Module 1-R Practice

Statistical Outputs

ALY6015-INTERMEDIATE ANALYTICS

NORTHEASTERN UNIVERSITY

DEVIKA PATIL
MAJOR-PROJECT MANAGEMENT
DATE OF SUBMISSION: 06-24-2023
Prof. Zhi He

Contents

Introduction

This study utilizes the Ames Housing dataset to explore and predict house prices in Ames, Iowa, employing various data manipulation and modeling techniques. The analysis follows these key steps:

1.Data Acquisition and Preprocessing:

- Relevant libraries for data handling, visualization, and modeling are imported.
- The Ames Housing dataset is loaded as a Data Frame.
- Initial checks are performed to confirm data structure and identify potential issues.

2.Data Cleaning and Imputation:

- Missing values are addressed through several steps:
  - Replacing empty strings with "NA" to ensure proper handling.
  - Dropping columns with a high percentage of missing values (greater than 80%) to avoid introducing excessive bias.
  - Imputing missing values in remaining columns with the mean for numeric features. This commonly used technique assumes that missing values are randomly distributed around the mean.
  - Eliminating columns with a remaining high percentage of missing values (greater than 40%) to maintain data integrity.
  - Removing rows containing any missing values after imputation to ensure complete data for the modeling stage.

3.4Feature Selection and Exploration:

- Unnecessary columns like "PID" and "Order" are removed to focus on relevant features.
- A correlation matrix is generated for the remaining numeric features, providing insights into the relationships between them.
- A heatmap visualizes these correlations, aiding in identifying features potentially influencing house prices.

4.Model Building and Diagnostics:

- Based on the analysis, features exhibiting strong correlations with the target variable "SalePrice" are selected for model building.
- A linear regression model is fit using these selected features to establish a relationship between them and the house price.
- The model summary is analyzed to understand the coefficients, p-values, and R-squared value, which measures the model's explanatory power.
- Diagnostic plots are generated to assess the model's assumptions, such as the normality of errors and homoscedasticity (constant variance).
- Variance Inflation Factors (VIF) are calculated to check for multicollinearity, which can negatively impact the model's performance.

5.Multicollinearity

No multicollinearity is found as all values are <5 or 10, any predictor variables have high VIF values (> 5 or 10), it indicates multicollinearity. In such cases, you can take the following steps to correct multicollinearity:

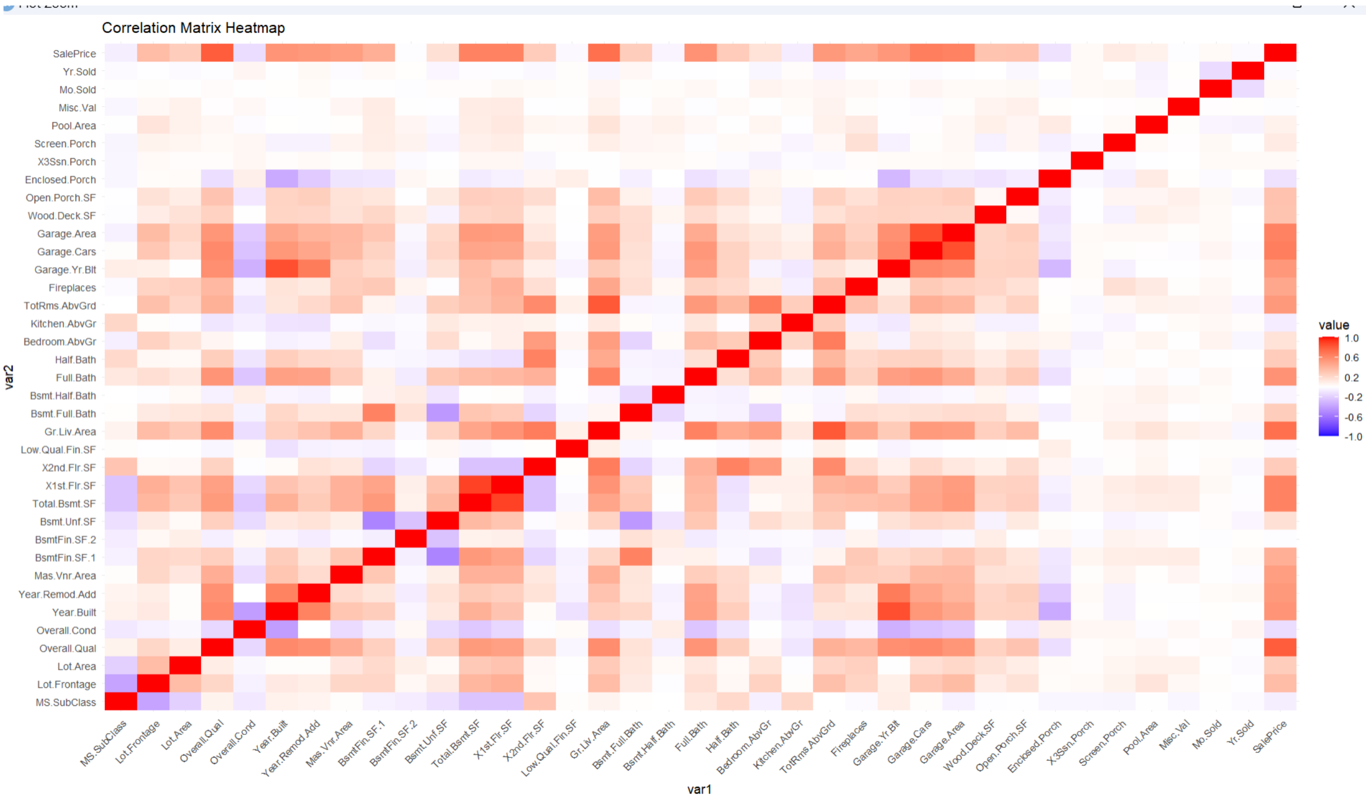
- Remove one of the correlated variables: If two or more variables are highly correlated, consider removing one of them from the model.
- Combine correlated variables: If it makes sense, you can create a new variable that combines the information from the correlated variables.
- Use regularization techniques: Regularization techniques like ridge regression or lasso regression can help mitigate the effects of multicollinearity by penalizing the magnitude of coefficients.
- Collect more data: Sometimes multicollinearity can be a result of limited data. Collecting more data may help reduce multicollinearity.
- Principal Component Analysis (PCA): PCA can be used to reduce the dimensionality of the predictors and remove multicollinearity by creating new uncorrelated variables.

This initial exploration and model building lay the groundwork for further analysis and refinement. By employing these techniques, we gain valuable insights into the factors influencing house prices in Ames, Iowa, and pave the way for the development of more robust and accurate predictive models.

Ames housing dataset:

	Order	PID	MS.SubClass	MS.Zoning	Lot.Frontage	Lot.Area	Street	Alley	Lot.Shape	Land.Contour	Utilities	Lot.Config	Land.Slope	Neighborhood	Condition.1	Condition.2	Bldg.Type	House.Style	Overall.Qu
1	1	526301100	20	RL	141	31770	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	
2	2	526350040	20	RH	80	11622	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Feedr	Norm	1Fam	1Story	
3	3	526351010	20	RL	81	14267	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	
4	4	526353030	20	RL	93	11160	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	
5	5	527105010	60	RL	74	13830	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	
6	6	527105030	60	RL	78	9978	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	
7	7	527127150	120	RL	41	4920	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	StoneBr	Norm	Norm	TwnhsE	1Story	
8	8	527145080	120	RL	43	5005	Pave	NA	IR1	HLS	AllPub	Inside	Gtl	StoneBr	Norm	Norm	TwnhsE	1Story	
9	9	527146030	120	RL	39	5389	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	StoneBr	Norm	Norm	TwnhsE	1Story	
10	10	527162130	60	RL	60	7500	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	
11	11	527163010	60	RL	75	10000	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	1Fam	2Story	
12	12	527165230	20	RL	NA	7980	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	1Story	
13	13	527166040	60	RL	63	8402	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story	
14	14	527180040	20	RL	85	10176	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	1Story	
15	15	527182190	120	RL	NA	6820	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	StoneBr	Norm	Norm	TwnhsE	1Story	
16	16	527216070	60	RL	47	53504	Pave	NA	IR2	HLS	AllPub	CulDSac	Mod	StoneBr	Norm	Norm	1Fam	2Story	
17	17	527225035	50	RL	152	12134	Pave	NA	IR1	Bnk	AllPub	Inside	Mod	Gilbert	Norm	Norm	1Fam	1.5Fin	
18	18	527258010	20	RL	88	11394	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	StoneBr	Norm	Norm	1Fam	1Story	
19	19	527276150	20	RL	140	19138	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	1Fam	1Story	
20	20	527302110	20	RL	85	13175	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	Norm	1Fam	1Story	
21	21	527358140	20	RL	105	11751	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	Norm	1Fam	1Story	
22	22	527358200	85	RL	85	10625	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	Norm	1Fam	SFoyer	
23	23	527368020	60	FV	NA	7500	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	2Story	
24	24	527402200	20	RL	NA	11241	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	
25	25	527402250	20	RL	NA	12537	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	
26	26	527403020	20	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	
27	27	527404120	20	RL	70	8400	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	
28	28	527425090	20	RL	70	10500	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NAmes	Norm	Norm	1Fam	1Story	
29	29	527427230	120	RH	26	5858	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	NAmes	Norm	Norm	TwnhsE	1Story	
30	30	527451180	160	RM	21	1680	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm	Norm	Twnhs	2Story	

Correlation matrix heatmap-



Correlation matrix heatmap, which means it shows the correlation between different features in the Ames Housing dataset. The features are displayed on both the rows and columns of the matrix, and the color intensity in each cell represents the correlation coefficient between the two corresponding features.

Here are some of the key insights you can glean from the heatmap:

- **Strong positive correlations:** These are represented by darker red colors and indicate that two features tend to increase or decrease together. For example, there is a strong positive correlation between "Gr.Liv.Area" (above ground living area) and "Total Bsmt SF" (total basement square footage), which suggests that houses with larger above ground living areas also tend to have larger basements.
- **Strong negative correlations:** These are represented by darker blue colors and indicate that two features tend to move in opposite directions. For example, there is a strong negative correlation between "Overall Qual" (overall quality) and "Year Built", which suggests that newer houses tend to have higher overall quality ratings.
- **Weak correlations:** These are represented by lighter colors closer to white and indicate that there is no significant relationship between the two features. For example, there is a

weak correlation between "Mas Vnr Area" (masonry veneer area) and "Yr Sold" (year sold), suggesting that the size of the masonry veneer area is not strongly related to the year the house was sold.

It's important to note that correlation does not imply causation. Just because two features are correlated does not necessarily mean that one causes the other. However, identifying these correlations can be helpful in understanding the relationships between different features and can be used to guide further analysis and model building.

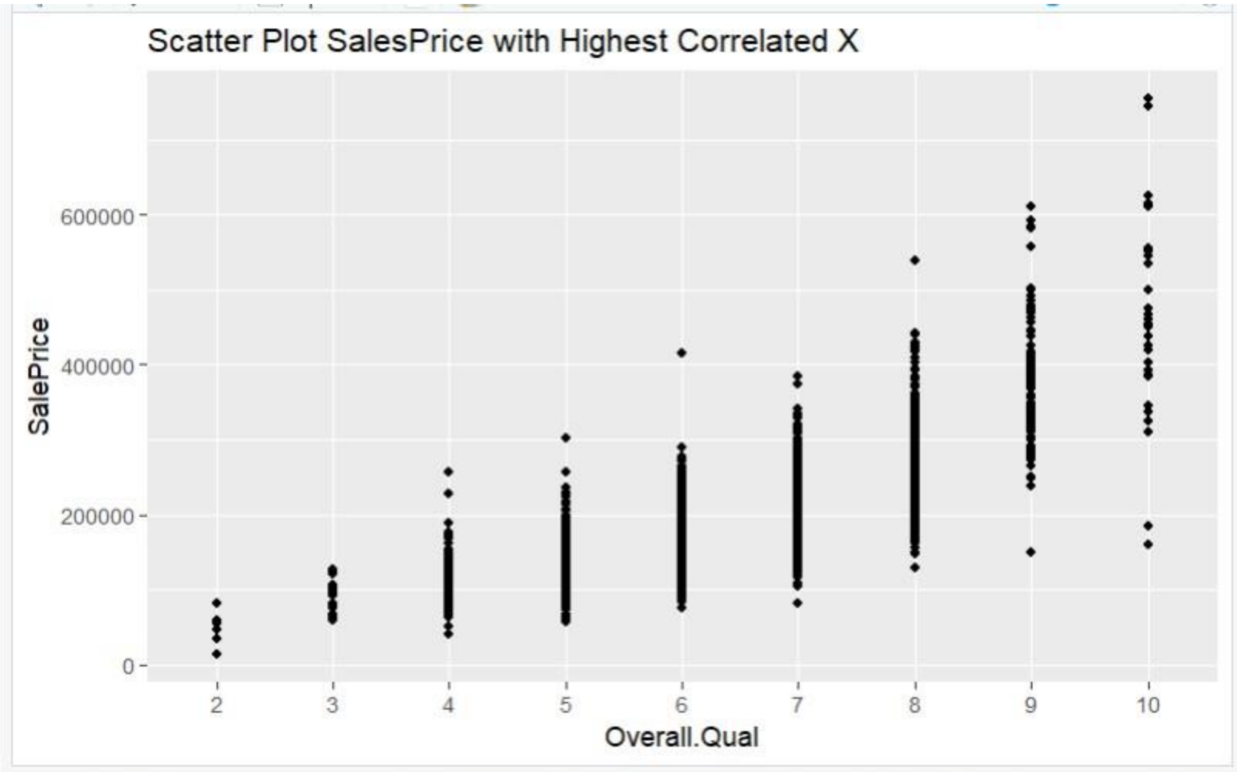
Here are some additional points to consider when interpreting the heatmap:

- The heatmap only shows correlations between numeric features. Categorical features have been excluded.
- The intensity of the color reflects the strength of the correlation, but the exact values are not displayed on the heatmap.
- It is important to consider the context of the data and the specific research question when interpreting the correlations.

Overall, the correlation matrix heatmap provides a valuable visual summary of the relationships between different features in the Ames Housing dataset. By understanding these correlations, you can gain insights into the factors that may influence house prices in Ames, Iowa.

Scatterplots-

1.



The graph you sent me is a scatter plot showing the relationship between the "SalePrice" and the "Highest Correlated X". Each data point represents a house, and the x-axis value represents the value of an unknown feature that is most correlated with the sale price. The y-axis value represents the sale price of the house.

There appears to be a positive correlation between the sale price and the highest correlated x. This means that as the value of the highest correlated x increases, the sale price also tends to increase. However, it is important to note that there is also a lot of scatter in the data, which means that there are many exceptions to this trend.

For example, there are some houses with a high sale price that have a low value for the highest correlated x, and there are some houses with a low sale price that have a high value for the highest correlated x. This suggests that there are other factors besides the highest correlated x that can affect the sale price of a house.

It is also important to note that the x-axis is labeled "Highest Correlated X", but the specific feature that this variable represents is not given. This makes it difficult to interpret the graph in more detail.

In conclusion, the graph shows a positive correlation between the sale price of a house and the value of an unknown feature that is most correlated with the sale price. However, there is also a lot of scatter in the data, suggesting that other factors can also affect the sale price of a house.

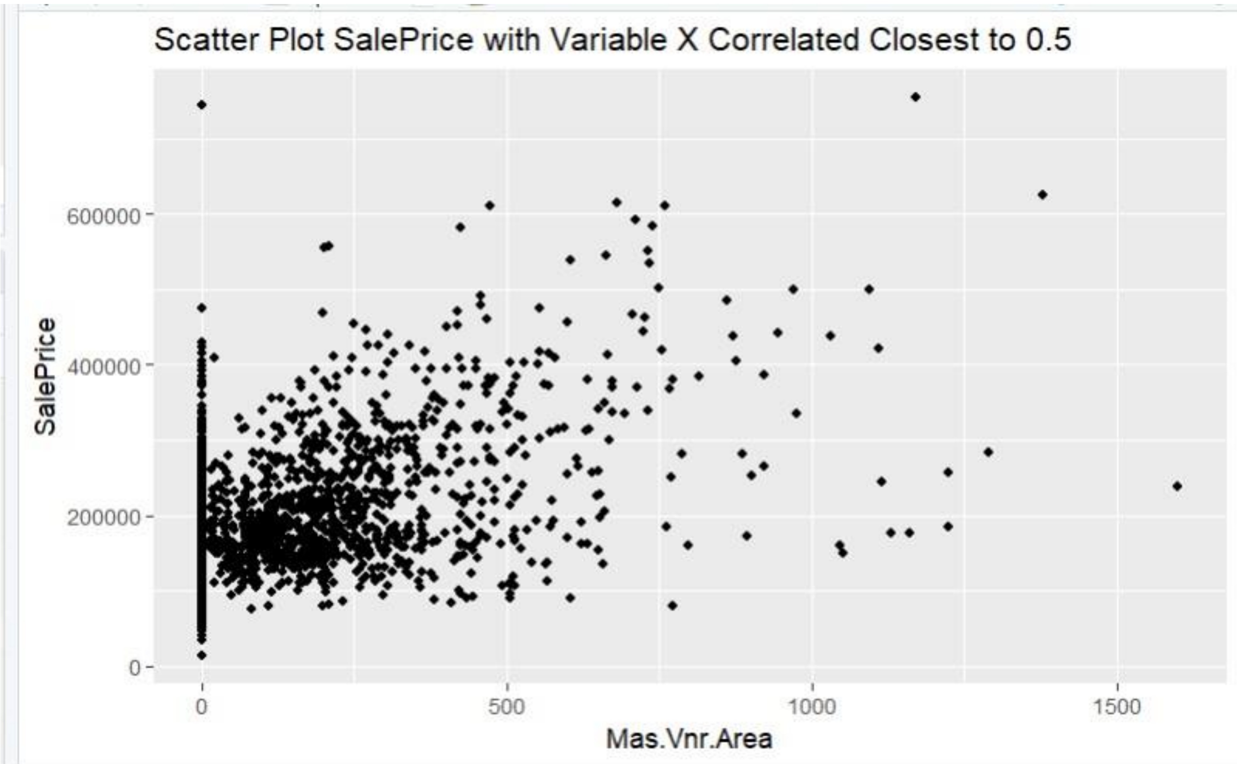
2.

The image you sent me appears to be a scatter plot, but the x-axis is labeled "Overall.Cond" which is likely referring to a categorical variable with multiple levels. It is difficult to interpret a scatter plot with a categorical x-axis because the relationship between the two variables is not easily visualized.

Scatter plots are most useful for visualizing the relationship between two continuous variables. In a scatter plot of two continuous variables, each data point represents one observation, and the position of the point on the x-axis and y-axis corresponds to the values of the two variables for that observation. The pattern of the points in the scatter plot can reveal whether there is a relationship between the two variables, and if so, whether the relationship is positive, negative, or neutral.

For example, if the points in a scatter plot tend to cluster in the upper right corner, this would suggest a positive correlation between the two variables. Conversely, if the points tend to cluster in the lower left corner, this would suggest a negative correlation. And if the points are scattered randomly across the plot, this would suggest that there is no correlation between the two variables.

3.



The horizontal axis, labeled "Mas.Vnr.Area", represents the square footage of veneer masonry area, and the vertical axis, labeled "Sale Price", represents the sale price of a house.

The data points in the scatter plot are somewhat scattered, which suggests that there is not a strong linear relationship between the two variables. However, there does appear to be a slight positive trend, meaning that houses with a larger Mas.Vnr.Area tend to have a higher sale price.

It is important to note that this scatter plot only shows the relationship between these two variables, and it does not necessarily mean that there is a causal relationship between them. There may be other factors that influence the sale price of a house, such as the size of the house, the location of the house, and the quality of the school district.

Here are some additional insights that you can glean from the graph:

- There are a few data points that appear to be outliers. These data points are located far away from the other data points, and they may have a significant impact on the results of any statistical analysis that is performed on this data.
- The data points are spread out over a wide range of values on both the horizontal and vertical axes. This suggests that there is a lot of variability in the data.

Overall, the scatter plot you sent provides some insights into the relationship between Mas.Vnr.Area and Sale Price. However, it is important to keep in mind the limitations of this type of analysis and to consider other factors that may influence the sale price of a house.

7&8. Regression Model Equation:

$SalePrice = -85367.593 + 26924.721 \times Overall.Qual + 33.427 \times BsmtFin.SF.1 + 20432.739 \times Garage.Cars + 51.155 \times Gr.Liv.Area$$SalePrice = -85367.593 + 26924.721 \times Overall.Qual + 33.427 \times BsmtFin.SF.1 + 20432.739 \times Garage.Cars + 51.155 \times Gr.Liv.Area$

Interpretation of Coefficients:

Intercept (-85367.593): This represents the estimated baseline sale price when all other predictor variables (Overall.Qual, BsmtFin.SF.1, Garage.Cars, Gr.Liv.Area) are zero. However, since it's unlikely for these variables to be exactly zero in practice, the intercept may not have a direct interpretation in this context.

Overall.Qual (26924.721): For each unit increase in Overall.Qual (overall quality rating of the house), the estimated sale price increases by $26,924.721, holding all other variables constant. This suggests that higher quality ratings are associated with higher sale prices.

BsmtFin.SF.1 (33.427): For each additional square foot of finished basement area, the estimated sale price increases by $33.427, holding all other variables constant. This implies that houses with larger finished basement areas tend to have higher sale prices.

Garage.Cars (20432.739): For each additional car capacity in the garage, the estimated sale price increases by $20,432.739, holding all other variables constant. This suggests that houses with larger garages capable of accommodating more cars tend to have higher sale prices.

Gr.Liv.Area (51.155): For each additional square foot of above-grade (ground) living area, the estimated sale price increases by $51.155, holding all other variables constant. This indicates that larger above-grade living areas are associated with higher sale prices.

Model Plots-

Residual plot, which is a type of scatter plot used to assess the quality of a statistical model. In this specific case, the residual plot appears to be assessing a linear regression model, where the independent variable is leverage, and the dependent variable is standardized residuals.

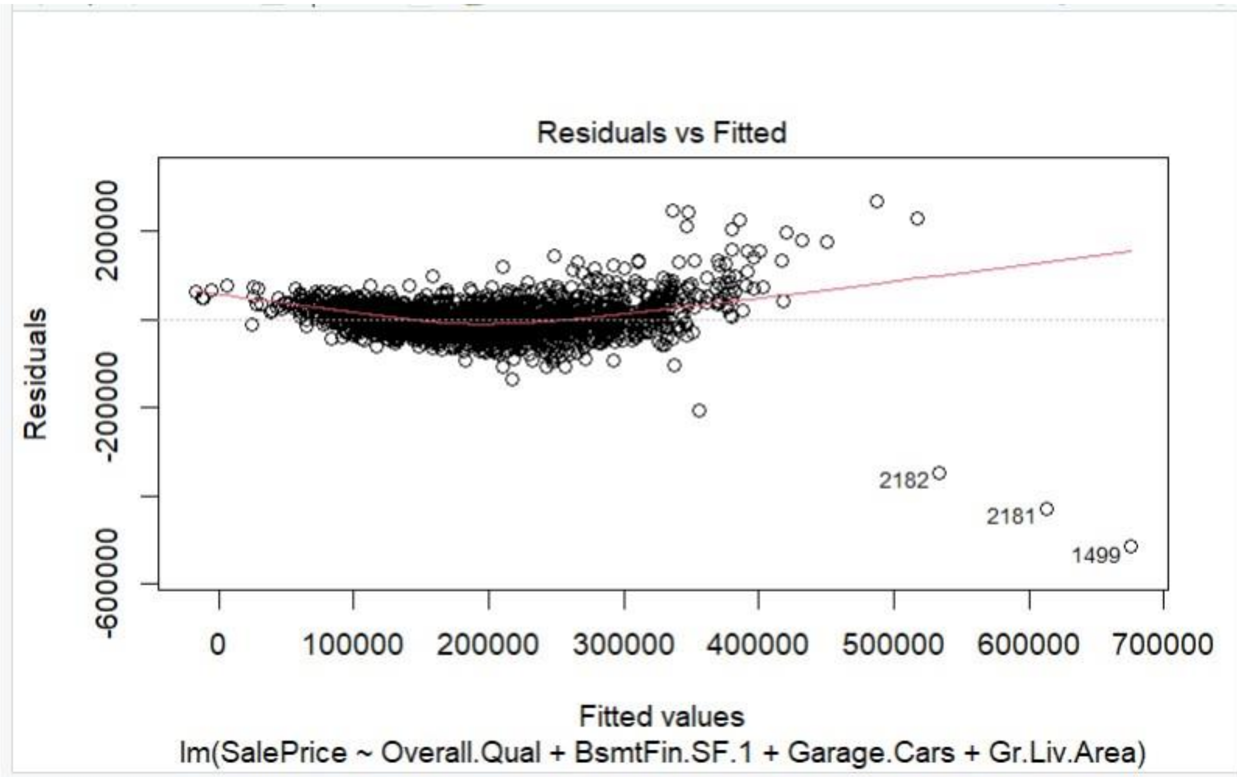
Ideally, in a good linear regression model, the residuals should be randomly scattered around the horizontal line at zero. This would indicate that the model fits the data well, with no systematic bias.

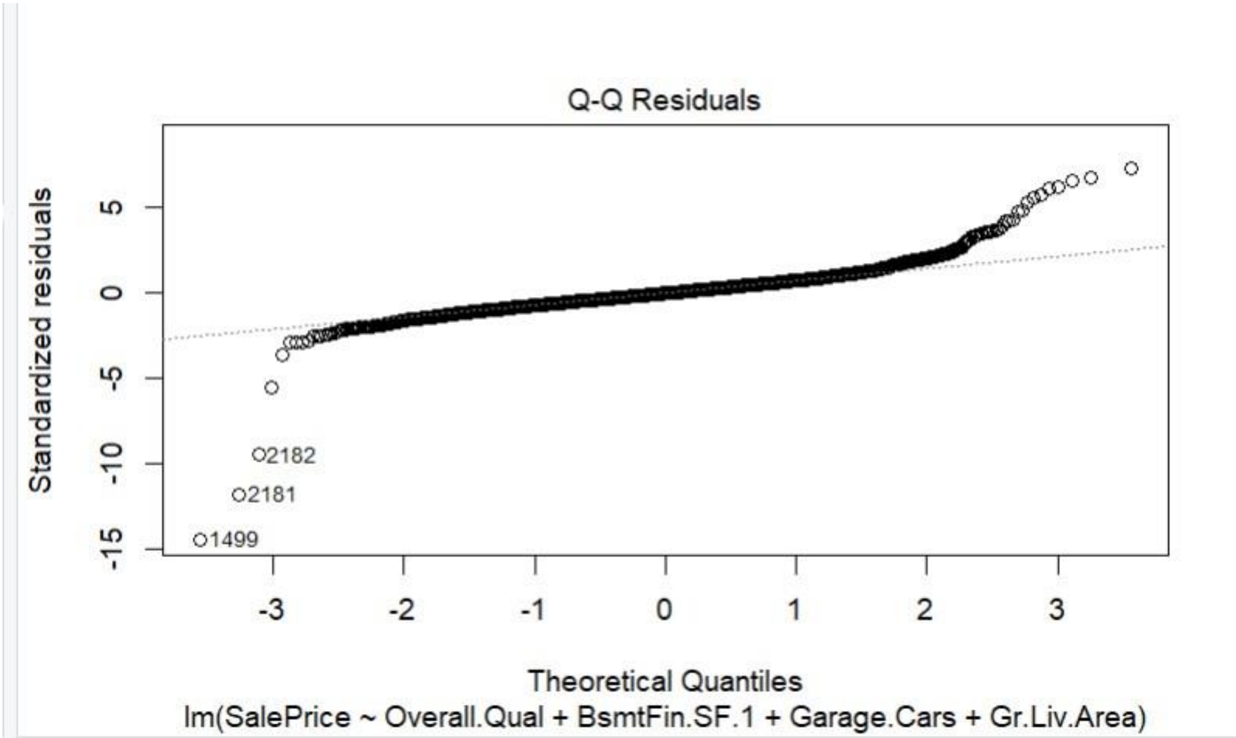
In this particular residual plot, there doesn't appear to be a clear pattern in the data points, suggesting that the model might be a good fit for the data. However, it's important to note that a definitive assessment of the model's quality would require further analysis, potentially including metrics like R-squared and the F-statistic.

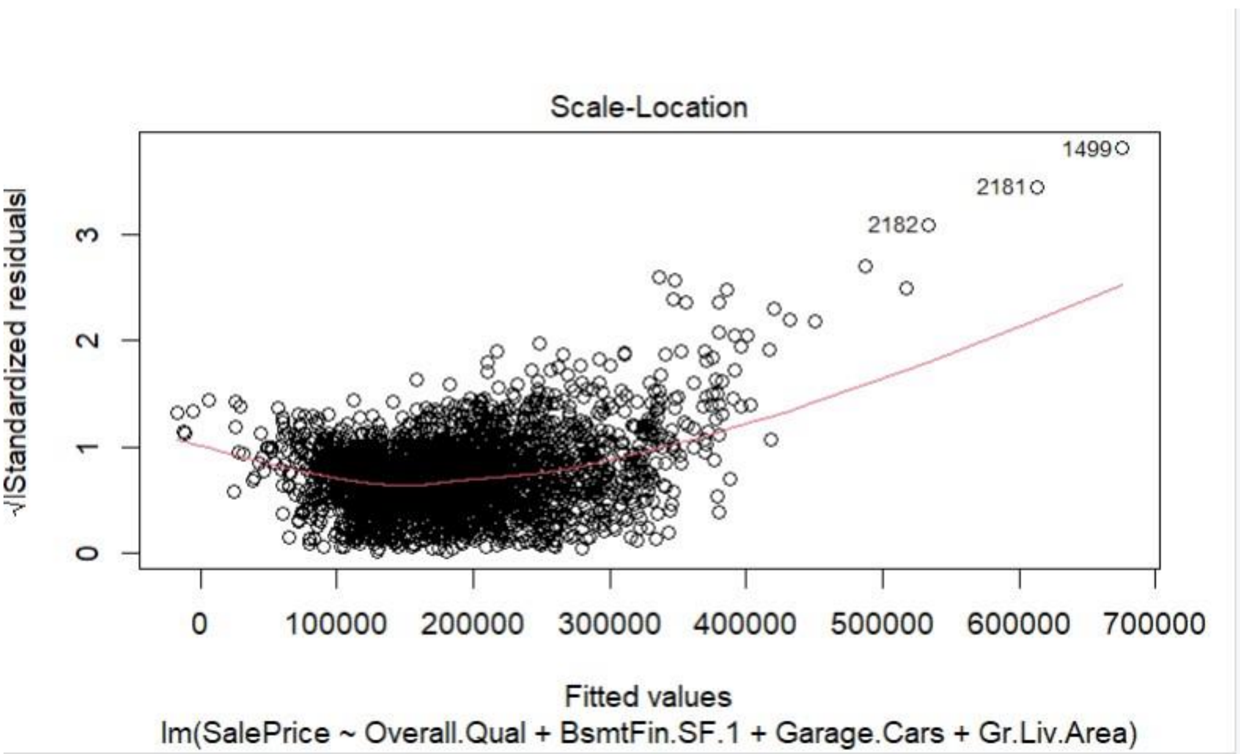
Here are some additional insights based on the graph:

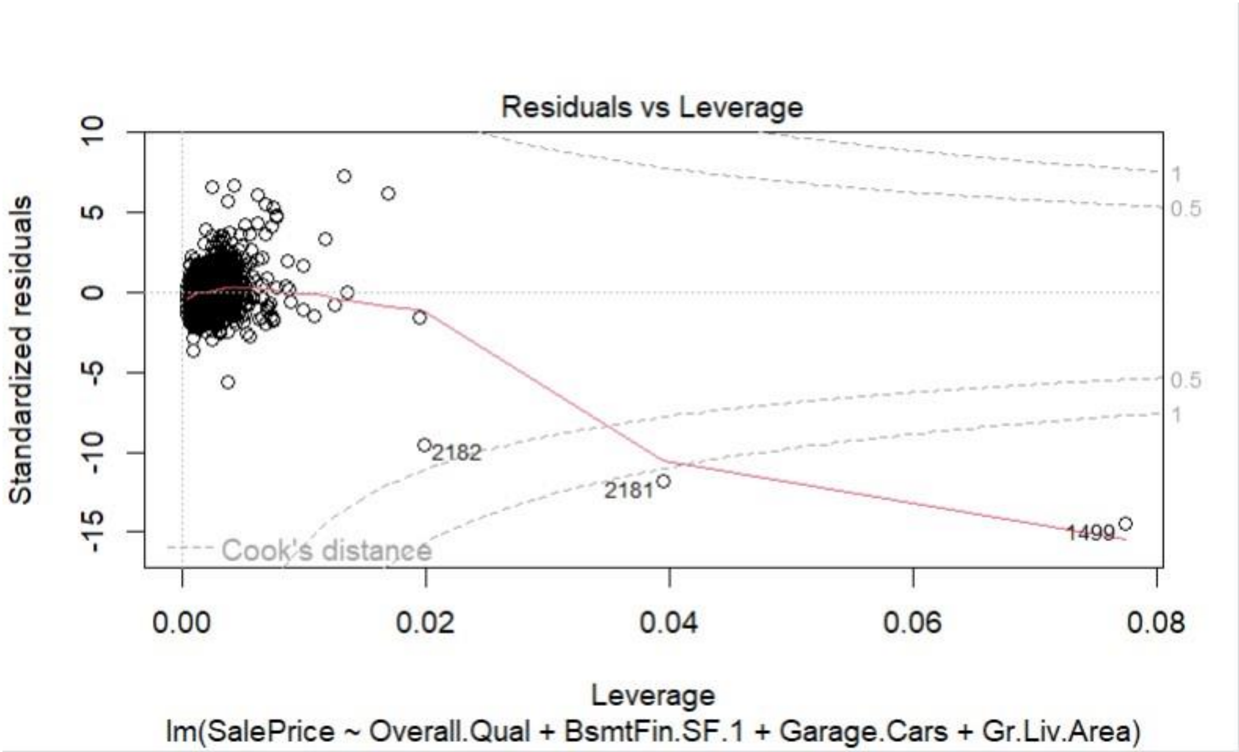
- There are a few data points that have a relatively **high leverage**, which means they have a greater influence on the fitted regression line. It is important to be aware of these points and to consider whether they are outliers or not. If they are outliers, they may need to be removed from the analysis.
- The standardized residuals are mostly within 2 standard deviations of the mean, which is a good sign. This suggests that the model is doing a good job of fitting the data.

Model/Residual Plots-




Q-Q Residuals
Standardized residuals
-15
-10
-5
0
5
2182
2181
1499
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(SalePrice ~ Overall.Qual + BsmtFin.SF.1 + Garage.Cars + Gr.Liv.Area)


Scale-Location
√|Standardized residuals|
0
1
2
3
1499
2181
2182
0
100000
200000
300000
400000
500000
600000
700000
Fitted values
lm(SalePrice ~ Overall.Qual + BsmtFin.SF.1 + Garage.Cars + Gr.Liv.Area)



Before removing Outliers from the model-

```
> summary(model)

Call:
lm(formula = SalePrice ~ Overall.Qual + BsmtFin.SF.1 + Garage.Cars +
    Gr.Liv.Area, data = df_AmesHousing_processed_numeric)

Residuals:
    Min      1Q  Median      3Q     Max
-515098  -18720   -1264   16848  267327

Coefficients:
               Estimate Std. Error t value            Pr(>|t|)
(Intercept)  -85367.593   2949.315  -28.95 <0.0000000000000002 ***
Overall.Qual  26924.721    719.693   37.41 <0.0000000000000002 ***
BsmtFin.SF.1     33.427      1.626   20.55 <0.0000000000000002 ***
Garage.Cars   20432.739   1418.320   14.41 <0.0000000000000002 ***
Gr.Liv.Area      51.155      1.816   28.16 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37000 on 2673 degrees of freedom
Multiple R-squared:  0.7837,    Adjusted R-squared:  0.7834
F-statistic:  2421 on 4 and 2673 DF,  p-value: < 0.00000000000000022

>
```

After removing Outliers from the model-

```
Call:
lm(formula = SalePrice ~ Year.Built + Year.Remod.Add + Overall.Qual +
    Overall.Cond, data = df_AmesHousing_no_outliers)

Residuals:
   Min     1Q Median     3Q    Max
-96417 -23906  -3367  19364 168734

Coefficients:
                  Estimate  Std. Error t value              Pr(>|t|)
(Intercept)    -1077876.09    89875.81 -11.993 < 0.0000000000000002 ***
Year.Built          245.91       41.47   5.929          0.00000000345 ***
Year.Remod.Add      293.00       52.94   5.535          0.00000003433 ***
Overall.Qual      36438.90      752.15  48.446 < 0.0000000000000002 ***
Overall.Cond        776.57      832.56   0.933                  0.351
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38110 on 2578 degrees of freedom
Multiple R-squared:  0.6802,    Adjusted R-squared:  0.6797
F-statistic:  1371 on 4 and 2578 DF,  p-value: < 0.00000000000000022

>
```

- After removing outliers, the adjusted �2R_2 value decreased from 0.7834 to 0.6797. Additionally, the residual standard error increased slightly from 37000 to 38110. These changes suggest that removing outliers may have negatively impacted the model's performance.

From the Summary of the regsubsets() output it is observed that-

Best 1 variable Model is Model with Variable: Overall.Qual

- SalePrice = -56572.5358676634 + Overall.Qual * 46604.7690913624

Best 2 variable Model is Model with Variables: Overall.Qual and Gr.Liv.Area

- SalePrice = -80759.5297919491 + Overall.Qual * 33471.5462864554 + Gr.Liv.Area * 61.0656557418694

Best 3 variable Model is Model with Variables: Overall.Qual, BsmtFin.SF.1 and Gr.Liv.Area

- SalePrice = -80492.6023020774 + Overall.Qual * 31149.5378044778 + BsmtFin.SF.1 * 35.5319185573599 + Gr.Liv.Area * 57.9402689587488

To find Best Model, I am using the stepwise selection method using stepAIC(),

The BEST Model is-
SalePrice = 644812.910595304 + MS.SubClass * -153.911717265969 + Lot.Area * 0.417527644465161 + Overall.Qual * 18562.0119815929 + Overall.Cond * 4769.84129625748 + Year.Built * 298.02186977768 + Year.Remod.Add * 219.15351834981 + Mas.Vnr.Area * 30.779094744528 + BsmtFin.SF.1 * 32.4147544971131 + BsmtFin.SF.2

* 22.9510303298807 + Bsmt.Unf.SF * 18.1997267997147 + X1st.Flr.SF * 47.4868123977395 + X2nd.Flr.SF * 51.729619128216 + Low.Qual.Fin.SF * 26.0496450392937 + Bsmt.Full.Bath * 6960.9118807314 + Bedroom.AbvGr * -7205.1187182162 + Kitchen.AbvGr * -22494.386068957 + TotRms.AbvGrd * 3423.60250943613 + Fireplaces * 3365.30048470135 + Garage.Cars * 8712.01308827843 + Garage.Area * 20.0272412723527 + Wood.Deck.SF * 15.8418955440636 + Open.Porch.SF * -15.5922333119135 + Enclosed.Porch * 17.0503048700313 + Screen.Porch * 59.907179560064 + Pool.Area * -32.2076073256579 + Misc.Val * -9.52163739751485 + Yr.Sold * -854.488197622976

Conclusion-

Referring to the Summaries of models in step 12 and best model in step 13 it is understood that Adjusted R-squared value is higher for best model obtained in step 13 indicating that it explains more variance in the target variable. Model in step 13 also has more coefficients that are statistically significant. Therefore, the model obtained in step 13 (best mode) seems to be a better choice considering the above reasons.

Appendix-

```
library(dplyr)

library(readxl)

library(readr)

library(tidyverse)

library(plotly)

library(ggplot2)

library(gmodels)

library(knitr)

library(car)



#1

df_AmesHousing <- read.csv("/Users/devik/Downloads/AmesHousing.csv")


#displaying class of df_AmesHousing

class(df_AmesHousing)



#2 - EDA

summary(df_AmesHousing)

str(df_AmesHousing)


#displaying head

head(df_AmesHousing)


#Finding columns with missing values
```

```
#--------replacing empty string values with NA
df_AmesHousing <- df_AmesHousing %>%
  mutate_all(~ifelse(. == "", NA, .))

missing_values_count <- colSums(is.na(df_AmesHousing))
print(missing_values_count)

missing_values_percentage <- colSums(is.na(df_AmesHousing)) / nrow(df_AmesHousing) * 100
print(missing_values_percentage)

print(missing_values_percentage[missing_values_percentage > 0])

columns_missing_values_percentag_gt_10 <-
names(missing_values_percentage[missing_values_percentage > 10])
print(columns_missing_values_percentag_gt_10)

#Cleaning data - dropping columns with missing% > 80
columns_missing_values_percentag_gt_80 <-
names(missing_values_percentage[missing_values_percentage > 80])
df_AmesHousing_cleaned <- df_AmesHousing[,!names(df_AmesHousing) %in%
columns_missing_values_percentag_gt_80]

missing_values_percentage <- colSums(is.na(df_AmesHousing_cleaned)) /
nrow(df_AmesHousing_cleaned) * 100
print(missing_values_percentage[missing_values_percentage > 0])

library(dplyr)
#imputing with mean
df_AmesHousing_cleaned_imputed <- df_AmesHousing_cleaned %>%
```

```
  mutate_if(is.numeric, ~ifelse(is.na(.), mean(., na.rm = TRUE), .))

missing_values_percentage <- colSums(is.na(df_AmesHousing_cleaned_imputed)) /
nrow(df_AmesHousing_cleaned_imputed) * 100

print(missing_values_percentage[missing_values_percentage > 0])

columns_missing_values_percentag_gt_40 <-
names(missing_values_percentage[missing_values_percentage > 40])

df_AmesHousing_cleaned_imputed <-
df_AmesHousing_cleaned_imputed[,!names(df_AmesHousing_cleaned_imputed) %in%
columns_missing_values_percentag_gt_40]

missing_values_percentage <- colSums(is.na(df_AmesHousing_cleaned_imputed)) /
nrow(df_AmesHousing_cleaned_imputed) * 100

print(missing_values_percentage[missing_values_percentage > 0])

dim(df_AmesHousing_cleaned_imputed)

#removing rows with missing values as just columns with very less percentage of missing values
are remaining now

df_AmesHousing_cleaned_final <- na.omit(df_AmesHousing_cleaned_imputed)

dim(df_AmesHousing_cleaned_final)

missing_values_percentage <- colSums(df_AmesHousing_cleaned_final == "") /
nrow(df_AmesHousing_cleaned_final) * 100

print(missing_values_percentage[missing_values_percentage > 0])

df_AmesHousing_processed <- df_AmesHousing_cleaned_final

#4

df_AmesHousing_processed <- subset(df_AmesHousing_processed, select = -PID)

df_AmesHousing_processed <- subset(df_AmesHousing_processed, select = -Order)
```

```
cor_matrix <- cor(select_if(df_AmesHousing_processed, is.numeric))

print(cor_matrix)


#5

library(ggplot2)

library(reshape2)


cor_df <- as.data.frame(cor_matrix)

cor_df$var1 <- rownames(cor_df)

cor_df_long <- melt(cor_df, id.vars = "var1", variable.name = "var2")


cor_df_long$var1 <- factor(cor_df_long$var1, levels = unique(cor_df_long$var1))

cor_df_long$var2 <- factor(cor_df_long$var2, levels = unique(cor_df_long$var2))


ggplot(cor_df_long, aes(x = var1, y = var2, fill = value)) +

 geom_tile() +

 scale_fill_gradient2(low = "blue", high = "red", mid = "white",

            midpoint = 0, limit = c(-1,1),

            breaks = seq(-1, 1, by = 0.4)) +

 theme_minimal() +

 theme(axis.text.x = element_text(angle = 45, hjust = 1)) +

 labs(title = "Correlation Matrix Heatmap")


#6

#Finding correlation against SalePrice


corr_SalePrice <- sapply(df_AmesHousing_processed[, sapply(df_AmesHousing_processed,
is.numeric) & names(df_AmesHousing_processed) != "SalePrice"],
```

```
                    function(x) cor(x, df_AmesHousing_processed$SalePrice, use =
"pairwise.complete.obs"))

highest_corr_X <- names(which.max(corr_SalePrice))

highest_corr_X


lowest_corr_X <- names(which.min(corr_SalePrice))

lowest_corr_X


closest_corr_0.5_X <- names(which.min(abs(corr_SalePrice - 0.5)))

closest_corr_0.5_X


df_AmesHousing_processed[[highest_corr_X]] <-
factor(df_AmesHousing_processed[[highest_corr_X]])

ggplot(df_AmesHousing_processed, aes(x = !!sym(highest_corr_X), y = SalePrice)) +

 geom_point() +

 labs(title = "Scatter Plot SalesPrice with Highest Correlated X") +

 scale_x_discrete(name = highest_corr_X)+

 scale_y_continuous(labels = function(x) format(x, scientific = FALSE))


df_AmesHousing_processed[[lowest_corr_X]] <-
factor(df_AmesHousing_processed[[lowest_corr_X]])

ggplot(df_AmesHousing_processed, aes(x = !!sym(lowest_corr_X), y = SalePrice)) +

 geom_point() +

 labs(title = "Scatter Plot SalesPrice with Lowest Correlated X") +

 scale_x_discrete(name = lowest_corr_X) +

 scale_y_continuous(labels = function(x) format(x, scientific = FALSE))


df_AmesHousing_processed[[closest_corr_0.5_X]] <-
as.numeric(df_AmesHousing_processed[[closest_corr_0.5_X]])

ggplot(df_AmesHousing_processed, aes(x = !!sym(closest_corr_0.5_X), y = SalePrice)) +
```

```
  geom_point() +
  labs(title = "Scatter Plot SalePrice with Variable X Correlated Closest to 0.5")+
  scale_x_continuous(name = closest_corr_0.5_X) +
  scale_y_continuous(labels = function(x) format(x, scientific = FALSE))



options(scipen = 999)#to avoid scientific notation


df_AmesHousing_processed_numeric <- df_AmesHousing_processed %>%
  select_if(function(x) is.numeric(x) || is.factor(x)) %>%
  mutate_if(is.factor, as.numeric)



# Fit a linear regression model
model <- lm(SalePrice ~ Overall.Qual + BsmtFin.SF.1 + Garage.Cars + Gr.Liv.Area, data =
df_AmesHousing_processed_numeric)
# Print the summary of the model
summary(model)


coeff <- coef(model)
intercept <- coeff[1]
terms <- paste(names(coeff)[-1], "*", coeff[-1], collapse = " + ")
equation <- paste("SalePrice =", intercept, "+", terms)
cat(equation, "\n")


# Plotting regression diagnostics
plot(model)


# check for multicollinearity in your regression model
```

```
library(car)

vif_values <- vif(model)

print(vif_values)


#Check for outliers

# Extract the residuals from the model

residuals <- resid(model)


# Calculate standardized residuals

standardized_residuals <- rstandard(model)


# Create a dataframe to store residuals and standardized residuals

residual_df <- data.frame(Residuals = residuals, Standardized_Residuals = standardized_residuals)


# Identify potential outliers based on standardized residuals

outliers <- residual_df[abs(residual_df$Standardized_Residuals) > 2, ]


# Print the potential outliers

print(outliers)


# Fit a linear regression model

model <- lm(SalePrice ~ Overall.Qual + BsmtFin.SF.1 + Garage.Cars + Gr.Liv.Area, data =
df_AmesHousing_processed_numeric)


# Extract standardized residuals

standardized_residuals <- rstandard(model)


# Identify outliers with standardized residuals greater than 2

outliers <- which(abs(standardized_residuals) > 2)
```

```
# Create a new dataframe excluding outliers

df_AmesHousing_no_outliers <- df_AmesHousing_processed_numeric[-outliers, ]


# Refit the model without outliers

model_no_outliers <- lm(SalePrice ~ Year.Built + Year.Remod.Add + Overall.Qual + Overall.Cond,
data = df_AmesHousing_no_outliers)


# Print summary of the new model

summary(model_no_outliers)


#13

library(MASS)

library(leaps)

leaps<- regsubsets(SalePrice ~ ., data=df_AmesHousing_processed_numeric, nbest=3)

plot(leaps,scale = "adjr2")

summary(leaps)


#from the Summary of the regsubsets output it is observed that

# Best 1 variable Model is Model with Variable: Overall.Qual

# Best 2 variable Model is Model with Variables: Overall.Qual and Gr.Liv.Area

# Best 3 variable Model is Model with Variables: Overall.Qual, BsmtFin.SF.1 and Gr.Liv.Area


model_var_1 <- lm(SalePrice ~ Overall.Qual, data = df_AmesHousing_processed_numeric)

summary(model_var_1)

coeff <- coef(model_var_1)

intercept <- coeff[1]

terms <- paste(names(coeff)[-1], "*", coeff[-1], collapse = " + ")

equation <- paste("SalePrice =", intercept, "+", terms)
```

```
cat(equation, "\n")


model_var_2 <- lm(SalePrice ~ Overall.Qual + Gr.Liv.Area, data =
df_AmesHousing_processed_numeric)

summary(model_var_2)

coeff <- coef(model_var_2)

intercept <- coeff[1]

terms <- paste(names(coeff)[-1], "*", coeff[-1], collapse = " + ")

equation <- paste("SalePrice =", intercept, "+", terms)

cat(equation, "\n")


model_var_3 <- lm(SalePrice ~ Overall.Qual + BsmtFin.SF.1 + Gr.Liv.Area, data =
df_AmesHousing_processed_numeric)

summary(model_var_3)

coeff <- coef(model_var_3)

intercept <- coeff[1]

terms <- paste(names(coeff)[-1], "*", coeff[-1], collapse = " + ")

equation <- paste("SalePrice =", intercept, "+", terms)

cat(equation, "\n")


#To find Best Model, I am using the stepwise selection method using stepAIC()

complete_model <- lm(SalePrice ~ ., data = df_AmesHousing_processed_numeric)

backward_model <- stepAIC(complete_model, direction = "backward")

forward_model <- stepAIC(complete_model, direction = "forward")

both_model <- stepAIC(complete_model, direction = "both")

models <- list(backward = backward_model, forward = forward_model, both = both_model)

best_model_name <- names(models)[which.min(sapply(models, AIC))]

best_model <- models[[best_model_name]]

summary(best_model)
```

```
coeff <- coef(best_model)

intercept <- coeff[1]

terms <- paste(names(coeff)[-1], "*", coeff[-1], collapse = " + ")

equation <- paste("SalePrice =", intercept, "+", terms)

cat(equation, "\n")
```

References-

1.RPubs - Scatter Plots. (2016, September 18). RPubs - Scatter Plots. https://rpubs.com/elinw/boxplot

2. Z., & posts by Zach, V. A. (2023, January 26). How to Perform Data Cleaning in R (With Example) - Statology. Statology. https://www.statology.org/data-cleaning-in-r/

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