**ALY6015 CRN 80403**

**Intermediate Analytics**

**NORTHEASTERN UNIVERSITY**

**SANA YASIN Date :19 April 2023**

**MODULE 1 -Regression Diagnostic with R**

**Professor: Richard He**

**INTRODUCTION: -**

In this project we are going to fit, interpret and evaluate linear regression model on “AmesHousing” dataset. We are also going to implement diagnostic techniques to identify and correct issues with the model. We are also going to test that our model is following the assumption of linear regression.

**LINEAR REGRESION MODEL: -**

Linear regression is a predictive analysis method used to model the relationship between a dependent variable and one or more independent variables. The basic idea behind linear regression is to find the best-fit straight line that determines the relationship between the variables.

A linear regression model can be written as

Y=βo+β1X1+β2X2+β3X3+…….βNXN +εwhere

Y=target variable

X1, X2,X3….XN:-Predictor variables

β0 is intercept, β1,β2,β3….βN are coefficients and ε<- it is the error

Chart, scatter chart

Description automatically generated

**ORDINARY LEAST SQUARE METHOD:-**

The Ordinary Least Squares (OLS) method is a commonly used technique in linear regression analysis. It is used to estimate the parameters (coefficients) like β0,β1 of a linear regression model that minimizes the sum of the squared differences between the predicted value and actual values of the target variable.

Residuals ε=Y(Observed)-Y(Predicted)

**ASSUMPTIONS OF LINEAR REGRESSION: -**

**1.Linearity of the data**: - The linear relationship between the predictor and the target (y) .We have assumed that there is a linear relationship between our target variables and predictor variables . This means that change in independent variable should be proportional to dependent variable.

**2.Normality of residuals:-** The residuals should be distributed normally. This means that distribution should follow the bell curve.

**3.Homogeneity of residuals variance**:- The residuals should have a constant variance.

**4.No Multicollinearity**: - The predictor variables should be independent of each other .

**LOADING PACKAGES AND DATASETS: -**

## Rows: 5  
## Columns: 81  
## $ `MS SubClass` <chr> "020", "020", "020", "020", "060"  
## $ `MS Zoning` <chr> "RL", "RH", "RL", "RL", "RL"  
## $ `Lot Frontage` <dbl> 141, 80, 81, 93, 74  
## $ `Lot Area` <dbl> 31770, 11622, 14267, 11160, 13830  
## $ Street <chr> "Pave", "Pave", "Pave", "Pave", "Pave"  
## $ Alley <chr> NA, NA, NA, NA, NA  
## $ `Lot Shape` <chr> "IR1", "Reg", "IR1", "Reg", "IR1"  
## $ `Land Contour` <chr> "Lvl", "Lvl", "Lvl", "Lvl", "Lvl"  
## $ Utilities <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub"  
## $ `Lot Config` <chr> "Corner", "Inside", "Corner", "Corner", "Inside"  
## $ `Land Slope` <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl"  
## $ Neighborhood <chr> "NAmes", "NAmes", "NAmes", "NAmes", "Gilbert"  
## $ `Condition 1` <chr> "Norm", "Feedr", "Norm", "Norm", "Norm"  
## $ `Condition 2` <chr> "Norm", "Norm", "Norm", "Norm", "Norm"  
## $ `Bldg Type` <chr> "1Fam", "1Fam", "1Fam", "1Fam", "1Fam"  
## $ `House Style` <chr> "1Story", "1Story", "1Story", "1Story", "2Story"  
## $ `Overall Qual` <dbl> 6, 5, 6, 7, 5  
## $ `Overall Cond` <dbl> 5, 6, 6, 5, 5  
## $ `Year Built` <dbl> 1960, 1961, 1958, 1968, 1997  
## $ `Year Remod/Add` <dbl> 1960, 1961, 1958, 1968, 1998  
## $ `Roof Style` <chr> "Hip", "Gable", "Hip", "Hip", "Gable"  
## $ `Roof Matl` <chr> "CompShg", "CompShg", "CompShg", "CompShg", "CompShg"  
## $ `Exterior 1st` <chr> "BrkFace", "VinylSd", "Wd Sdng", "BrkFace", "VinylSd"  
## $ `Exterior 2nd` <chr> "Plywood", "VinylSd", "Wd Sdng", "BrkFace", "VinylSd"  
## $ `Mas Vnr Type` <chr> "Stone", "None", "BrkFace", "None", "None"  
## $ `Mas Vnr Area` <dbl> 112, 0, 108, 0, 0  
## $ `Exter Qual` <chr> "TA", "TA", "TA", "Gd", "TA"  
## $ `Exter Cond` <chr> "TA", "TA", "TA", "TA", "TA"  
## $ Foundation <chr> "CBlock", "CBlock", "CBlock", "CBlock", "PConc"  
## $ `Bsmt Qual` <chr> "TA", "TA", "TA", "TA", "Gd"  
## $ `Bsmt Cond` <chr> "Gd", "TA", "TA", "TA", "TA"  
## $ `Bsmt Exposure` <chr> "Gd", "No", "No", "No", "No"  
## $ `BsmtFin Type 1` <chr> "BLQ", "Rec", "ALQ", "ALQ", "GLQ"  
## $ `BsmtFin SF 1` <dbl> 639, 468, 923, 1065, 791  
## $ `BsmtFin Type 2` <chr> "Unf", "LwQ", "Unf", "Unf", "Unf"  
## $ `BsmtFin SF 2` <dbl> 0, 144, 0, 0, 0  
## $ `Bsmt Unf SF` <dbl> 441, 270, 406, 1045, 137  
## $ `Total Bsmt SF` <dbl> 1080, 882, 1329, 2110, 928  
## $ Heating <chr> "GasA", "GasA", "GasA", "GasA", "GasA"  
## $ `Heating QC` <chr> "Fa", "TA", "TA", "Ex", "Gd"  
## $ `Central Air` <chr> "Y", "Y", "Y", "Y", "Y"  
## $ Electrical <chr> "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr"  
## $ `1st Flr SF` <dbl> 1656, 896, 1329, 2110, 928  
## $ `2nd Flr SF` <dbl> 0, 0, 0, 0, 701  
## $ `Low Qual Fin SF` <dbl> 0, 0, 0, 0, 0  
## $ `Gr Liv Area` <dbl> 1656, 896, 1329, 2110, 1629  
## $ `Bsmt Full Bath` <dbl> 1, 0, 0, 1, 0  
## $ `Bsmt Half Bath` <dbl> 0, 0, 0, 0, 0  
## $ `Full Bath` <dbl> 1, 1, 1, 2, 2  
## $ `Half Bath` <dbl> 0, 0, 1, 1, 1  
## $ `Bedroom AbvGr` <dbl> 3, 2, 3, 3, 3  
## $ `Kitchen AbvGr` <dbl> 1, 1, 1, 1, 1  
## $ `Kitchen Qual` <chr> "TA", "TA", "Gd", "Ex", "TA"  
## $ `TotRms AbvGrd` <dbl> 7, 5, 6, 8, 6  
## $ Functional <chr> "Typ", "Typ", "Typ", "Typ", "Typ"  
## $ Fireplaces <dbl> 2, 0, 0, 2, 1  
## $ `Fireplace Qu` <chr> "Gd", NA, NA, "TA", "TA"  
## $ `Garage Type` <chr> "Attchd", "Attchd", "Attchd", "Attchd", "Attchd"  
## $ `Garage Yr Blt` <dbl> 1960, 1961, 1958, 1968, 1997  
## $ `Garage Finish` <chr> "Fin", "Unf", "Unf", "Fin", "Fin"  
## $ `Garage Cars` <dbl> 2, 1, 1, 2, 2  
## $ `Garage Area` <dbl> 528, 730, 312, 522, 482  
## $ `Garage Qual` <chr> "TA", "TA", "TA", "TA", "TA"  
## $ `Garage Cond` <chr> "TA", "TA", "TA", "TA", "TA"  
## $ `Paved Drive` <chr> "P", "Y", "Y", "Y", "Y"  
## $ `Wood Deck SF` <dbl> 210, 140, 393, 0, 212  
## $ `Open Porch SF` <dbl> 62, 0, 36, 0, 34  
## $ `Enclosed Porch` <dbl> 0, 0, 0, 0, 0  
## $ `3Ssn Porch` <dbl> 0, 0, 0, 0, 0  
## $ `Screen Porch` <dbl> 0, 120, 0, 0, 0  
## $ `Pool Area` <dbl> 0, 0, 0, 0, 0  
## $ `Pool QC` <chr> NA, NA, NA, NA, NA  
## $ Fence <chr> NA, "MnPrv", NA, NA, "MnPrv"  
## $ `Misc Feature` <chr> NA, NA, "Gar2", NA, NA  
## $ `Misc Val` <dbl> 0, 0, 12500, 0, 0  
## $ `Mo Sold` <dbl> 5, 6, 6, 4, 3  
## $ `Yr Sold` <dbl> 2010, 2010, 2010, 2010, 2010  
## $ `Sale Type` <chr> "WD", "WD", "WD", "WD", "WD"  
## $ `Sale Condition` <chr> "Normal", "Normal", "Normal", "Normal", "Normal"  
## $ SalePrice <dbl> 215000, 105000, 172000, 244000, 189900  
## $ Property\_age <dbl> 63, 62, 65, 55, 26

The dataset contains **2930** rows and **81** columns.

**SUMMARY TABLES FOR CATEGORICAL VARIABLES**

In the Dataset, we have 80 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers which we removed earlier).

To create summary tables, we have separated the categorical variables and created tables for each of the variables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MS SubClass** | **Street** | **MS Zoning** | **Lot Frontage** | **Alley** | **Lot Shape** | **Land Contour** |
| 020 :1079 | Grvl: 12 | A (agr): 2 | Grvl: 12 | Grvl: 120 | IR1: 979 | Bnk: 117 |
| 060 : 575 | Pave:2918 | C (all): 25 | Pave:2918 | N/A :2732 | IR2: 76 | HLS: 120 |
| 050 : 287 |  | FV : 139 |  | Pave: 78 | IR3: 16 | Low: 60 |
| 120 : 192 |  | I (all): 2 |  |  | Reg:1859 | Lvl:2633 |
| 030 : 139 |  | RH : 27 |  |  |  |  |
| 160 : 129 |  | RL :2273 |  |  |  |  |
| (Other): 529 |  | RM : 462 |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Utilities** | **Lot Config** | **Land Slope** | **Condition** 1 | **Condition 2** | **House** **Style** | **Roof Style** |
| AllPub:2927 | Corner : 511 | Gtl:2789 | Norm :2522 | Norm :2900 | 1Story :1481 | Flat : 20 |
| NoSeWa: 1 | CulDSac: 180 | Mod: 125 | Feedr : 164 | Feedr : 13 | 2Story : 873 | Gable :2321 |
| NoSewr: 2 | FR2 : 85 | Sev: 16 | Artery : 92 | Artery : 5 | 1.5Fin : 314 | Gambrel: 22 |
|  | FR3 : 14 |  | RRAn : 50 | PosA : 4 | SLvl : 128 | Hip : 551 |
|  | Inside :2140 |  | PosN : 39 | PosN : 4 | SFoyer : 83 | Mansard: 11 |
|  |  |  | RRAe : 28 | RRNn : 2 | 2.5Unf : 24 | Shed : 5 |
|  |  |  |  | (Other): 35 | (Other): 2 | (Other): 27 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Roof Matl** | **Exterior 1st** | **Exterior 2nd** | **Mas Vnr Type** | **Exter Qual** | **Exter Cond** | **Foundation** |
| CompShg:2887 | VinylSd:1026 | VinylSd:1015 | BrkCmn : 25 | Ex: 107 | Ex: 12 | BrkTil: 311 |
| Tar&Grv: 23 | MetalSd: 450 | MetalSd: 447 | BrkFace: 880 | Fa: 35 | Fa: 67 | CBlock:1244 |
| WdShake: 9 | HdBoard: 442 | HdBoard: 406 | CBlock : 1 | Gd: 989 | Gd: 299 | PConc :1310 |
| WdShngl: 7 | Wd Sdng: 420 | Wd Sdng: 397 | N/A : 23 | TA:1799 | Po: 3 | Slab : 49 |
| ClyTile: 1 | Plywood: 221 | Plywood: 274 | None :1752 |  | TA:2549 | Stone : 11 |
| Membran: 1 | CemntBd: 126 | CmentBd: 126 | Stone : 249 |  |  | Wood : 5 |
| (Other): 2 | (Other): 245 | (Other): 265 |  |  |  |  |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Bsmt Qual** | **Bsmt Exposure** | **BsmtFin Type 1** | **BsmtFin Type 2** | **Overall**  **Qual** | **Overall Cond** | **Bsmt Full Bath** |  |
| Ex : 258 | Av : 418 | ALQ:429 | ALQ: 53 | 5 :825 | 5 :1654 | 0 :1707 |  |
| Fa : 88 | Gd : 284 | BLQ:269 | BLQ: 68 | 6 :732 | 6 : 533 | 1 :1181 |  |
| Gd :1219 | Mn : 239 | GLQ:859 | GLQ: 34 | 7 :602 | 7 : 390 | 2 : 38 |  |
| N/A: 80 | N/A: 83 | LwQ:154 | LwQ: 89 | 8 :350 | 8 : 144 | 3 : 2 |  |
| Po : 2 | No :1906 | N/A: 80 | N/A: 81 | 4 :226 | 4 : 101 | NA’s: 2 |  |
| TA :1283 |  | Rec:288 | Rec: 106 | 9 :107 | 3 : 50 |  |  |
|  |  | Unf:851 | Unf:2499 | (Other): 88 | (Other): 58 |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Bsmt Half Bath** | **Bedroom AbvGr** | **Kitchen AbvGr** | **Heating** | **Heating QC** | **Central Air** | **Electrical** |
| 0 :2753 | 3 :1597 | 00:03 | Floor: 1 | Ex:1495 | N: 196 | FuseA: 188 |
| 1 : 171 | 2 : 743 | 1:2796 | GasA :2885 | Fa: 92 | Y:2734 | FuseF: 50 |
| 2 : 4 | 4 : 400 | 2: 129 | GasW : 27 | Gd: 476 |  | FuseP: 8 |
| NA’s: 2 | 1 : 112 | 03:02 | Grav : 9 | Po: 3 |  | Mix : 1 |
|  | 05:48 |  | OthW : 2 | TA: 864 |  | SBrkr:2682 |
|  | 06:21 |  | Wall : 6 |  |  | NA’s : 1 |
|  | (Other): 9 |  |  |  |  |  |

**SUMMARY TABLES FOR NUMERICAL VARIABLES**

To create summary tables, we have separated the numerical variables and created tables for each of the variables. We have replaced the Null values with zeros.

Summary of `Lot Area`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1300 7440 9436 10148 11555 215245

Summary of ` Mas Vnr Area`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 0.0 101.1 162.8 1600.0

Summary of ` BsmtFin SF 1`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 370.0 442.5 734.0 5644.0

Summary of ` BsmtFin SF 2`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 49.71 0.00 1526.00

Summary of ` Bsmt Unf SF`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 219.0 465.5 559.1 801.8 2336.0

Summary of ` Total Bsmt SF`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 793 990 1051 1302 6110

Summary of ` 1st Flr SF`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334.0 876.2 1084.0 1159.6 1384.0 5095.0

Summary of ` 2nd Flr SF`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 0.0 335.5 703.8 2065.0

Summary of ` Low Qual Fin SF`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 4.677 0.000 1064.000

Summary of `Gr Liv Area`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 1126 1442 1500 1743 5642

Summary of ` Garage Area`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 320.0 480.0 472.7 576.0 1488.0

Summary of ` Wood Deck SF`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 93.75 168.00 1424.00

Summary of ` Open Porch SF`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 27.00 47.53 70.00 742.00

Summary of ` Enclosed Porch`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 23.01 0.00 1012.00

Summary of ` 3Ssn Porch`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 2.592 0.000 508.000

Summary of ` Screen Porch`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 16 0 576

Summary of Pool Area`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 2.243 0.000 800.000

Summary of Misc Val`

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 50.63 0.00 17000.00

Summary of Property\_age

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 13.00 22.00 50.00 51.64 69.00 151.00

Summary of SalePrice

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12789 129500 160000 180796 213500 755000

**EXPLORATORY DATA ANALYSIS**

The objective of this project is to create linear regression model. For this we need to establish the relation between our target variable “SalesPrice” and predictors variables.Before creating the model we need to examine the predictors graphically . So we will using:-

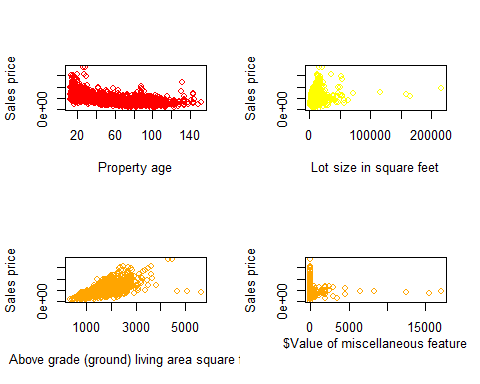
**1.Scatter plot**:- To visualize correlation between target and predictor variables.

**2.Box Plot**:-To examine outliers and variation with respect to categorical variables.

**3.Density Plot** :- To check the normality of Distributions.

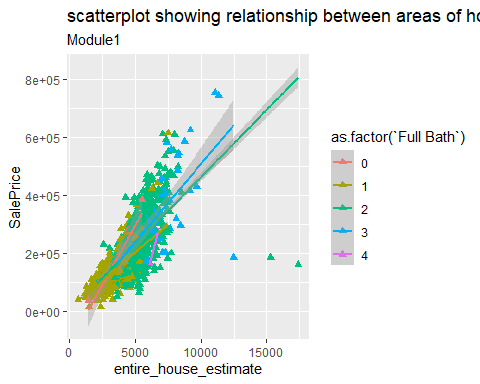
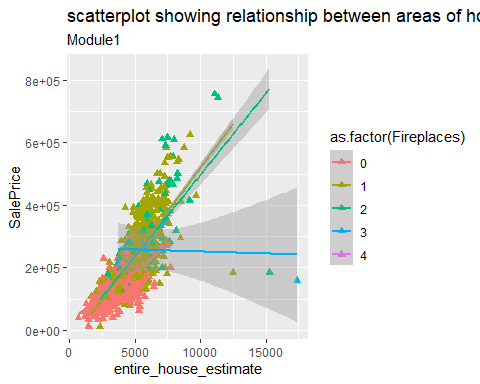
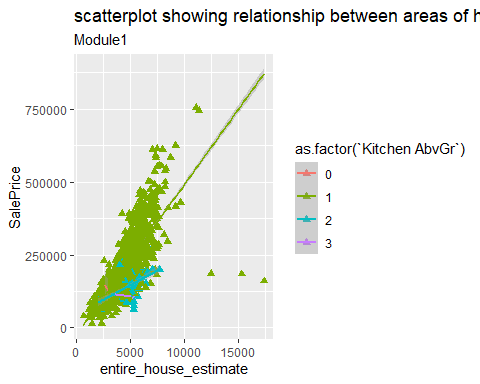
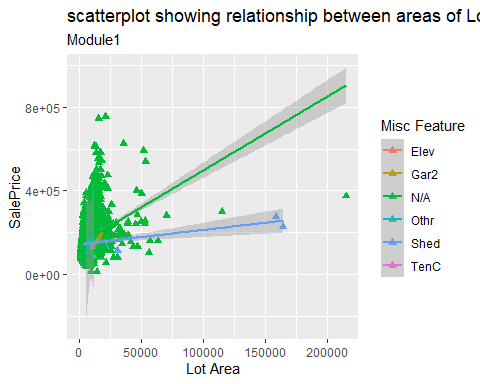
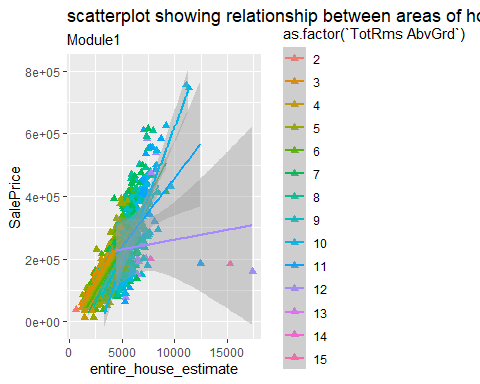
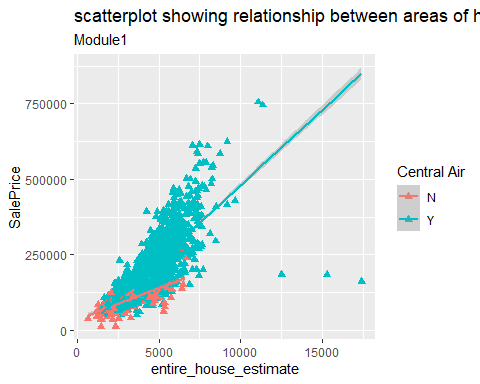
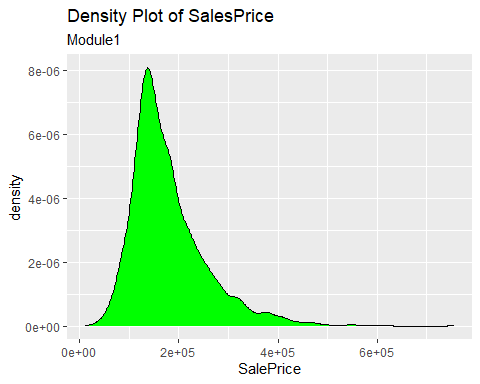
Graphical user interface, diagram, application

Description automatically generated



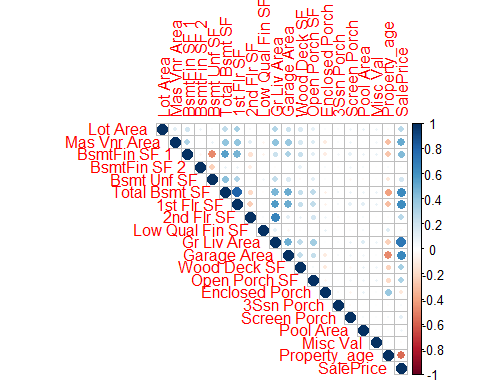
**OBSERVATION**

The variables Total Bsmt SF,1st Flr SF,2nd Flr SF,Wood Deck SF,Gr Liv Area vary directly with target variable SalesPrice. The variable Property\_age vary negatively with target variable.



**CORRELATION MATRIX:-**

Correlation is statistical measure of linear dependence between 2 variables . It takes values between +1 to -1. We need to test and plot the correlation matrix for our predictor variables and target variables.



##   
## Pearson's product-moment correlation  
##   
## data: AmesHousing\_cor$SalePrice and AmesHousing\_cor$`Total Bsmt SF`  
## t = 44.19, df = 2928, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6102961 0.6537660  
## sample estimates:  
## cor   
## 0.6325288

##   
## Pearson's product-moment correlation  
##   
## data: AmesHousing\_cor$SalePrice and AmesHousing\_cor$`Gr Liv Area`  
## t = 54.061, df = 2928, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6881814 0.7244502  
## sample estimates:  
## cor   
## 0.7067799

##   
## Pearson's product-moment correlation  
##   
## data: AmesHousing\_cor$SalePrice and AmesHousing\_cor$Property\_age  
## t = -36.426, df = 2928, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.5828515 -0.5329925  
## sample estimates:  
## cor   
## -0.5584261

##   
## Pearson's product-moment correlation  
##   
## data: AmesHousing\_cor$SalePrice and AmesHousing\_cor$`1st Flr SF`  
## t = 42.947, df = 2928, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5989480 0.6434034  
## sample estimates:  
## cor   
## 0.6216761

##   
## Pearson's product-moment correlation  
##   
## data: AmesHousing\_cor$SalePrice and AmesHousing\_cor$`2nd Flr SF`  
## t = 15.136, df = 2928, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2354586 0.3026329  
## sample estimates:  
## cor   
## 0.2693734

##   
## Pearson's product-moment correlation  
##   
## data: AmesHousing\_cor$SalePrice and AmesHousing\_cor$`Open Porch SF`  
## t = 17.83, df = 2928, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2799111 0.3452495  
## sample estimates:  
## cor   
## 0.3129505

##   
## Pearson's product-moment correlation  
##   
## data: AmesHousing\_cor$SalePrice and AmesHousing\_cor$`Wood Deck SF`  
## t = 18.733, df = 2928, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2944194 0.3591007  
## sample estimates:  
## cor   
## 0.3271432

**LINEAR REGRESSION MODEL**

##   
## Call:  
## lm(formula = SalePrice ~ `Total Bsmt SF` + `Gr Liv Area` + Property\_age +   
## `1st Flr SF` + `2nd Flr SF` + `Open Porch SF` + `Wood Deck SF`,   
## data = AmesHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -624356 -20484 -2851 16797 272113   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 48033.723 3665.006 13.106 < 2e-16 \*\*\*  
## `Total Bsmt SF` 42.088 3.101 13.572 < 2e-16 \*\*\*  
## `Gr Liv Area` 19.469 17.118 1.137 0.255480   
## Property\_age -813.519 29.306 -27.760 < 2e-16 \*\*\*  
## `1st Flr SF` 66.637 17.480 3.812 0.000141 \*\*\*  
## `2nd Flr SF` 56.589 17.283 3.274 0.001072 \*\*   
## `Open Porch SF` 19.116 12.530 1.526 0.127203   
## `Wood Deck SF` 44.497 6.553 6.791 1.35e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 42390 on 2922 degrees of freedom  
## Multiple R-squared: 0.7191, Adjusted R-squared: 0.7184   
## F-statistic: 1068 on 7 and 2922 DF, p-value: < 2.2e-16

## [1] 70761.78

## [1] 70815.63

**MODEL INTERPRETATION**

1.The R-squared values is 0.7184.

2.The AIC value is 7.0761785^{4}.<BR>

3.The BIC value is 7.081563\times 10^{4}.

**DIAGNOSTIC PLOTS**

we are going to plot 4 diagnostic plots to check the below assumptions:-

1. Residual vs Fitted :- This plot shows the whether our model follows the assumpltion that there is linear relationship between predictor and target variable.
2. Normal Q-Q plot :- This plot shows us whether the residuals have normal distribution or not.
3. Scale-location plot :-It is used to check the variance of residuals i.e homoscedasity.
4. Residual vs leverage plot:- Used to identify if there is any influential value or outlier present in target varaiable which is affecting our model .

par(mfrow=c(2,2))  
plot(model\_ames)

Graphical user interface, chart

Description automatically generated

**OBSERVATION**

1. From the residual vs fitted plot we can see that the line id not constant and that the relationship between target and predictors is not linear.
2. From Q-Q plot we can see that it does not follow the straight line .
3. From Scale -location plot we can say that variance of residuals is evenly speard. We have detected heteroscedasity problem.

**CHECKING MULTICOLINEARITY: -**

As our model is not performing well and not following the above assumptions we need to remove the problem of multicollinearity and ouliers to fit our model.

**VARIANCE INFLATION FACTOR: -**

It is a statistical measure used to detect multicollinearity in regression model . We need to remove the variables which having VIF score greater than 5 .

**VIF=1/(1-R^2)**

vif(model\_ames)

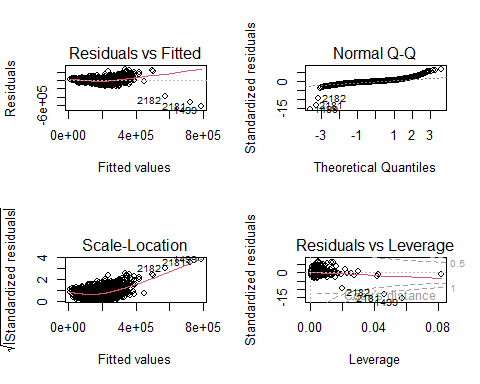
## `Total Bsmt SF` `Gr Liv Area` Property\_age `1st Flr SF` `2nd Flr SF`   
## 3.047621 122.033664 1.280422 76.477880 89.346416   
## `Open Porch SF` `Wood Deck SF`   
## 1.165174 1.117406

From the above we can say that , `Gr Liv Area` having very high VIF score so we will remove it from our model and rerun the model.

model\_ames\_1<-lm(SalePrice~`Total Bsmt SF`+`Property\_age`+`1st Flr SF`+`2nd Flr SF`+`Open Porch SF`+`Wood Deck SF`,data = AmesHousing)  
summary(model\_ames\_1)

##   
## Call:  
## lm(formula = SalePrice ~ `Total Bsmt SF` + Property\_age + `1st Flr SF` +   
## `2nd Flr SF` + `Open Porch SF` + `Wood Deck SF`, data = AmesHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -624822 -20548 -2828 16723 272105   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 47721.903 3654.920 13.057 < 2e-16 \*\*\*  
## `Total Bsmt SF` 42.159 3.101 13.597 < 2e-16 \*\*\*  
## Property\_age -808.510 28.975 -27.904 < 2e-16 \*\*\*  
## `1st Flr SF` 86.135 3.414 25.233 < 2e-16 \*\*\*  
## `2nd Flr SF` 76.116 1.988 38.284 < 2e-16 \*\*\*  
## `Open Porch SF` 19.323 12.529 1.542 0.123   
## `Wood Deck SF` 44.560 6.553 6.800 1.26e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 42400 on 2923 degrees of freedom  
## Multiple R-squared: 0.7189, Adjusted R-squared: 0.7184   
## F-statistic: 1246 on 6 and 2923 DF, p-value: < 2.2e-16

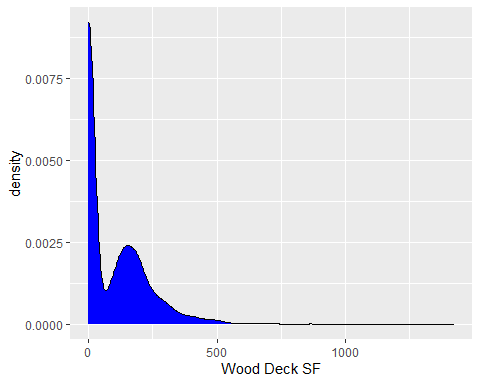
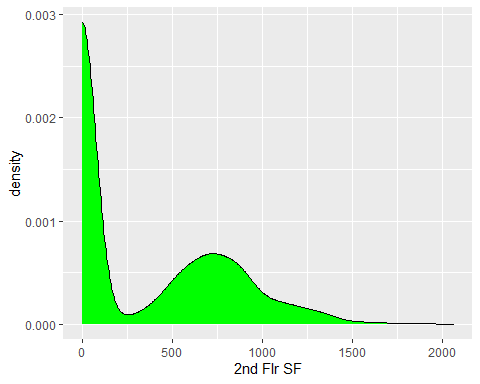
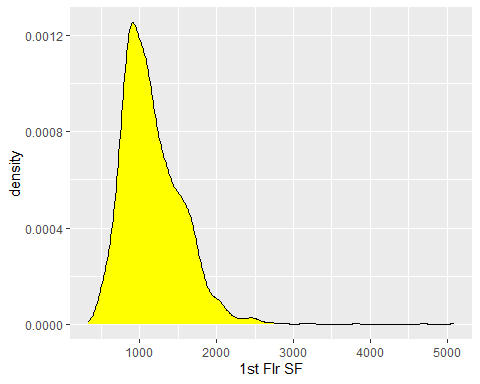
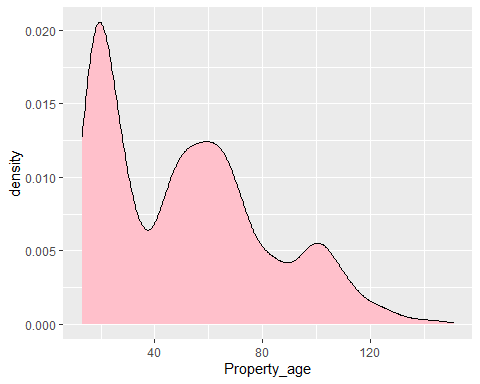
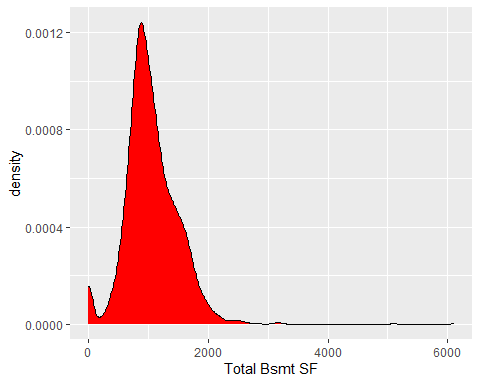
par(mfrow=c(2,2))  
plot(model\_ames\_1)



vif(model\_ames\_1)

## `Total Bsmt SF` Property\_age `1st Flr SF` `2nd Flr SF` `Open Porch SF`   
## 3.046368 1.251500 2.916340 1.182219 1.164927   
## `Wood Deck SF`   
## 1.117326

**CHECKING NORMALITY OF TARGET AS WELL AS PREDICTORS VARIABLES**



Diagram

Description automatically generated

**OBSERVATION AND CORRECTIONS**

We can see that there is no linear relationship between the target and predictor variables because there are outliers present and we do not have perfect normally distributed variables.

To overcome this problem, we will use log (y) in our model.

model\_ames\_2<-lm(log(SalePrice)~`Total Bsmt SF`+(`Property\_age`)+`1st Flr SF`+(`2nd Flr SF`)+(`Wood Deck SF`),data = AmesHousing)  
summary(model\_ames\_2)

##   
## Call:  
## lm(formula = log(SalePrice) ~ `Total Bsmt SF` + (Property\_age) +   
## `1st Flr SF` + (`2nd Flr SF`) + (`Wood Deck SF`), data = AmesHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.97727 -0.08320 0.01149 0.10422 0.86386   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.145e+01 1.777e-02 644.041 < 2e-16 \*\*\*  
## `Total Bsmt SF` 2.088e-04 1.504e-05 13.885 < 2e-16 \*\*\*  
## Property\_age -5.116e-03 1.402e-04 -36.477 < 2e-16 \*\*\*  
## `1st Flr SF` 4.051e-04 1.648e-05 24.580 < 2e-16 \*\*\*  
## `2nd Flr SF` 3.860e-04 9.336e-06 41.344 < 2e-16 \*\*\*  
## `Wood Deck SF` 2.250e-04 3.177e-05 7.081 1.78e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2062 on 2924 degrees of freedom  
## Multiple R-squared: 0.7444, Adjusted R-squared: 0.744   
## F-statistic: 1703 on 5 and 2924 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(model\_ames\_2)

Graphical user interface, diagram

Description automatically generated

vif(model\_ames\_2)

## `Total Bsmt SF` Property\_age `1st Flr SF` `2nd Flr SF` `Wood Deck SF`   
## 3.028849 1.239243 2.873467 1.101678 1.109966

AIC(model\_ames\_2)

## [1] -928.6907

BIC(model\_ames\_2)

## [1] -886.8114

**INTERPRETATIONS**: -

After using Log(y) in our model we can see that our R-squared value has improved and the linearity problem is solved.

Furthermore, we can use the pre-build feature selection model to get our best fit model.

## Start: AIC=62444.49  
## SalePrice ~ `Total Bsmt SF` + Property\_age + `1st Flr SF` + `2nd Flr SF` +   
## `Wood Deck SF`  
##   
## Df Sum of Sq RSS AIC  
## <none> 5.2580e+12 62444  
## - `Wood Deck SF` 1 8.0611e+10 5.3386e+12 62487  
## - `Total Bsmt SF` 1 3.4000e+11 5.5980e+12 62626  
## - `1st Flr SF` 1 1.1787e+12 6.4367e+12 63035  
## - Property\_age 1 1.4289e+12 6.6869e+12 63147  
## - `2nd Flr SF` 1 2.8867e+12 8.1447e+12 63725

##   
## Call:  
## lm(formula = SalePrice ~ `Total Bsmt SF` + Property\_age + `1st Flr SF` +   
## `2nd Flr SF` + `Wood Deck SF`, data = AmesHousing)  
##   
## Coefficients:  
## (Intercept) `Total Bsmt SF` Property\_age `1st Flr SF`   
## 47555.78 42.52 -812.93 86.77   
## `2nd Flr SF` `Wood Deck SF`   
## 76.92 43.74

**CONCLUSION :-**

1.We learnt about the assumptions of linear regression and OLS method .

2.We have performed linear regression of dataset and used diagnostic plot to validate ourmodel.

3.We learnt to remove outlier and multicollinearity in our data to improve the performance of our model

4.We learnt to build model based on featureselection.

**REFERENCES**: -

(Agrawal, Published On May 19, 2021 and Last Modified On July 21st, 2022; "Linear Regression Assumptions and Diagnostics in R: Essentials," ; R-bloggers, October 3, 2021; Science, May 23, 2020

)

Agrawal, R. (Published On May 19, 2021 and Last Modified On July 21st, 2022). Know The Best Evaluation Metrics for Your Regression Model ! <https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/>

Linear Regression Assumptions and Diagnostics in R: Essentials. *STHDA*. <http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/>

R-bloggers. (October 3, 2021). Multiple linear regression made simple. <https://www.r-bloggers.com/2021/10/multiple-linear-regression-made-simple/>

Science, T. D. (May 23, 2020

). 3 Best metrics to evaluate Regression Model? <https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b>

**APPENDIX:-**

---

title: "Sana\_module1\_19042023"

output:

html\_document: default

word\_document: default

date: "2023-04-19"

---

<CENTER>

<B>

<FONT size=4, color="#66B2FF">

ALY6015 CRN 80403<BR>

Intermediate Analytics<BR>

NORTHEASTERN UNIVERSITY<BR>

</B>

SANA YASIN<BR>

Date :`r format(Sys.time(), "%d %B, %Y")`<BR>

<B>MODULE 1 -Regression Diagnostic with R</B><BR>

Professor: Richard He<BR><P>

</FONT>

<B>

</CENTER>

<FONT size=4, color="#66B2FF">

<B>INTRODUCTION :-</B><BR>

</FONT>

<FONT size=2, color="#000000">

In this project we are going fit, interpret and evaluate linear regression model on AmesHousing dataset. We are also going to implement diagnostic techniques to identify and correct issues with the model.<BR><P>

<FONT size=4, color="#66B2FF">

<B>LOADING PACKAGES AND DATASETS:-</B><BR>

</FONT>

```{r libraries and datasets , echo=FALSE, message=FALSE, warning=FALSE}

library(tidyverse)

library(readxl)

library(magrittr)

library(dplyr)

library(RColorBrewer)

library(kableExtra)

library(ggplot2)

library(forcats)

library(car)

#dataset

AmesHousing <- read\_csv("C:/Users/LENOVO/Documents/ALY6015/AmesHousing.csv")

rownames(AmesHousing)<- AmesHousing$PID

AmesHousing<-AmesHousing[,3:82]

current\_year=as.integer(format(Sys.Date(),"%Y"))

AmesHousing$Property\_age<-current\_year-AmesHousing$`Year Built`

dplyr::glimpse(AmesHousing[1:5,])

```

<FONT size=2, color="#000000">

The dataset contains `r nrow(AmesHousing)` rows and `r ncol(AmesHousing)` columns. <BR><P>

</FONT>

<FONT size=4, color="#66B2FF">

<B>SUMMARY TABLES FOR CATEGORICAL VARIABLES</B><BR></FONT>

<FONT size=2, color="#000000">

In the Dataset, we have 80 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers which we removed earlier).<P>

To create summary tables we have seperated the categorical variables and created tables for each of the variables .<P>

We have replaced the Null values with "N/A".<P></FONT>

```{r categorical summary, echo=FALSE, message=FALSE, warning=FALSE}

AmesHousing$`Alley`<-replace(AmesHousing$`Alley`,is.na(AmesHousing$`Alley`),"N/A")

AmesHousing$`Mas Vnr Type`<-replace(AmesHousing$`Mas Vnr Type`,is.na(AmesHousing$`Mas Vnr Type`),"N/A")

AmesHousing$`Bsmt Qual`<-replace(AmesHousing$`Bsmt Qual`,is.na(AmesHousing$`Bsmt Qual`),"N/A")

AmesHousing$`Bsmt Cond`<-replace(AmesHousing$`Bsmt Cond`,is.na(AmesHousing$`Bsmt Cond`),"N/A")

AmesHousing$`Bsmt Exposure`<-replace(AmesHousing$`Bsmt Exposure`,is.na(AmesHousing$`Bsmt Exposure`),"N/A")

AmesHousing$`Bsmt Exposure`<-replace(AmesHousing$`Bsmt Exposure`,is.na(AmesHousing$`Bsmt Exposure`),"N/A")

AmesHousing$`BsmtFin Type 1`<-replace(AmesHousing$`BsmtFin Type 1`,is.na(AmesHousing$`BsmtFin Type 1`),"N/A")

AmesHousing$`BsmtFin Type 2`<-replace(AmesHousing$`BsmtFin Type 2`,is.na(AmesHousing$`BsmtFin Type 2`),"N/A")

AmesHousing$`Garage Type`<-replace(AmesHousing$`Garage Type`,is.na(AmesHousing$`Garage Type`),"N/A")

AmesHousing$`Garage Finish`<-replace(AmesHousing$`Garage Finish`,is.na(AmesHousing$`Garage Finish`),"N/A")

AmesHousing$`Garage Qual`<-replace(AmesHousing$`Garage Qual`,is.na(AmesHousing$`Garage Qual`),"N/A")

AmesHousing$`Garage Cond`<-replace(AmesHousing$`Garage Cond`,is.na(AmesHousing$`Garage Cond`),"N/A")

AmesHousing$`Pool QC`<-replace(AmesHousing$`Pool QC`,is.na(AmesHousing$`Pool QC`),"N/A")

AmesHousing$`Fence`<-replace(AmesHousing$`Fence`,is.na(AmesHousing$`Fence`),"N/A")

AmesHousing$`Misc Feature`<-replace(AmesHousing$`Misc Feature`,is.na(AmesHousing$`Misc Feature`),"N/A")

AmesHousing$`Fireplace Qu`<-replace(AmesHousing$`Fireplace Qu`,is.na(AmesHousing$`Fireplace Qu`),"N/A")

AmesHousing$`Electrical`<-replace(AmesHousing$`Electrical `,is.na(AmesHousing$`Electrical`),"N/A")

AmesHousing<-AmesHousing%>%mutate(`Street`=as.factor(`Street`))

AmesHousing<-AmesHousing%>%mutate(`MS SubClass`=as.factor(`MS SubClass`))

AmesHousing<-AmesHousing%>%mutate(`MS Zoning`=as.factor(`MS Zoning`))

AmesHousing<-AmesHousing%>%mutate(`Lot Frontage`=as.factor(`Lot Frontage`))

AmesHousing<-AmesHousing%>%mutate(`Lot Frontage`=as.factor(`Lot Area`))

AmesHousing<-AmesHousing%>%mutate(`Lot Frontage`=as.factor(`Street`))

AmesHousing<-AmesHousing%>%mutate(`Alley`=as.factor(`Alley`))

AmesHousing<-AmesHousing%>%mutate(`Lot Shape`=as.factor(`Lot Shape`))

AmesHousing<-AmesHousing%>%mutate(`Land Contour`=as.factor(`Land Contour`))

AmesHousing<-AmesHousing%>%mutate(`Utilities`=as.factor(`Utilities`))

AmesHousing<-AmesHousing%>%mutate(`Lot Config`=as.factor(`Lot Config`))

AmesHousing<-AmesHousing%>%mutate(`Land Slope`=as.factor(`Land Slope`))

AmesHousing<-AmesHousing%>%mutate(`Condition 1`=as.factor(`Condition 1`))

AmesHousing<-AmesHousing%>%mutate(`Condition 2`=as.factor(`Condition 2`))

AmesHousing<-AmesHousing%>%mutate(`Bldg Type`=as.factor(`Bldg Type`))

AmesHousing<-AmesHousing%>%mutate(`House Style`=as.factor(`House Style`))

AmesHousing<-AmesHousing%>%mutate(`Roof Style`=as.factor(`Roof Style`))

AmesHousing<-AmesHousing%>%mutate(`Roof Matl`=as.factor(`Roof Matl`))

AmesHousing<-AmesHousing%>%mutate(`Exterior 1st`=as.factor(`Exterior 1st`))

AmesHousing<-AmesHousing%>%mutate(`Exterior 2nd`=as.factor(`Exterior 2nd`))

AmesHousing<-AmesHousing%>%mutate(`Mas Vnr Type`=as.factor(`Mas Vnr Type`))

AmesHousing<-AmesHousing%>%mutate(`Exter Qual`=as.factor(`Exter Qual`))

AmesHousing<-AmesHousing%>%mutate(`Exter Cond`=as.factor(`Exter Cond`))

AmesHousing<-AmesHousing%>%mutate(`Foundation`=as.factor(`Foundation`))

AmesHousing<-AmesHousing%>%mutate(`Bsmt Qual`=as.factor(`Bsmt Qual`))

AmesHousing<-AmesHousing%>%mutate(`Bsmt Cond`=as.factor(`Bsmt Cond`))

AmesHousing<-AmesHousing%>%mutate(`Bsmt Exposure`=as.factor(`Bsmt Exposure`))

AmesHousing<-AmesHousing%>%mutate(`BsmtFin Type 1`=as.factor(`BsmtFin Type 1`))

AmesHousing<-AmesHousing%>%mutate(`BsmtFin Type 2`=as.factor(`BsmtFin Type 2`))

AmesHousing<-AmesHousing%>%mutate(`Heating`=as.factor(`Heating`))

AmesHousing<-AmesHousing%>%mutate(`Heating QC`=as.factor(`Heating QC`))

AmesHousing<-AmesHousing%>%mutate(`Central Air`=as.factor(`Central Air`))

AmesHousing<-AmesHousing%>%mutate(`Electrical`=as.factor(`Electrical`))

AmesHousing<-AmesHousing%>%mutate(`Kitchen Qual`=as.factor(`Kitchen Qual`))

AmesHousing<-AmesHousing%>%mutate(`Functional`=as.factor(`Functional`))

AmesHousing<-AmesHousing%>%mutate(`Fireplace Qu`=as.factor(`Fireplace Qu`))

AmesHousing<-AmesHousing%>%mutate(`Garage Type`=as.factor(`Garage Type`))

AmesHousing<-AmesHousing%>%mutate(`Garage Finish`=as.factor(`Garage Finish`))

AmesHousing<-AmesHousing%>%mutate(`Garage Qual`=as.factor(`Garage Qual`))

AmesHousing<-AmesHousing%>%mutate(`Garage Cond`=as.factor(`Garage Cond`))

AmesHousing<-AmesHousing%>%mutate(`Paved Drive`=as.factor(`Paved Drive`))

AmesHousing<-AmesHousing%>%mutate(`Pool QC`=as.factor(`Pool QC`))

AmesHousing<-AmesHousing%>%mutate(`Fence`=as.factor(`Fence`))

AmesHousing<-AmesHousing%>%mutate(`Misc Feature`=as.factor(`Misc Feature`))

AmesHousing<-AmesHousing%>%mutate(`Sale Type`=as.factor(`Sale Type`))

AmesHousing<-AmesHousing%>%mutate(`Sale Condition`=as.factor(`Sale Condition`))

AmesHousing\_cat<-AmesHousing%>%select(`Street`,`MS SubClass`,

`MS Zoning`,

`Lot Frontage`,

`Lot Frontage`,

`Lot Frontage`,

`Alley`,

`Lot Shape`,

`Land Contour`,

`Utilities`,

`Lot Config`,

`Land Slope`,

`Condition 1`,

`Condition 2`,

`Bldg Type`,

`House Style`,

`Roof Style`,

`Roof Matl`,

`Exterior 1st`,

`Exterior 2nd`,

`Mas Vnr Type`,

`Exter Qual`,

`Exter Cond`,

`Foundation`,

`Bsmt Qual`,

`Bsmt Cond`,

`Bsmt Exposure`,

`BsmtFin Type 1`,

`BsmtFin Type 2`,

`Heating`,

`Heating QC`,

`Central Air`,

`Electrical`,

`Kitchen Qual`,

`Functional`,

`Fireplace Qu`,

`Garage Type`,

`Garage Finish`,

`Garage Qual`,

`Garage Cond`,

`Paved Drive`,

`Pool QC`,

`Fence`,

`Misc Feature`,

`Sale Type`,

`Sale Condition`,

`Overall Qual`,

`Overall Cond`,

`Bsmt Full Bath`,

`Bsmt Half Bath`,

`Full Bath`,

`Half Bath`,

`Bedroom AbvGr`,

`Kitchen AbvGr`,

`TotRms AbvGrd`,

Fireplaces,

`Garage Cars`

)

#dplyr::glimpse(AmesHousing\_cat[1:5,])

AmesHousing\_cat<-AmesHousing\_cat%>%mutate(`Overall Qual`=as.factor(`Overall Qual`))

AmesHousing\_cat<-AmesHousing\_cat%>%mutate(`Overall Cond`=as.factor(`Overall Cond`))

AmesHousing\_cat<-AmesHousing\_cat%>%mutate(`Bsmt Full Bath`=as.factor(`Bsmt Full Bath`))

AmesHousing\_cat<-AmesHousing\_cat%>%mutate(`Bsmt Half Bath`=as.factor(`Bsmt Half Bath`))

AmesHousing\_cat<-AmesHousing\_cat%>%mutate(`Bedroom AbvGr`=as.factor(`Bedroom AbvGr`))

AmesHousing\_cat<-AmesHousing\_cat%>%mutate(`Kitchen AbvGr`=as.factor(`Kitchen AbvGr`))

#AmesHousing\_cat<-AmesHousing\_cat%>%mutate(`TotRms AbvGr`=as.factor(`TotRms AbvGr`))

AmesHousing\_cat<-AmesHousing\_cat%>%mutate(`Fireplaces`=as.factor(`Fireplaces`))

AmesHousing\_cat<-AmesHousing\_cat%>%mutate(`Garage Cars`=as.factor(`Garage Cars`))

AmesHousing\_cat%>%select(`MS SubClass`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Street`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`MS Zoning`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Lot Frontage`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Lot Frontage`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Lot Frontage`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Alley`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Lot Shape`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Land Contour`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Utilities`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Lot Config`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Land Slope`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Condition 1`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Condition 2`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Bldg Type`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`House Style`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Roof Style`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Roof Matl`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Exterior 1st`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Exterior 2nd`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Mas Vnr Type`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Exter Qual`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Exter Cond`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Foundation`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Bsmt Qual`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Bsmt Cond`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Bsmt Exposure`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`BsmtFin Type 1`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Overall Qual`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`BsmtFin Type 2`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Overall Cond`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Bsmt Full Bath`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Bsmt Half Bath`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Bedroom AbvGr`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Kitchen AbvGr`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

#AmesHousing\_cat%>%select(`TotRms AbvGr`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Heating`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Heating QC`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Central Air`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Electrical`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Kitchen Qual`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Functional`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Fireplace Qu`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Garage Type`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Garage Finish`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Garage Qual`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Garage Cond`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Paved Drive`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Pool QC`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Fence`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Misc Feature`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Sale Type`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

AmesHousing\_cat%>%select(`Sale Condition`) %>%summary()%>%kable(align = "c",format = "markdown",table.attr="style='width:30%;'")

```

<FONT size=4, color="#66B2FF">

<B>SUMMARY TABLES FOR NUMERICAL VARIABLES</B><BR></FONT>

<FONT size=2, color="#000000">

To create summary tables we have separated the numerical variables and created tables for each of the variables .<BR>

We have replaced the Null values with zeros.<BR><P>

</FONT>

```{r numerical summary, echo=FALSE, message=FALSE, warning=FALSE}

AmesHousing$`Mas Vnr Area`<-replace(AmesHousing$`Mas Vnr Area`,is.na(AmesHousing$`Mas Vnr Area`),0)

AmesHousing$`BsmtFin SF 1`<-replace(AmesHousing$`BsmtFin SF 1`,is.na(AmesHousing$`BsmtFin SF 1`),0)

AmesHousing$`BsmtFin SF 2`<-replace(AmesHousing$`BsmtFin SF 2`,is.na(AmesHousing$`BsmtFin SF 2`),0)

AmesHousing$`Total Bsmt SF`<-replace(AmesHousing$`Total Bsmt SF`,is.na(AmesHousing$`Total Bsmt SF`),0)

AmesHousing$`Garage Area`<-replace(AmesHousing$`Garage Area`,is.na(AmesHousing$`Garage Area`),0)

AmesHousing$`Bsmt Unf SF`<-replace(AmesHousing$`Bsmt Unf SF`,is.na(AmesHousing$`Bsmt Unf SF`),0)

AmesHousing\_num<-AmesHousing%>%

select(`Lot Frontage`,

`Lot Area`,

`Mas Vnr Area`,

`BsmtFin SF 1`,

`BsmtFin SF 2`,

`Bsmt Unf SF`,

`Total Bsmt SF`,

`1st Flr SF`,

`2nd Flr SF`,

`Low Qual Fin SF`,

`Gr Liv Area`,

`Garage Area`,

`Wood Deck SF`,

`Open Porch SF`,

`Enclosed Porch`,

`3Ssn Porch`,

`Screen Porch`,

`Pool Area` ,

`Misc Val`,

`Property\_age`,

SalePrice)

AmesHousing\_num%>%pull(`Lot Frontage` )%>%summary()

AmesHousing\_num%>%pull(`Lot Area` )%>%summary()

AmesHousing\_num%>%pull(`Mas Vnr Area` )%>%summary()

AmesHousing\_num%>%pull(`BsmtFin SF 1` )%>%summary()

AmesHousing\_num%>%pull(`BsmtFin SF 2` )%>%summary()

AmesHousing\_num%>%pull(`Bsmt Unf SF` )%>%summary()

AmesHousing\_num%>%pull(`Total Bsmt SF` )%>%summary()

AmesHousing\_num%>%pull(`1st Flr SF` )%>%summary()

AmesHousing\_num%>%pull(`2nd Flr SF` )%>%summary()

AmesHousing\_num%>%pull(`Low Qual Fin SF` )%>%summary()

AmesHousing\_num%>%pull(`Gr Liv Area` )%>%summary()

AmesHousing\_num%>%pull(`Garage Area` )%>%summary()

AmesHousing\_num%>%pull(`Wood Deck SF` )%>%summary()

AmesHousing\_num%>%pull(`Open Porch SF` )%>%summary()

AmesHousing\_num%>%pull(`Enclosed Porch` )%>%summary()

AmesHousing\_num%>%pull(`3Ssn Porch` )%>%summary()

AmesHousing\_num%>%pull(`Screen Porch` )%>%summary()

AmesHousing\_num%>%pull(`Pool Area` )%>%summary()

AmesHousing\_num%>%pull(`Misc Val` )%>%summary()

AmesHousing\_num%>%pull(`Property\_age` )%>%summary()

AmesHousing\_num%>%pull(SalePrice )%>%summary()

```

<FONT size=4, color="#66B2FF">

<B>EXPLORATORY DATA ANALYSIS</B><BR>

</FONT>

<FONT size=2, color="#000000">

The objective of this project is to create linear regression model. For this we need to establish the relation between our target variable "SalesPrice" and predictors variables.Before creating the model we need to examine the predictors graphically . So we will using:-<BR><P>

1.Scatter plot:- To visualize correlation between target and predictor variables.<BR><P>

2.Box Plot:-To examine outliers and variation with respect to categorical variables.<BR><P>

3.Density Plot :- To check the normality of Distributions.<BR><P>

```{r scatterplots, echo=FALSE, message=FALSE, warning=FALSE}

par(mfrow=c(3,3))

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Total Bsmt SF`,

col="red",xlab="Total square feet of basement area",ylab="sales Price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Garage Area`,

col="green",xlab="Size of garage in square feet",ylab="Sales Price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`1st Flr SF`,

color="blue",xlab="First Floor square feet",ylab="Sales Price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`2nd Flr SF`,

col="yellow",xlab="Second floor square feet",ylab="Sales price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Wood Deck SF`,

col="pink",xlab="Wood deck area in square feet",ylab="Sales price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Pool Area`,

col="orange",xlab="Pool area in square feet",ylab="Sales price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Enclosed Porch`,

col="gray",xlab="Enclosed porch area in square feet",ylab="Sales price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Screen Porch`,

col="brown",xlab="Screen porch area in square feet",ylab="Sales price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Low Qual Fin SF`,

col="purple",xlab="Low quality finished square feet (all floors)",ylab="Sales price")

par(mfrow=c(2,2))

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Property\_age`,

col="red",xlab="Property age",ylab="Sales price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Lot Area`,

col="yellow",xlab="Lot size in square feet",ylab="Sales price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Gr Liv Area`,

col="orange",xlab="Above grade (ground) living area square feet",ylab="Sales price")

plot(AmesHousing\_num$SalePrice~AmesHousing\_num$`Misc Val`,

col="orange",xlab="$Value of miscellaneous feature

",ylab="Sales price")

```

<FONT size=4, color="#66B2FF">

<B>OBSERVATION</B><BR>

</FONT>

<FONT size=2, color="#000000">

The variables `Total Bsmt SF`,`1st Flr SF`,`2nd Flr SF`,`Wood Deck SF`,`Gr Liv Area` vary directly with target variable `SalesPrice`.<BR>

The variable `Property\_age` vary negatively with target variable.<BR>

</FONT>

```{r geom plots, echo=FALSE, message=FALSE, warning=FALSE}

AmesHousing%>%ggplot(aes(`SalePrice`))+geom\_density(fill="green")+labs(title="Density Plot of SalesPrice",subtitle="Module1")

AmesHousing$entire\_house\_estimate=AmesHousing$`Total Bsmt SF`+AmesHousing$`1st Flr SF`+AmesHousing$`2nd Flr SF`+AmesHousing$`Gr Liv Area`

AmesHousing%>% ggplot(aes(y=`SalePrice`,x=`entire\_house\_estimate`,color=`Central Air`))+geom\_point(size=2,pch=17,cex=2)+geom\_smooth(method=lm)+ labs(title="scatterplot showing relationship between areas of house and Price w.r.t Centrally air conditioned",subtitle="Module1")

AmesHousing%>% ggplot(aes(y=`SalePrice`,x=`entire\_house\_estimate`,color=as.factor(`TotRms AbvGrd`)))+

geom\_point(size=2,pch=17,cex=2)+geom\_smooth(method=lm)+ labs(title="scatterplot showing relationship between areas of house and Price w.r.t total rooms above groud",subtitle="Module1")

AmesHousing%>% ggplot(aes(y=`SalePrice`,x=`Lot Area`,color=`Misc Feature`))+

geom\_point(size=2,pch=17,cex=2)+geom\_smooth(method=lm)+ labs(title="scatterplot showing relationship between areas of Lot and Price w.r.t to Misc Features",subtitle="Module1")

AmesHousing%>% ggplot(aes(y=`SalePrice`,x=`entire\_house\_estimate`,color=as.factor(`Kitchen AbvGr`)))+

geom\_point(size=2,pch=17,cex=2)+geom\_smooth(method=lm)+ labs(title="scatterplot showing relationship between areas of house and Price w.r.t no of Kitchens above groud",subtitle="Module1")

AmesHousing%>% ggplot(aes(y=`SalePrice`,x=`entire\_house\_estimate`,color=as.factor(`Fireplaces`)))+

geom\_point(size=2,pch=17,cex=2)+geom\_smooth(method=lm)+ labs(title="scatterplot showing relationship between areas of house and Price w.r.t no of fireplaces ",subtitle="Module1")

AmesHousing%>% ggplot(aes(y=`SalePrice`,x=`entire\_house\_estimate`,color=as.factor(`Full Bath`)))+

geom\_point(size=2,pch=17,cex=2)+geom\_smooth(method=lm)+ labs(title="scatterplot showing relationship between areas of house and Price w.r.t no of full bathrooms above groud",subtitle="Module1")

```

<FONT size=4, color="#66B2FF">

<B>CORRELATION MATRIX:-</B><BR>

</FONT>

<FONT size=2, color="#000000">

Correlation is statistical measure of linear dependence between 2 variables . It takes values between +1 to -1. We need to test and plot the correlation matrix for our predictor variables and target variables.<BR></FONT>

```{r correlation matrix, echo=FALSE, message=FALSE, warning=FALSE, paged.print=TRUE}

AmesHousing\_cor<-AmesHousing%>%

select(`Lot Area`,

`Mas Vnr Area`,

`BsmtFin SF 1`,

`BsmtFin SF 2`,

`Bsmt Unf SF`,

`Total Bsmt SF`,

`1st Flr SF`,

`2nd Flr SF`,

`Low Qual Fin SF`,

`Gr Liv Area`,

`Garage Area`,

`Wood Deck SF`,

`Open Porch SF`,

`Enclosed Porch`,

`3Ssn Porch`,

`Screen Porch`,

`Pool Area` ,

`Misc Val`,

`Property\_age`,

SalePrice)

c<-AmesHousing\_cor%>%cor()%>%round(digits=2)

corrplot::corrplot(c,method=c("circle"),type="upper")

cor.test(AmesHousing\_cor$SalePrice,AmesHousing\_cor$`Total Bsmt SF`)

cor.test(AmesHousing\_cor$SalePrice,AmesHousing\_cor$`Gr Liv Area`)

cor.test(AmesHousing\_cor$SalePrice,AmesHousing\_cor$`Property\_age`)

cor.test(AmesHousing\_cor$SalePrice,AmesHousing\_cor$`1st Flr SF`)

cor.test(AmesHousing\_cor$SalePrice,AmesHousing\_cor$`2nd Flr SF`)

cor.test(AmesHousing\_cor$SalePrice,AmesHousing\_cor$`Open Porch SF`)

cor.test(AmesHousing\_cor$SalePrice,AmesHousing\_cor$`Wood Deck SF`)

```

<FONT size=4, color="#66B2FF">

<B>LINEAR REGRESSION MODEL</B><BR>

</FONT>

```{r model, echo=FALSE, message=FALSE, warning=FALSE}

model\_ames<-lm(SalePrice~`Total Bsmt SF`+`Gr Liv Area`+`Property\_age`+`1st Flr SF`+`2nd Flr SF`+`Open Porch SF`+`Wood Deck SF`,data = AmesHousing)

summary(model\_ames)

AIC(model\_ames)

BIC(model\_ames)

```

<FONT size=4, color="#66B2FF">

<B>MODEL INTERPRETATION</B><BR>

</FONT>

<FONT size=2, color="#000000">

1.The R-squared values is 0.7184.<BR>

2.The AIC value is `r AIC(model\_ames)``.<BR>

3.The AIC value is `r BIC(model\_ames)``.<BR>

<FONT>

<FONT size=4, color="#66B2FF">

<B>DIAGNOSTIC PLOTS</B><BR>

</FONT>

<FONT size=2, color="#000000">

we are going to plot 4 diagnostic plots to check the below assumptions :-<P>

1.Linearity of the data:- The linear relationship between the predictor and the target (y) .This can be checked by residual vs fitted values plot.<BR>

Normality of residuals.The Q-Q plot is used to check normality of residual errors .<BR>

Homogeneity of residuals variance. The residuals are assumed to have a constant variance (homoscedasticity).we need to check this assumption using Scale-location plot.<BR>

Independence of residuals error terms.<BR><P>

</FONT>

```{r diagnostic plots, message=FALSE, warning=FALSE}

par(mfrow=c(2,2))

plot(model\_ames)

```

```{r check multicolinearity , message=FALSE, warning=FALSE}

vif(model\_ames)

model\_ames\_1<-lm(SalePrice~`Total Bsmt SF`+`Property\_age`+`1st Flr SF`+`2nd Flr SF`+`Open Porch SF`+`Wood Deck SF`,data = AmesHousing)

summary(model\_ames\_1)

par(mfrow=c(2,2))

plot(model\_ames\_1)

vif(model\_ames\_1)

```

```{r normality of predictors, echo=FALSE, message=FALSE, warning=FALSE}

AmesHousing%>%ggplot(aes(`Total Bsmt SF`))+geom\_density(fill="red")

AmesHousing%>%ggplot(aes(`Property\_age`))+geom\_density(fill="pink")

AmesHousing%>%ggplot(aes(`1st Flr SF`))+geom\_density(fill="yellow")

AmesHousing%>%ggplot(aes(`2nd Flr SF`))+geom\_density(fill="green")

AmesHousing%>%ggplot(aes(`Wood Deck SF`))+geom\_density(fill="blue")

#AmesHousing$Property\_age\_new<-bcPower(AmesHousing$Property\_age,2)

#AmesHousing%>%ggplot(aes(Property\_age\_new))+geom\_density(fill="pink")

```

```{r}

par(mfrow=c(2,2))

boxplot(AmesHousing$`Property\_age`,col="green",main="boxplot of Property age",ylab="Property\_age")

boxplot(AmesHousing$`SalePrice`,col="red",main="boxplot of Salesprice",ylab="SalesPrice")

boxplot(AmesHousing$`2nd Flr SF`,col="yellow",main="boxplot of 2nd Flr SF",ylab="2nd Flr SF")

boxplot(AmesHousing$`Wood Deck SF`,col="pink",main="boxplot of Wood Deck SF",ylab="Wood Deck SF")

AmesHousing$`SalePrice`%>%boxplot.stats()

AmesHousing\_new<-AmesHousing%>%filter(`SalePrice`<=338931)

AmesHousing\_new%>%ggplot(aes(`Total Bsmt SF`))+geom\_density(fill="red")

AmesHousing\_new%>%ggplot(aes((`Property\_age`)^2))+geom\_density(fill="pink")

AmesHousing\_new%>%ggplot(aes(`1st Flr SF`))+geom\_density(fill="yellow")

AmesHousing\_new%>%ggplot(aes((`2nd Flr SF`)^2))+geom\_density(fill="green")

AmesHousing\_new%>%ggplot(aes((`Wood Deck SF`)^2))+geom\_density(fill="blue")

```

```{r}

vif(model\_ames)

model\_ames\_2<-lm(log(SalePrice)~`Total Bsmt SF`+(`Property\_age`)+`1st Flr SF`+(`2nd Flr SF`)+(`Wood Deck SF`),data = AmesHousing)

summary(model\_ames\_2)

par(mfrow=c(2,2))

plot(model\_ames\_2)

vif(model\_ames\_2)

AIC(model\_ames\_2)

BIC(model\_ames\_2)

```

```{r echo=FALSE, message=FALSE, warning=FALSE}

library(MASS)

fit\_model= lm(SalePrice~`Total Bsmt SF`+`Property\_age`+`1st Flr SF`+`2nd Flr SF`+`Wood Deck SF`,data = AmesHousing)

stepAIC(fit\_model,direction="both")

```