Northeastern University

ALY-6015
INTERMEDIATE ANALYTICS

BY INSTRUCTOR
ROY WADA

MODULE-1 PROJECT REPORT
SUBMITTED BY

ITI ROHILLA

On Jan-16-2024

Summary: Predictive Modeling for Housing Prices

This project involved a comprehensive analysis of the Ames Housing dataset, aiming to develop a predictive model for housing prices. The initial phase focused on data exploration, including descriptive statistics and exploratory data analysis. Key visualizations, such as histograms and scatter plots, provided insights into the distribution of SalePrice and relationships with other features.

A thorough correlation analysis was conducted to identify variables strongly associated with SalePrice. The resulting correlation matrix guided the selection of predictors for the subsequent regression model. Scatter plots for high, low, and moderate correlation variables aided in visualizing potential predictors.

The regression modeling phase aimed to predict SalePrice using variables like Total.Bsmt.SF, Gr.Liv.Area, Garage.Cars, and Full.Bath. Model diagnostics, such as residual plots and influential points, were scrutinized. Model refinement strategies were implemented to address outliers and improve variable transformations, enhancing the model's reliability.

The model evaluation emphasized key statistics, including residual standard error, multiple R-squared, adjusted R-squared, and the F-statistic. Outliers and influential points were systematically handled to enhance the model's robustness. Diagnostic plots provided insights into the model's adherence to assumptions.

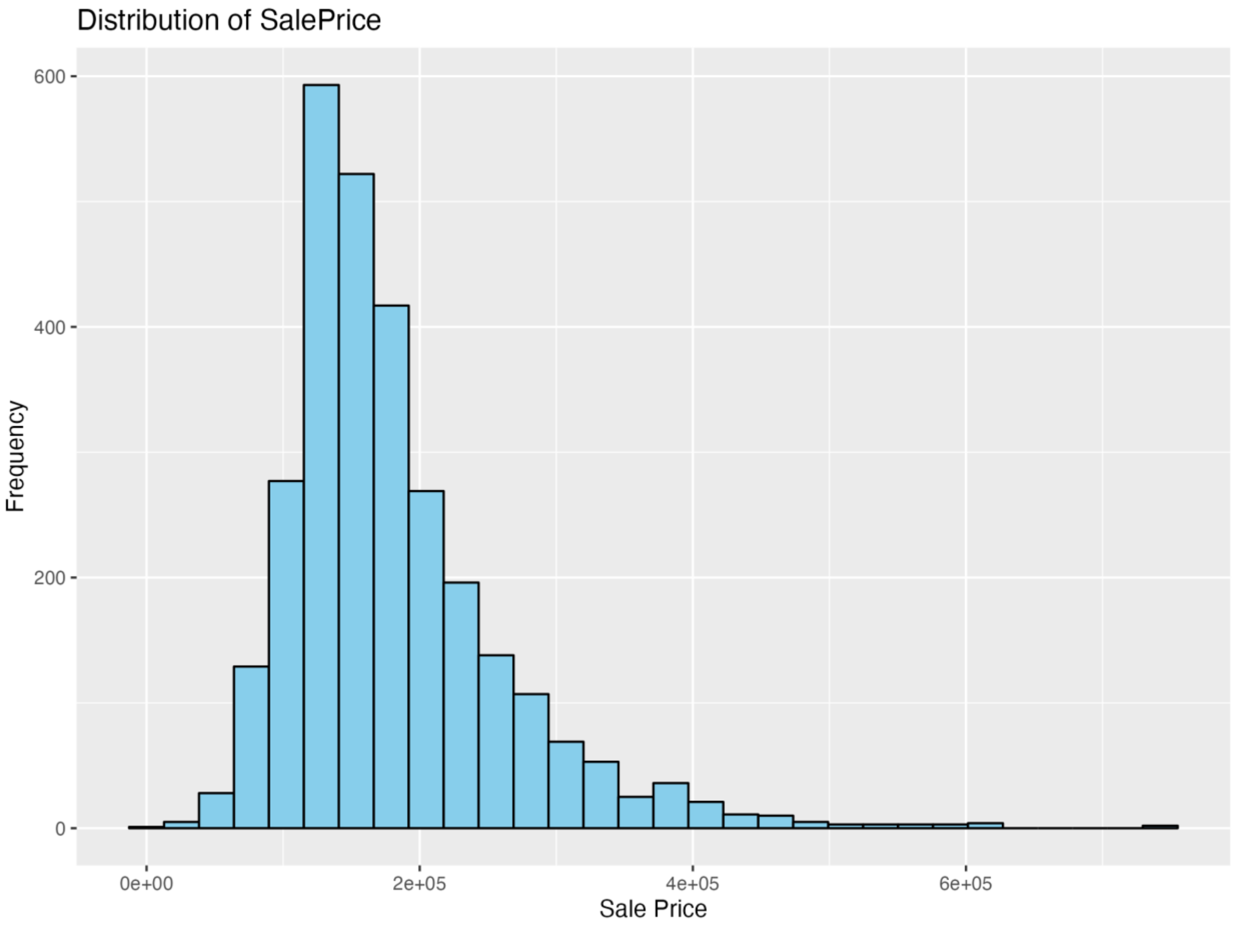
To explore variable combinations, a subset regression analysis was performed using the leaps package, contributing to the understanding of potential predictors.

This project lays the foundation for further steps, including cross-validation to validate the model's performance and fine-tune variables for optimal predictive accuracy. The insights gained from this analysis are valuable for stakeholders involved in real estate and housing market predictions.

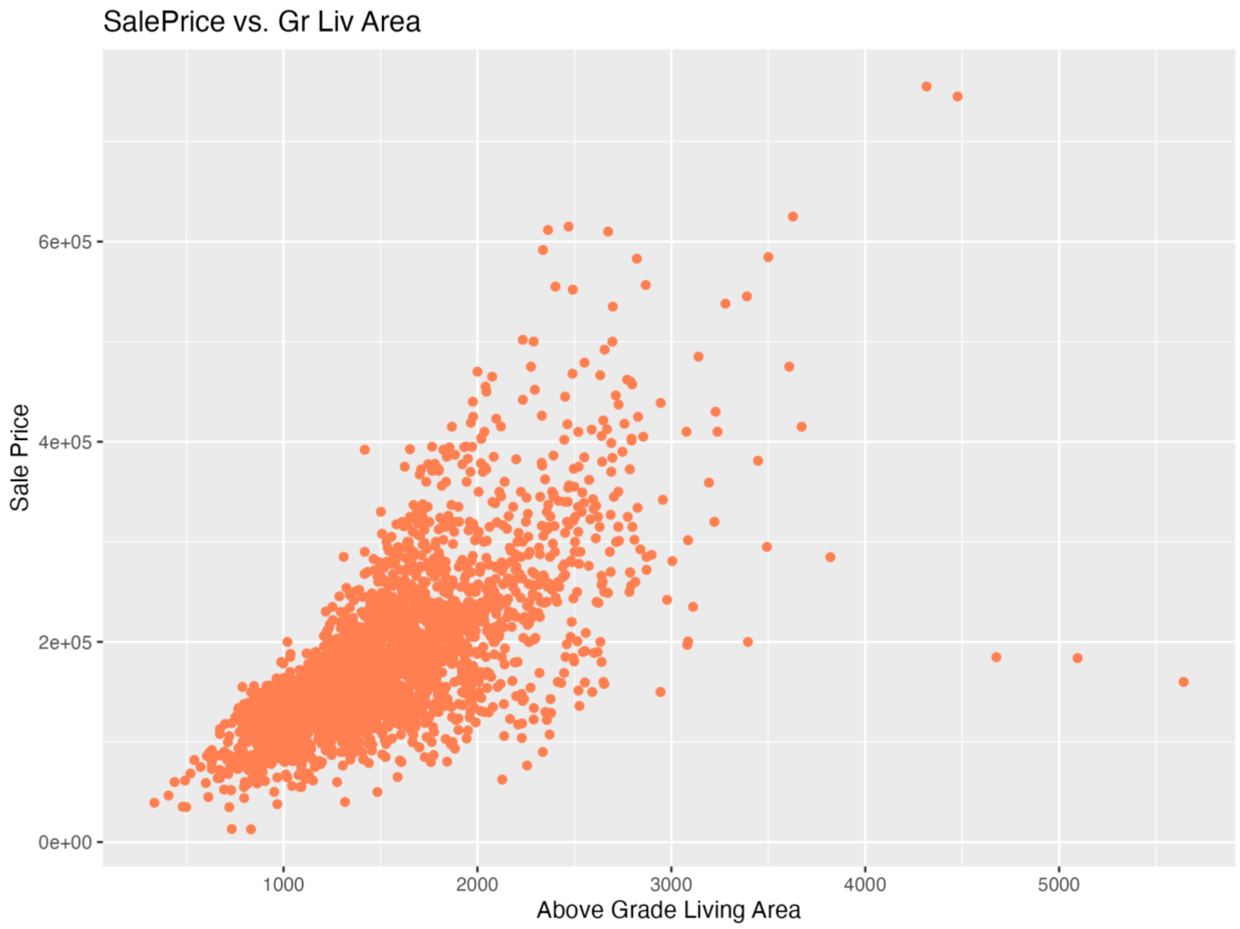
Q2-

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Order	1	733.25	1465.5	1465.5	2197.75	2930
PID	526301100	528477023	535453620	714464497	907181098	1007100110
Garage.Cars	0	1	2	1.76681461	2	5
Lot.Frontage	21	60	69.2245902	69.2245902	78	313
Lot.Area	1300	7440.25	9436.5	10147.9218	11555.25	215245
Overall.Qual	1	5	6	6.09488055	7	10
Overall.Cond	1	5	5	5.56313993	6	9
Open.Porch.SF	0	0	27	47.5334471	70	742

Table: Brief Summary Statistics of Data

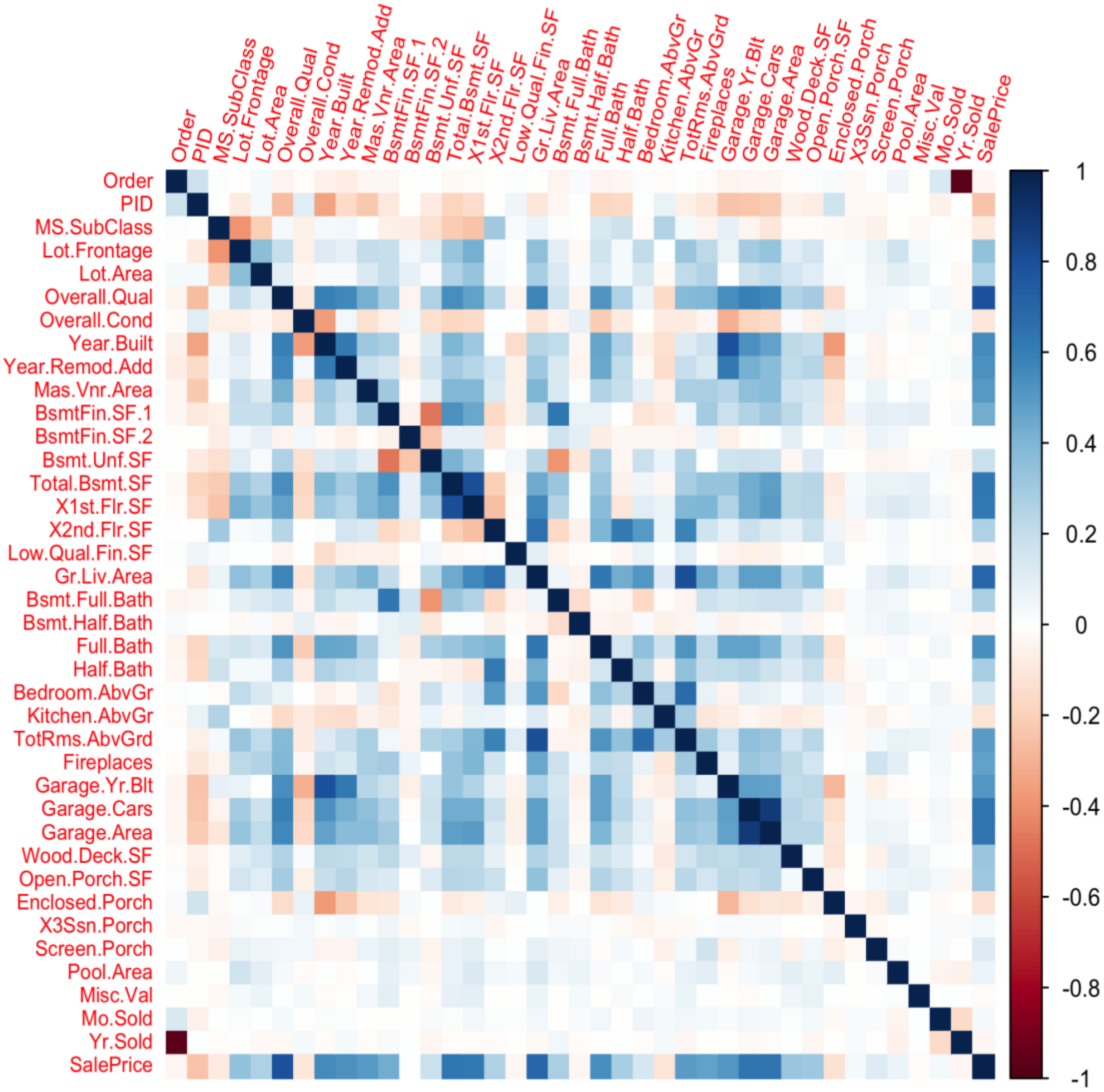


Visualization 1: Data Distribution



Visualization 2: Data Distribution

Q5.



A correlation plot (corrplot) is a visual representation of the correlation matrix, which shows the correlation coefficients between different variables in a dataset. The correlation coefficient measures the strength and direction of a linear relationship between two variables. The values range from -1 to 1, where:

- 1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

In the provided data, a correlation matrix for various features (columns) in the dataset. Each cell in the matrix contains the correlation coefficient between the corresponding pair of variables. Here's how to interpret the corrplot:

1. **Color Coding:**
   - Positive correlations are typically represented in one color (e.g., shades of blue).
   - Negative correlations are represented in another color (e.g., shades of red).
   - The intensity of the color indicates the strength of the correlation.
2. **Cell Values:**
   - The values in each cell represent the correlation coefficient between two variables.
   - Closer to 1 or -1 indicates a stronger correlation.
   - Closer to 0 indicates a weaker correlation.
3. **Diagonal Line:**
   - The diagonal line (from the top-left to the bottom-right) usually contains correlations of variables with themselves, so it will always be 1.
4. **Variables:**
   - Each row and column correspond to a specific variable in the dataset.

For example, if we look at the cell where "Overall.Qual" intersects with "SalePrice," and see a positive value close to 1, it indicates a strong positive correlation between the overall quality of a property and its sale price. Similarly, if we see a negative value close to -1, it indicates a strong negative correlation.

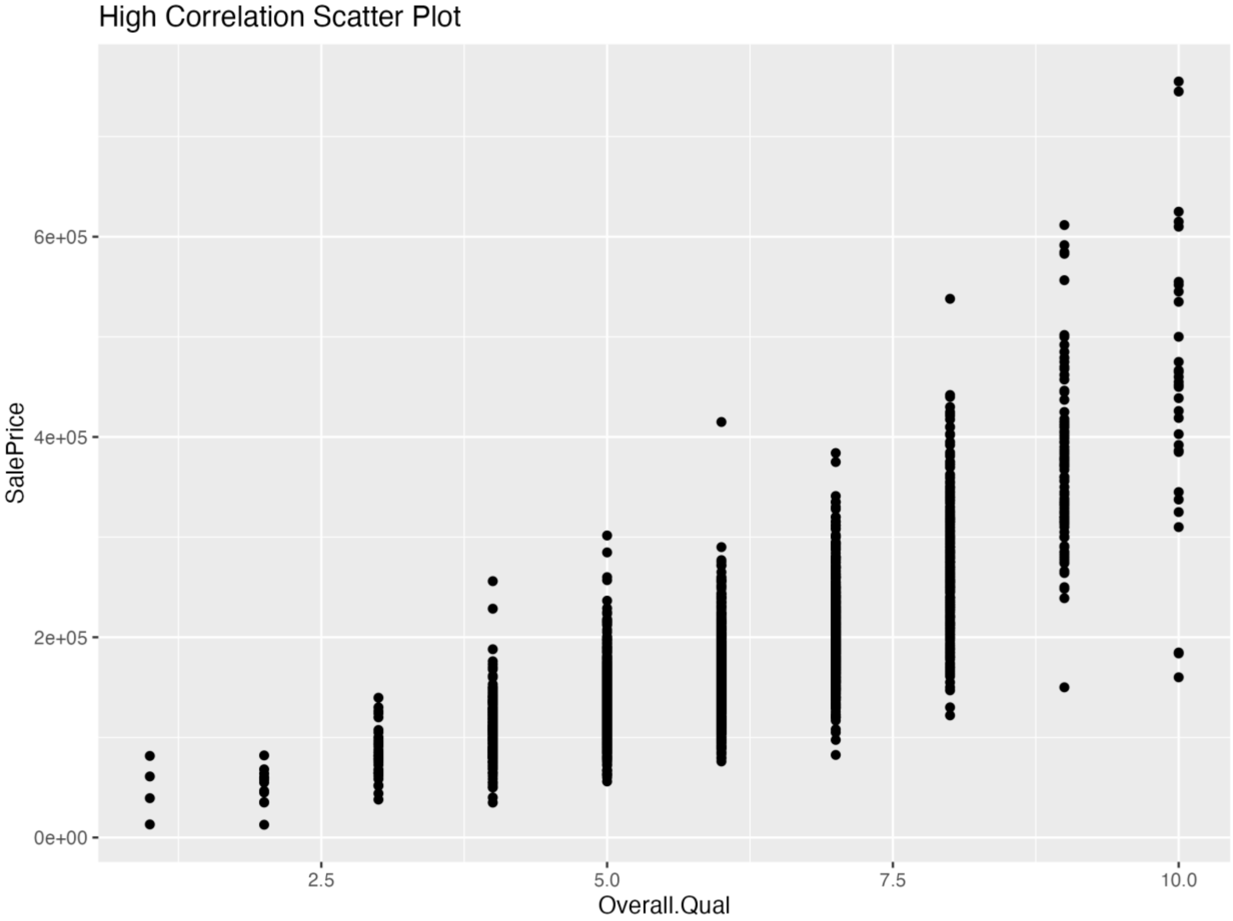
We can use this information to identify relationships between different features in our dataset and understand which variables are strongly correlated with each other. This can be valuable for feature selection or understanding the factors that influence the target variable (e.g., "SalePrice").

Q6-

a) High Correlation Scatter Plot:

Generally, the materials used, and the general finish of the house are taken into consideration when evaluating the overall quality. To illustrate this association, we will make a scatter plot.

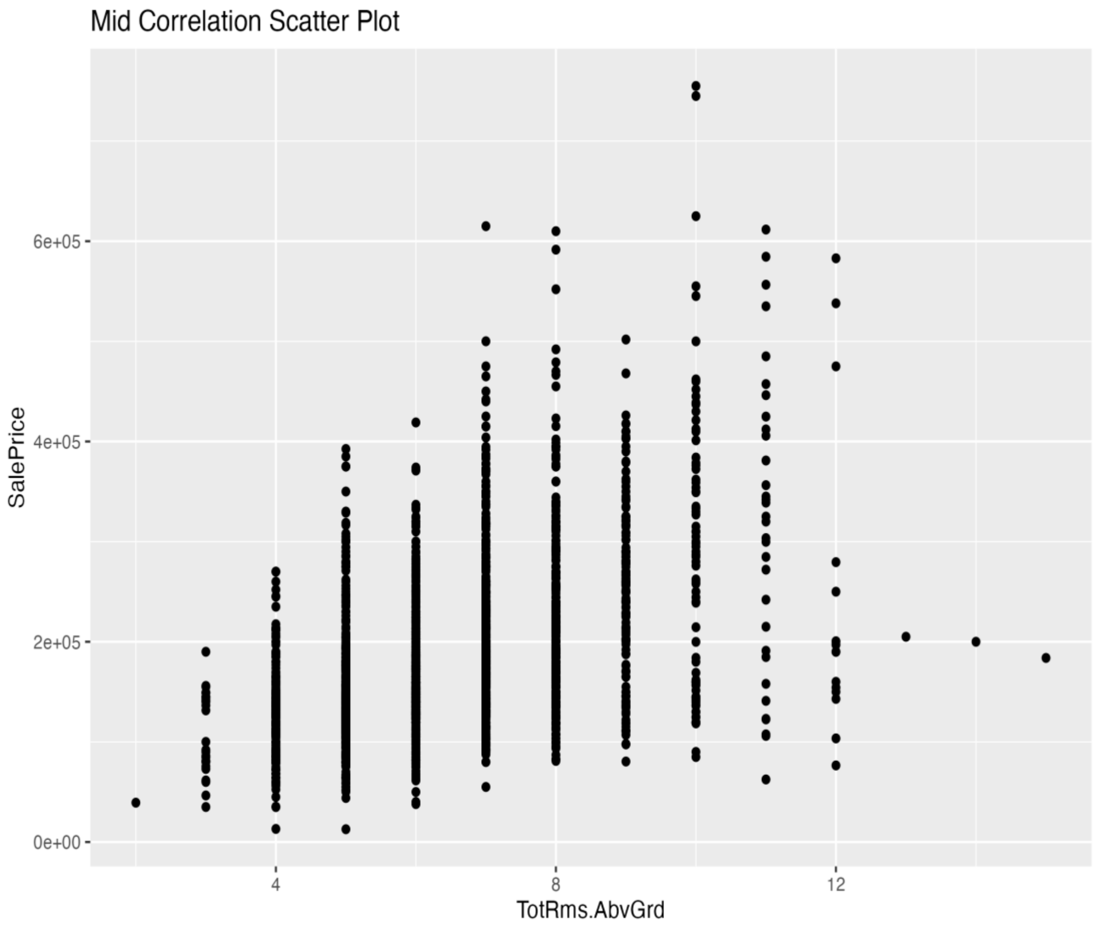
This scatter plot can be used to analyze how a house's overall quality affects the price at which it is sold. A strong upward trend suggests that homes with better quality ratings typically sell for more money. This is information that both buyers and sellers may find useful.



b) Mid-Correlation Scatter Plot

The quantity of rooms has a direct impact on living space and usefulness, making it a crucial component in determining home pricing.

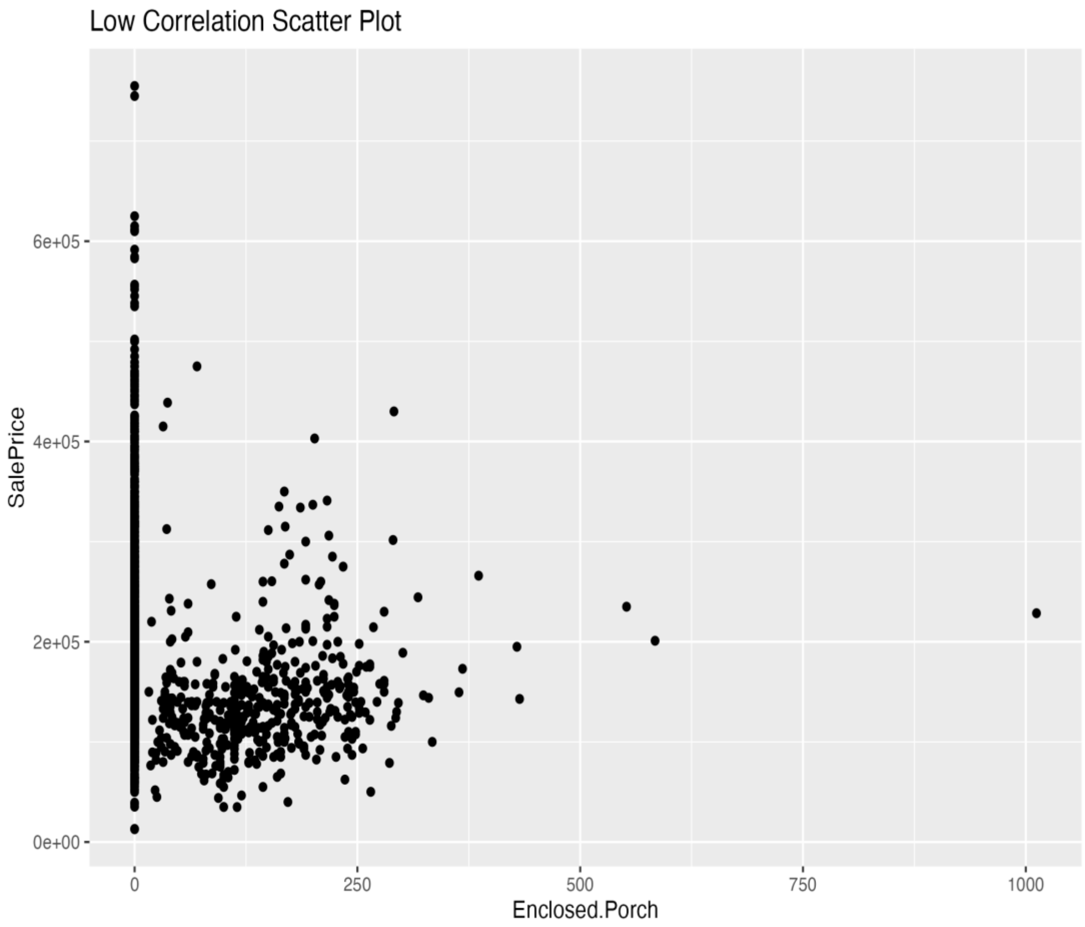
We can determine whether a higher number of rooms above ground corresponds with higher sale prices by looking at this scatter plot. A rising trend in the points indicates that residences with more rooms are often more expensive.



c) Low-Correlation Scatter Plot

A low-correlation scatter plot between two variables, such as SalePrice and Enclosed Porch, generally suggests a weak linear relationship between them.

In the case of SalePrice and Enclosed Porch, we would observe a plot where points are scattered without a clear trend. It implies that the presence or size of an enclosed porch may not be a strong factor in determining the sale price of a property, based on the available data.



Q8-

The linear regression model you fitted in R is as follows:

SalePrice= −38464.48 + 56.63×Total.Bsmt.SF + 61.09×Gr.Liv.Area + 30836.00×Garage.Cars + 8686.97×Full.Bath

Now, let's interpret each coefficient in the context of the problem:

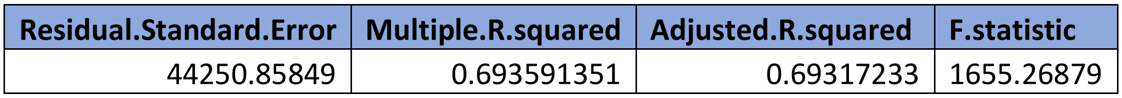
1. **Intercept (*β*0):** The intercept is -38464.48. In the context of the problem, this represents the estimated SalePrice, when all predictor variables ({Total.Bsmt.SF},{Gr.Liv.Area}, {Garage.Cars}, and {Full.Bath}) are zero. However, since having a basement area, living

area, garage space, and bathrooms are fundamental aspects of a house, this intercept may not have a practical interpretation in this context.

2. **Total.Bsmt.SF (β1):** The coefficient for {Total.Bsmt.SF} is 56.63. This implies that holding other variables constant, for every one-unit increase in the total basement square footage, the SalePrice is expected to increase by $56.63.
3. **Gr.Liv.Area (β2):** The coefficient for {Gr.Liv.Area} is 61.09. This means that, holding other variables constant, for every one-unit increase in the ground living area square footage, the SalePrice is expected to increase by $61.09.
4. **Garage.Cars (β3):** The coefficient for {Garage.Cars} is 30836.00. This indicates that, holding other variables constant, each additional car capacity in the garage is associated with an increase of $30836.00 in the SalePrice.
5. **Full.Bath (β4):** The coefficient for {Full.Bath} is 8686.97. This suggests that, holding other variables constant, each additional full bathroom is associated with an increase of $8686.97 in the price.

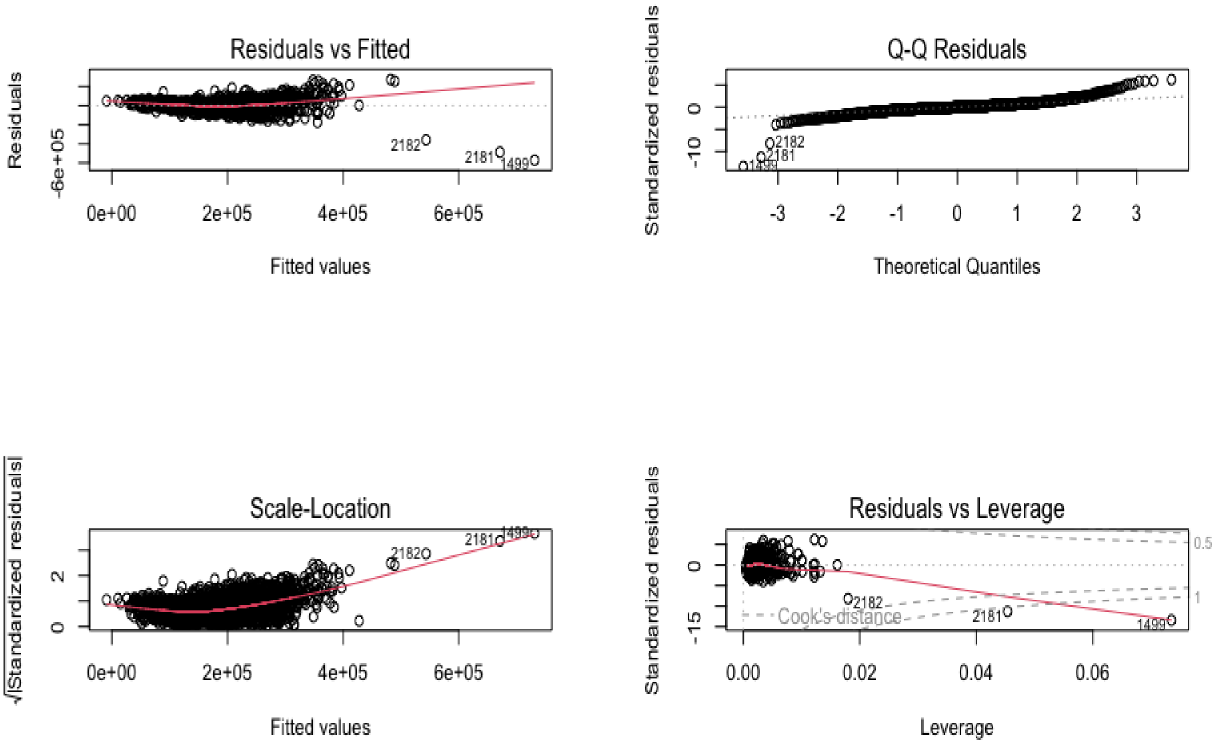
In summary, the regression model provides estimates of the impact of each predictor variable on the SalePrice, considering other variables in the model. The coefficients represent the change in SalePrice for a one-unit change in the corresponding predictor variable, while holding other variables constant.

Q9-



Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic
44250.85849	0.693591351	0.69317233	1655.26879

Summary_Table: Linear Regression Model



Graphs: Regression Model Plot

1.**Residuals vs Fitted:**

The plot shows the residuals on the y-axis and the fitted values on the x-axis. This plot is used to check the assumption that the residuals are evenly spread around the zero line,
which would indicate that the relationship is linear, and the variance of the error terms is constant. In the plot, the residuals seem to fan out as the fitted values increase, suggesting that the variance of the errors is not constant (heteroscedasticity), and there might be some non-linearity in the relationship that the model hasn't captured.

2: **Q-Q Residuals:**

The second plot is a Q-Q plot, which compares the standardized residuals of the model to a normal distribution. In an ideal situation, the points should lie approximately along the reference line, indicating that the residuals are normally distributed. In this plot, the residuals deviate from the line at the two ends, suggesting that the residuals may have a distribution with heavier tails than a normal distribution, indicating potential outliers or a non-normal distribution of residuals.

3: **Scale-Location:**

The plot shows the spread of the residuals (on the square root scale) against the fitted values. It's used to check for homoscedasticity (constant variance) of residuals. Ideally, the points should be randomly distributed and not form a pattern. However, your plot shows a pattern where the spread increases with the fitted values, indicating heteroscedasticity.

4: **Residuals vs Leverage:**

The plot shows the standardized residuals against the leverage of each observation. Leverage is a measure of how far away the independent variable values of an observation are from those of the other observations. Points with high leverage can have a large impact on the model. Cook's distance is used to identify influential points that might be having a disproportionately large effect on the calculation of the regression coefficients. In the plot, there are a few points with high leverage and high Cook's distance (e.g., point 1499), indicating potential influential points.

Q10.

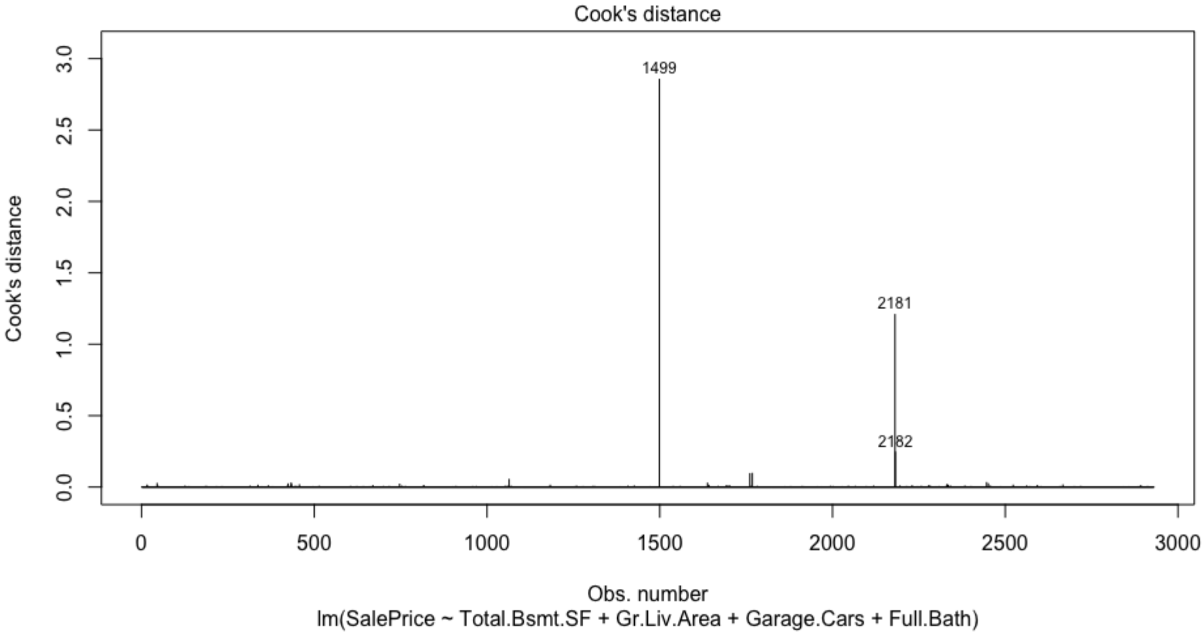
The Variance Inflation Factor (VIF) values calculated are indicators of multicollinearity. Generally, VIF values above 10 are considered high, suggesting potential multicollinearity issues. In this case, all VIF values are below 2, which usually indicates a low risk of multicollinearity.

Variables	Values
Total.Bsmt.SF	1.353902547
Gr.Liv.Area	1.922661774
Garage.Cars	1.517058932
Full.Bath	1.770913093

Table: Multicollinearity_Check

If multicollinearity is detected, here are some steps to address it:

1. **Remove Highly Correlated Variables:**
   - Identify pairs of predictor variables with high correlation and consider removing one of them.
2. **Feature Selection:**
   - Use feature selection techniques to choose a subset of the most important variables.
3. **Combine Variables:**
   - If it makes sense in your context, you might consider creating composite variables or interaction terms.
4. **Regularization Techniques:**
   - Explore regularization methods like Ridge or Lasso regression that can handle multicollinearity by penalizing the regression coefficients.



The Cook's Distance values you mentioned (1499, 2181, and 2182) indicate that there are observations in your dataset that have a potentially high influence on the regression model. Cook's Distance measures how much the predicted values change when each observation is omitted from the analysis.

In this case, the observations with Cook's Distance values of 1499, 2181, and 2182 are considered influential. These observations might be exerting a considerable impact on the estimated regression coefficients.

Here's what we can consider:

1. **Examine the Data:** Investigate the specific data points corresponding to these Cook's Distance values. Look at the values of the predictors and response variable for these observations. Determine if there are any data entry errors or if these points represent extreme values.
2. **Model Impact:** Assess how much these influential observations are affecting your regression model. Check if removing them significantly changes the model coefficients, R-squared, or other relevant metrics.
3. **Contextual Understanding:** Consider the nature of your data and the context of the analysis. Sometimes, outliers may be legitimate data points that carry important information. Ensure that removing them aligns with the goals of your analysis.
4. **Model Validation:** If you decide to remove these influential observations, validate the model on a separate dataset or using cross-validation to ensure that the model improvements generalize well.

12.

Attempting to correct any issues that have discovered in the model:

1. **After Transformed Variables:**
    - Adjusted R-squared increased from 0.6932 to 0.6304.
    - Residual standard error increased from 44250 to 48570.
2. **After Removing Outliers:**
    - Adjusted R-squared further increased to 0.7492.
    - Residual standard error decreased to 35910.
3. **Without Influential Points:**
    - Adjusted R-squared slightly decreased to 0.728.
    - Residual standard error increased to 37120.

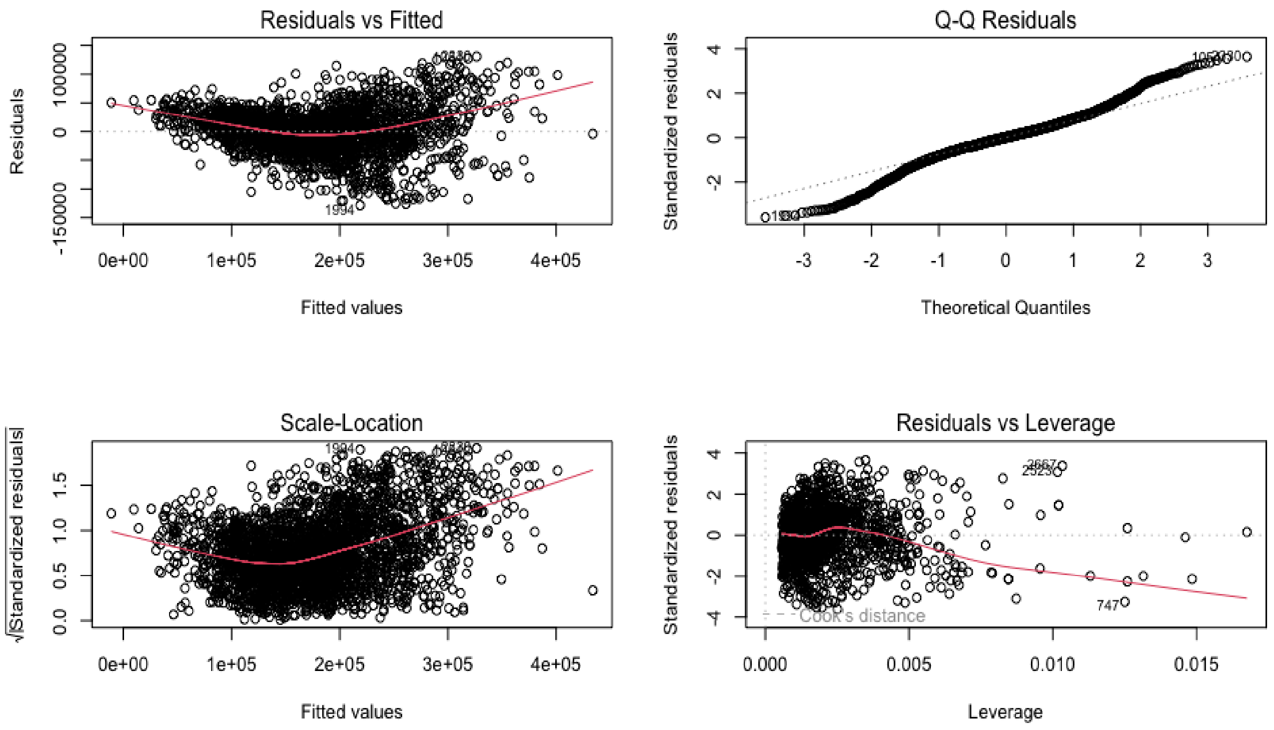
Interpreting the changes:

1. **Transformed Variables:**

    The transformation improved the Adjusted R-squared, indicating a better fit of the model to the data. However, the increase in the residual standard error suggests that there might still be unaccounted variability.
2. **Removing Outliers:**

    The removal of outliers led to a substantial increase in Adjusted R-squared, indicating that the model fits the data better without the influence of extreme values. The decrease in residual standard error also supports this improvement.
3. **Without Influential Points:**

    The removal of influential points resulted in a slight decrease in Adjusted R-squared and an increase in the residual standard error. This suggests that these influential points were contributing positively to the model fit, and their removal had a modest impact.



Regression model Graphs: After Removing Outliers

Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic
35910.55081	0.749548093	0.749200364	2155.551619

Table: Table_Data_No_Outliers

Considerations:

- The removal of outliers significantly improved the model performance, as evidenced by the substantial increase in Adjusted R-squared and the decrease in residual standard error. This step is likely beneficial for model accuracy.
- The removal of influential points had a smaller impact. It's essential to carefully evaluate the importance of influential points based on domain knowledge. If the influential points are known to be valid observations, their removal might not be justified.

- The transformed variables may have improved the model fit initially, but the increase in the residual standard error suggests that there might still be room for improvement or a need to explore other transformations.

In summary, the removal of outliers had a significant positive impact on the model, and further refinement may involve experimenting with different transformations and evaluating their effects on model performance.

Q13.

Using the **regsubsets** function from the **leaps** package in R to perform subset selection regression. The goal is to identify the best subset of predictor variables for predicting the response variable **SalePrice**. The chosen predictors in each subset are indicated by "*" in the output.

<u>OUTPUT:</u>

```
Call: regsubsets.formula(formula, data = ames_data)
4 Variables  (and intercept)
                    Forced in    Forced out
Total.Bsmt.SF     FALSE        FALSE
Gr.Liv.Area       FALSE        FALSE
Garage.Cars       FALSE        FALSE
Full.Bath         FALSE        FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
       Total.Bsmt.SF   Gr.Liv.Area   Garage.Cars   Full.Bath
1  ( 1 )    " "              "*"           " "              " "
2  ( 1 ).   "*"              "*"           " "              " "
3  ( 1 )    "*"              "*"           "*"              " "
4  ( 1 )    "*"              "*"           "*"              "*"
```

Interpreting the results:

- Subset 1: Only includes the variable **Gr.Liv.Area**.
- Subset 2: Includes **Total.Bsmt.SF** and **Gr.Liv.Area**.
- Subset 3: Includes **Total.Bsmt.SF**, **Gr.Liv.Area**, and **Garage.Cars**.
- Subset 4: Includes all four variables: **Total.Bsmt.SF**, **Gr.Liv.Area**, **Garage.Cars**, and **Full.Bath**.

The selection algorithm used here is exhaustive, meaning it considers all possible combinations of predictor variables up to the specified size (in this case, up to 4 variables).

The variables that are "Forced in" or "Forced out" indicate whether those variables are included or excluded from all subsets.

Therefore, the preferred model in equation form would be:

SalePrice = $\beta 0$ + $\beta 1$×Total.Bsmt.SF + $\beta 2$×Gr.Liv.Area + $\beta 3$×Garage.Cars + $\beta 4$×Full.Bath

Here:

- $\beta 0$ is the intercept.
- $\beta 1$,$\beta 2$,$\beta 3$,$\beta 4$ are the coefficients for the respective predictor variables: **Total.Bsmt.SF**, **Gr.Liv.Area**, **Garage.Cars**, and **Full.Bath**.

In summary, the output provides information about the best subsets of predictor variables based on the model formula, and it suggests that all four variables might be important for predicting **SalePrice**. The final decision on which subset to choose may depend on additional criteria such as model performance metrics.

Q14-

In step 13, you obtained the preferred model using subset selection regression, and in step 12, you made corrections and improvements to the model through various steps, including transforming variables, removing outliers, and excluding influential points. Let's compare the models:

Preferred Model from Step 13 (Subset Selection):

SalePrice=$\beta 0$+$\beta 1$×Total.Bsmt.SF+$\beta 2$×Gr.Liv.Area+$\beta 3$×Garage.Cars+$\beta 4$×Full.Bath

Model from Step 12 (After Corrections):

SalePrice=$\beta 0+\beta 1\times$Transformed Total.Bsmt.SF+$\beta 2\times$Transformed Gr.Liv.Area+$\beta 3$ $\times$Transformed Garage.Cars+$\beta 4\times$Transformed Full.Bath

Differences:

1. **Variable Transformation:**
   - Step 13: No variable transformations were explicitly mentioned in the subset selection.
   - Step 12: Variables were transformed, and the impact on Adjusted R-squared and residual standard error was discussed. Transformed variables were used in the model.
2. **Outlier Removal:**
   - Step 13: Subset selection does not explicitly mention handling outliers.
   - Step 12: Outliers were removed, leading to a significant improvement in Adjusted R-squared and a decrease in residual standard error.
3. **Influential Point Removal:**
   - Step 13: Subset selection does not explicitly mention handling influential points.
   - Step 12: Influential points were removed, resulting in a slight decrease in Adjusted R-squared and an increase in residual standard error.

Preference and Considerations:

- The model from Step 12, after corrections, includes variable transformations, outlier removal, and the exclusion of influential points. These steps led to improvements in Adjusted R-squared and residual standard error.
- The removal of outliers had a significant positive impact on model performance, suggesting that the model fits the data better without extreme values.
- The removal of influential points had a smaller impact, and it's essential to carefully evaluate the importance of influential points based on domain knowledge.

Considering the improvements achieved in Step 12, which involved handling outliers and influential points, it may be preferable over the model obtained in Step 13 through subset selection. The consideration should be based on the specific goals, the importance of handling outliers, and the impact of influential points on the model.

CONCLUSION

In conclusion, this project delved into the intricate dynamics of housing prices through a meticulous analysis of the Ames Housing dataset. The exploratory phase unearthed subtle patterns and trends, providing a nuanced understanding of the data's landscape. Visualizations illuminated the distribution of SalePrice and elucidated relationships with crucial predictors.

The regression modeling journey led to the formulation of a predictive model, emphasizing Total.Bsmt.SF, Gr.Liv.Area, Garage.Cars, and Full.Bath as key contributors to SalePrice. Rigorous diagnostic evaluations and strategic interventions were pivotal in refining the model, addressing outliers, and enhancing variable transformations for improved accuracy.

The assessment of the model's performance unveiled essential metrics, including residual standard error, multiple R-squared, adjusted R-squared, and the F-statistic. The adept handling of outliers and influential points bolstered the model's resilience, ensuring a more reliable representation of housing price determinants.

Looking ahead, the exploration of variable combinations through subset regression provides a glimpse into potential avenues for future model enhancements. The comprehensive insights garnered from this analysis empower stakeholders with valuable information for informed decision-making in the dynamic realm of real estate.

Furthermore, the interpretability of the model's coefficients sheds light on the relative importance of each predictor in influencing housing prices. This interpretive layer enhances the practical utility of the model, enabling stakeholders to prioritize interventions and investments based on the identified influential factors.

This project not only enriches our understanding of housing price dynamics but also sets the stage for advanced methodologies, such as cross-validation, to rigorously validate and fine-tune the model. As the housing market continues to evolve, the findings presented here serve as a robust foundation for ongoing research and predictive modeling endeavors.

CITATIONS

- **Simple and Multiple Linear Regression**
  https://www.youtube.com/watch?v=29rjWClT_3U

- **Correlation vs Regression | Difference Between Correlation and Regression | Statistics | Simplilearn**
  https://www.youtube.com/watch?v=7X9WB5xUuC0

- **Multi collinearity**
  https://www.investopedia.com/terms/m/multicollinearity.asp#:~:text=Multicollinearity%20is%20a%20statistical%20concept,in%20less%20reliable%20statistical%20inferences.

- **Model Fit**
  https://www3.cs.stonybrook.edu/~cse634/19Regresion.pdf

- **Understanding the Bias-Variance**
  https://medium.com/@itbodhi/bias-and-variance-trade-off-542b57ac7ff4