Module 1: Regression Diagnostics with R

Mohit Ravindra Kamble

College of Professional Studies – Northeastern University

ALY6015: Intermediate Analytics

Prof. Roy Wada

January 16, 2024

01. Overview:

This assignment involved exploring the Ames housing dataset, using descriptive statistics, creating a correlation matrix, fitting a regression model, addressing multicollinearity and outliers, and comparing the best model with the previous one. The data was analyzed using the corrplot library, scatter plots, and plot() functions. The model was re-fitted after removing observations and comparing it with the previous one.

02. Analysis:

2.1: Load the Ames housing dataset.

In the Ames Housing dataset, I explore information about 2930 homes in Ames, Iowa, with 82 different metrics. These metrics encompass a wide range of factors, including price, location, house style, overall quality, garage capacity, construction year, exterior condition, and more. Out of the 82 variables, twenty are continuous numerical, fourteen are discrete numerical, twenty are nominal categorical, and 23 are ordinal categorical. My objective was to conduct exploratory data analysis on this diverse dataset to gain comprehensive insights.

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

I checked the first few rows of the Ames dataset with head(ames) and viewed the variable names using names(ames). The View(ames) function allowed me to explore the entire dataset interactively. To get summary statistics and understand the data structure, I used summary(ames) and str(ames) respectively. These steps help me familiarize myself with the dataset before deeper analysis.

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

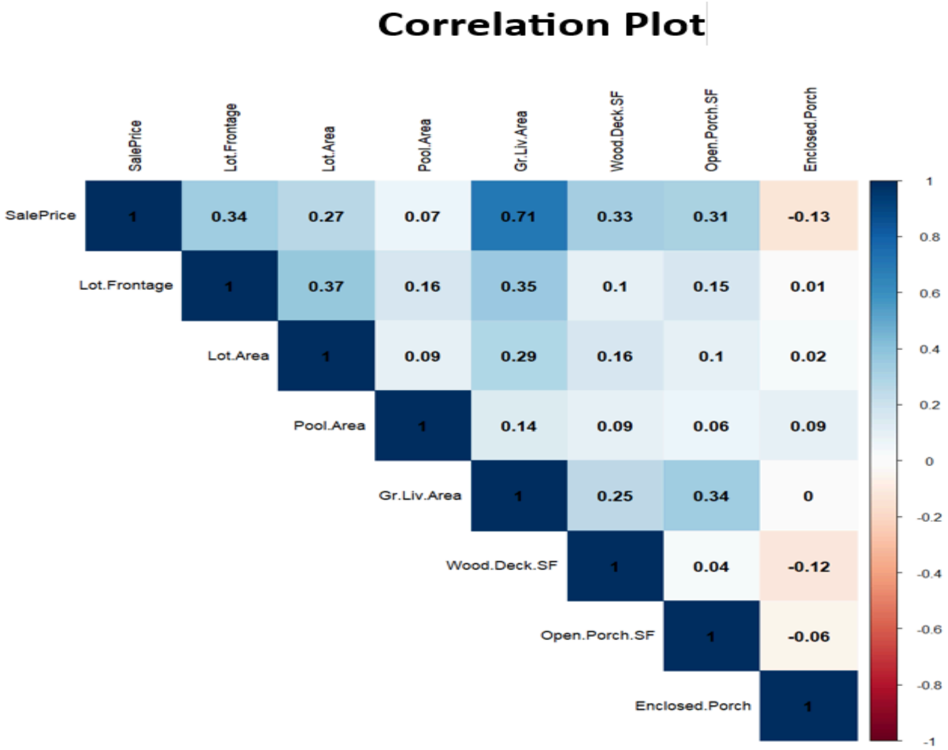
I loaded the dplyr library to make data manipulation easier in R. Then, I used the %>% operator to chain operations together. I applied the mutate_all function to the Ames dataset, replacing any missing values with the mean of their respective columns. This helps ensure that missing data doesn't hinder further analysis, and it's a common approach to handling missing values in a dataset.

2.4: Use the "cor()" function to produce a correlation matrix of the numeric values.

I selected specific continuous numeric variables from the Ames dataset, including "SalePrice," "Lot.Frontage," "Lot.Area," and others of interest. This helps me focus on relevant features for analysis. Then, I calculated the correlation matrix (cm) for these selected variables using the cor() function. The correlation matrix shows how each pair of variables is related, which is crucial for understanding patterns and relationships in the data.

2.5: Produce a plot of the correlation matrix, and explain how to interpret it. (hint - check the corrplot or ggcorrplot plot libraries)

I loaded the corrplot library, which helps create visualizations for correlation matrices. Then, I used the corrplot() function to generate a colored plot (method = "color") of the correlation matrix (cm). The specified parameters, such as type = "upper" and text size adjustments (tl.cex, tl.col), enhance the clarity of the plot. The added coefficients (addCoef.col = "black") provide numerical information on the correlations between variables. This visual representation makes it easier for me to interpret and understand the relationships within the selected numeric variables.



	SalePrice	Lot.Frontage	Lot.Area	Pool.Area	Gr.Liv.Area	Wood.Deck.SF	Open.Porch.SF	Enclosed.Porch
SalePrice	1	0.34	0.27	0.07	0.71	0.33	0.31	-0.13
Lot.Frontage		1	0.37	0.16	0.35	0.1	0.15	0.01
Lot.Area			1	0.09	0.29	0.16	0.1	0.02
Pool.Area				1	0.14	0.09	0.06	0.09
Gr.Liv.Area					1	0.25	0.34	0
Wood.Deck.SF						1	0.04	-0.12
Open.Porch.SF							1	-0.06
Enclosed.Porch								1

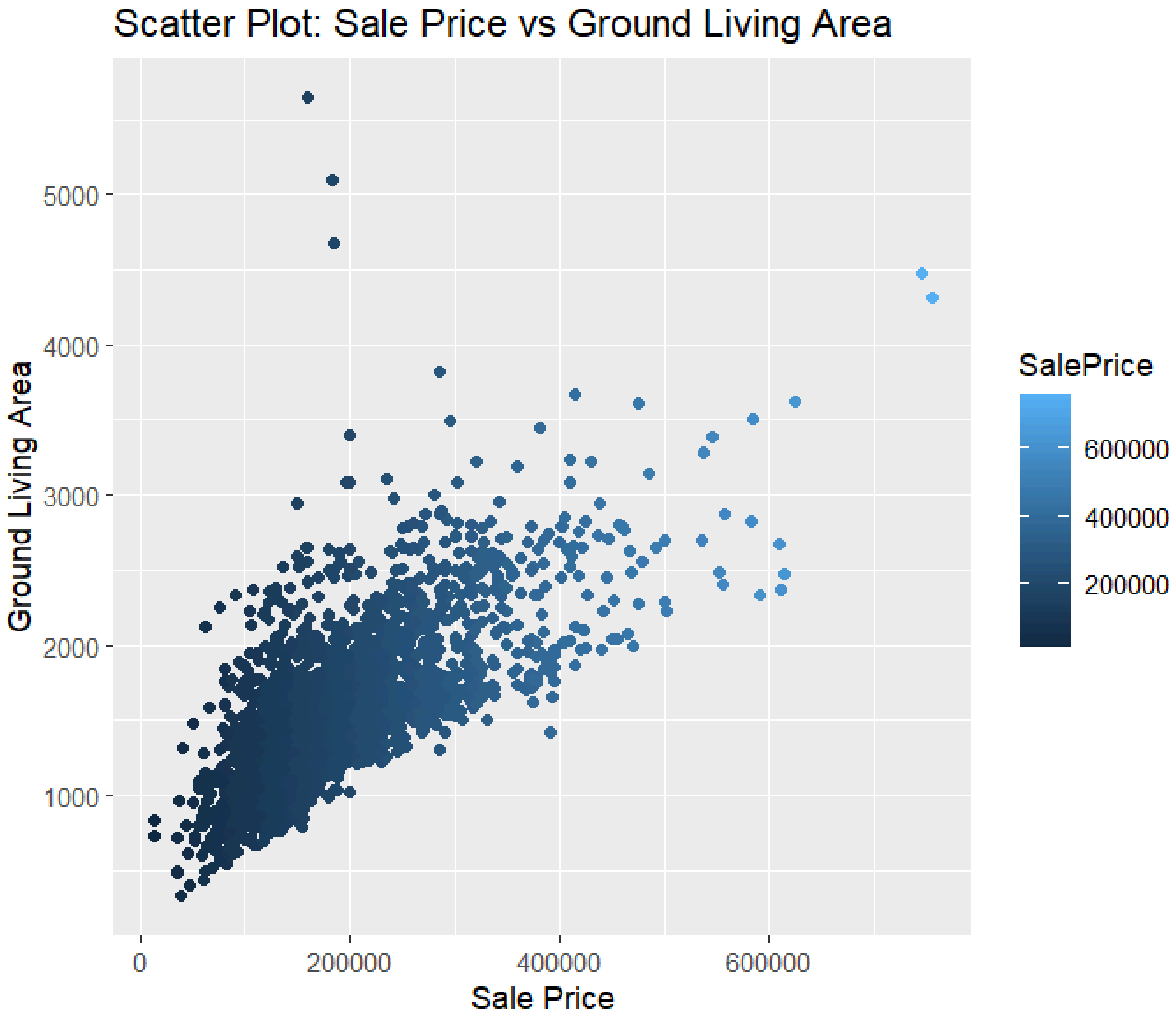
Insights:

- ➔ The correlation plot reveals numeric feature relationships in a housing dataset. Gr.Liv.Area strongly correlates (0.71) with SalePrice, indicating a significant impact on house price.
- ➔ Conversely, Enclosed.Porch has the lowest correlation (-0.13). Variables like Lot.Frontage (0.34) and Open.Porch.SF (0.31) moderately correlate with the sale price.
- ➔ Overall, house and lot size notably influence price, while factors like pool presence or enclosed porch size show weaker correlations.

2.6: 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.

I created three scatter plots using ggplot in R to visualize relationships between Sale Price and different variables in the Ames dataset. Each plot represents a different variable. The first plot compares Sale Price and Ground Living Area. The x-axis represents Sale Price, the y-axis represents Ground Living Area, and points are color-coded by Sale Price. The title is "Scatter Plot: Sale Price vs Ground Living Area." The second plot explores the relationship between Sale Price and Pool Area. Similar to the first, the x-axis is Sale Price, the y-axis is Pool Area, and points are color-coded by Sale Price. The title is "Scatter Plot: Sale Price vs Pool Area." The third plot examines Sale Price and Lot Frontage. The x-axis is Sale Price, the y-axis is Lot Frontage, and points are color-coded by Sale Price. The title is "Scatter Plot: Sale Price vs Lot Frontage."

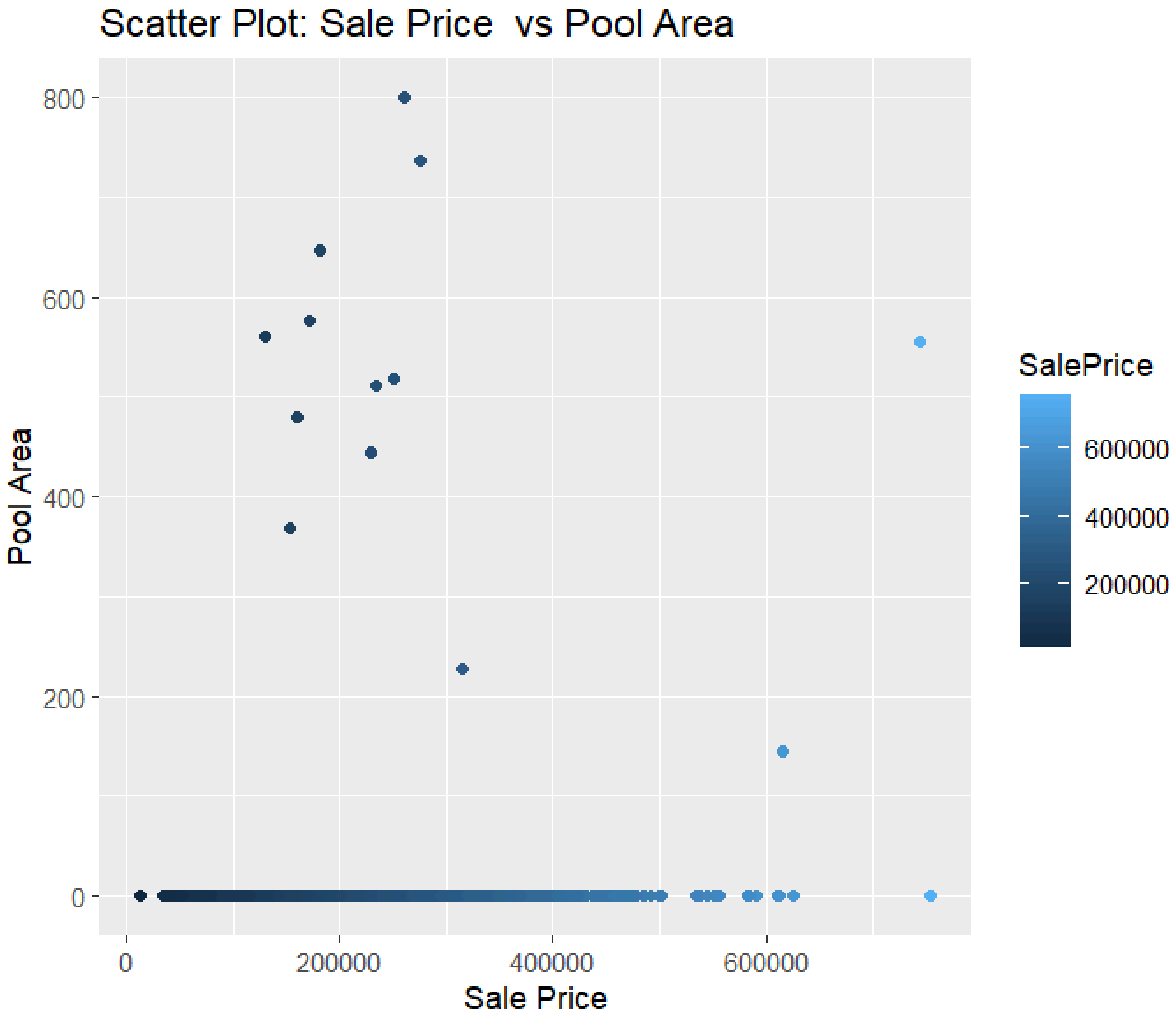
2.6.1: Scatter Plot: Sale Price vs Ground Living Area



Insights:

- ➔ The scatter plot of SalePrice against Ground Living Area shows a clear positive correlation, indicating higher sale prices with larger living spaces.
- ➔ However, a slight curve suggests a non-linear relationship, potentially tapering in higher living areas.
- ➔ The presence of a few outliers prompts further investigation for analysis integrity.

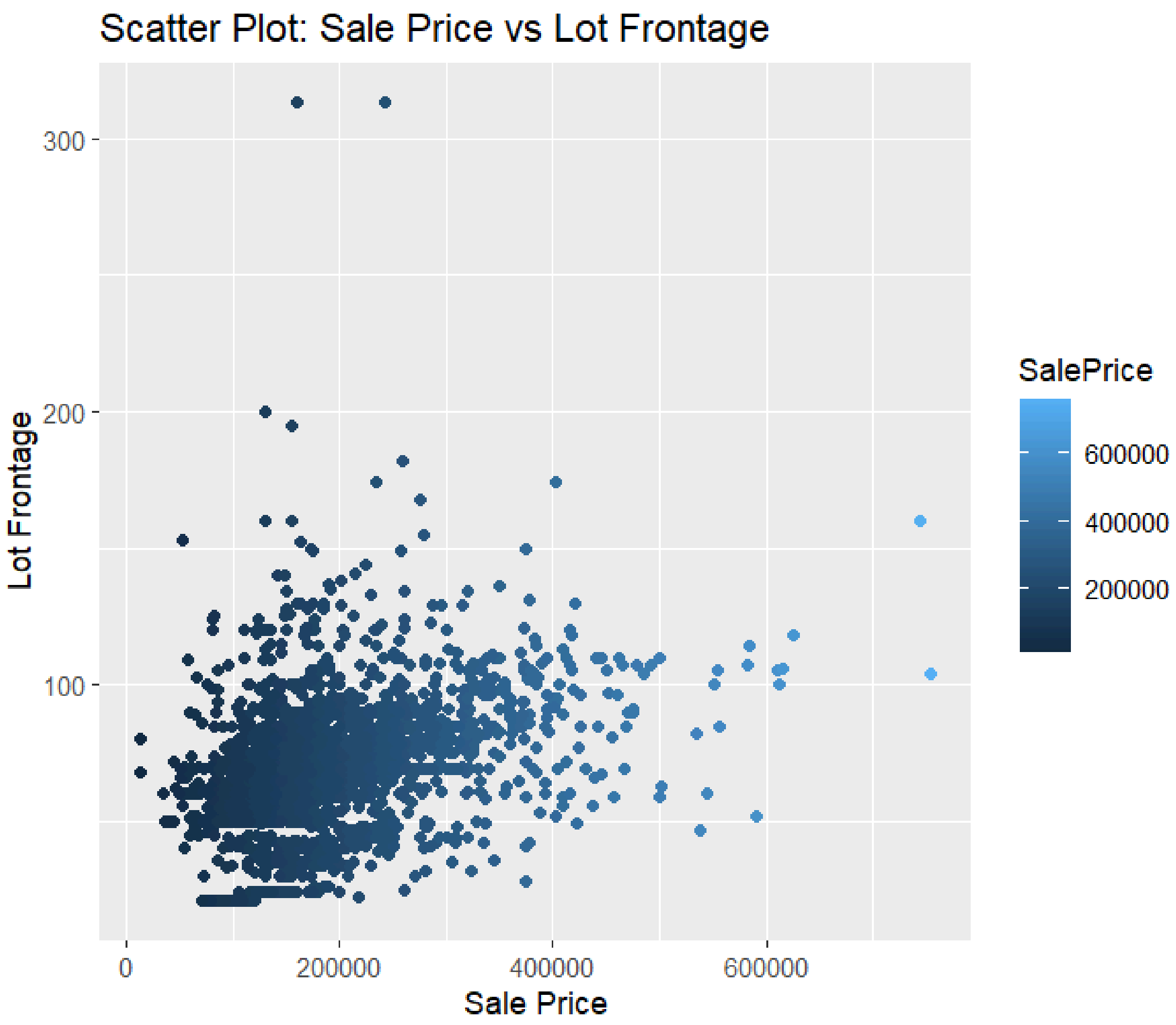
2.6.2: Scatter Plot: Sale Price vs Pool Area



Insights:

- ➔ The scatter plot for SalePrice and Pool Area displays a scattered pattern, indicating a weak correlation between these variables.
- ➔ The pool size doesn't strongly impact home sale prices. Other factors like house size, location, or overall condition likely play a more significant role in determining prices.
- ➔ Despite the weak correlation, exploring distinct clusters or patterns in the plot may reveal nuanced relationships and insights into price differences for homes with and without pools.

2.6.3: Scatter Plot: Sale Price vs Lot Frontage



Insights:

- ➔ The scatter plot of SalePrice and Lot Frontage reveals a moderately positive correlation, with wider frontages associated with somewhat higher sale prices.
- ➔ However, the correlation is not as strong as with other variables, like living area. Beyond frontage, factors contribute to price variability.
- ➔ There's a potential ceiling effect at higher frontage values, indicating a limited impact on sale prices.
- ➔ Examining points towards lower frontage values and outliers provides nuanced insights into the relationship.

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

I've constructed a regression model, utilizing SalePrice as the response variable and three continuous predictors: BsmtFin.SF.1 (representing Type 1 finished square feet), BsmtFin.SF.2 (capturing Type 2 finished square feet), and Bsmt.Unf.SF (depicting the unfinished square footage of the basement). The primary objective of this model is to delve into the intricate relationship between the sale price of homes and the diverse components constituting the square footage of the basement. Through this analysis, I aim to uncover insights into how specific aspects of basement space contribute to variations in home prices.

O/P:

```
Call:
lm(formula = SalePrice ~ BsmtFin.SF.1 + BsmtFin.SF.2 + Bsmt.Unf.SF,
    data = ames)

Coefficients:
 (Intercept)  BsmtFin.SF.1  BsmtFin.SF.2   Bsmt.Unf.SF
    63608.42        125.09         84.86        102.99
```

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

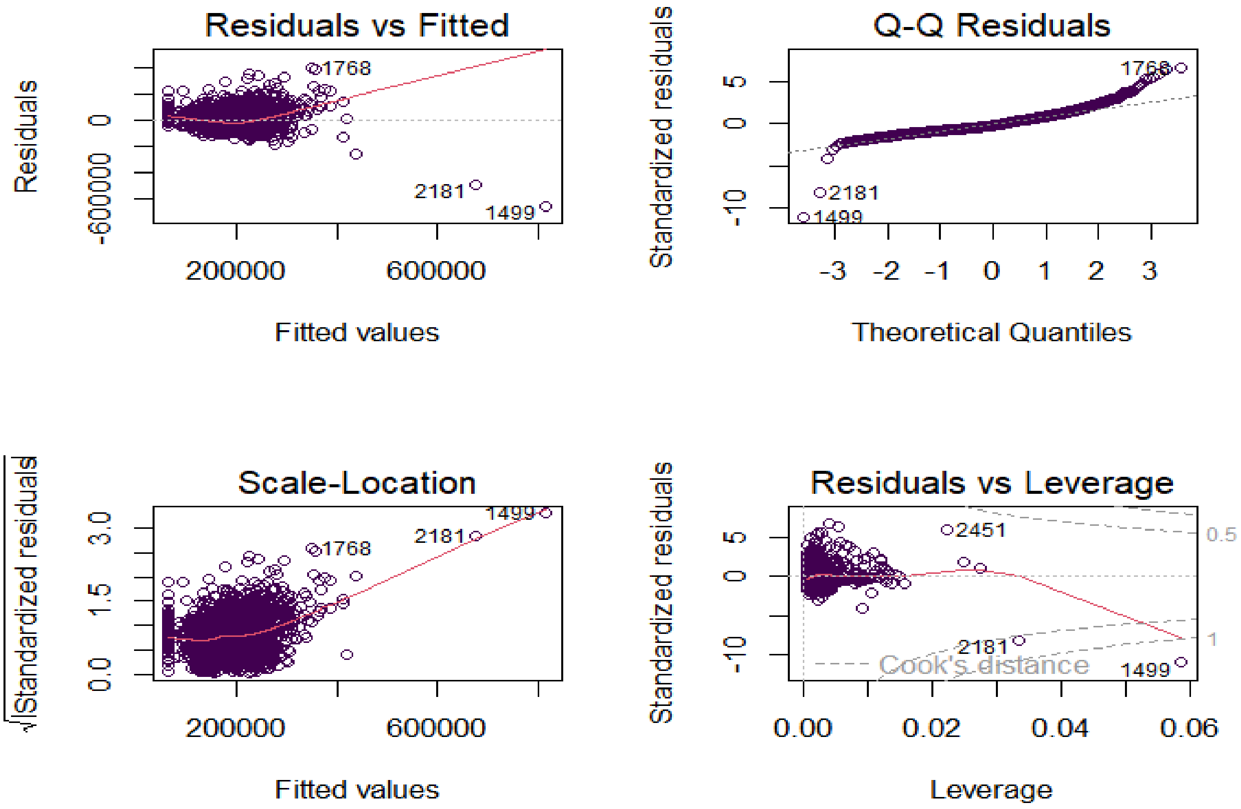
The regression model, with BsmtFin.SF.1, BsmtFin.SF.2, and Bsmt.Unf.SF as predictors for SalePrice, shows meaningful insights. Each extra square foot of Type 1 finished basement increases the sale price by $125.09, Type 2 finished basement adds $84.86, and unfinished basement adds $102.99 (all p < 0.0001). The model explains 41% of sale price variability, indicating moderate explanatory power with potential for additional influencing factors.

```
Call:
lm(formula = SalePrice ~ BsmtFin.SF.1 + BsmtFin.SF.2 + Bsmt.Unf.SF,
    data = ames)

Residuals:
    Min      1Q  Median      3Q     Max
-657596  -39052  -12617   32632  407529

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
```

```
(Intercept)   63608.422   2957.041   21.51 <0.0000000000000002 ***
BsmtFin.SF.1    125.087      2.884   43.38 <0.0000000000000002 ***
BsmtFin.SF.2     84.864      7.025   12.08 <0.0000000000000002 ***
Bsmt.Unf.SF     102.994      3.074   33.51 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61220 on 2926 degrees of freedom
Multiple R-squared:  0.4133,   Adjusted R-squared:  0.4127
F-statistic: 687.1 on 3 and 2926 DF,  p-value: < 0.00000000000000022
```

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

I installed and loaded the "viridis" package for color palettes. Using Viridis, I created a color palette. Setting up a 2x2 layout, I used the palette to add visual elements to diagnostic plots of the regression model (reg_mod).



Insights:

- ➔ The regression model, including BsmtFin.SF.1, BsmtFin.SF.2, and Bsmt.Unf.SF as predictors, shows limitations in its fit.
- ➔ Residuals vs. Fitted and Scale-Location plots indicate a "funnel-shaped" pattern, suggesting potential non-linearity and homoscedasticity violations.
- ➔ While the Normal Q-Q plot shows deviations from normality, they are not alarming.
- ➔ No clear outliers or high-leverage points are apparent, but examining extreme residuals may be beneficial.
- ➔ In summary, the model captures relationships, but observed residual patterns suggest potential issues requiring further investigation or model adjustments for improved accuracy and reliability.

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

The Variance Inflation Factor (VIF) values for the predictors, including BsmtFin.SF.1 (1.35), BsmtFin.SF.2 (1.10), and Bsmt.Unf.SF (1.43), indicate that multicollinearity is at low levels. All VIF values comfortably stay below the threshold of 5, alleviating concerns about collinearity issues among the continuous variables. This reassuring observation contributes to the enhanced stability and interpretability of the model, ensuring its robustness in capturing the relationships between predictors and the response variable.

O/P:

```
BsmtFin.SF.1 BsmtFin.SF.2  Bsmt.Unf.SF
    1.348342     1.103246     1.425560
```

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

The outlier test flags top outliers (rows 1499, 2181, 1768) with markedly low p-values, suggesting their substantial deviation from the model's expected behavior. Further investigation is recommended, and excluding them may enhance the analysis's robustness.

O/P:

```
     rstudent                 unadjusted p-value                    Bonferroni p
1499 -11.308823 0.00000000000000000000000000004716 0.00000000000000000000000013818
2181  -8.286361 0.00000000000000017564000000000001 0.00000000000051464000000000002
1768   6.720072 0.00000000002174800000000000143631 0.00000006372200000000000639274
1761   6.403442 0.00000000017637000000000000719293 0.00000051676000000000002956163
2451   6.002155 0.00000000218679999999999993441288 0.00000640739999999999984255442
434    5.602530 0.00000002308800000000000133626790 0.00006764800000000000012364415
2446   5.455110 0.00000005302799999999999679221180 0.00015536999999999999812552720
2331   5.326339 0.00000010780999999999999725041125 0.00031588999999999998590655137
2667   4.878603 0.00000112560000000000007356476539 0.00329800000000000015018541966
2333   4.638301 0.00000366710000000000000785864429 0.01074499999999999934330308093
```

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

Removing outliers improved the regression model, lowering the residual standard error and increasing the adjusted R-squared from 0.41 to 0.46. The coefficients for key predictors remain significant, resulting in a more robust model.

O/P:

```
Call:
lm(formula = SalePrice ~ BsmtFin.SF.1 + BsmtFin.SF.2 + Bsmt.Unf.SF,
    data = ames)

Residuals:
    Min      1Q  Median      3Q     Max
-657596  -39052  -12617   32632  407529

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)  63608.422   2957.041   21.51 <0.0000000000000002 ***
BsmtFin.SF.1   125.087      2.884   43.38 <0.0000000000000002 ***
BsmtFin.SF.2    84.864      7.025   12.08 <0.0000000000000002 ***
Bsmt.Unf.SF    102.994      3.074   33.51 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61220 on 2926 degrees of freedom
Multiple R-squared:  0.4133,   Adjusted R-squared:  0.4127
F-statistic: 687.1 on 3 and 2926 DF,  p-value: < 0.00000000000000022
```

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

Using the leaps package, I conducted subset regression analysis, examining various combinations of continuous variables (BsmtFin.SF.1, BsmtFin.SF.2, Bsmt.Unf.SF) to predict SalePrice. The models incorporating all three variables and the combination of BsmtFin.SF.1 and Bsmt.Unf.SF exhibit the highest selection frequency, underscoring their significance. This assists in selecting features for a more concise and effective model.

```
Subset selection object
Call: regsubsets.formula(SalePrice ~ BsmtFin.SF.1 + BsmtFin.SF.2 + 
    Bsmt.Unf.SF, data = ames, nbest = 3)
3 Variables  (and intercept)
             Forced in Forced out
BsmtFin.SF.1     FALSE      FALSE
BsmtFin.SF.2     FALSE      FALSE
Bsmt.Unf.SF      FALSE      FALSE
3 subsets of each size up to 3
Selection Algorithm: exhaustive
         BsmtFin.SF.1 BsmtFin.SF.2 Bsmt.Unf.SF
1  ( 1 ) "*"          " "          " "        
1  ( 2 ) " "          " "          "*"        
1  ( 3 ) " "          "*"          " "        
2  ( 1 ) "*"          " "          "*"        
2  ( 2 ) "*"          "*"          " "        
2  ( 3 ) " "          "*"          "*"        
3  ( 1 ) "*"          "*"          "*"        
```

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

I lean towards the subset regression model for its simplicity, achieving comparable performance with fewer variables, enhancing interpretability, and potential generalization. The model I favor from subset regression, including BsmtFin.SF.1 and Bsmt.Unf.SF, corresponds to the full regression model, underscoring the significance of these variables. Both models emphasize the importance of these predictors in elucidating SalePrice.

03. Conclusion:

In this assignment, I thoroughly explored the Ames housing dataset, employing various statistical techniques and regression modeling in R. Beginning with descriptive statistics and a correlation matrix, I progressed to fitting a regression model and addressing issues like multicollinearity and outliers. Visualizations aided in interpreting relationships, and the analysis culminated in comparing models using subset regression. Notably, the removal of outliers enhanced the robustness of the regression model. The subset regression model, emphasizing simplicity and interpretability, aligned with the full model, providing valuable insights into predictors' significance. This comprehensive analysis showcased the iterative and dynamic nature of data exploration and model refinement in statistical analysis.

04. Citations:

- ➢ Replacing Na values with Mean: source.
- ➢ Elementary Statistics : source.
- ➢ Corrplot: source.
- ➢ Regression Model: source.

05. Appendix:

```
cat("\014") # clears console
rm(list = ls()) # clears global environment
try(dev.off(dev.list()["RStudioGD"]), silent = TRUE) # clears plots
try(p_unload(p_loaded(), character.only = TRUE), silent = TRUE) # clears packages
options(scipen = 100) # disables scientific notation for entire R session


#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>Week - 1<<<<<<<<<<<<<<<<<<<<<<

# Q1)
ames <- read.csv("D:/NEU STUDY/2nd Quarter/Intermediate Analytics (ALY 6015)/Week 1/AmesHousing.csv")

# Q2)
head(ames)
names(ames)
View(ames)
summary(ames)
str(ames)


# Q3)
library(dplyr)

# Imputing missing values with mean
ames <- ames %>%
  mutate_all(funs(ifelse(is.na(.), mean(., na.rm = TRUE), .)))

# Q4)
# Continuous numeric variables of interest
contd_vars <- ames[c("SalePrice", "Lot.Frontage", "Lot.Area", "Pool.Area", "Gr.Liv.Area",
                "Wood.Deck.SF", "Open.Porch.SF", "Enclosed.Porch")]

cm <- cor(contd_vars)

# Q5)
library(corrplot)
corrplot(cm, method = "color", type = "upper", tl.cex = 0.9, tl.col = "black",
    addCoef.col = "black", title = "Correlation plot")



# Q6)
library(ggplot2)

ggplot(ames, aes(x = SalePrice, y = Gr.Liv.Area, color = SalePrice)) +
  geom_point() +
  labs(title = "Scatter Plot: Sale Price vs Ground Living Area",
       x = "Sale Price",
       y = "Ground Living Area")

ggplot(ames, aes(x = SalePrice, y = Pool.Area, color = SalePrice)) +
  geom_point() +
  labs(title = "Scatter Plot: Sale Price  vs Pool Area",
       x = "Sale Price",
       y = "Pool Area")

ggplot(ames, aes(x = SalePrice, y = Lot.Frontage, color = SalePrice)) +
  geom_point() +
  labs(title = "Scatter Plot: Sale Price vs Lot Frontage",
       x = "Sale Price",
       y = "Lot Frontage")



# Q7)
reg_mod <- lm(SalePrice ~ BsmtFin.SF.1 + BsmtFin.SF.2 + Bsmt.Unf.SF, data = ames)

# Q8)
summary(reg_mod)


# Q9)
install.packages("viridis")
library(viridis)

col_palette <- viridis(1)
par(mfrow = c(2, 2))
plot(reg_mod, col = col_palette)
```

```
# Q10)
install.packages("car")
library(car)

vif(reg_mod)

# Q11)
ot <- outlierTest(reg_mod)


# Q12)
row_del <- c(1499,2181,1768,1761,2451,434,2446,2331,2667,2333)
ames <- ames[-row_del, ]
reg_mod <- lm(SalePrice ~ BsmtFin.SF.1 + BsmtFin.SF.2 + Bsmt.Unf.SF, data = ames)
summary(reg_mod)


# Q13)
install.packages('leaps')
library(leaps)
mod_sub = regsubsets(SalePrice ~ BsmtFin.SF.1 + BsmtFin.SF.2 + Bsmt.Unf.SF, data =
ames,
                     nbest = 3)
summary(mod_sub)
```