames$SalePrice)


#histograms of Sale price

hist(ames$SalePrice, main = "Distribution of SalePrice", xlab = "SalePrice", col = "skyblue", breaks = 30)
```

```
#Boxplot of sale price
boxplot(ames$SalePrice, main = "Boxplot of SalePrice", col = "skyblue")


#View scatter plots of Ground living area and sale price
scatterplot(SalePrice ~ Gr.Liv.Area, data = ames,
        main = "Scatterplot: SalePrice vs GrLivArea",
        col = "skyblue", pch = 19)

#View scatter plots of total basement square feet area and sale price
scatterplot(SalePrice ~ Total.Bsmt.SF, data = ames,
        main = "Scatterplot: SalePrice vs TotalBsmtSF",
        col = "green", pch = 19)


#View scatter plots of garage area and sale price
scatterplot(SalePrice ~ Garage.Area, data = ames,
        main = "Scatterplot: SalePrice vs GarageArea",
        col = "purple", pch = 19)

#counting missing values
missing_counts <- colSums(is.na(ames))
missing_counts[missing_counts > 0]

#percentage of missing values inside their column
missing_percent <- colMeans(is.na(ames)) * 100
missing_percent[missing_percent > 0]

#analyzing histogram form to validate which is the best input to use (mean, median, mode or zero)
```

```
#histogram of Lot. Frontage

hist(ames$Lot.Frontage, main = "Distribution of Lot.Frontage",

    xlab = "Lot.Frontage", col = "skyblue", breaks = 30) #imputing the median because it is right skewed distributed and it is 16% of missing values so we need to be careful by selecting the best selection in this case is median, because choosing mean will be too simplistic

# Garage.Yr.Blt

hist(ames$Garage.Yr.Blt, main = "Distribution of Garage.Yr.Blt",

    xlab = "Garage.Yr.Blt", col = "lightgreen", breaks = 30) #imputing the median because it is normally distributed and it is 5.4% of missing values


summary(ames$Garage.Yr.Blt)

boxplot(ames$Garage.Yr.Blt, main = "Boxplot of Garage.Yr.Blt", col = "lightblue")



# Mas.Vnr.Area

hist(ames$Mas.Vnr.Area, main = "Distribution of Mas.Vnr.Area",

    xlab = "Mas.Vnr.Area", col = "salmon", breaks = 30) #even though it is right extremely right skewed, its missing values is less than 1% this make safe to input with zero


# Scatterplot Matrix for selected variables

scatterplotMatrix(~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area + SalePrice,

              data = ames,

              spread = FALSE,

              smoother.args = list(lty = 2),

              main = "Scatterplot Matrix: Ames Housing Variables")


#step 3 - imputing missing values


#removing observations greater than 4000 based on the documentation

ames <- ames[ames$Gr.Liv.Area <= 4000, ]
```

```
#filtering to exclude foreclosures and family sales based on the documentation
ames <- ames[ames$Sale.Condition %in% c("Normal", "Alloca", "AdjLand"), ]


# Impute Lot.Frontage using median because it's right-skewed and has 16.7% missing values.
# Mean would be too sensitive to outliers.
ames$Lot.Frontage[is.na(ames$Lot.Frontage)] <- median(ames$Lot.Frontage, na.rm = TRUE)


# Impute Garage.Yr.Blt using median.
ames$Garage.Yr.Blt[is.na(ames$Garage.Yr.Blt)] <- median(ames$Garage.Yr.Blt, na.rm = TRUE)


# Impute Mas.Vnr.Area using 0.
# This variable is extremely right-skewed and most values are already 0 (indicating no veneer).
# Since missingness is <1%, it's safe to impute with 0.
ames$Mas.Vnr.Area[is.na(ames$Mas.Vnr.Area)] <- 0



#impute the other numerical variables
ames$BsmtFin.SF.1[is.na(ames$BsmtFin.SF.1)] <- median(ames$BsmtFin.SF.1, na.rm = TRUE)
ames$BsmtFin.SF.2[is.na(ames$BsmtFin.SF.2)] <- median(ames$BsmtFin.SF.2, na.rm = TRUE)
ames$Bsmt.Unf.SF[is.na(ames$Bsmt.Unf.SF)]   <- median(ames$Bsmt.Unf.SF, na.rm = TRUE)
ames$Total.Bsmt.SF[is.na(ames$Total.Bsmt.SF)] <- median(ames$Total.Bsmt.SF, na.rm = TRUE)

ames$Bsmt.Full.Bath[is.na(ames$Bsmt.Full.Bath)] <- median(ames$Bsmt.Full.Bath, na.rm = TRUE)
```

```
ames$Bsmt.Half.Bath[is.na(ames$Bsmt.Half.Bath)] <- median(ames$Bsmt.Half.Bath, na.rm =
TRUE)


ames$Garage.Cars[is.na(ames$Garage.Cars)] <- median(ames$Garage.Cars, na.rm = TRUE)

ames$Garage.Area[is.na(ames$Garage.Area)] <- median(ames$Garage.Area, na.rm = TRUE)


#step 4 and step 5

cor_vars <- ames_numeric[, c("SalePrice", "Overall.Qual", "Gr.Liv.Area", "Garage.Area",
"Total.Bsmt.SF",

                  "X1st.Flr.SF", "Garage.Cars", "Year.Built", "Lot.Area", "Full.Bath")]


# Create cleaner correlation matrix

cor_subset <- cor(cor_vars, use = "complete.obs")


#correlation matrix

corrplot(cor_subset,

     type = "upper",

     col = brewer.pal(n = 8, name = "RdYlBu"),

     addCoef.col = "black",

     tl.cex = 0.8,

     number.cex = 0.7,

     tl.col = "black",

     mar = c(0,0,1,0),

     title = "Focused Correlation Matrix")


#step 6

# Highest correlation

scatterplot(SalePrice ~ Overall.Qual, data = ames,

      main = "SalePrice vs Overall Quality",
    col = "steelblue", pch = 19)

# Lowest correlation
scatterplot(SalePrice ~ Lot.Area, data = ames,
    main = "SalePrice vs Lot Area",
    col = "tomato", pch = 19)

# Correlation closest to 0.5
scatterplot(SalePrice ~ Full.Bath, data = ames,
    main = "SalePrice vs Full Bathrooms",
    col = "darkgreen", pch = 19)

#step 7 and step 8
# Fit the multiple linear regression model
model <- lm(SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area, data = ames)
summary(model)

#step 9 - Plot the regression model
# Step 9
par(mfrow = c(2, 2))
plot(model)
dev.off()

# Q-Q Plot for normality of residuals
qqPlot(model, labels = row.names(ames), simulate = TRUE, main = "Q-Q Plot of Residuals")

# Component + Residual (Partial Residual) Plots to assess linearity
crPlots(model, main = "Component + Residual Plots")
```

```
# Spread-Level Plot to check constant variance (homoscedasticity)
spreadLevelPlot(model, main = "Spread-Level Plot")

#step 10 -  Check VIF for multicollinearity
vif(model)

#step 11
# 1. Outliers (Extreme residuals)
outlierTest(model)

# 2. High Leverage Observations
hat.plot <- function(fit) {
  p <- length(coefficients(fit))
  n <- length(fitted(fit))
  plot(hatvalues(fit), main = "Index Plot of Hat Values")
  abline(h = c(2, 3) * p / n, col = "red", lty = 2)
  identify(1:n, hatvalues(fit), names(hatvalues(fit)))
}
hat.plot(model)

# 3. Influential Observations using Cook's Distance
cutoff <- 4 / (nrow(ames) - length(model$coefficients) - 2)
plot(model, which = 4, cook.levels = cutoff)
abline(h = cutoff, lty = 2, col = "red")

# Step 12: Remove influential observations
ames_cleaned <- ames[-c(433, 1064, 2446), ]
# Refit the model
model_cleaned <- lm(SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area, data = ames_cleaned)
```

```
# Compare with original model

summary(model)

summary(model_cleaned)


#step 13

# Run all-subsets regression

leaps <- regsubsets(SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +

           Overall.Qual + Year.Built + X1st.Flr.SF + Full.Bath + Garage.Cars,

          data = ames_cleaned)


# Visualize the performance of subsets using Adjusted R^2

plot(leaps, scale = "adjr2")


#step 14

#linear regression model with the 7 predictors

model7 <- lm(SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area +

       Overall.Qual + Year.Built + X1st.Flr.SF + Full.Bath,

      data = ames_cleaned)

summary(model7)

summary(model_cleaned)


# Compare AIC and BIC

AIC(model_cleaned)

BIC(model_cleaned)


AIC(model7)

BIC(model7)
```