Intermediate Analytics

ALY6015, Spring 2022

Module 1 Regression Diagnostics – R

Week-1

Submitted by: Abhinav Jain

NUID: 002938209

Submitted To: Roy Wada

Date: 04/17/2022

Introduction: Regression Diagnostics in R for Ames Housing

In this project, our objective is to analyze the dataset through a regression model by investigating different parameters while implementing the regression model. The Ames Housing dataset consist of Ames assessor's office which is taken form the Ames, Iowa Assessor's Office. It contains various nominal, continuous, discrete, and ordinal variables with 2930 Observations and 82 variables includes (23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables)

This report discusses descriptive statistics to understand the data the dimensions, about the variables which will be used while analyzing the investigating the data. After analyzing the data, we have prepared the dataset name as df_hou which we got it while importing the .csv. To do the regression analyses sales price is the dependent variable and other are independent variables like "Lot.Frontage","Lot.Area",Mas.Vnr.Area","BsmtFin.SF.1","Bsmt.Unf.SF",Total.Bsmt.SF","X1st.Flr.SF","X2nd.Flr.SF","Low.Qual.Fin.SF","Gr.Liv.Area","Garage.Area","Wood.Deck.SF","Open.Porch.SF","Enclosed.Porch","X3Ssn.Porch","Screen.Porch","Pool.Area","Misc.Val" these are the selected variable.

Data Cleaning/ imputing missing values: Prepare the dataset to fill the missing values by the mean value which will help to aggregate the NA values in the dataset variable which are used while selecting the data for correlation analysis are Lot. Frontage, Mas. Vnr. Area, BsmtFin.SF1, Bsmt.Unf.Sg, Tota. Bsmt SF will be aggregated by using the na. aggregate function.

Mostly, variables are selected based on requirements as per the regression modeling which we are going to implement in R. While performing the correlation analysis and regression analysis mainly focuses on the dependent variable and independent variables which will help in the dataset. Firstly, need to understand the correlation between the variables which has the highest, lowest, or nearest to the 0.5 value for implementing the regression model.

To create the new dataset to experiment with the regression model which has the highest correlation with the sale price as a dependent variable and other desired 3 continuous variables. After fitting the regression model and understanding the finding to get the patterns and problems with the model. Then check the multicollinearity finding and correct the problem with the model. Understanding the outliers to make the model more accurate and by using the subsets regression method to check the highest accuracy of the new model. Lastly, check the best-fitted model or the preferred model best for this dataset.

Analysis

Task 1 : Import the dataset of Ames Housing

In this task after setting the environment in R studio imported the dataset to understand the variables and dimensions of about the dataset which has 2930 observations and 82 variables.

sumtable {vtable} Summary Statistics

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
ï..Order	2930	1465.5	845.962	1	733.25	2197.75	2930
PID	2930	714464496.989	188730844.649	526301100	528477022.5	907181097.5	1007100110
MS.SubClass	2930	57.387	42.638	20	20	70	190
Lot.Frontage	2440	69.225	23.365	21	58	80	313
Lot.Area	2930	10147.922	7880.018	1300	7440.25	11555.25	215245
Street	2930						
... Grvl	12	0.4%					
... Pave	2918	99.6%					
Alley	198						
... Grvl	120	60.6%					
... Pave	78	39.4%					
Lot.Shape	2930						
... IR1	979	33.4%					
... IR2	76	2.6%					
... IR3	16	0.5%					
... Reg	1859	63.4%					
Land.Contour	2930						
... Bnk	117	4%					
... HLS	120	4.1%					
... Low	60	2%					
... Lvl	2633	89.9%					
Utilities	2930						
... AllPub	2927	99.9%					
... NoSeWa	1	0%					
... NoSewr	2	0.1%					
Lot.Config	2930						
... Corner	511	17.4%					
... CulDSac	180	6.1%					

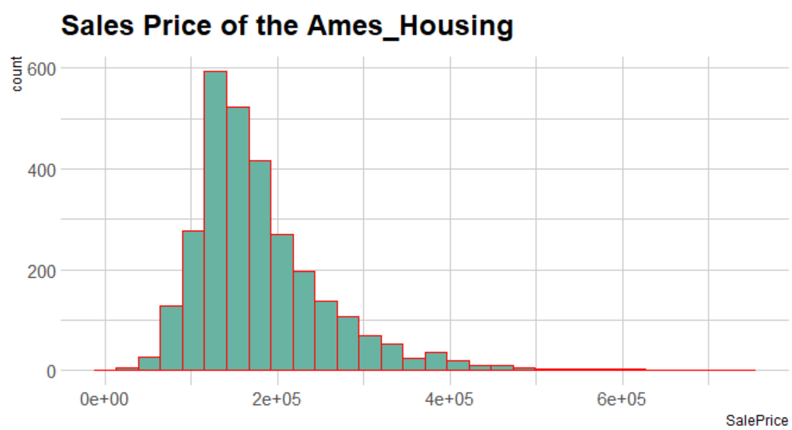
Task 2: Descriptive Analysis and EDA

This describes the variance, mean, median, and other statistical data points to understand the constructive and meaningful patterns in the data

```
> describe(df_hou$SalePrice)
   vars    n     mean       sd median  trimmed     mad   min    max  range skew kurtosis      se
X1    1 2930 180796.1 79886.69 160000 170429.1 54856.2 12789 755000 742211 1.74      5.1 1475.84
> psych::describe(df_hou$Bsmt.Unf.SF)
   vars    n   mean     sd median trimmed    mad min  max range skew kurtosis   se
X1    1 2930 559.26 439.42    466  510.69 415.13   0 2336  2336 0.92     0.41 8.12
> describe(df_hou$Total.Bsmt.SF)
   vars    n    mean     sd median trimmed    mad min  max range skew kurtosis   se
X1    1 2930 1051.61 440.54    990    1035 349.89   0 6110  6110 1.16     9.11 8.14
> describe(df_hou$Gr.Liv.Area)
   vars    n    mean     sd median trimmed    mad min  max range skew kurtosis   se
X1    1 2930 1499.69 505.51   1442 1452.25 461.09 334 5642  5308 1.27     4.12 9.34
```

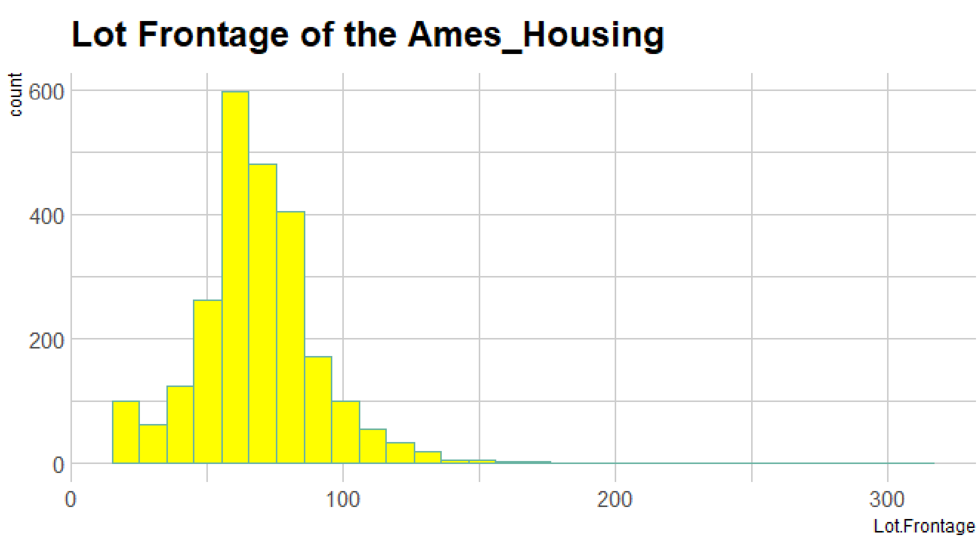
Created a Histogram Sales Price of the Ames_Housing

A histogram which is the graphical analysis of the sales price of the Ames housing dataset interprets the range shown in the dataset as 600. Whereas the sales price is shown in exponential price in the dataset.



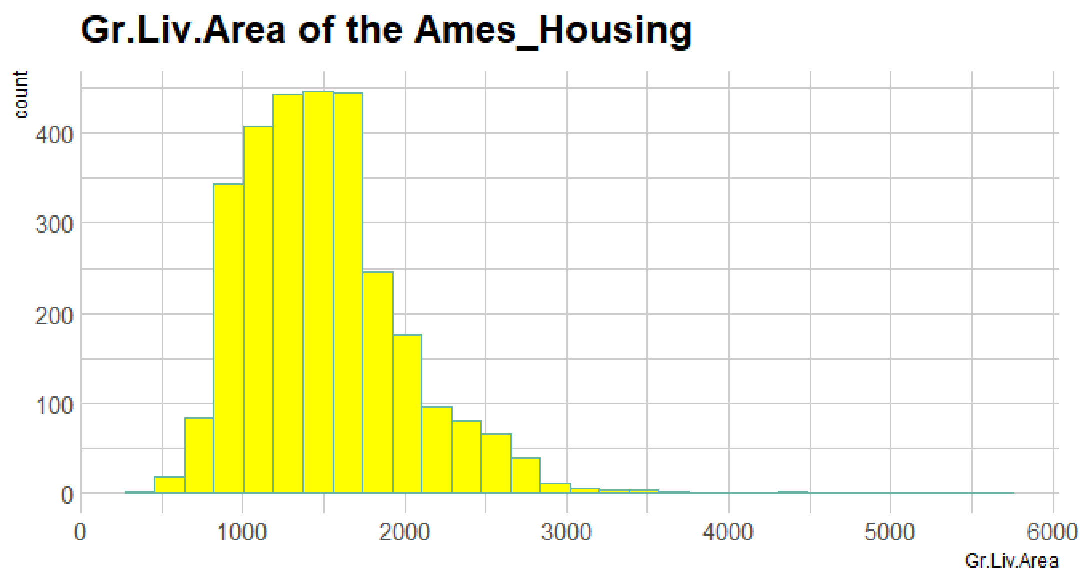
Created a Histogram Lot Frontage of the Ames_Housing

A histogram which is the graphical analysis of the Lot Frontage of the Ames housing dataset interprets the range shown in the dataset as 600. Whereas the Lot Frontage is shown the 300 but this variable have low frequency of dataset to justify the counts.



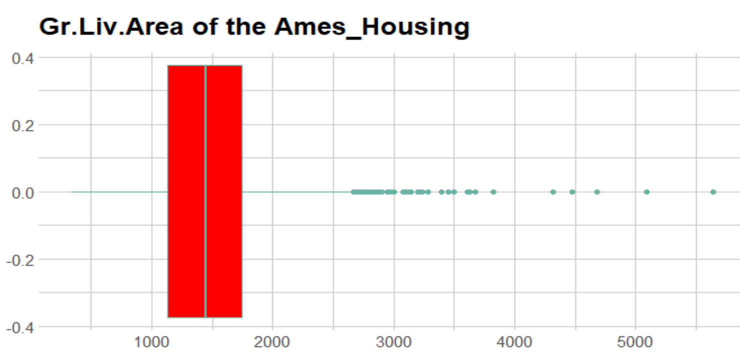
Created a Histogram Gr. Liv. Area of the Ames_Housing

A histogram which is the graphical analysis of the Gr. Liv. Area of the Ames housing dataset interprets the range shown in the dataset as 450. Whereas the Gr. Liv. Area is shown in exponential price in the dataset.



Created a Boxplot Gr.Liv.Area of the Ames_Housing

To analyze the outlier to get an accurate result by understanding the outlier's data value falling in the range. Boxplot help researchers to analyzing the mean value of the variables.



```
> unique(df_hou$Street)
[1] "Pave" "Grvl"
> unique(df_hou$MS.Zoning)
[1] "RL"       "RH"       "FV"       "RM"       "C (all)" "I (all)" "A (agr)"
```

Task 3: In this task preparing the data for modeling and imputing the missing values with the mean of the variable in the required fields.

 # Lot.Frontage

```
> df_hou$Lot.Frontage<- na.aggregate(df_hou$Lot.Frontage)
> summary(df_hou$Lot.Frontage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  21.00   60.00   69.22   69.22   78.00  313.00 
```

 # Mas.Vnr.Area

```
> df_hou$Mas.Vnr.Area<- na.aggregate(df_hou$Mas.Vnr.Area)
> summary(df_hou$Mas.Vnr.Area)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0     0.0     0.0   101.9   162.8  1600.0 
```

 # Total.Bsmt.SF

```
> df_hou$Bsmt.Unf.SF<-na.aggregate(df_hou$Bsmt.Unf.SF)
> summary(df_hou$Bsmt.Unf.SF)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   219.0   466.0   559.3   801.8  2336.0 
```

 # Garage.Area

```
> df_hou$Garage.Area<-na.aggregate(df_hou$Garage.Area)
> summary(df_hou$Garage.Area)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   320.0   480.0   472.8   576.0  1488.0 
```

Task 4: Created a dataset for implementing the correlation function of the numerical variables

In this task, the correlation matrix will help in expressing the relation between the coefficient between different variables.

Implementing the correlation function to form a correlation matrix

Correlation matrix for the data frame df_cor: In this task correlation with value 1 shows the correlated variables which will help in understanding the how the variables are correlated.

```
> round(cor(df_cor), 2)
```

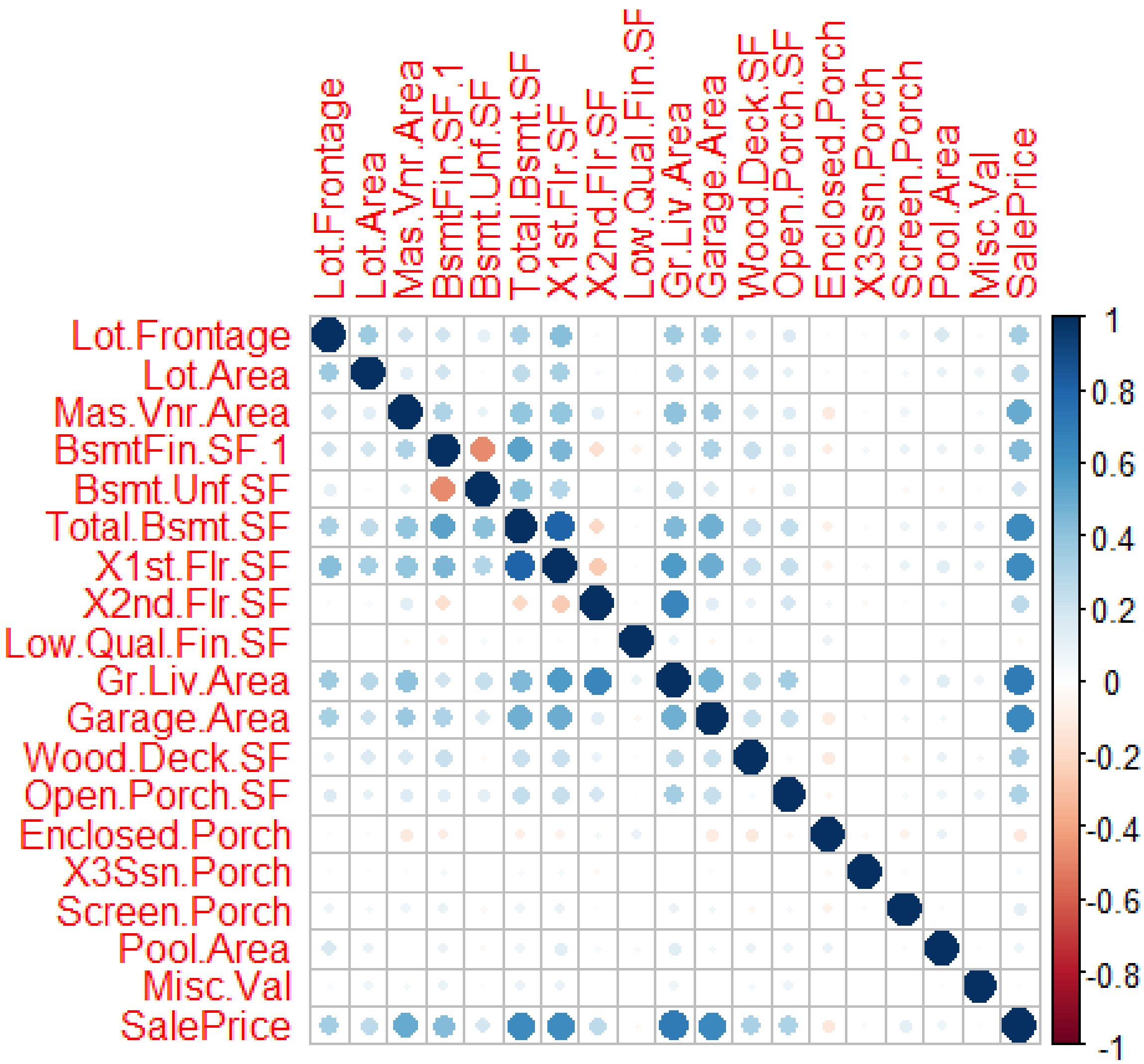
	Lot.Frontage	Lot.Area	Mas.Vnr.Area	BsmtFin.SF.1	Bsmt.Unf.SF	Total.Bsmt.SF	X1st.Flr.SF	X2nd.Flr.SF	Low.Qual.Fin.SF
Lot.Frontage	1.00	0.37	0.20	0.20	0.11	0.33	0.42	0.03	0.01
Lot.Area	0.37	1.00	0.13	0.19	0.02	0.25	0.33	0.03	0.00
Mas.Vnr.Area	0.20	0.13	1.00	0.30	0.09	0.40	0.39	0.12	-0.06
BsmtFin.SF.1	0.20	0.19	0.30	1.00	-0.48	0.54	0.46	-0.16	-0.07
Bsmt.Unf.SF	0.11	0.02	0.09	-0.48	1.00	0.41	0.30	0.00	0.05
Total.Bsmt.SF	0.33	0.25	0.40	0.54	0.41	1.00	0.80	-0.21	-0.02
X1st.Flr.SF	0.42	0.33	0.39	0.46	0.30	0.80	1.00	-0.25	-0.01
X2nd.Flr.SF	0.03	0.03	0.12	-0.16	0.00	-0.21	-0.25	1.00	0.02
Low.Qual.Fin.SF	0.01	0.00	-0.06	-0.07	0.05	-0.02	-0.01	0.02	1.00
Gr.Liv.Area	0.35	0.29	0.40	0.21	0.24	0.44	0.56	0.66	0.10
Garage.Area	0.34	0.21	0.37	0.31	0.16	0.49	0.49	0.13	-0.05
Wood.Deck.SF	0.10	0.16	0.17	0.22	-0.04	0.23	0.23	0.09	-0.02
Open.Porch.SF	0.15	0.10	0.14	0.12	0.12	0.25	0.24	0.18	0.00
Enclosed.Porch	0.01	0.02	-0.11	-0.10	0.01	-0.09	-0.07	0.06	0.09
X3Ssn.Porch	0.03	0.02	0.01	0.05	-0.01	0.04	0.04	-0.03	0.00
Screen.Porch	0.07	0.06	0.07	0.10	-0.05	0.08	0.10	0.01	0.01
Pool.Area	0.16	0.09	0.00	0.08	-0.03	0.07	0.12	0.04	0.04
Misc.Val	0.04	0.07	0.04	0.09	-0.01	0.08	0.09	-0.01	-0.01
SalePrice	0.34	0.27	0.51	0.43	0.18	0.63	0.62	0.27	-0.04

	Gr.Liv.Area	Garage.Area	Wood.Deck.SF	Open.Porch.SF	Enclosed.Porch	X3Ssn.Porch	Screen.Porch	Pool.Area	Misc.Val	SalePrice
Lot.Frontage	0.35	0.34	0.10	0.15	0.01	0.03	0.07	0.16	0.04	0.34
Lot.Area	0.29	0.21	0.16	0.10	0.02	0.02	0.06	0.09	0.07	0.27
Mas.Vnr.Area	0.40	0.37	0.17	0.14	-0.11	0.01	0.07	0.00	0.04	0.51
BsmtFin.SF.1	0.21	0.31	0.22	0.12	-0.10	0.05	0.10	0.08	0.09	0.43
Bsmt.Unf.SF	0.24	0.16	-0.04	0.12	0.01	-0.01	-0.05	-0.03	-0.01	0.18
Total.Bsmt.SF	0.44	0.49	0.23	0.25	-0.09	0.04	0.08	0.07	0.08	0.63
X1st.Flr.SF	0.56	0.49	0.23	0.24	-0.07	0.04	0.10	0.12	0.09	0.62
X2nd.Flr.SF	0.66	0.13	0.09	0.18	0.06	-0.03	0.01	0.04	-0.01	0.27
Low.Qual.Fin.SF	0.10	-0.05	-0.02	0.00	0.09	0.00	0.01	0.04	-0.01	-0.04
Gr.Liv.Area	1.00	0.48	0.25	0.34	0.00	0.01	0.09	0.14	0.07	0.71
Garage.Area	0.48	1.00	0.24	0.23	-0.11	0.03	0.06	0.05	0.01	0.64
Wood.Deck.SF	0.25	0.24	1.00	0.04	-0.12	0.00	-0.05	0.09	0.06	0.33
Open.Porch.SF	0.34	0.23	0.04	1.00	-0.06	-0.01	0.05	0.06	0.08	0.31
Enclosed.Porch	0.00	-0.11	-0.12	-0.06	1.00	-0.03	-0.06	0.09	0.01	-0.13
X3Ssn.Porch	0.01	0.03	0.00	-0.01	-0.03	1.00	-0.03	-0.01	0.00	0.03
Screen.Porch	0.09	0.06	-0.05	0.05	-0.06	-0.03	1.00	0.03	0.01	0.11
Pool.Area	0.14	0.05	0.09	0.06	0.09	-0.01	0.03	1.00	0.01	0.07
Misc.Val	0.07	0.01	0.06	0.08	0.01	0.00	0.01	0.01	1.00	-0.02
SalePrice	0.71	0.64	0.33	0.31	-0.13	0.03	0.11	0.07	-0.02	1.00

Task5: Created a scatter plot with after plotting the correlation matrix.

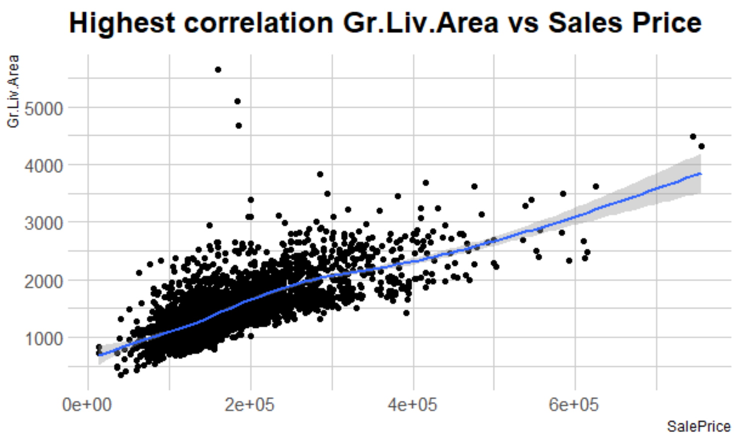
 # Correlation matrix for the continuous variables

Below the matrix shows that the dark blue color indicates the highest correlation between the two variables while darker the red color will be the lowest correlated variables.

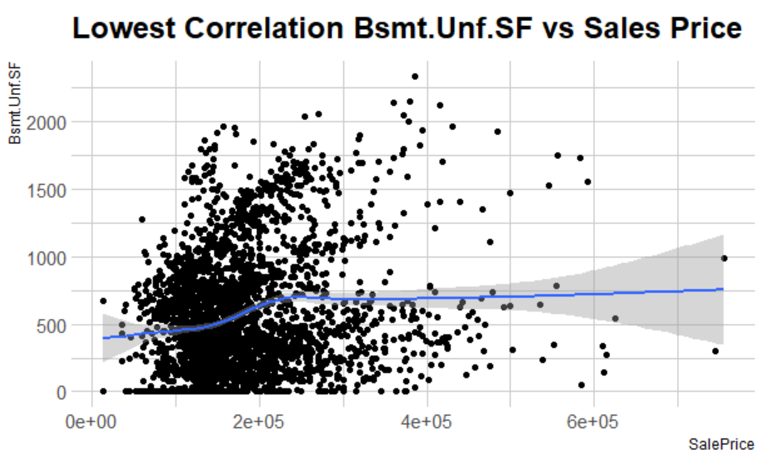


Task 6: Created a scatter plot with the highest, lowest and nearest to 0.5 correlation of the variable while using the sale price

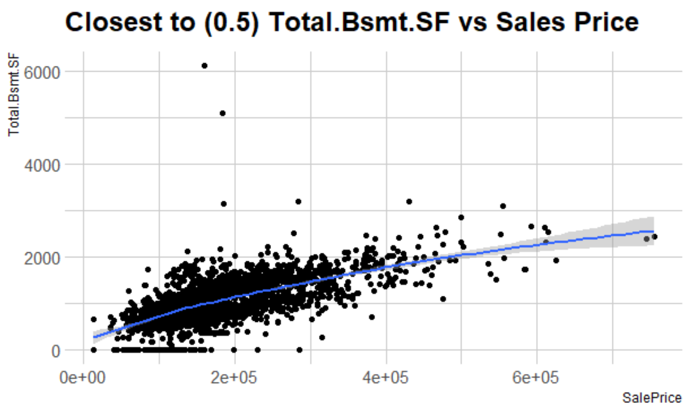
 # Highest Correlation: In the below graph directly shows that the maximum points are near the line which proves the Saleprice is highly correlated with Gr. live Area.



 # Lowest Correlation In the below graph most of the data points are scattered from the line which shows that sales price and Bsmt Unf.SF is less correlated.



#Closed to 0.5 : In this graph Sales price and Total Bsmt SF price is perfectly correlated



Task 7: By using 3 variables implemented the regression model in the dataset to analyze the model to check the accuracy.

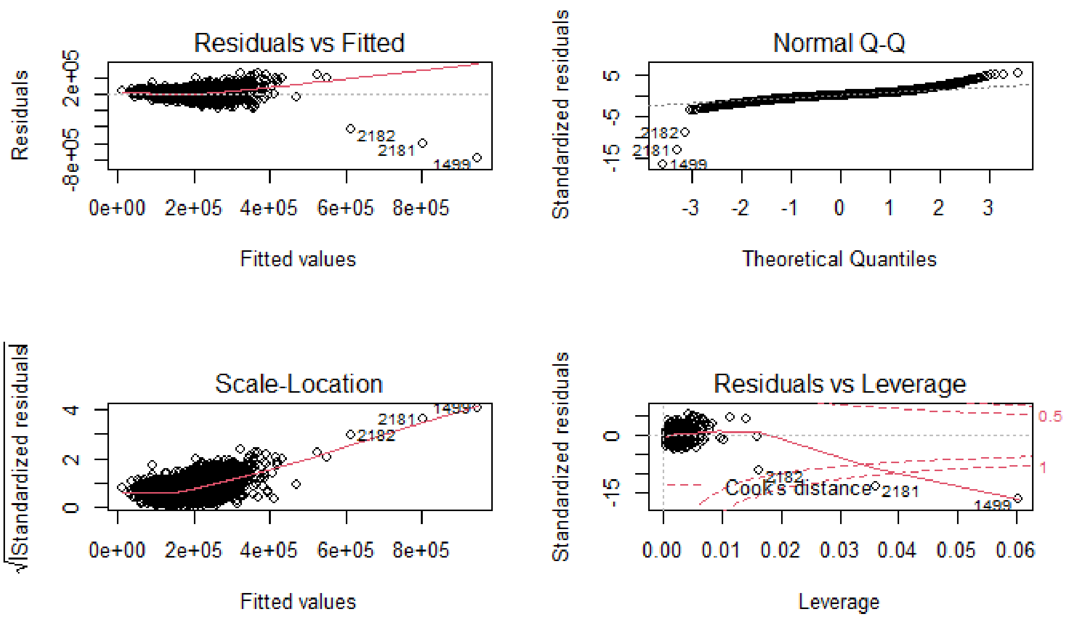
This model shows the 64% accuracy which shows the sales price might be dependent on other independent variable as the accuracy should be 80% to 90% is acceptance in terms of statistics by giving the strength between the two variables. Whereas the Akaike information criterion AIC and BIC provides the better likelihood of the model. Below the data shows the 64% fitted R square and adujusted Rsquare is also 64%.

```
> #Model 1
> df_reg <- as.data.frame(df_cor)
> fit <- lm(SalePrice ~ Gr.Liv.Area+ Bsmt.Unf.SF+ Total.Bsmt.SF, data = df_reg)
> summ(fit)
MODEL INFO:
Observations: 2930
Dependent Variable: SalePrice
Type: OLS linear regression

MODEL FIT:
F(3,2926) = 1725.52, p = 0.00
R² = 0.64
Adj. R² = 0.64

Standard errors: OLS
------------------------------------------------------------
                           Est.       S.E.    t val.       p
------------------- ----------- ---------- --------- ------
(Intercept)           -18837.76    2977.55     -6.33    0.00
Gr.Liv.Area               85.17       1.96     43.37    0.00
Bsmt.Unf.SF              -23.13       2.22    -10.42    0.00
Total.Bsmt.SF             80.68       2.40     33.57    0.00
------------------------------------------------------------
> summary(fit)$adj.r.squared
[1] 0.6385088
> AIC(fit)
[1] 71489.47
> BIC(fit)
[1] 71519.38
```

Task 8: Implementation of the equation of coefficient of the model

```
> SalePrice <- (-18837.76) + (85.17 * df_reg$Gr.Liv.Area) + (-23.13 * df_reg$Bsmt.Unf.SF) + (80.68 *df_reg$Total.Bsmt.SF)
> summary(fit)$coefficient
                Estimate  Std. Error    t value      Pr(>|t|)
(Intercept)  -18837.75786 2977.553538  -6.326589  2.889876e-10
Gr.Liv.Area      85.17140    1.964020  43.365848 1.007165e-317
Bsmt.Unf.SF     -23.13385    2.220830 -10.416759  5.651381e-25
Total.Bsmt.SF    80.67688    2.403497  33.566455 2.981637e-209
```

Task 9: Using plot() the regression model by four graphs and produced



Task 10:Check the multicollinearity and findings about existing multicollinearity

Variance inflation factor shows the estimated regression coefficient the value is about > 0.7 this indicate existence of multi collinearity among the 3 continious variables.

```
> vif(fit)
  Gr.Liv.Area   Bsmt.Unf.SF Total.Bsmt.SF
     1.251475      1.209098      1.423411
```

Task 11: Applied the outlier and findings of existing the observations from the fit regression model

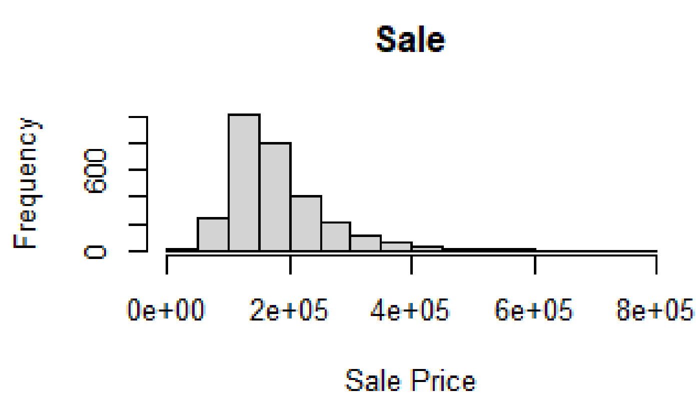
d

```
> outlierTest(model = fit)
       rstudent unadjusted p-value Bonferroni p
1499 -17.712259         9.1528e-67   2.6818e-63
2181 -13.487911         2.8727e-40   8.4171e-37
2182  -9.099272         1.6389e-19   4.8018e-16
434     5.485918         4.4645e-08   1.3081e-04
45      5.124923         3.1702e-07   9.2887e-04
1638    4.877380         1.1325e-06   3.3183e-03
1768    4.875346         1.1442e-06   3.3524e-03
1064    4.723684         2.4253e-06   7.1062e-03
2333    4.416007         1.0419e-05   3.0527e-02
```

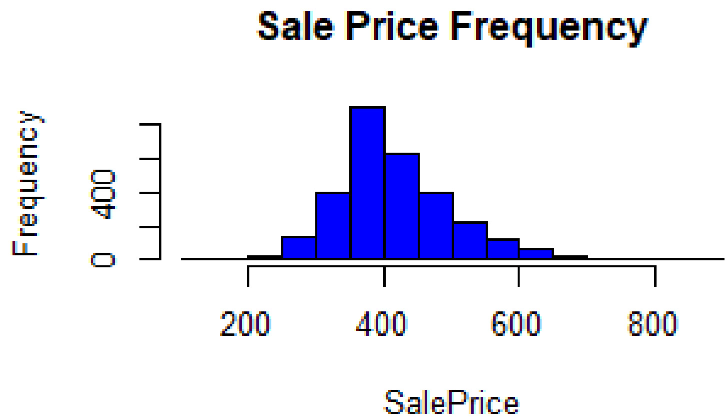
Task 12: Model 2 - After discovering the low accuracy implemented the regression model after making changes.

#Created a histogram of Sales Price before fitting

```
> hist(df_cor$SalePrice,xlab = "Sale Price",main = "Sale")
> summary(powerTransform(df_cor$SalePrice))
bcPower Transformation to Normality
                 Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
df_cor$SalePrice    0.0076           0      -0.0501       0.0654

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                             LRT df     pval
LR test, lambda = (0) 0.06741106  1 0.79514

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 966.3636  1 < 2.22e-16
```



Sales Price after fitting the regression MODEL




```
> df_cor$SalePrice_sqrt <- sqrt(df_cor$SalePrice)
> hist(df_cor$SalePrice_sqrt,xlab = "SalePrice",main = "Sale Price Frequency",
+      col="BLue")
> fit_model2<-lm(SalePrice_sqrt~ Gr.Liv.Area+ Bsmt.Unf.SF+ Total.Bsmt.SF,data=df_cor)
> summary(fit_model2)

Call:
lm(formula = SalePrice_sqrt ~ Gr.Liv.Area + Bsmt.Unf.SF + Total.Bsmt.SF,
    data = df_cor)

Residuals:
    Min      1Q  Median      3Q     Max
-847.55  -24.82    3.20   27.97  191.89

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   197.589342   3.194450  61.854   <2e-16 ***
Gr.Liv.Area     0.093074   0.002107  44.172   <2e-16 ***
Bsmt.Unf.SF    -0.023466   0.002383  -9.849   <2e-16 ***
Total.Bsmt.SF   0.087687   0.002579  34.006   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.53 on 2926 degrees of freedom
Multiple R-squared:  0.6475,    Adjusted R-squared:  0.6471
F-statistic:  1791 on 3 and 2926 DF,  p-value: < 2.2e-16
```

The above graphs show after finding the model with low accuracy show the 64% fitted but after refining the model it shows that there is no change in the values. Just have the scope to improve the model by removing outliers.

Task 13: Created the subset of the regression model to take the best outcomes and run the equation

#Regression Model after creating the subset

```
> df_hou_sub = subset(df_reg, select = c(SalePrice,Gr.Liv.Area,Bsmt.Unf.SF,Total.Bsmt.SF))
> fit_sub<-lm(SalePrice ~ Gr.Liv.Area+ Bsmt.Unf.SF+ Total.Bsmt.SF,data=df_hou_sub)
> summary(fit_sub)

Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Bsmt.Unf.SF + Total.Bsmt.SF,
    data = df_hou_sub)

Residuals:
    Min      1Q  Median      3Q     Max
-783855  -22259     715   20235  261637

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -18837.758   2977.554  -6.327 2.89e-10 ***
Gr.Liv.Area       85.171      1.964  43.366  < 2e-16 ***
Bsmt.Unf.SF      -23.134      2.221 -10.417  < 2e-16 ***
Total.Bsmt.SF     80.677      2.403  33.566  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48030 on 2926 degrees of freedom
Multiple R-squared:  0.6389,    Adjusted R-squared:  0.6385
F-statistic:  1726 on 3 and 2926 DF,  p-value: < 2.2e-16
```

 # Implementing the Backward Selection :

```
> stepAIC(fit_sub,direction="backward")
Start:  AIC=63172.49
SalePrice ~ Gr.Liv.Area + Bsmt.Unf.SF + Total.Bsmt.SF

                Df  Sum of Sq        RSS   AIC
<none>                        6.7503e+12 63172
- Bsmt.Unf.SF    1 2.5033e+11 7.0006e+12 63277
- Total.Bsmt.SF  1 2.5993e+12 9.3496e+12 64125
- Gr.Liv.Area    1 4.3385e+12 1.1089e+13 64625

Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Bsmt.Unf.SF + Total.Bsmt.SF,
    data = df_hou_sub)

Coefficients:
  (Intercept)    Gr.Liv.Area    Bsmt.Unf.SF  Total.Bsmt.SF
    -18837.76          85.17         -23.13          80.68
```

#Implementing the Forward Selection

```
> stepAIC(fit_sub,direction="forward")
Start:  AIC=63172.49
SalePrice ~ Gr.Liv.Area + Bsmt.Unf.SF + Total.Bsmt.SF


Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Bsmt.Unf.SF + Total.Bsmt.SF,
    data = df_hou_sub)

Coefficients:
  (Intercept)    Gr.Liv.Area    Bsmt.Unf.SF  Total.Bsmt.SF
    -18837.76          85.17         -23.13          80.68
```

Implementing the Both selections

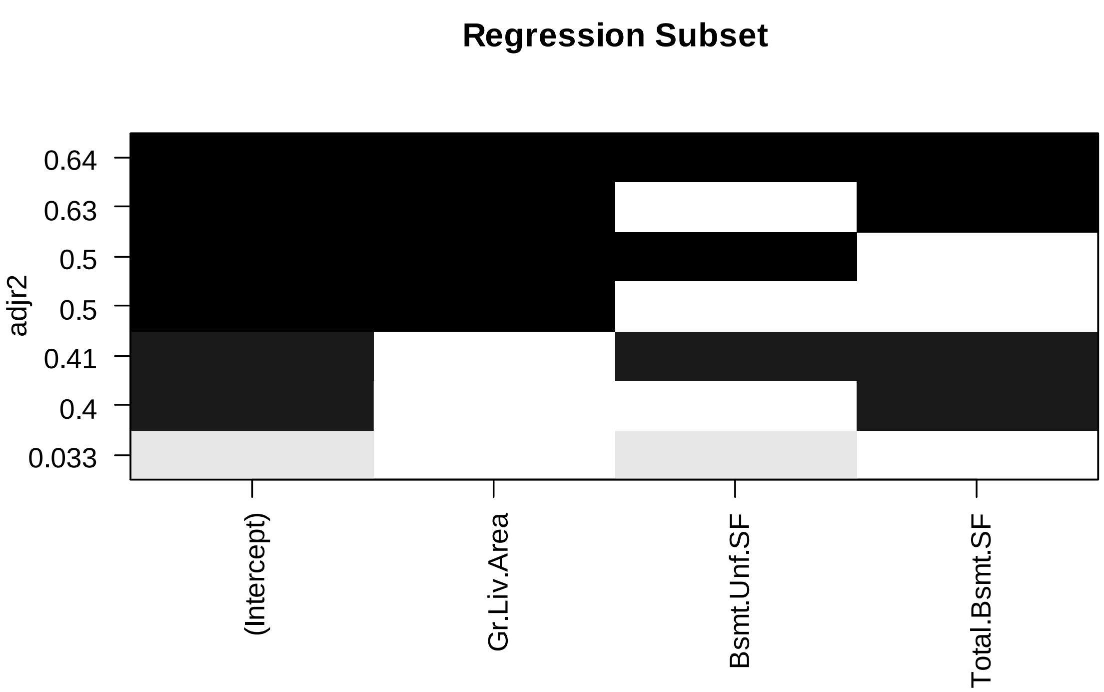
```
> stepAIC(fit_sub,direction="both")
Start:  AIC=63172.49
SalePrice ~ Gr.Liv.Area + Bsmt.Unf.SF + Total.Bsmt.SF

                Df  Sum of Sq        RSS   AIC
<none>                         6.7503e+12 63172
- Bsmt.Unf.SF    1 2.5033e+11 7.0006e+12 63277
- Total.Bsmt.SF  1 2.5993e+12 9.3496e+12 64125
- Gr.Liv.Area    1 4.3385e+12 1.1089e+13 64625

Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Bsmt.Unf.SF + Total.Bsmt.SF,
    data = df_hou_sub)

Coefficients:
  (Intercept)    Gr.Liv.Area    Bsmt.Unf.SF  Total.Bsmt.SF
    -18837.76          85.17         -23.13          80.68
```

Task 14: In this task created the model to check the difference and understand the accuracy of both using the leap function and selecting the correct variables this shows that

1. Gr. live Area is the best the one predictor variable
2. Gr.live area and Total Bsmt.SF is the second-best two predictor model.
3. All three of the predictor model are best with three Gr.Live, Bsmt.Unf.SF and Total Bsmt.Sf

```
> leap<-regsubsets(SalePrice~ Gr.Liv.Area+ Bsmt.Unf.SF+ Total.Bsmt.SF,data=df_hou_sub,nbest=4)
> plot(leap,scale="adjr2")
> summary(leap)
Subset selection object
Call: regsubsets.formula(SalePrice ~ Gr.Liv.Area + Bsmt.Unf.SF + Total.Bsmt.SF,
    data = df_hou_sub, nbest = 4)
3 Variables  (and intercept)
              Forced in Forced out
Gr.Liv.Area       FALSE      FALSE
Bsmt.Unf.SF       FALSE      FALSE
Total.Bsmt.SF     FALSE      FALSE
4 subsets of each size up to 3
Selection Algorithm: exhaustive
         Gr.Liv.Area Bsmt.Unf.SF Total.Bsmt.SF
1  ( 1 ) "*"         " "         " "
1  ( 2 ) " "         " "         "*"
1  ( 3 ) " "         "*"         " "
2  ( 1 ) "*"         " "         "*"
2  ( 2 ) "*"         "*"         " "
2  ( 3 ) " "         "*"         "*"
3  ( 1 ) "*"         "*"         "*"
```

Below the matrix show the Bsmt.Unf.Sf shows the low accuracy with a value of 0.33 whereas, Gr.Liv.area and Total Bsmt.SF creates the does not contain the low intercept value. If we check the adjusted R-square .64 which is 64% accuracy for all the three variables it suggested that all three predictors best subset for the model.



Conclusion and Interpretation

After configuring the environment in R studio, import the dataset to learn about the variables and dimensions of the dataset, which has 2930 observations and 82 variables. To grasp the constructive and meaningful patterns in the data, this describes the variance, mean, median, and other statistical data points. The most crucial phase is preparing the data for modeling and filling in the missing values in the appropriate fields using the variable's mean. The correlation matrix will assist you in expressing the relationship between the coefficients of distinct variables in this work. The correlated variables are shown in this task as correlation with value 1, which will aid in understanding how the variables are connected. Whereas, In the correlation matrix, the dark blue hue denotes the best correlation between the two variables, while the darker red color indicates the lowest correlation between the two variables.

To check the correlation within the matrix- Highest Correlation: demonstrates that the greatest points are close to the line, indicating that Saleprice is highly associated with Gross Live Area, it is observed Minimum Correlation The majority of the data points are dispersed from the line, indicating that sales price and Bsmt Unf.SF is less connected moreover 0.5 or less: the price of sales and the price of total Bsmt SF are exactly associated in this graph. The regression model evaluated the 64 percent accuracy, indicating that the sales price may be influenced by other independent factors, while the accuracy should be between 80 and 90 percent in terms of statistics, indicating the strength of the relationship between the two variables. The Akaike information criterion AIC and BIC, on the other hand, offer a greater probability of the model.

In conclusion, the sales price is 64% dependent on the Ground living area, basement on the second floor, and total basement in the second floor. The data below indicates that the fitted R square is 64 percent and that the adjusted R-square is similarly 64 percent. If we can see the estimated regression coefficient is shown by the variance inflation factor, which has a value of about > 0.7, indicating the presence of multicollinearity among the three continuous variables. It is more to discover about a model with poor accuracy, the values are 64 percent fitted, but after improving the model, the values are unchanged. You only need to remove outliers from the model to enhance it. The best predictive variable is Gr. live Area. The second-best two predictor model is Gr.live area and Total Bsmt.SF. With three Gr.Live, Bsmt.Unf.SF, and Total Bsmt.Sf, all three prediction models perform best. The Bsmt Unf. Sf is shown below the matrix has a poor precision of 0.33, whereas Gr.Liv.area and Total Bsmt. SF generates that does not have a low intercept value. If we look at the adjusted R-square(.64), which is 64 percent accuracy for all three variables, we can conclude that all three predictors are the optimal subset for the model.

References:

[1] What is a Correlation Matrix?

https://www.displayr.com/what-is-a-correlation-matrix/#:~:text=A%20correlation%20matrix%20is%20a,a%20diagnostic%20for%20advanced%20analyses.

[2] Lesson 15 - Correlation and Simple Linear Regression in R ...

View Lesson 15 - Correlation and Simple Linear Regression in R _Notes_.pdf from ACMS 10145 at University of Notre Dame. R: Correlation and Simple Linear Regression ITAO 20200: Statistical Inference

https://www.coursehero.com/file/66402542/Lesson-15-Correlation-and-Simple-Linear...

[3]The Complete Guide: How to Report Regression Results

Zach

https://www.statology.org/how-to-report-regression-results/

[4]Stepwise Regression Essentials in R

Kassambara et al.

http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/

[5]Best Subsets Regression Essentials in R

Kassambara et al.

http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/

Appendix: R Script for Ames_Housing data

```
#++++++++++++++++++++++++++++++
#+Regression Diagnostics
#+and Features Selection in R
#+Module 1 - ALY6015
#+Author : Abhinav Jain
#++++++++++++++++++++++++++++++


#importing the libraries
library(tidyverse)
library(hrbrthemes)
library(dplyr)
library(ggplot2)
library(corrplot)
library(RColorBrewer)
library(caTools)
library(car)
library(BSDA)
library(vtable)
library(zoo)
library(Hmisc)
library(psych)
library(fs)
library(corrplot)
library(PerformanceAnalytics)
library(jtools)
library(car)
library(BSDA)
library(MASS)
library(leaps)
#===============
#EDA
#===============
#Import the Ames housing dataset.

df_hou <- read.csv("AmesHousing.csv")
df_hou
summary(df_hou)
st(df_hou)

 #===============

# Perform Exploratory Data Analysis and use descriptive statistics to describe the data.
describe(df_hou$SalePrice)
psych::describe(df_hou)
ggplot(df_hou, aes(x= SalePrice))+
  geom_histogram(fill="#69b3a2", color="#FF0000")+
  ggtitle("Sales Price of the Ames_Housing")+
  theme_ipsum()

ggplot(df_hou, aes(x= Lot.Frontage))+
  geom_histogram(fill="#FFFF00", color="#69b3a2")+
  ggtitle("Lot Frontage of the Ames_Housing")+
  theme_ipsum()
```

```
#===============

# Perform Exploratory Data Analysis and use descriptive statistics to describe the data.
describe(df_hou$SalePrice)
psych::describe(df_hou)
ggplot(df_hou, aes(x= SalePrice))+
  geom_histogram(fill="#69b3a2", color="#FF0000")+
  ggtitle("Sales Price of the Ames_Housing")+
  theme_ipsum()

ggplot(df_hou, aes(x= Lot.Frontage))+
  geom_histogram(fill="#FFFF00", color="#69b3a2")+
  ggtitle("Lot Frontage of the Ames_Housing")+
  theme_ipsum()

Gr.Liv.Area
ggplot(df_hou, aes(x= Gr.Liv.Area))+
  geom_histogram(fill="#FFFF00", color="#69b3a2")+
  ggtitle("Gr.Liv.Area of the Ames_Housing")+
  theme_ipsum()


ggplot(df_hou, aes(x= Gr.Liv.Area))+
  geom_boxplot(fill="#FF0000", color="#69b3a2")+
  ggtitle("Boxplot Gr.Liv.Area of the Ames_Housing")+
  theme_ipsum()


unique(df_hou$Street)

unique(df_hou$MS.Zoning)

# Prepare the dataset for modeling by imputing missing values with the variable's mean value or any other value that you prefer.
#Aggregate the NA(Missing values) values of Lot.Frontage

df_hou$Lot.Frontage<- na.aggregate(df_hou$Lot.Frontage)
summary(df_hou$Lot.Frontage)

df_hou$Mas.Vnr.Area<- na.aggregate(df_hou$Mas.Vnr.Area)
summary(df_hou$Mas.Vnr.Area)

df_hou$BsmtFin.SF.1<- na.aggregate(df_hou$BsmtFin.SF.1)
summary(df_hou$BsmtFin.SF.1)

df_hou$Bsmt.Unf.SF<-na.aggregate(df_hou$Bsmt.Unf.SF)
summary(df_hou$Bsmt.Unf.SF)

df_hou$Total.Bsmt.SF<-na.aggregate(df_hou$Total.Bsmt.SF)
summary(df_hou$Total.Bsmt.SF)

df_hou$Garage.Area<-na.aggregate(df_hou$Garage.Area)
summary(df_hou$Garage.Area)
```

```
df_hou$Garage.Area<-na.aggregate(df_hou$Garage.Area)
summary(df_hou$Garage.Area)

df_cor<- df_hou[,c("Lot.Frontage", "Lot.Area", "Mas.Vnr.Area", "BsmtFin.SF.1",
                   "Bsmt.Unf.SF", "Total.Bsmt.SF","X1st.Flr.SF","X2nd.Flr.SF",
                   "Low.Qual.Fin.SF","Gr.Liv.Area","Garage.Area","Wood.Deck.SF","Open.Porch.SF","Enclosed.Porch",
                   "X3Ssn.Porch","Screen.Porch","Pool.Area","Misc.Val","SalePrice")]
summary(df_cor)

df_cor$Lot.Frontage<-as.numeric(df_cor$Lot.Frontage)
df_cor$Lot.Area<-as.numeric(df_cor$Lot.Area)
df_cor$Mas.Vnr.Area<-as.numeric(df_cor$Mas.Vnr.Area)
df_cor$BsmtFin.SF.1<-as.numeric(df_cor$BsmtFin.SF.1)
df_cor$Bsmt.Unf.SF<-as.numeric(df_cor$Bsmt.Unf.SF)
df_cor$Total.Bsmt.SF<-as.numeric(df_cor$Total.Bsmt.SF)
df_cor$X1st.Flr.SF<-as.numeric(df_cor$X1st.Flr.SF)
df_cor$X2nd.Flr.SF<-as.numeric(df_cor$X2nd.Flr.SF)
df_cor$Low.Qual.Fin.SF<-as.numeric(df_cor$Low.Qual.Fin.SF)
df_cor$Gr.Liv.Area<-as.numeric(df_cor$Gr.Liv.Area)
df_cor$Garage.Area<-as.numeric(df_cor$Garage.Area)
df_cor$Wood.Deck.SF<-as.numeric(df_cor$Wood.Deck.SF)
df_cor$Open.Porch.SF<-as.numeric(df_cor$Open.Porch.SF)
df_cor$Enclosed.Porch<-as.numeric(df_cor$Enclosed.Porch)
df_cor$X3Ssn.Porch<-as.numeric(df_cor$X3Ssn.Porch)
df_cor$Screen.Porch<-as.numeric(df_cor$Screen.Porch)
df_cor$Pool.Area<-as.numeric(df_cor$Pool.Area)
df_cor$Misc.Val<-as.numeric(df_cor$Misc.Val)
df_cor$SalePrice<-as.numeric(df_cor$SalePrice)

# Use the "cor()" function to produce a correlation matrix of the numeric values.
cor(df_cor)
round(cor(df_cor), 2)

# Produce a plot of the correlation matrix, and explain how to interpret it. (hint - check the corrplot or ggcorrplot
corrplot(cor(df_cor), method = "circle")

pairs(df_cor[,1:4], pch = 19)
my_cols <- c("SalePrice", "Gr.Liv.Area", "Bsmt.Unf.SF","Total.Bsmt.SF")
pairs(my_cols, pch = 19,  cex = 0.5,
      col = my_cols[df_cor$SalePrice],
      lower.panel=NULL)
#
plot(df_cor$SalePrice)


# Make a scatter plot for the X continuous variable with the highest correlation with SalePrice. Do the same for the
#Finally, make a scatter plot between X and SalePrice with the correlation closest to 0.5. Interpret the scatter plot
#HighestCorrelation
ggplot(df_cor, aes(x= SalePrice, y= Gr.Liv.Area ))+
  geom_point()+
  ggtitle("Highest correlation Gr.Liv.Area vs Sales Price")+
  theme_ipsum()+
  stat_smooth()
```

```
#Lowest Correlation
ggplot(df_cor, aes(x= SalePrice, y= Bsmt.Unf.SF ))+
  geom_point()+
  ggtitle("Lowest Correlation Bsmt.Unf.SF vs Sales Price")+
  theme_ipsum()+
  stat_smooth()

#Correlation with Closest to 0.5
ggplot(df_cor, aes(x= SalePrice, y= Total.Bsmt.SF ))+
  geom_point()+
  ggtitle("Closest to (0.5) Total.Bsmt.SF vs Sales Price")+
  theme_ipsum()+
  stat_smooth()


# Using at least 3 continuous variables, fit a regression model in R.
#Model 1
df_reg <- as.data.frame(df_cor)
fit <- lm(SalePrice ~ Gr.Liv.Area+ Bsmt.Unf.SF+ Total.Bsmt.SF, data = df_reg)
summ(fit)
summary(fit)$adj.r.squared
AIC(fit)
BIC(fit)
par(mfrow=c(2,2))
plot(fit)



# Report the model in equation form and interpret each coefficient of the model in the context of this problem.
SalePrice <- (-18837.76) + (85.17 * df_reg$Gr.Liv.Area) + (-23.13 * df_reg$Bsmt.Unf.SF) + (80.68 *df_reg$Total.Bsmt.SF)
summary(fit)$coefficient



# Use the "plot()" function to plot your regression model. Interpret the four graphs that are produced.
par(mfrow=c(2,2))
plot(fit)
#
crPlots(model=fit)
qqnorm(df_reg$SalePrice)
qqline(df_reg$SalePrice)
qqPlot(df_reg$Gr.Liv.Area)
sd(df_reg$SalePrice)

spreadLevelPlot(fit)
# Check your model for multicollinearity and report your findings. What steps would you take to correct multicollinearity if it exists?

vif(fit)
# Check your model for outliers and report your findings. Should these observations be removed from the model?
```


```
spreadLevelPlot(fit)
# Check your model for multicollinearity and report your findings. What steps would you take to correct multicollinearity if it exists?

vif(fit)
# Check your model for outliers and report your findings. Should these observations be removed from the model?
```


```
outlierTest(model = fit)
# Attempt to correct any issues that you have discovered in your model. Did your changes improve the model, why or why not?
hist(df_cor$SalePrice,xlab = "Sale Price",main = "Sale")

summary(powerTransform(df_cor$SalePrice))

df_cor$SalePrice_sqrt <- sqrt(df_cor$SalePrice)

hist(df_cor$SalePrice_sqrt,xlab = "SalePrice",main = "Sale Price Frequency",
     col="BLue")

fit_model2<-lm(SalePrice_sqrt~ Gr.Liv.Area+ Bsmt.Unf.SF+ Total.Bsmt.SF,data=df_cor)

summary(fit_model2)



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

df_hou_sub = subset(df_reg, select = c(SalePrice,Gr.Liv.Area,Bsmt.Unf.SF,Total.Bsmt.SF))

fit_sub<-lm(SalePrice ~ Gr.Liv.Area+ Bsmt.Unf.SF+ Total.Bsmt.SF,data=df_hou_sub)
summary(fit_sub)

stepAIC(fit_sub,direction="backward")
stepAIC(fit_sub,direction="forward")
stepAIC(fit_sub,direction="both")

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

leap<-regsubsets(SalePrice~ Gr.Liv.Area+ Bsmt.Unf.SF+ Total.Bsmt.SF,data=df_hou_sub,nbest=4)
plot(leap,scale="adjr2")

summary(leap)
```