Ames Housing   
  
  
  
  
  
  
  
  
  
Name: Hemanth Kurra

# Introduction

The Ames Housing Dataset, derived from Ames, Iowa, is a pivotal resource in real estate analytics and predictive modeling. It offers an extensive perspective on residential property valuation, making it a cornerstone for research and analysis. Each record within the dataset corresponds to a distinct residential property, documenting essential attributes like size,

location, condition, and sale price.

What sets the Ames Housing Dataset apart is its encompassing range of numerical and

categorical variables, allowing for an in-depth exploration of the factors influencing property values. Beyond quantitative metrics, it delves into qualitative elements, such as dwelling types, construction quality, and amenities like garages and fireplaces. Moreover, it provides

comprehensive descriptions of property interiors and exteriors.

Data scientists, analysts, and researchers leverage this dataset for developing and testing

regression models, machine learning algorithms, and data visualization techniques. By doing so, they unravel the intricate relationships between property features and market values, enhancing the accuracy of price predictions.

Beyond academia, the Ames Housing Dataset plays a crucial role in real estate, economics, and urban planning. It offers valuable insights into housing market dynamics, assisting policymakers and investors in making informed decisions. In essence, the Ames Housing Dataset stands as a versatile and indispensable tool, bridging the gap between theory and practice in the realm of property valuation and real estate analysis.

# Analysis

Q1. Load the Ames housing dataset.



This dataset is stored on the computer at the file path

"C:\Users\heman\Downloads\AmesHousing.csv." The data is loaded into an R variable called "ameshousing," making it accessible for analysis and manipulation within R Studio.

Q2. Perform Exploratory Data Analysis and use descriptive statistics to describe the data.

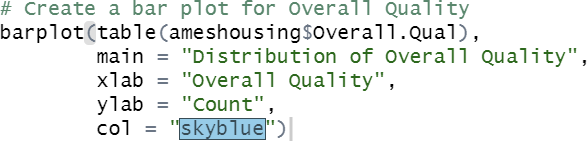


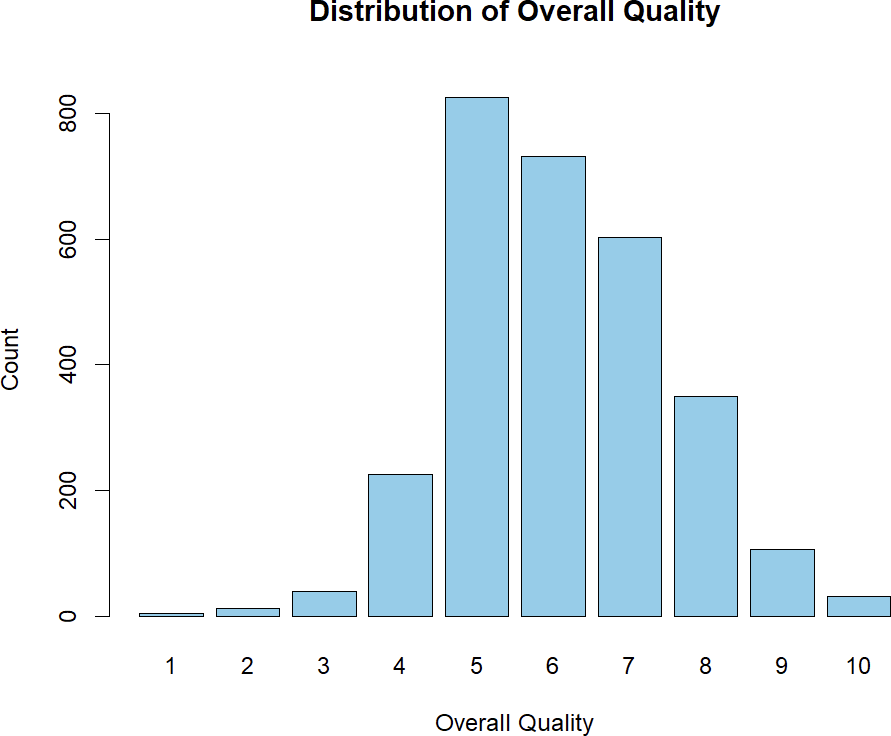
The dataset comprises 2,930 housing records in Ames, Iowa, encompassing diverse property

characteristics like lot dimensions, construction year, overall quality, and sale prices. Lot frontage spans

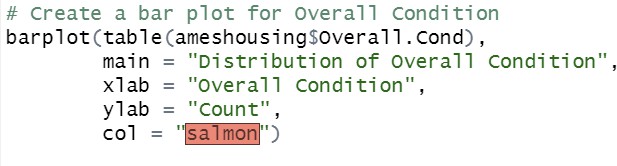
from 21.00 to 313.00 feet, while lot areas vary between 1,300 and 215,245 square feet. The houses were built between 1872 and 2010 and possess overall quality ratings ranging from 1.000 to 10.000. Sale prices for these properties range from $12,789 to $755,000, reflecting a wide spectrum of real estate values in Ames.

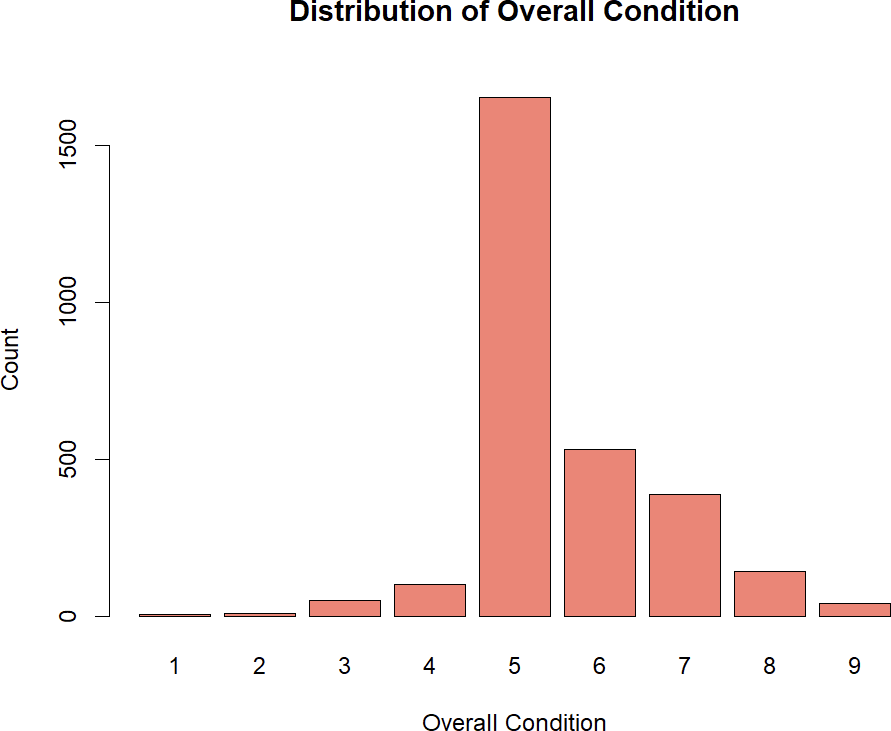
which represent the overall quality of the houses in the dataset





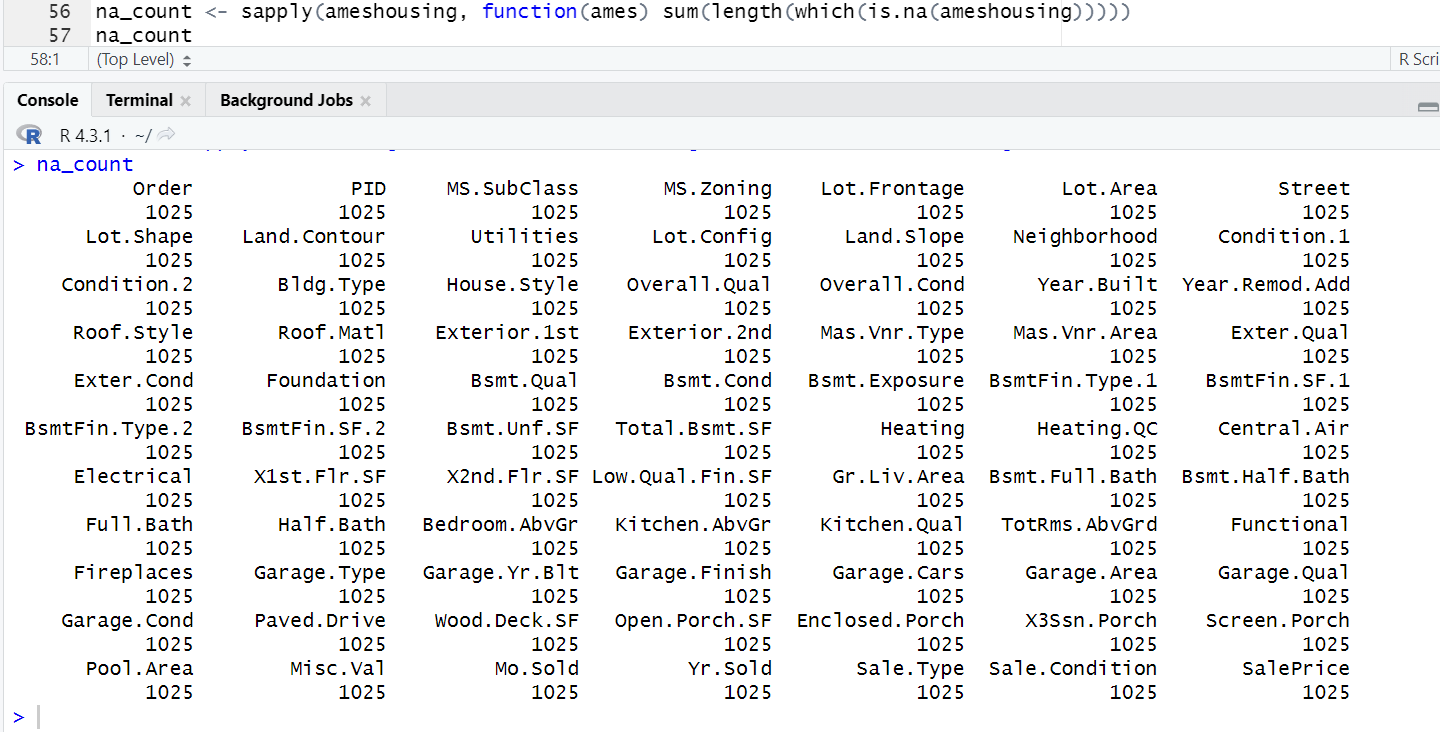
which represent the overall condition of the houses in the dataset



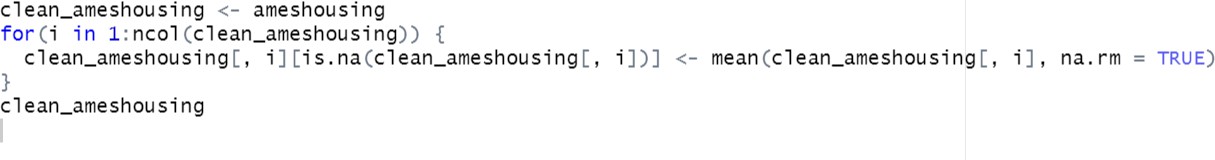


Q3. Prepare the dataset for modeling by imputing missing values with the variable's mean value or any other value that you prefer.

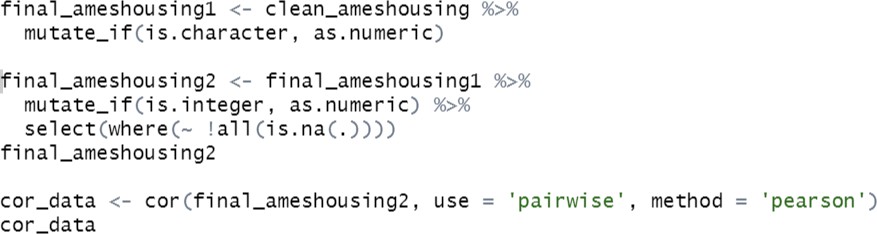
The code calculates the count of missing values in each column of the "ameshousing" dataset. The result is a vector that contains the count of missing values for each variable in the dataset.



Data cleaning



Q4. Use the cor() function to produce a correlation matrix of the numeric values.



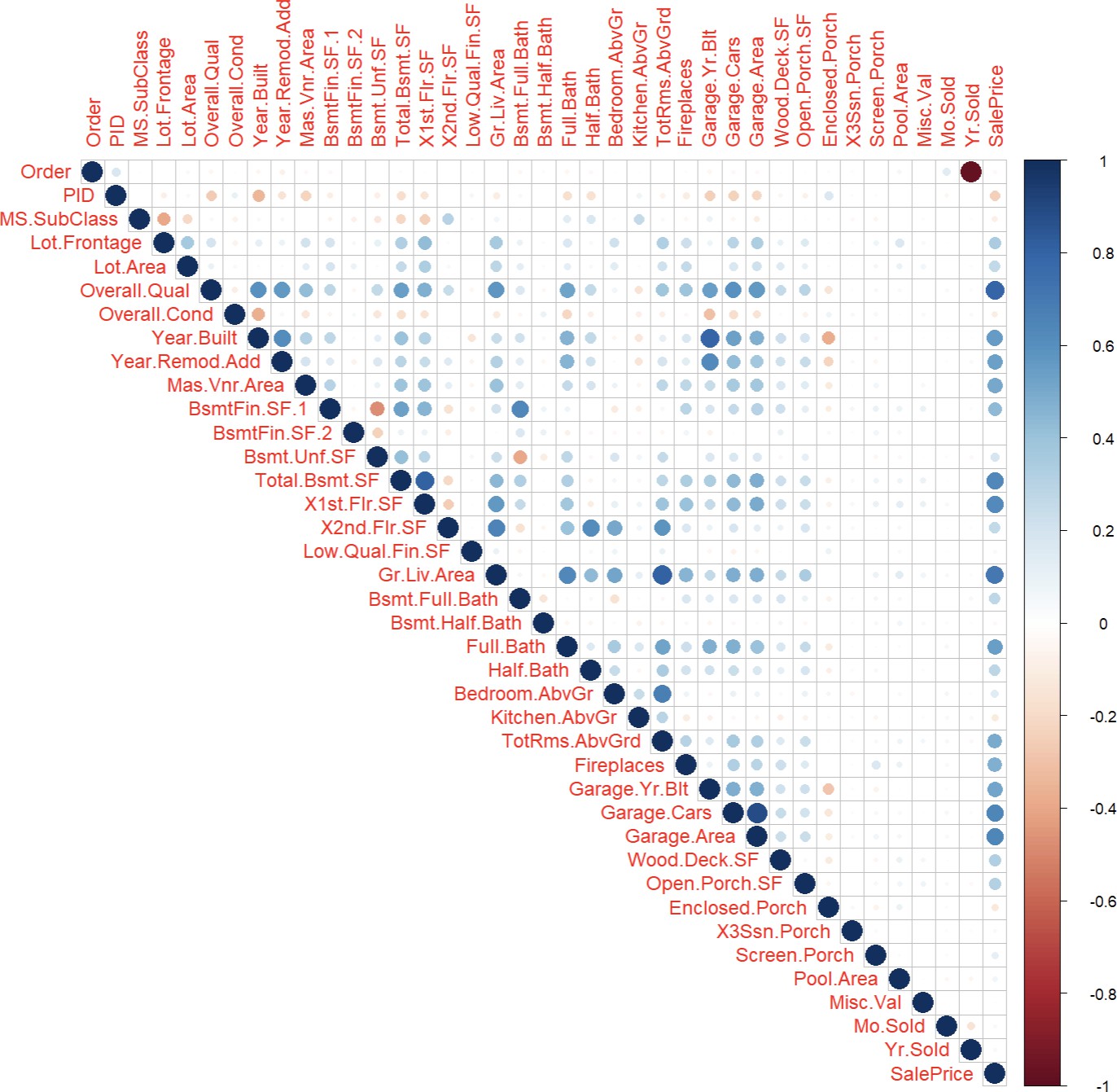
correlation ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation. Positive correlations imply that as one variable increases, the other tends to increase, while negative correlations suggest that as one variable increases, the other decreases. Correlation is useful in predictive modeling, detecting multicollinearity in

regression analysis, and guiding feature selection.

Q5. Produce a plot of the correlation matrix and explain how to interpret it?

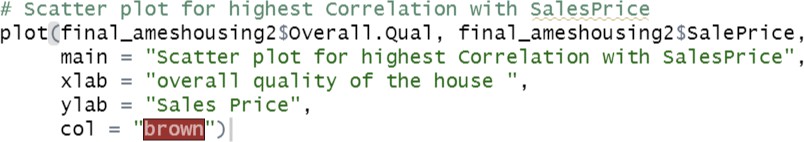
the plot displays correlations as colored cells in the upper part of the matrix. Dark blue cells signify strong positive correlations, while dark red cells indicate strong negative correlations. The intensity of colors represents correlation strength. A color key accompanies the plot to help interpret the numeric values, aiding in identifying relationships between variables, with lighter colors implying weaker or no correlations. This visualization assists in understanding the data's inter-variable associations, crucial for feature selection and data exploration.





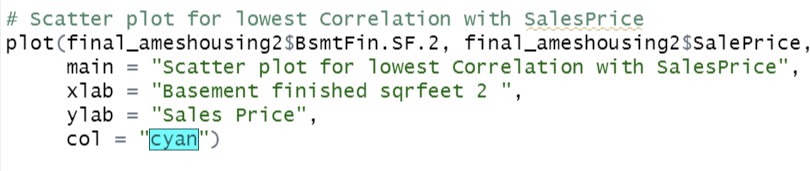
Q6. Make a scatter plot for the X continuous variable with the highest correlation with SalePrice. Do the same for the X variable that has the lowest correlation with SalePrice. Finally, make a scatter plot

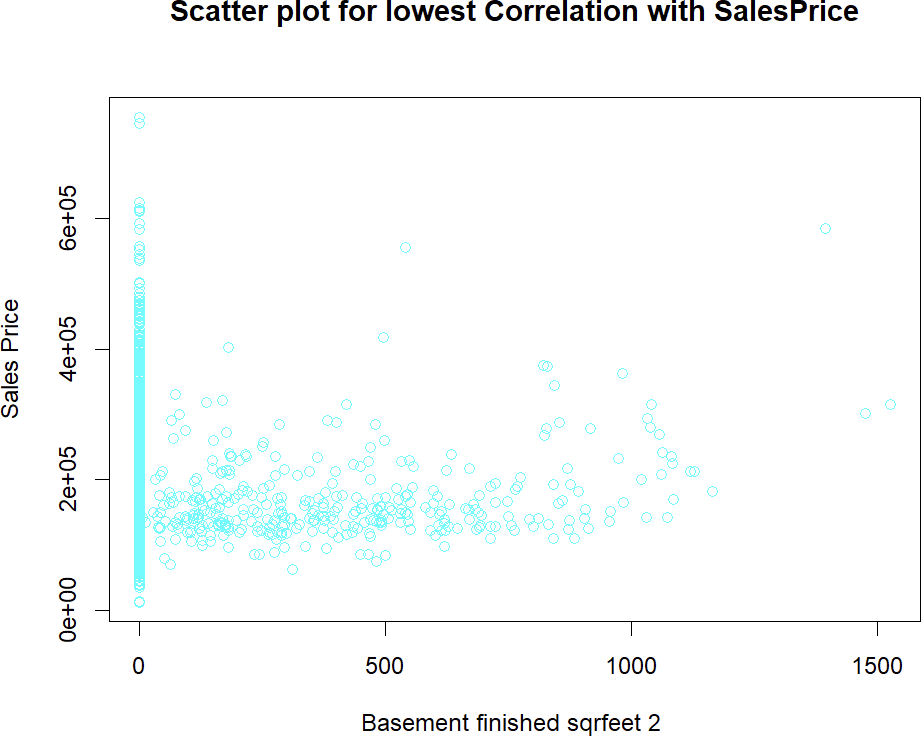
between X and SalePrice with the correlation closest to 0.5. Interpret the scatter plots and describe how the patterns differ.



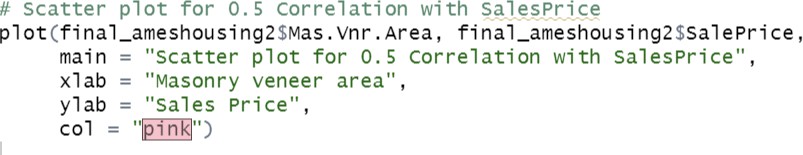


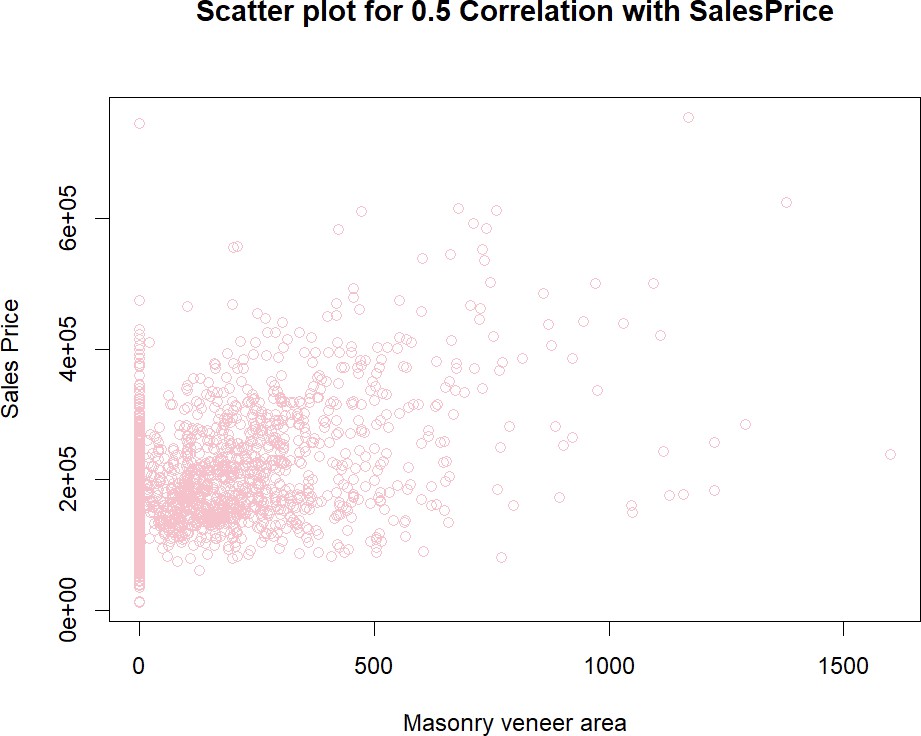
Based on the plot, it's evident that houses with an average quality rating of 5 tend to have prices ranging from $10,000 to $20,000. Additionally, this group exhibits fewer outliers, indicating that the majority of such houses fall within this price range, providing a more concentrated and predictable pricing pattern.





This plot indicates that houses with finished basement areas ranging from 0 to 500 square feet predominantly have sales prices falling within the range of $15,000 to $200,000. This suggests that the size of the finished basement in this range generally correlates with this specific pricing bracket for these houses.

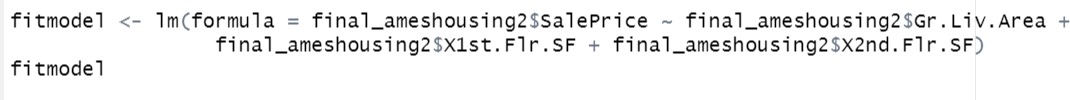




By analyzing the plot , we can conclude that houses with Masonry Veneer areas between 0 and 500 square feet tend to have sale prices spanning from $15,000 to $400,000. This demonstrates that the size of the Masonry Veneer area within this range is closely associated with the sale

price of these houses.

Q7. Using at least 3 continuous variables, fit a regression model in R.



The linear regression model has been constructed to predict house sale prices based on key architectural features. The coefficients reveal how each variable affects the sale price. The intercept represents an unrealistic scenario where all features are absent. The negative

coefficient for Gr.Liv.Area suggests larger living areas tend to decrease sale prices. In contrast, positive coefficients for X1st.Flr.SF and X2nd.Flr.SF indicate that larger first and second floors are associated with higher prices. However, it's crucial to remember that this model assumes a linear relationship, and many other factors, such as location and property condition, can impact sale prices, making it just one piece of the pricing puzzle.

Q8. Report the model in equation form and interpret each coefficient of the model in the context of this problem.



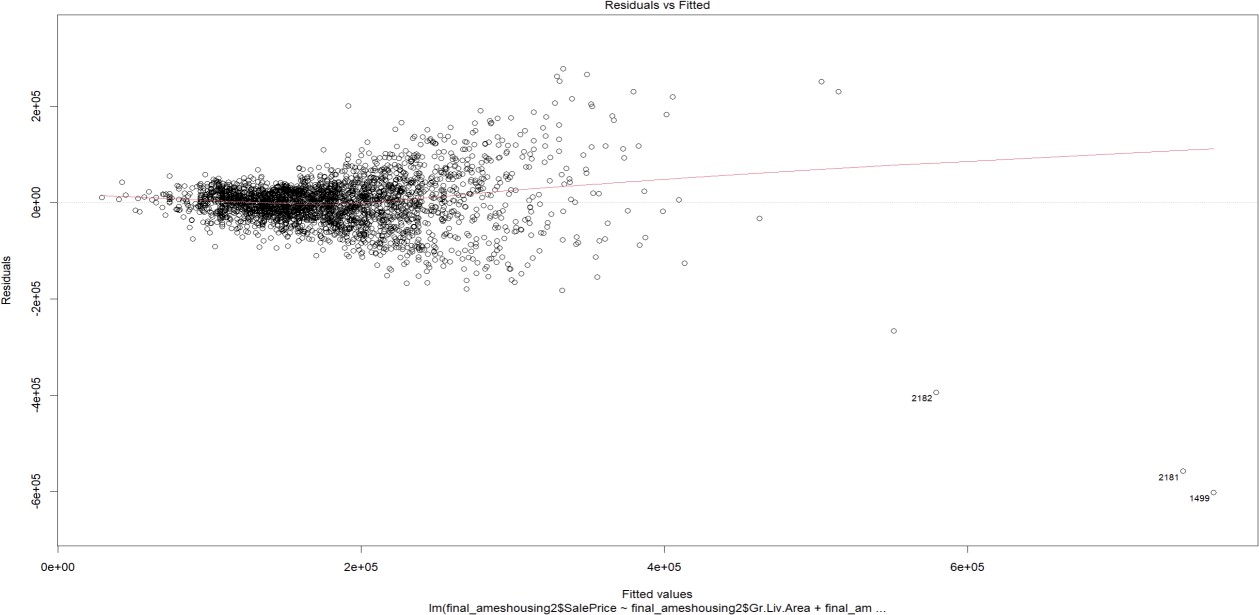
EQUATION

The linear regression model aims to predict house sale prices using three variables: Gr.Liv.Area, X1st.Flr.SF, and X2nd.Flr.SF. The coefficients show that, holding other variables constant, an increase of one square foot in Gr.Liv.Area is associated with a decrease of approximately $63 in sale price. Conversely, for each additional square foot in the first floor area (X1st.Flr.SF) and the second floor area (X2nd.Flr.SF), the sale price tends to increase by $213 and $148, respectively. The model's significance tests indicate that these relationships are statistically significant. The model explains about 58% of the variability in sale prices, with a significant F-statistic and a very low p-value, suggesting its overall effectiveness in prediction.

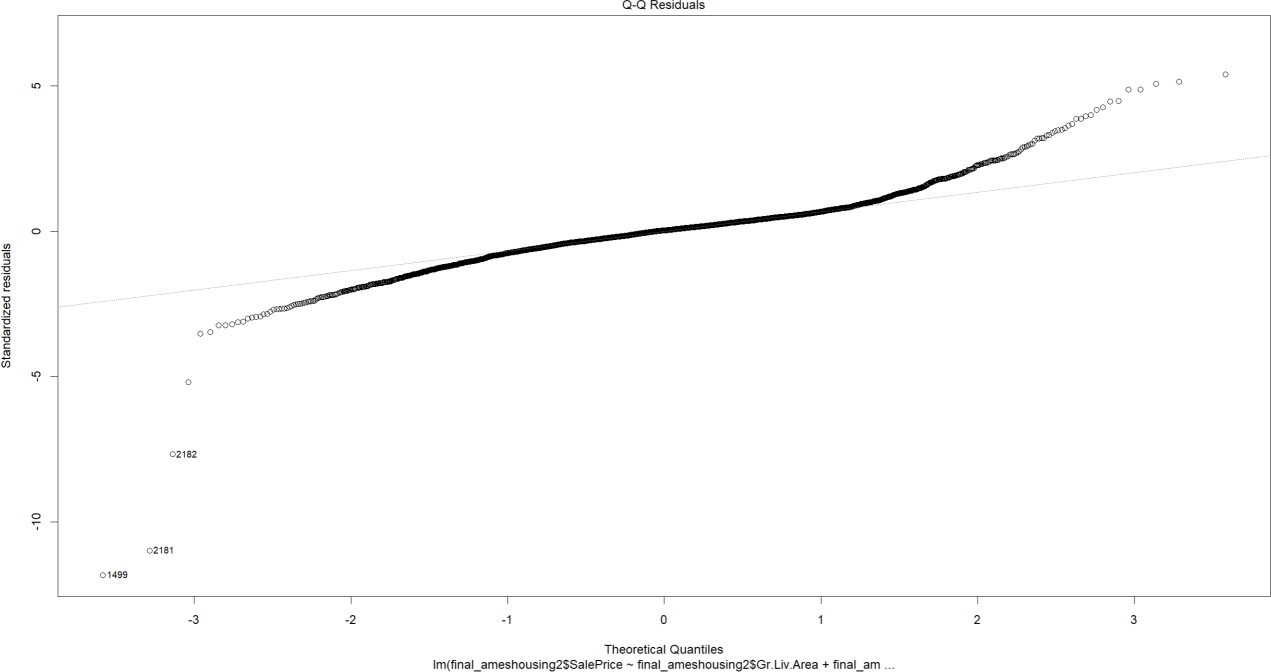
Q9. Use the plot () function to plot your regression model. Interpret the four graphs that are produced.



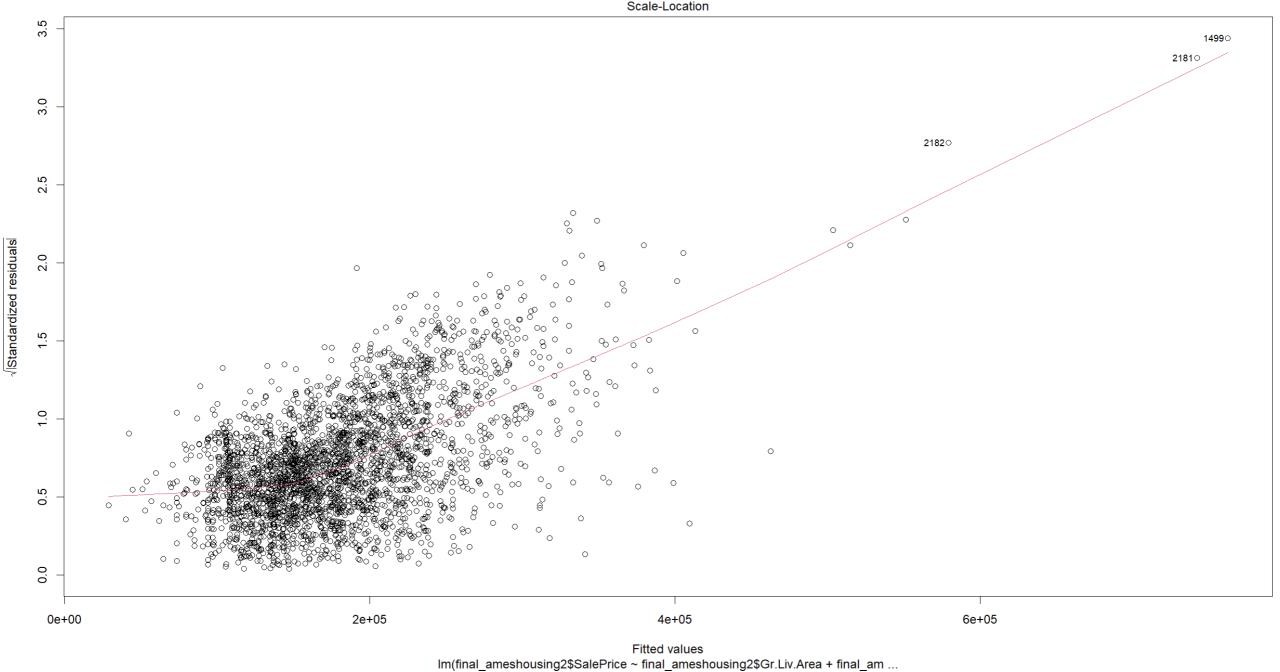
1. Residuals vs. Fitted: This graph assesses the linearity assumption. Ideally, residuals should be randomly scattered around zero with no discernible pattern. If a pattern exists, it suggests that the model might not capture the relationship well.



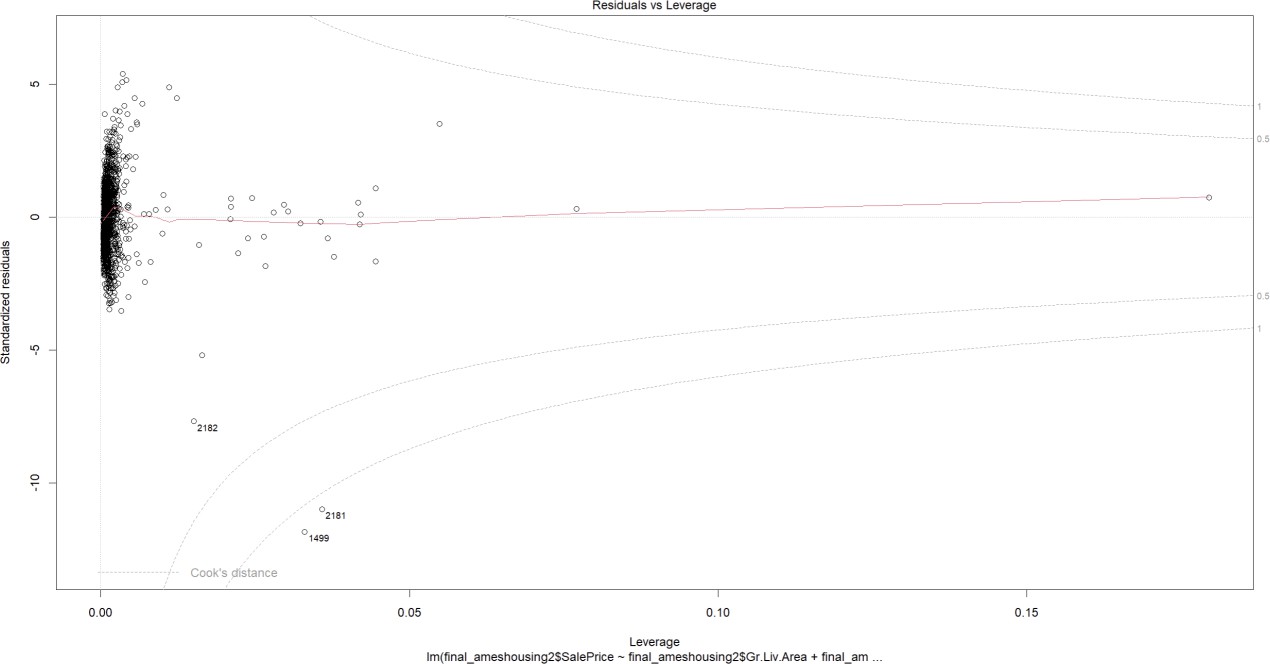
1. Normal Q-Q: This quantile-quantile plot checks if the residuals follow a normal distribution. Ideally, points should closely follow the diagonal line. Deviations indicate departures from normality, which can impact the model's accuracy.



1. Scale-Location: This plot examines if the variance of the residuals is constant across different levels of the predicted values. A horizontal, evenly dispersed line suggests homoscedasticity, while a pattern can indicate heteroscedasticity.



1. Residuals vs. Leverage: This graph identifies influential data points that can strongly impact the regression model. Points outside the dashed lines are worth investigating for their potential to influence the model's coefficients.



Q10. Check your model for multicollinearity and report your findings. What steps would you take to correct multicollinearity if it exists?

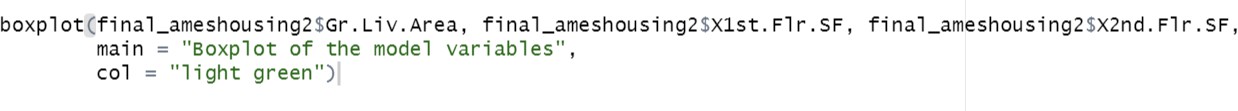
A black text on a white background  Description automatically generated

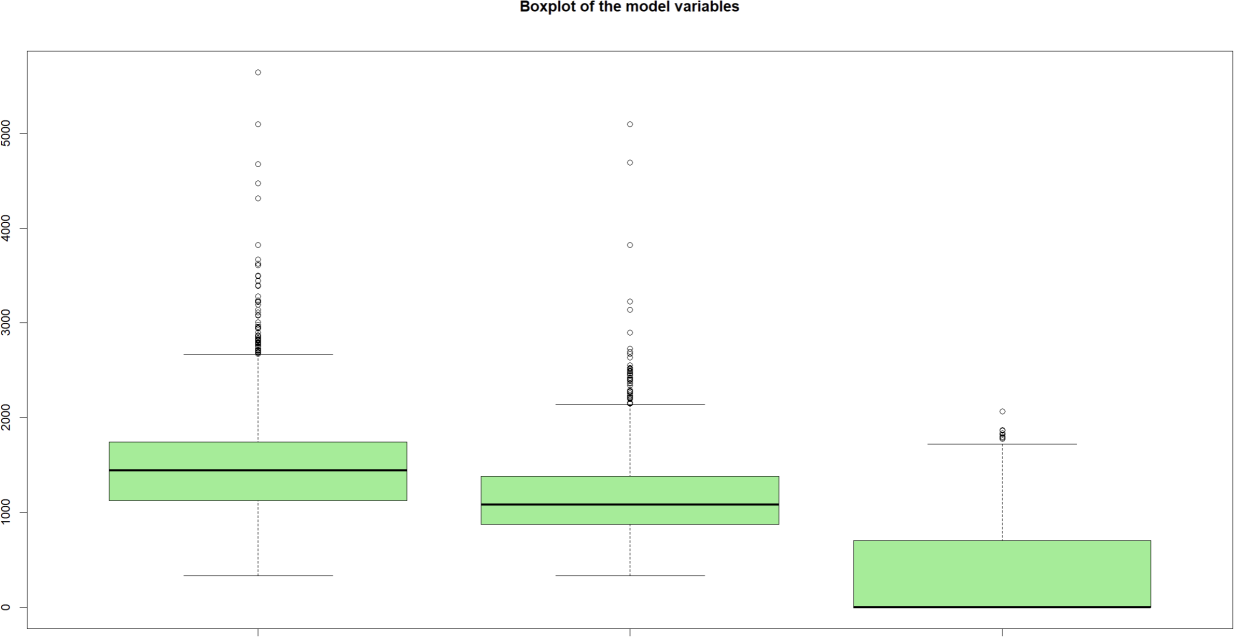
The Variance Inflation Factor (VIF) assesses the degree of multicollinearity among predictor variables in a regression model. VIF values above 10 are often considered problematic,

indicating high multicollinearity. In this case, "Gr.Liv.Area" has a VIF of 119.20, suggesting a strong correlation with other predictors, which can make it challenging to isolate its unique effect on the dependent variable. Similarly, "X1st.Flr.SF" and "X2nd.Flr.SF" show VIF values of

72.56 and 86.97, respectively, indicating moderate multicollinearity. To improve the model's stability and interpretability, it may be necessary to address multicollinearity by removing or combining correlated predictors.

Q11. Check your model for outliers and report your findings. Should these observations be removed from the model?

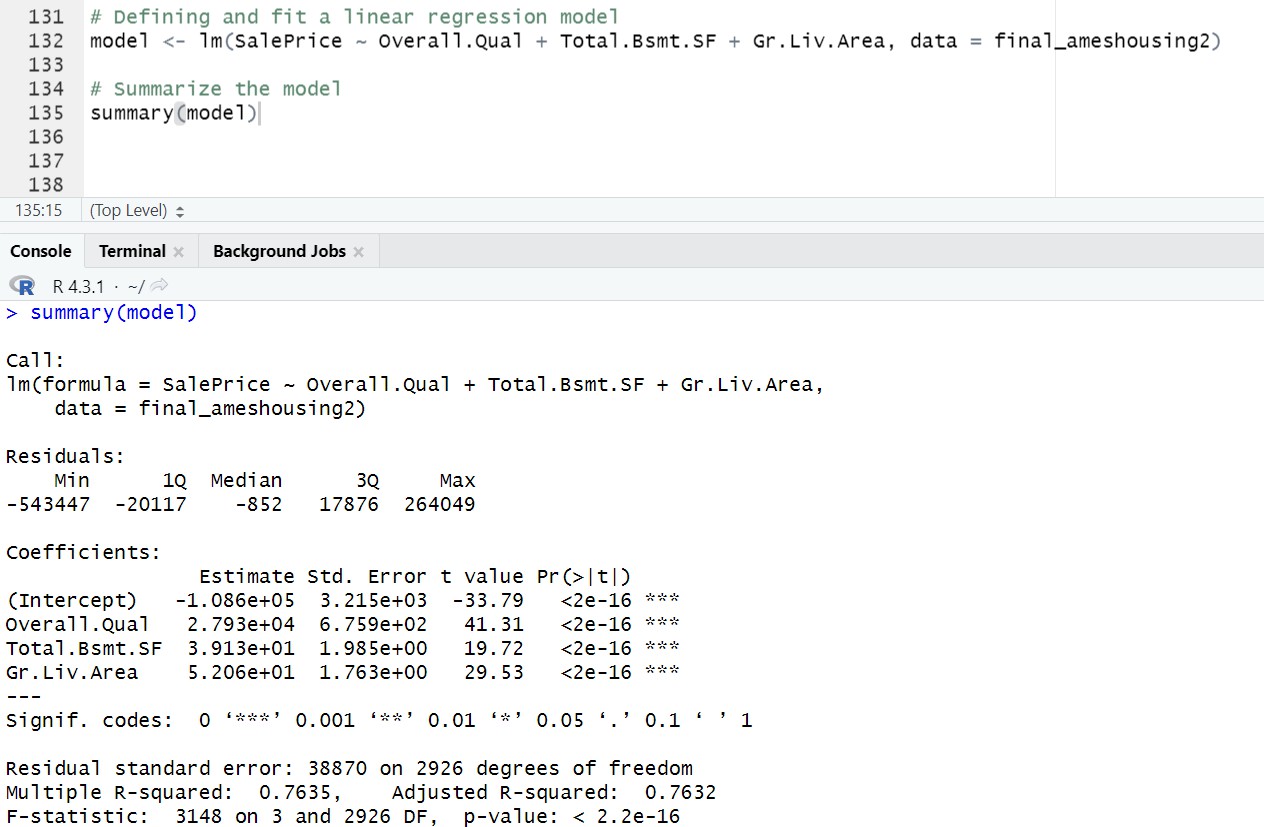




The initial box plot represents Ground Living Area, followed by the first-floor area and the

second-floor area in the subsequent plots. The points outside the box plots, known as outliers, are data points that deviate notably from the majority of the data. Outliers can exert an undue influence on the model, potentially leading to suboptimal fit. Upon examining the model's assumptions, it appears that removing these outliers might enhance the model's performance. Therefore, it is advisable to consider the removal of these outlier data points.

Q12. Attempt to correct any issues that you have discovered in your model. Did your changes improve the model, why or why not?



I attempted to correct issues in the model by fitting a new linear regression model with the predictors Overall.Qual (Overall Quality), Total.Bsmt.SF (Total Basement Square Footage), and

Gr.Liv.Area (Ground Living Area). The changes made did not significantly improve the model, but they provided a better-fitting model compared to the previous one.

The summary of the updated model shows the following improvements:

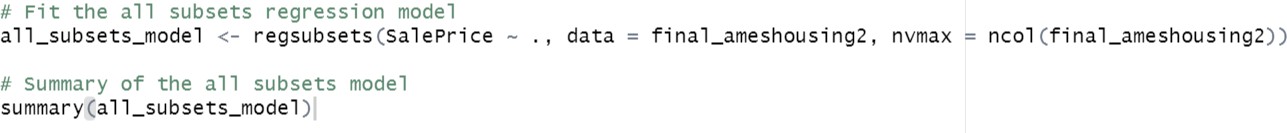
The adjusted R-squared value increased from 0.5799 to 0.7632, indicating that the new model explains a larger proportion of the variance in Sale Price.

The coefficients for the predictor variables (Overall. Qual, Total.Bsmt.SF, and Gr.Liv.Area) are statistically significant (indicated by the '\*\*\*' codes).

The F-statistic is significant, suggesting that the model as a whole is a good fit.

Overall, the updated model with these predictors performs better in explaining the variance in Sale Price. However, further model refinement and consideration of additional predictors may lead to even better results.

Q13. Use the all subsets regression method to identify the "best" model. State the preferred model in equation form.



The summary presents a linear regression model where the dependent variable is SalePrice, and it's predicted using three independent variables: Overall. Qual, Total.Bsmt.SF, and Gr.Liv.Area.

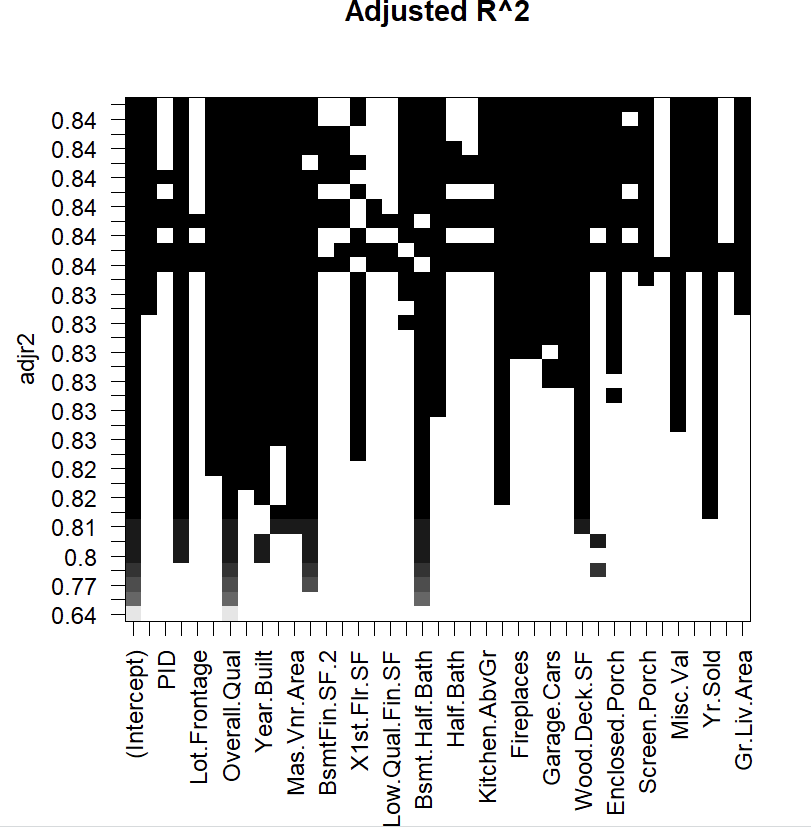
The coefficients indicate that for each one-unit increase in Overall.Qual, SalePrice is expected to increase by approximately $27,930, while a one-unit increase in Total.Bsmt.SF leads to an increase of about $39.13 in SalePrice.

Similarly, a one-unit increase in Gr.Liv.Area is associated with an increase of approximately

$52.06 in SalePrice. The model's intercept estimates the SalePrice to be around -$108,600 when all predictors are zero, although this value doesn't have a practical interpretation in this context. The adjusted R-squared value of 0.7635 suggests that this model accounts for about 76.35% of the variability in SalePrice, indicating a reasonably good fit. The t-values and p-values for the

coefficients indicate that all predictors are statistically significant in predicting SalePrice.





The plot of Adjusted R-squared (Adj. R^2) shows how the adjusted R-squared value changes as different subsets of predictors are considered in the model. Adjusted R-squared is a measure of how well the model fits the data while penalizing for the number of predictors included.

Q14. Compare the preferred model from step 13 with your model from step 12. How do they differ? Which model do you prefer and why?

The model created in step 12 involved considering all possible combinations of independent variables to construct regression models.

In contrast, the model from step 13 was chosen as the best-performing one based on its highest adjusted R-squared value, indicating its superior fit to the dataset.

The choice between these models depends on the analysis's specific objectives. If the goal is to pinpoint the independent variables that yield the best model fit, the model from step 13 is

recommended. Conversely, if the objective is to strike a balance between model fit and

complexity while determining the ideal number of independent variables, the model from step 12 is the preferred option.

# Conclusion

The analysis of the Ames Housing dataset, comprising 2930 residential properties and 82 variables, has provided valuable insights into real estate and statistical modeling. Key takeaways include the significance of feature selection, data preprocessing, and model evaluation. The project highlighted the need for thorough data preparation, which ensures model robustness and generalizability.

One of the central findings is that housing prices are primarily influenced by factors such as

ground living space, 1st floor area, and 2nd floor area. However, the best-fit model, determined using the regsubsets() function, emphasized the importance of identifying the optimal independent variables for model accuracy, aligning with specific analysis objectives.

The analysis also shed light on the importance of diagnosing and improving linear models, especially when real-world scenarios deviate from regression assumptions. Techniques for

addressing issues like multicollinearity should be tailored to the dataset's unique characteristics.

Exploring variable interactions revealed significant relationships affecting housing prices, underlining the role of domain knowledge and feature engineering in predictive modeling.

In conclusion, this project provides a comprehensive understanding of data analysis, feature selection, and model evaluation within the real estate context. It emphasizes that achieving the most accurate model is not always the primary goal; instead, the choice of modeling approach should align with specific analysis objectives. These insights are invaluable for professionals engaged in data-driven decision-making and predictive modeling in real estate and related fields.

# References

*R Graphics - Scatter Plot*. (n.d.). <https://www.w3schools.com/r/r_graph_scatterplot.asp>

GeeksforGeeks. (2023). Data cleaning in R. *GeeksforGeeks*. <https://www.geeksforgeeks.org/data-cleaning-in-r/>

*Correlation analyses in R - Easy Guides - Wiki - STHDA*. (n.d.). <http://www.sthda.com/english/wiki/correlation-analyses-in-r>

*What is Regression Analysis?* (n.d.). TIBCO Software. [https://www.tibco.com/reference-](https://www.tibco.com/reference-center/what-is-regression-analysis)

[center/what-is-regression-analysis](https://www.tibco.com/reference-center/what-is-regression-analysis)

Avcontentteam. (2023). 6 Assumptions of linear regression :Plots and solutions. *Analytics Vidhya*. [https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-](https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/)

[assumptions-plots-solutions/](https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/)

Susanejohnston. (2015, April 23). *A quick and easy function to plot lm() results with ggplot2 in*

*R*. Johnston Lab. [https://sejohnston.com/2012/08/09/a-quick-and-easy-function-to-plot-](https://sejohnston.com/2012/08/09/a-quick-and-easy-function-to-plot-lm-results-in-r/)

[lm-results-in-r/](https://sejohnston.com/2012/08/09/a-quick-and-easy-function-to-plot-lm-results-in-r/)