# SalePrice of Properties

Assignment 1: Regression Diagnostics with R

ALY6015

Prepared by: Heejae Roh
Presented to: Professor Paromita Guha
Date: Jan 15th, 2023

## INTRODUCTION

To solve the problem, effective steps are necessary. In this module, I will find the module for property price and interpret and evaluate the model I made. To make regression model, there is assumptions of Ordinary Least Square Estimators (OLS). While interpreting and evaluating my model, I will use OLS to check the fit of the model. I will briefly explain my research process and its contents.

## ABSTRACTION

### Data itself

Target variable (dependent variable) is SalePrice (price of properties). Dataset contains 77 variables which are related to SalePrice. There are 20 continuous and 14 discrete variables. To make regression model, I will concentrate on continuous model in this project.

### Data Preparation

There are many NA values in specific variables. I delete the column which has over 1,000 NA data to make my analysis more accurate.

### EDA

Histogram of SalePrice and Gr.Liv.Area (which is above grade living area square feet) is bell shape. There are many '0' values in specific variables.

### PREREQUISITE for MODELING

I make correlation matrix of continuous & discrete variables. While checking the correlation with SalePrice, I also checked the correlationship between other variables. I draw plots of correlation matrix and explained it.

### TESTING ASSUMPTIONS of OLS

I follow the steps for Assumptions of Ordinary Least Square Estimators in this dataset. Using specific method like Durbin-Watson test.

### MODEL on my own

I do normalization to compare the impact of different independent variables. After making the model, I draw a plots and evaluate the model with AIC & BIC.

### MULTICOLLINEARITY

To check multicollinearity, I used VIF values to my own model. I used whole continuous variables to find out steps to deal with multicollinearity.

### OUTLIERS

I find out outliers of my own model. By deleting them, I think about the effect of outliers to model. I also used AIC & BIC in this case.

### ALL SUBSETS REGRESSION MODEL

Using lmSubsets() function, I learn the all subsets regression method and think about the advantages of this method.

## CONCLUSION

I realize that understanding variables before making regression model is very important. I learn how to make model better with AIC & BIC and what's the assumption for Ordinary Least Square Estimateors (OLS). I practice every step of OLS with real-world dataset. I think about the specific situation that I prefer to use all subsets regression method.

## PART 0. INTRODUCTION

In this project I will think about the steps to analyze dataset. Checking Assumptions of Ordinary Least Square Estimators (OLS) is important to make regression model. In regression model, these OLS steps are needed to analyze dataset. Considering about OLS we can find out 'the best equation' with steps. The steps are very important to find out answers for the questions. These questions include all the questions that we can encounter in real world problem.

## PART 1. ANALYSIS

**Data Analysis by data explanation ('AmesHousingDataDocumentation.txt')**

1. Data set contains information from the Ames Assessor Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010
2. Tab characters are used to separate variables in the data file. The data has 82 columns which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers)
3. Dependent variable (Target variable) could be SalePrice (Continuous): Sale price
4. There are some categories like Bsmt, Mas, SF, etc. which can have multicollinearity according to the explanation

## PART 1-1. SUMMARY of EDA

**Headtail of Data 1**

| | SalePrice | Gr.Liv.Area | Total.Bsmt.SF | Garage.Area | Mas.Vnr.Area | Wood.Deck.SF |
|---|---|---|---|---|---|---|
| 1 | 215000 | 1956 | 1080 | 528 | 112 | 210 |
| 2 | 105000 | 896 | 882 | 730 | 0 | 140 |
| 3 | 172000 | 1329 | 1329 | 312 | 108 | 393 |
| … | | | | | | |
| 2928 | 132000 | 970 | 912 | 0 | 0 | 80 |
| 2929 | 170000 | 1389 | 1389 | 418 | 0 | 240 |
| 2930 | 188000 | 2000 | 996 | 650 | 94 | 190 |

**Headtail of Data 2**

| | Screen.Porch | BsmtFin.SF.1 | X1st.Flr.SF | Eclosed.Porch | Year.Built |
|---|---|---|---|---|---|
| 1 | 0 | 639 | 1656 | 0 | 1329 |
| 2 | 120 | 468 | 896 | 0 | 1961 |
| 3 | 0 | 923 | 1329 | 0 | 1958 |
| … | | | | | |
| 2928 | 0 | 337 | 970 | 0 | 1992 |
| 2929 | 0 | 1071 | 1389 | 0 | 1974 |
| 2930 | 0 | 758 | 996 | 0 | 1993 |

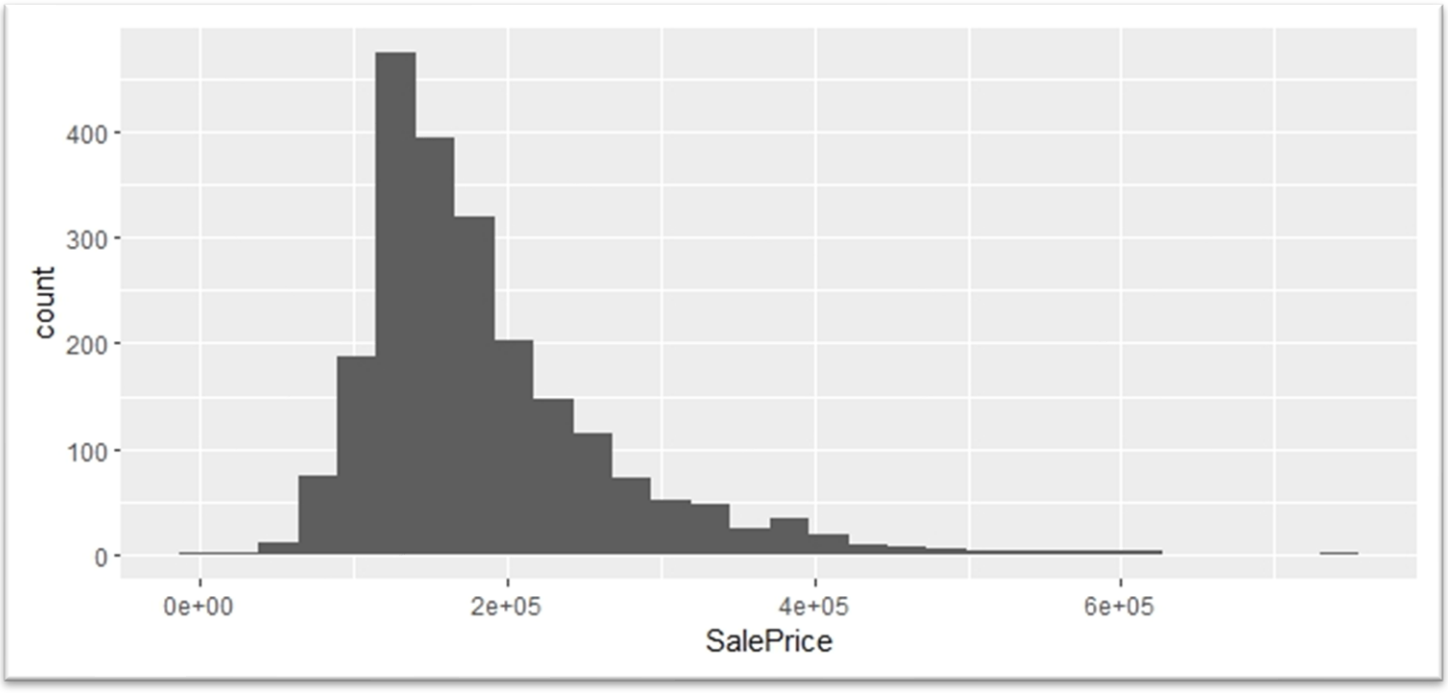
**Data Cleaning**

1. Checking summary(dataset)
2. Find out NA over 1,000 values: Alley, Fireplace.Qu, Fence, Pool.QC, Misc.Feature
3. Delete column which name is Alley, Fireplace.Qu, Fence, Pool.QC, Misc.Feature
4. Execute na.omit(dataset)
5. Finally, there are 2,223 observations and 77 variables

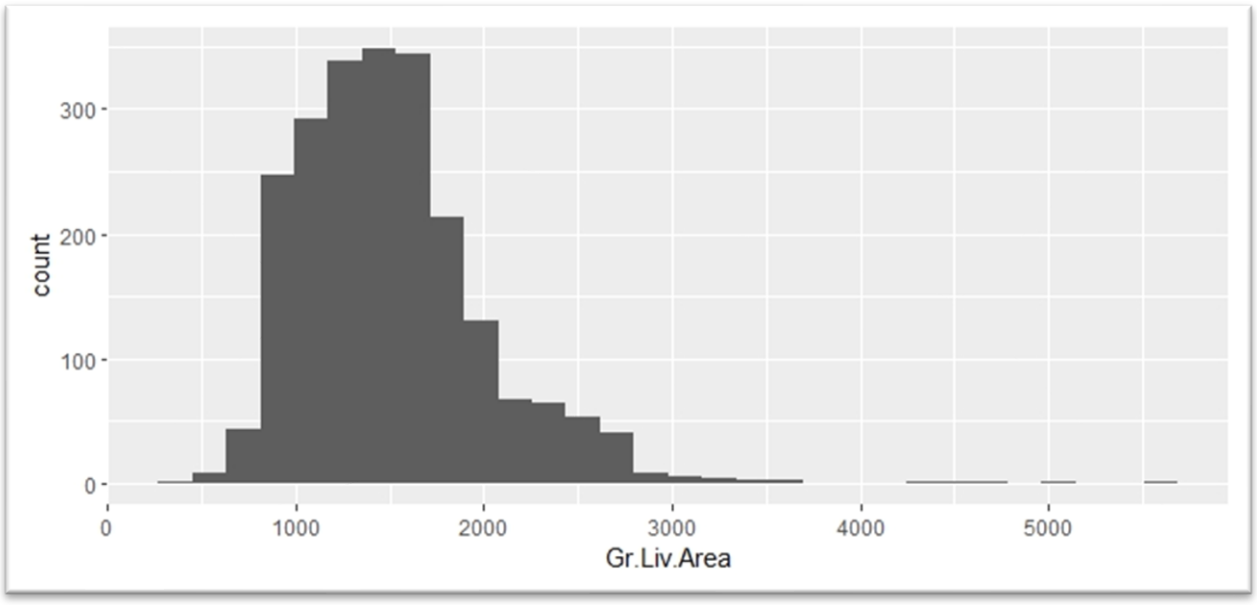
**Descriptive Analysis of primary data**

| | n | Mean | distinct | Info | Gmd | .25 | .50 | .75 | lowest | highest |
|---|---|---|---|---|---|---|---|---|---|---|
| SalePrice | 2223 | 186118 | 880 | 1 | 84683 | 131500 | 164000 | 217750 | 12789 | 755000 |
| Gr.Liv.Area | 2223 | 1507 | 1117 | 1 | 535.7 | 1144 | 1448 | 1734 | 407 | 5642 |
| Total.Bsmt.SF | 2223 | 1091 | 951 | 1 | 446.3 | 810 | 1008 | 1336 | 105 | 6110 |
| Garage.Area | 2223 | 499.8 | 556 | 1 | 214.6 | 350.5 | 484 | 588 | 100 | 1488 |
| Mas.Vnr.Area | 2223 | 107.1 | 398 | 0.793 | 161.6 | 0 | 0 | 172 | 0 | 1600 |
| Wood.Deck.SF | 2223 | 94.09 | 332 | 0.867 | 122.9 | 0 | 0 | 168 | 0 | 870 |
| Screen.Porch | 2223 | 17.01 | 104 | 0.254 | 31.52 | 0 | 0 | 0 | 0 | 576 |
| BsmtFin.SF.1 | 2223 | 450.2 | 871 | 0.972 | 494.9 | 0 | 375 | 737.5 | 0 | 5644 |
| X1st.Flr.SF | 2223 | 1165 | 966 | 1 | 426.2 | 876 | 1086 | 1392 | 407 | 5095 |
| Eclosed.Porch | 2223 | 22.71 | 153 | 0.396 | 40.6 | 0 | 0 | 0 | 0 | 1012 |
| Year.Built | 2223 | 1972 | 113 | 0.999 | 34.59 | 1954 | 1975 | 2003 | 1879 | 2010 |

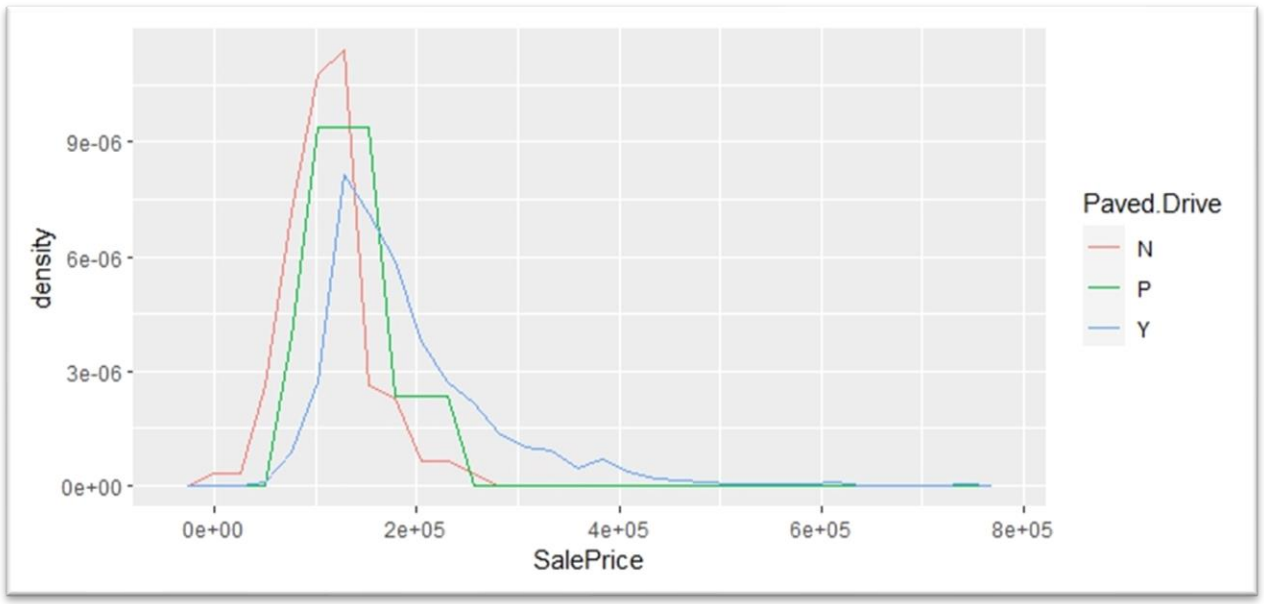
**Histogram of SalePrice**



**Histogram of Gr.Liv.Area**




**Histogram of SalePrice by Paved (Y: Paved, P: Partial Pavement, N: Dirt/Gravel)**

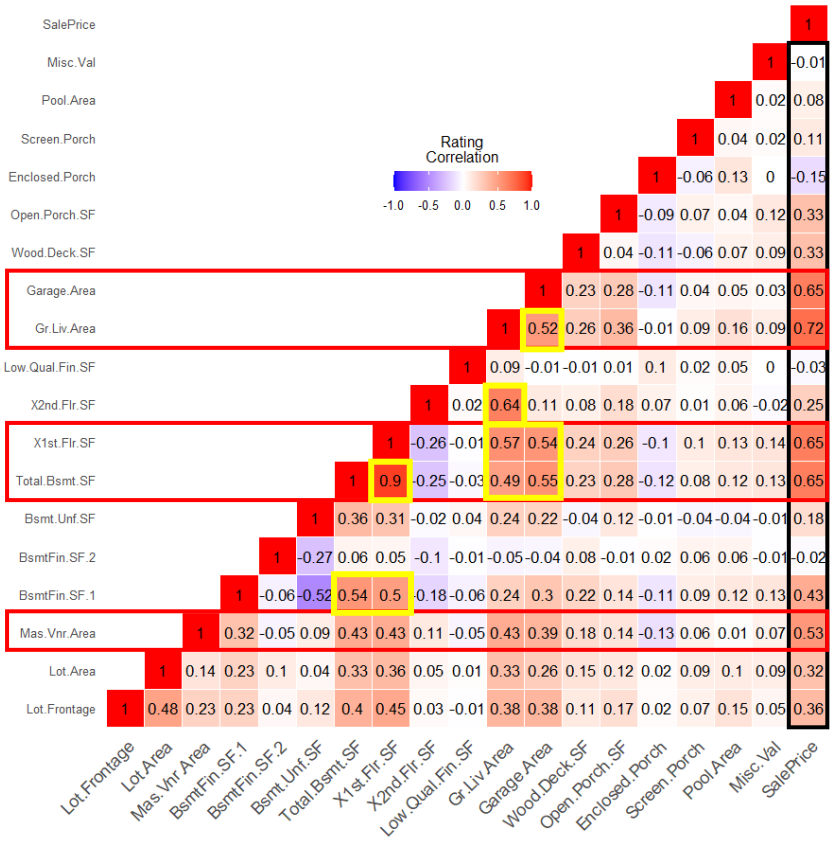



**Explanation:**

1. Start with the understanding explanation about dataset
2. 2223 observations, 77 variables after data cleaning
3. Target variable is SalePrice. Median of SalePrice is 186,118
4. SalePrice by paved is higher than partially paved & dirt/gravel. The number of observations by pave status is ‘N: Dirt/Gravel’ > ‘P: Partial Pavement’ > ‘Y: Paved’

## PART 1-2. PREREQUISITE for MODELING

### Correlation Matrix of continuous variables, Target variable=SalePrice




### Explanation (Continuous):

1. (RED) Highest Correlation with SalePrice: Gr.Liv.Area (0.72) > Garage.Area (0.65) = X1st.Flr.SF (0.65) = Total.Bsmt.SF (0.65) > Mas.Vnr.Area (0.53) …
2. (YELLOW) Highest Cor. between other variables: Total.Bsmt.SF-X1st.Flr.SF (0.9) Gr.Liv.Area-X1st.Flr.SF (0.57) > Garage.Area- Total.Bsmt.SF (0.55) > Garage.Area- X1st.Flr.SF (0.54) = BsmtFin.SF.1- Total.Bsmt.SF (0.54) > Gr.Liv.Area- Garage.Area (0.52) > BsmtFin.SF.1- X1st.flr.SF (0.5) > Gr.Liv.Area- Total.Bsmt.SF (0.49)

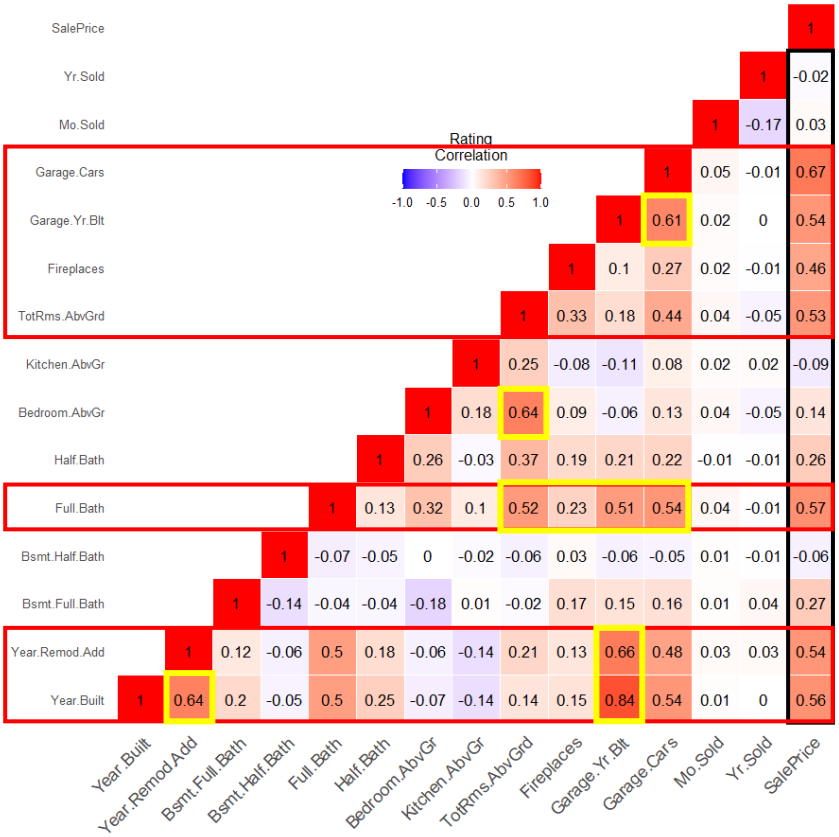
### Interpretation (Continuous):

1. SalePrice is correlated with Gr.Liv.Area (0.72), Garage.Area (0.65), X1st.Flr.SF (0.65), Total.Bsmt.SF (0.65), Mas.Vnr.Area (0.53) …
2. Might be multicollinearity between Total.Bsmt.SF-X1st.Flr.SF (0.9) Gr.Liv.Area-X1st.Flr.SF (0.57), Garage.Area- Total.Bsmt.SF (0.55),

Garage.Area- X1st.Flr.SF (0.54), BsmtFin.SF.1- Total.Bsmt.SF (0.54), Gr.Liv.Area- Garage.Area (0.52), BsmtFin.SF.1- X1st.flr.SF (0.5), Gr.Liv.Area- Total.Bsmt.SF (0.49)

3. Enclosed.Porch has highest negative correlation with SalePrice, But not huge.

**Correlation Matrix of discrete variables, Target variable=SalePrice**




**Explanation (Discrete):**

1. (RED) Highest Correlation with SalePrice: Garage.Cars (0.67) > Full.Bath (0.57) > Year.Built (0.56) > Garage.Yr.Blt(0.54) > Year.Remod.Add (0.54) …
2. (YELLOW) Highest Cor. between others: Year.Built - Garage.Yr.Blt (0.84) Garage.Yr.Blt -Year.Remod.Add (0.66) > Year.Built - Year.Remod.Add (0.64) …

**Interpretation (Discrete):**

1. SalePrice is correlated with Garage.Cars (0.67), Full.Bath (0.57), Year.Built (0.56), Garage.Yr.Blt(0.54), Year.Remod.Add (0.54), etc.
2. Might be multicollinearity between Year.Built - Garage.Yr.Blt (0.84), Garage.Yr.Blt -Year.Remod.Add (0.66), Year.Built - Year.Remod.Add (0.64), etc.

3. Kitchen.AbvGr has highest negative correlation with SalePrice, but not huge
4. There are many variables highly correlated to continuous variables like Garage.Cars (which means numbers of car available in Garage), etc.
5. Year.Built is highly correlated to other discrete variables, but it seems like not highly correlated to continuous variables

## Plot of the Correlation matrix of Continuous Variables PART 1


## Explanation:

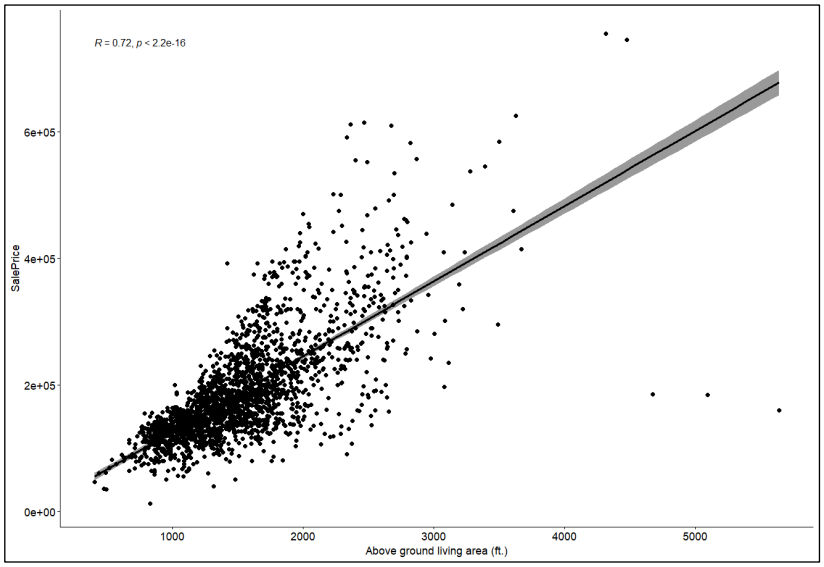
1. (RED) Correlation Between SalePrice and Total.Bsmt.SF is 0.646*** and scatter plot of them looks linear without extreme outliers
2. (YELLOW) Scatter plot Between X1st.Flr.SF and Total.Bsmt.SF looks like y=x, whose correlation coefficient was 0.9

## Plot of the Correlation matrix of Continuous Variables PART 2


## Explanation:

1. (RED) Correlation Between SalePrice and Gr.Liv.Area is 0.717*** and scatter plot of them looks linear without extreme outliers
2. (Limitation 1) These plots from ggpairs() is executed after divided continuous variables into two group. There is missing graph between other variables like Gr.Liv.Area-X1st.Flr.SF (0.57), which was 2nd highest correlation coefficient variables.

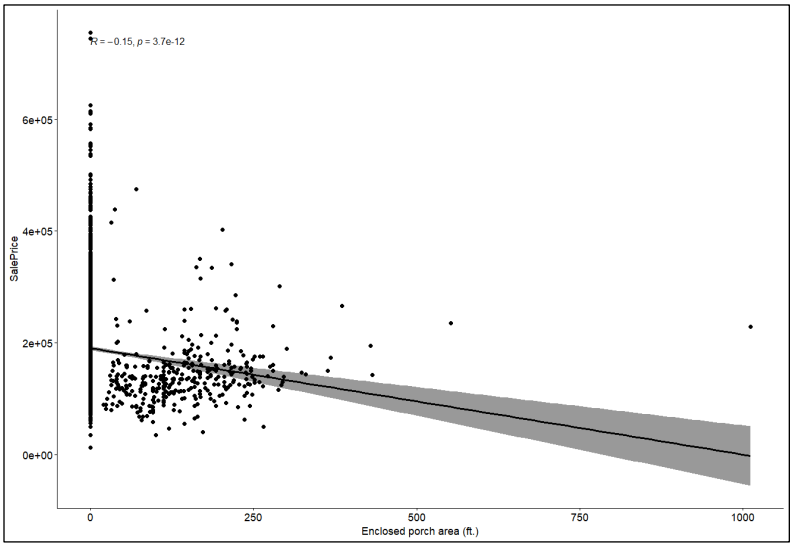
**Scatter plot for highest correlation with SalePrice: Gr.Liv.Area**




**Interpretation:**

1. R=0.72 between SalePrice and Gr.Liv.Area and p < 0.01
2. There is linear relation between two

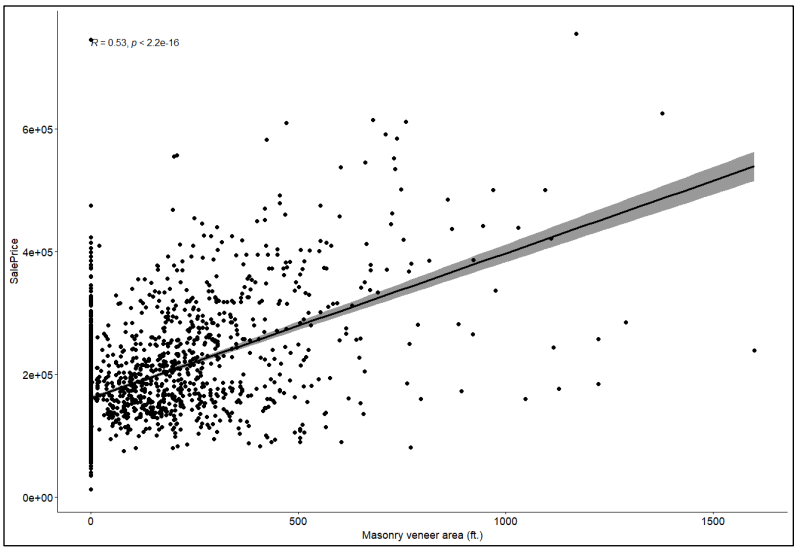
**Scatter plot for lowest correlation with SalePrice: Enclosed.Porch**




**Interpretation:**

1. R=-0.15 between SalePrice and Enclosed.Porch and p < 0.01
2. There are many 0 values in Enclosed.Porch, check needed for Enclosed.Porch residuals

**Scatter plot for lowest correlation with SalePrice: Mas.Vnr.Area**




**Interpretation:**

1. R=0.53 between SalePrice and Mas.Vnr.Area and p < 0.01
2. There are many 0 values in Mas.Vnr.Area, check needed for Enclosed.Porch residuals

**Prerequisite for model: normalization**

**Explanation:**

1. To compare correlation between different scaled variables, do normalization with preProcess()

**Making model with all continuous variables**

Coefficients: (2 not defined because of singularities)

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -1.318e-05 | 4.681e-03 | -0.003 | 0.997754 | |
| Lot.Frontage | -3.808e-02 | 1.946e-02 | -1.957 | 0.050431 | . |
| Lot.Area | 1.016e-01 | 4.874e-02 | 2.085 | 0.037188 | * |
| Mas.Vnr.Area | 1.213e-01 | 1.292e-02 | 9.388 | < 2e-16 | *** |
| BsmtFin.SF.1 | 4.252e-01 | 4.064e-02 | 10.461 | < 2e-16 | *** |
| BsmtFin.SF.2 | 7.228e-02 | 1.509e-02 | 4.790 | 1.78e-06 | *** |
| Bsmt.Unf.SF | 1.196e-01 | 1.674e-02 | 7.144 | 1.23e-12 | *** |
| Total.Bsmt.SF | NA | NA | NA | NA | |
| X1st.Flr.SF | 4.316e-01 | 3.578e-02 | 12.062 | < 2e-16 | *** |
| X2nd.Flr.SF | 1.883e-01 | 7.391e-03 | 25.482 | < 2e-16 | *** |
| Low.Qual.Fin.SF | -2.762e-03 | 3.092e-02 | -0.089 | 0.928828 | |
| Gr.Liv.Area | NA | NA | NA | NA | |
| Garage.Area | 1.814e-01 | 1.175e-02 | 15.432 | < 2e-16 | *** |
| Wood.Deck.SF | 6.674e-02 | 9.598e-03 | 6.954 | 4.67e-12 | *** |
| Open.Porch.SF | 3.649e-02 | 1.190e-02 | 3.067 | 0.002192 | ** |
| Enclosed.Porch | -8.007e-02 | 2.063e-02 | -3.882 | 0.000107 | *** |
| Screen.Porch | 4.003e-02 | 1.282e-02 | 3.123 | 0.001811 | ** |
| Pool.Area | -8.384e-02 | 2.911e-02 | -2.880 | 0.004015 | ** |
| Misc.Val | -4.466e-01 | 4.226e-02 | -10.567 | < 2e-16 | *** |

**Explanation:**

1. There are '2 not defined because of singularities', it means that two or more of your independent variables are perfectly collinear (Peter, 2012)
2. Execute alias(model) and find out Total.Bsmt.SF is related to BsmtFin.SF.1, BsmtFin.SF.2, and Bsmt.Unf.SF
3. Find out Gr.Liv.Area is related to X1st.Flr.SF, X2nd.Flr.SF, and Low.Qual.Fin.SF
4. Delete X1st.Flr.SF, X2nd.Flr.SF, Low.Qual.Fin.SF, BsmtFin.SF.1, BsmtFin.SF.2, Bsmt.Unf.SF
5. Making model with whole variables and check correlation coefficients by order
6. Norm_scale2 is without too many na values (Alley,Fireplace.Qu,Fence,Pool.QC,Misc.Feature) and without collinear variables (X1st.Flr.SF, X2nd.Flr.SF, Low.Qual.Fin.SF, BsmtFin.SF.1, BsmtFin.SF.2, Bsmt.Unf.SF)

**Finding correlation coefficient in descending order**

Call: lm(formula = SalePrice ~ ., data = norm_scale2

Residuals

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -0.88920 | -0.02714 | 0.00058 | 0.02627 | 0.42224 |

Coefficients:

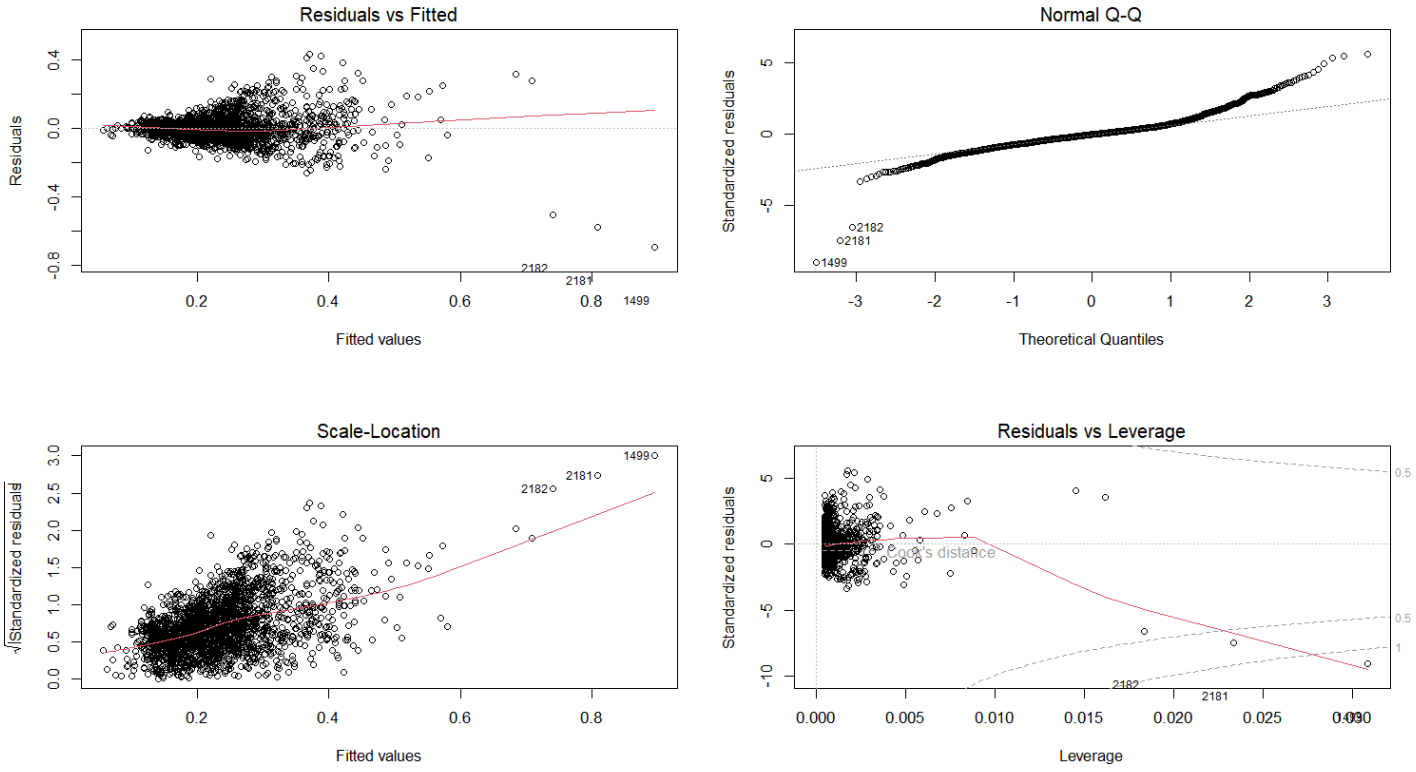
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.003859 | 0.004140 | 0.932 | 0.351391 |
| Gr.Liv.Area | 0.456368 | 0.017957 | 25.415 | < 2e-16 *** |
| Misc.Val | -0.426396 | 0.042730 | -9.979 | < 2e-16 *** |
| Total.Bsmt.SF | 0.389607 | 0.024271 | 16.053 | < 2e-16 *** |
| Garage.Area | 0.182397 | 0.011900 | 15.328 | < 2e-16 *** |
| Mas.Vnr.Area | 0.136858 | 0.012957 | 10.562 | < 2e-16 *** |
| Enclosed.Porch | -0.090898 | 0.020833 | -4.363 | 1.34e-05 *** |
| Wood.Deck.SF | 0.074695 | 0.009612 | 7.771 | 1.18e-14 *** |
| Screen.Porch | 0.044714 | 0.012927 | 3.459 | 0.000552 *** |
| Open.Porch.SF | 0.037573 | 0.012038 | 3.121 | 0.001824 ** |
| Lot.Area | 0.121534 | 0.049115 | 2.474 | 0.013418 * |
| Pool.Area | -0.068736 | 0.029351 | -2.342 | 0.019275 * |
| Lot.Frontage - | -0.035507 | 0.019504 | -1.820 | 0.068824 . |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06002 on 2210 degrees of freedom
Multiple R-squared: 0.716, Adjusted R-squared: 0.7145
F-statistic: 464.3 on 12 and 2210 DF p-value: < 2.2e-16

## PART 1-3. TESTING ASSUMPTIONS of OLS

1. Normality (Residuals vs fitted & Normal Q-Q): Checking residual normality of each variable
2. Linearity (Residuals vs fitted): Checking linear relationship between X and Y
3. Homoscedasticity (Scale-Location): Checking same variance of errors
   + identifying outliers (Residuals vs Leverage)
4. Independence of errors (Durbin-Watson Test): Cehcking autocorrelation and distribution of errors is not correlated to the errors in prior observations (Paro, 2023)
5. Will check No 5. Multicollinearity in PART 1-5 and 6. Overfit in Conclusion 7. Outliers in PART 1-6




**Explanation:**

1. This is plot() of Model lm(formula = SalePrice ~ Gr.Liv.Area, data = norm_scale2)
2. Checking every variable which has over **(** or more) with this process
3. Durbin-Watson Test result is 1.077 is between 0 and 4.

| **durbinWatsonTest(for_mod1)** | | | |
|---|---|---|---|
| lag | Autocorrelation | D-W Statistic | p-value |
| 1 | 0.4611415 | 1.077271 | 0 |

**Interpretation (Based on Pre-Assignment Lab):**

1. (Residuals vs fitted) Checking Linearity, It has to be random pattern. In this case, it has a particular shape, it's not good for the model assumption
2. (Normal Q-Q) The plot does not follow the diagonal line. There are many things that does not follow diagonal line which means that, there is unusual observation or if there are a few spots, so this might be violating the assumption of normality. There is increasing shape, Theoretical Quantiles > 1 and decreasing shape after Theoretical Quantiles < -1
3. (Scale-Location) Checking constant variance or Homoscedasticity, there is not-randomized shape in here
4. (Residuals vs Leverage) Identifying unusual observations, there are extreme 3 observations like 2182 and 2181. This let us know there might be errors very high either positive or negative residuals. We might consult with a subject matter expert or business user

## PART 1-4. MODEL with BACK & FORWARD SELECTION wit AIC and BIC

**Making model: Descending order of correlation coefficient 'and understanding of component**

SalePrice = – 1,312,000+663.4* Year.Built +69.51* Gr.Liv.Area-15.84* Misc.Val
+38.04*Total.Bsmt.SF+65*Garage.Area+50.59*Mas.Vnr.Area
+48.31*Wood.Deck.SF+92.93*Screen.Porch

**Explanation:**

1. This model is made by normalization analysis 'PART 1-2. PREREQUISITE for MODELING'. Including highest correlation coefficient in normalized analysis.
2. Adding Year.Built which is discrete variables in explanation of dataset. I think it has high correlation coefficient with SalePrice of properties and there is low risk of multicollinearity with others.

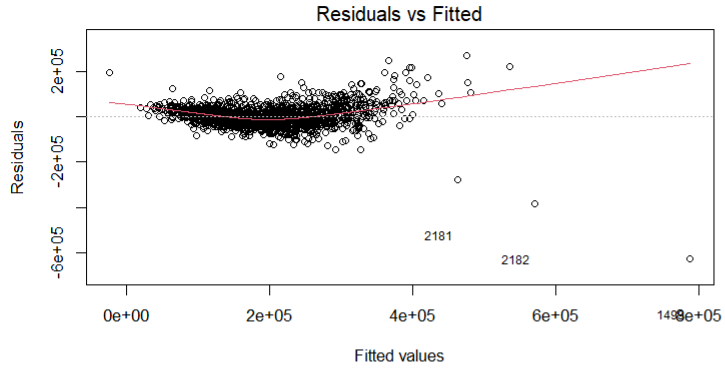
**Steps to make fitted model**

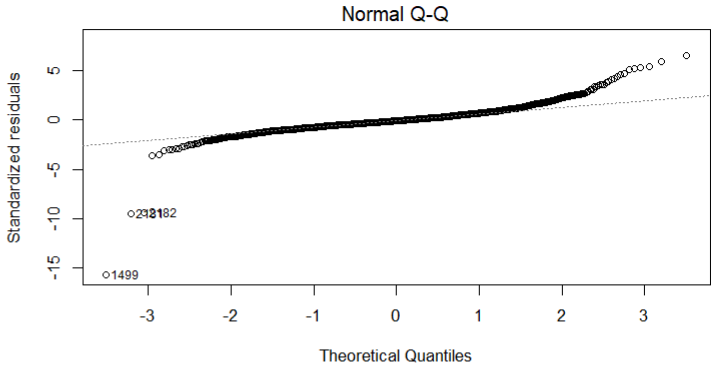
| | Model | AIC | BIC |
|---|---|---|---|
| 1 | Only with Continuous variables (SalePrice~Gr.Liv.Area+Misc.Val+Total.Bsmt.SF+Garage.Area+Mas.Vnr.Area+Enclosed.Porch+Wood.Deck.SF+Screen.Porch) | 53929 | 53986 |
| 2 | Add Year.Built in model 1 | 53606(-323) | 53669(-317) |
| 3 | Delete Enclosed.Porch (one of continuous variables), because P value is 0.07 in the model 2 | 53608(+2) | 53665(-4) |

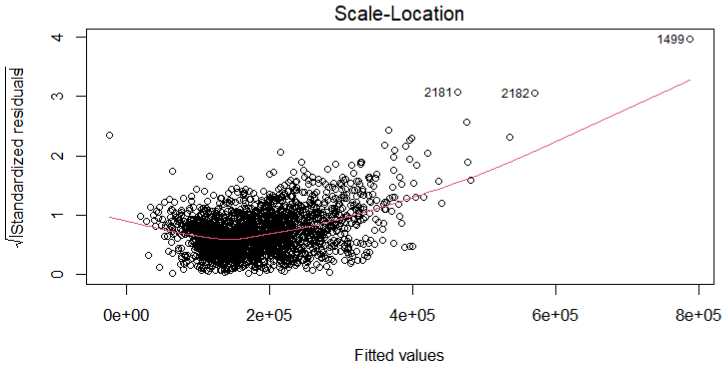
## R-squared with my model

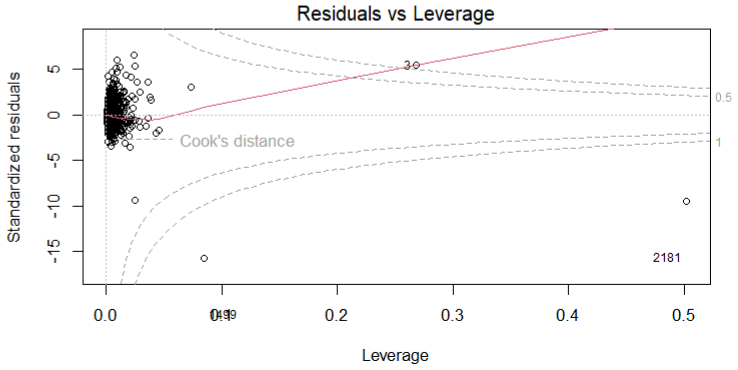
| | |
|---|---|
| Residual standard error: 41360 on 2210 degrees of freedom | |
| Multiple R-squared: 0.755, | Adjusted R-squared: 0.7541 |
| F-statistic: 851.1 on 8 and 2210 DF | p-value: < 2.2e-16 |

## 4 Plots of fitted model



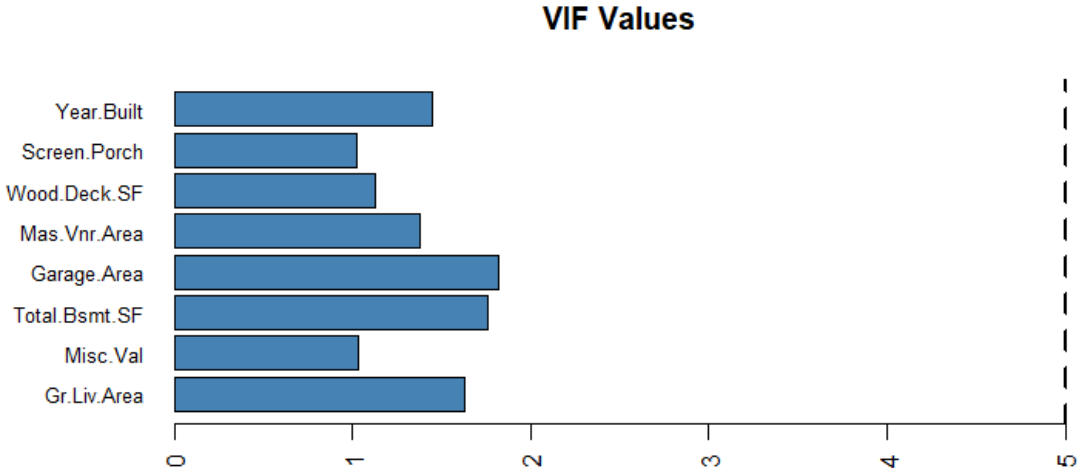




## Interpretation (Based on Pre-Assignment Lab):

1. (Residuals vs fitted) Checking Linearity, it has a particular shape
2. (Normal Q-Q) The plot does not follow the diagonal line perfectly
3. (Scale-Location) Checking constant variance or Homoscedasticity, it seems little bit randomized shape without some outliers
4. (Residuals vs Leverage) there are extreme 3 observations 3, 1499, and 2181.

## PART 1-5. MULTICOLLINEARITY

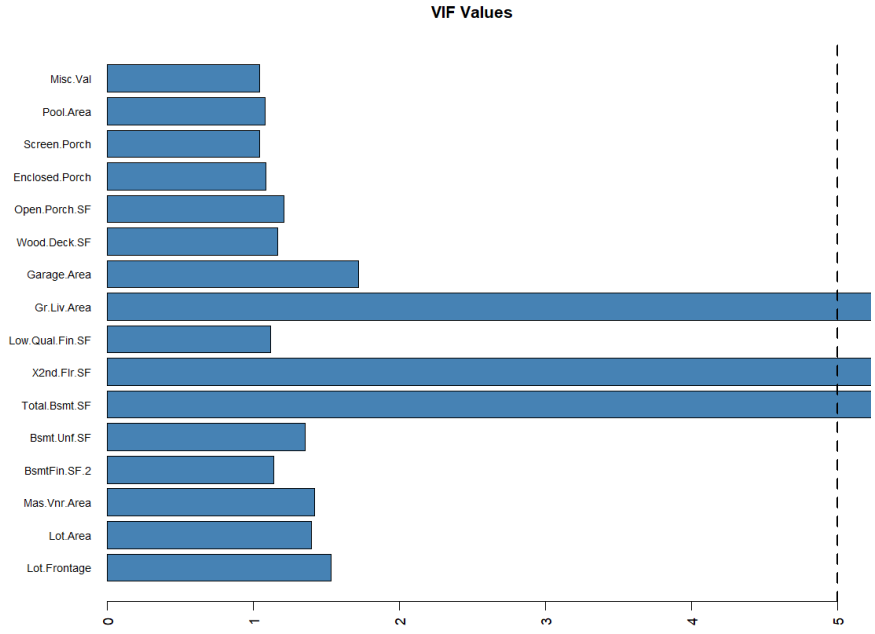
### Checking Multicollinearity of my model




### Interpretation:

1. The VIF value of each variable is between 1 and 2. It's lower than 5, so there is moderate multicollinearity
2. I think, this is because I already delete the variables while checking singularity with alias() function. At that time, I delete BsmtFin.SF.1, BsmtFin.SF.2, and Bsmt.Unf.SF which have multicollinearity with Total.Bsmt.SF
3. Also deleted X1st.Flr.SF, X2nd.Flr.SF, and Low.Qual.Fin.SF which have multicollinearity with Gr.Liv.Area

### What if, multicollinearity exists: This is VIF of model with all continuous variables



**Steps to correct multicollinearity:**

1. Checking multicollinearity with alias() function

| Complete: | | | | |
|---|---|---|---|---|
| | (Intercept) | BsmtFin.SF.1 | BsmtFin.SF.2 | Bsmt.Unf.SF |
| Total.Bsmt.SF | -21/1201 | 2095/2229 | 1474/6005 | 2336/6005 |

| Complete: | | | | |
|---|---|---|---|---|
| | (Intercept) | X1st.Flr.SF | X2nd.Flr.SF | Low.Qual.Fin.SF |
| Gr.Liv.Area | 0 | 4688/5235 | 299/758 | 1064/5235 |

2. Delete the variables that have multicollinearity

## PART 1-6. OUTLIERS

**Finding outliers in my model, Removed? Yes**

1. (Residuals vs Fitted) 2181, 2182, and 1499
2. (Normal Q-Q) 2181, 2182, and 1499
3. (Scale-Location) 2181, 2182, and 1499
4. (Residuals vs Leverage) 3, 1499, and 2181

**Checking dataset: 2181, 2182, 1499, 3 about our model**

| | Gr.Liv | Misc.Val | Total.Bsmt | Garage | Mas.Vnr | Wood.Deck | Screen | Year.Built |
|---|---|---|---|---|---|---|---|---|
| 3 | 1329 | 12500 | 1329 | 312 | 108 | 393 | 0 | 1958 |
| 1499 | 5642 | 0 | 6110 | 1418 | 796 | 214 | 0 | 2008 |
| 2181 | 5095 | 17000 | 5095 | 1154 | 1224 | 546 | 0 | 2008 |
| 2182 | 4676 | 0 | 3138 | 884 | 762 | 208 | 0 | 2007 |
| 4 | 2110 | 0 | 2110 | 522 | 0 | 0 | 0 | 1968 |
| 5 | 1629 | 0 | 928 | 482 | 0 | 212 | 0 | 1997 |
| 6 | 1604 | 0 | 926 | 470 | 20 | 360 | 0 | 1998 |
| 7 | 1338 | 0 | 1338 | 582 | 0 | 0 | 0 | 2001 |
| 8 | 1280 | 0 | 1280 | 506 | 0 | 0 | 144 | 1992 |
| 1000 | 1593 | 0 | 1593 | 682 | 0 | 0 | 224 | 2001 |

**Making model without outliers 3, 1499, 2181, 2182**

| | |
|---|---|
| Residual standard error: 0.06002 on 2210 degrees of freedom | |
| Multiple R-squared: 0.8033, | Adjusted R-squared: 0.8026 |
| F-statistic: 1128 on 8 and 2210 DF | p-value: < 2.2e-16 |

**Improvement of the model without outliers**

| | R-squared | Adjusted R-squared | AIC | BIC |
|---|---|---|---|---|
| previous | 0.755 | 0.7541(-0.0009) | 53484 | 53541 |
| without outliers | 0.8033 | 0.8026(-0.0007) | 52998(-486) | 53055(-486) |

**Interpretation:**

1. After deleting outliers in dataset, R-squared getting higher as 0.8033 from 0.755
2. Adjusted R-squared getting better as 0.0007 difference from 0.0009
3. There is no value for AIC that can be considered “good” or “bad” because we simply use AIC as a way to compare regression models. The model with the lowest AIC offers the best fit (ZACH, 2021).
4. AIC and BIC getting smaller as much as the difference of 486
5. Why? By deleting the outliers among the values used to create the model, the residuals decreased and the ability to explain the dependent variable, SalePrice, increased.

## PART 1-7. ALL SUBSETS REGRESSION MODEL

**All subsets regression method**

```
lmSubsets(SalePrice ~ ., data=norm_scale)
```

**Results (In all SIZE, BEST=1 and pval is < 2.22e-16)**

| SIZE | sigma | R2 | R2adj | cp | AIC | BIC |
|---|---|---|---|---|---|---|
| 2 | 0.07827 | 0.5146 | 0.5143 | 1661.70 | -5014 | -4997 |
| 3 | 0.06851 | 0.6283 | 0.6279 | 754.87 | -5605 | -5582 |
| 4 | 0.06427 | 0.6730 | 0.6726 | 398.81 | -5888 | -5860 |
| 5 | 0.06269 | 0.6891 | 0.6885 | 272.71 | -5998 | -5964 |
| 6 | 0.06159 | 0.7000 | 0.6994 | 187.07 | -6076 | -6036 |
| 7 | 0.06065 | 0.7092 | 0.7084 | 115.66 | -6143 | -6097 |
| 8 | 0.06002 | 0.7154 | 0.7145 | 68.47 | -6188 | -6137 |
| 9 | … | … | … | … | … | … |
| 16 | | | | | | |
| 17 | 0.05917 | 0.7244 | 0.7224 | 13.81 | -6243 | -6140 |
| 18 | 0.05919 | 0.7244 | 0.7223 | 15.80 | -6241 | -6132 |
| 19 | 0.05920 | 0.7244 | 0.7244 | 17.78 | -6239 | -6124 |

| [variable, best] | = size 1st |
|---|---|
| +(Intercept) | 2-19 |
| Lot.Frontage | 15-19 |
| Lot.Area | 14-19 |
| Mas.Vnr.Area | 5-19 |
| BsmtFin.SF.1 | 7-19 |
| BsmtFin.SF.2 | 17-19 |
| Bsmt.Unf.SF | 16-19 |
| Total.Bsmt.SF | 3-19 |
| X1st.Flr.SF | 10-17,19 |
| X2nd.Flr.SF | 10-19 |
| Low.Qual.Fin.SF | 18-19 |
| Gr.Liv.Area | 2-9,18-19 |
| Garage.Area | 4-19 |
| Wood.Deck.SF | 8-19 |
| Open.Porch.SF | 13-19 |
| Enclosed.Porch | 9-19 |
| Screen.Porch | 11-19 |
| Pool.Area | 12-19 |
| Misc.Val | 6-19 |

**Identifying "best" model with SIZE 8**

SalePrice= -49951.9*(Intercept)+76.2*Gr.Liv.Area+
52.3*Total.Bsmt.SF+85.9*Garage.Area
+55.7*Mas.Vnr.Area-2.0*Misc.Val
+26.9*BsmtFin.SF.1+43.8*Wood.Deck.SF

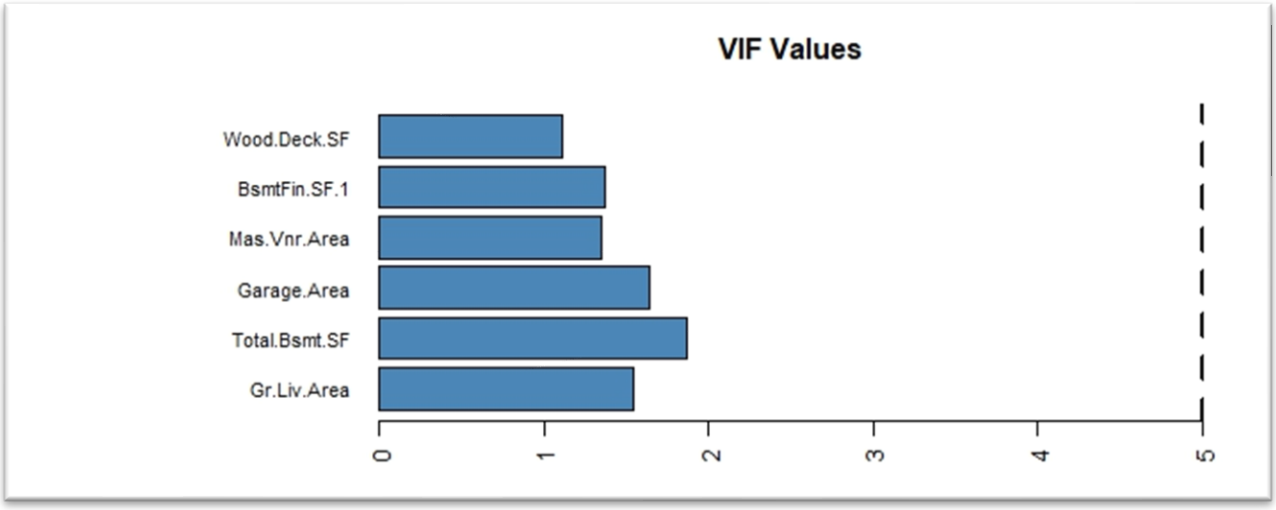
**Delete 'Misc.Val', because p-value is 0.535**

SalePrice= -50044.3*(Intercept)+76.2*Gr.Liv.Area+
52.4*Total.Bsmt.SF+85.9*Garage.Area
+55.7*Mas.Vnr.Area+26.9*BsmtFin.SF.1
+43.7*Wood.Deck.SF

**All subsets regression method vs Forward Selection method (or other)**

**SIZE 7 model without 'Misc.Val' & VIF**

Residual standard error: 38900 on 2212 degrees of freedom
Multiple R-squared: 0.7833, Adjusted R-squared: 0.7827
F-statistic: 1333 on 6 and 2212 DF p-value: < 2.2e-16




**+Year.Built with all subsets regression model**

SalePrice= -1,169,000*(Intercept)+80.0*Gr.Liv.Area+41.9*Total.Bsmt.SF
+55.4*Garage.Area+43.8*Mas.Vnr.Area+26.1*BsmtFin.SF.1
+28.0*Wood.Deck.SF+579.3*Year.Built

**Summary of +Year.Built with all subsets regression model**

| | |
|---|---|
| Residual standard error: 35920 on 2211 degrees of freedom | |
| Multiple R-squared: 0.8153, | Adjusted R-squared: 0.8147 |
| F-statistic: 1394 on 7 and 2211 DF | p-value: < 2.2e-16 |

**Comparing model of mine vs all subsets regression**

**Interpretation:**

1. I think we can find a model faster with all subsets regression method than my model from step 12. Especially, if there are many variables in dataset, it could be better to use all subsets regression model, because it informs the size and effect of the model considering all possibilities of various variables
2. However, my model from step 12 completed the model by going through the steps of analysis one by one. Therefore, it was possible to create a model with a higher understanding of the data, and it can be said that the accuracy is higher
3. As such, the advantage of the all subsets regression model and the step by step regression model is differ. For me, when time is short and there are many variables, I will first analyze the data using the 'all subsets regression model'. However, if I have a lot of time, I will analyze the data by following the analysis learned in this module step by step

## RESULT & INTERPRETATIONS

1. According to ‘TESTING ASSUMPTIONS of OLS’, I learn every step of assumptions. I use plot() function and check ‘Residuals vs fitted’ and ‘Normal Q-Q’ to check normality. With Durbin-Watson Test, I can check the independence of errors. To check out linearity ‘Residuals vs fitted’ works. With ‘Scale-Location’ plot, I can check the homoscedasticity to see the shape of graph.
2. No multicollinearity is also important thing to check. I use VIF value and alias() function to check this. I carefully added variables in my own model to avoid the overfit bias. To avoid overfit bias, I think the understanding of dataset itself is also important.
3. I learn how to find the outliers with plot() function. After checking outliers, my model improved, and I also learn all subsets regression method in this module.

## ANSWERS FOR THE QUESTIONS

1. Number 6 Deliverable, Interpret the scatter plots and describe how the patterns differ

   The highest correlation coefficient is 0.72 of Gr.Liv.Area and the angle on the x-axis is steeper (In other words, more closer to y-axis). The lowest correlation coefficient is -0.15 of Enclosed.Porch. It has even negative value but the absolute value is not huge, so is closer to x-axis. The closest to 0.5 correlation coefficient is 0.53 of Mas.Vnr.Area. Although there are many '0' values in Mas.Vnr.Area, 0.53 means the slope that the change in Y over the change in X.

   If X changed A amount, how much Y changed. In other words, X is independent variable and Y is dependent variable. If the absolute number of slope (correlation coefficient) is higher, there might be strong correlation between independent & dependent variables.

2. Number 8 Deliverable, Interpret each coefficient of the model

   The first model which use whole continuous variable has 0.16 R-squared, and adjusted 0.7145 R-squared. R-Squared ($R^2$ or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable (CFI Team, 2022). The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model(CFI Team, 2022). 0.7145 R-squared means that 71.45% of variance in the dependent variable can be explained by the independent variable (model).

   My last model (without outliers) has 0.8033 R-squared and 0.8026 adjusted R-squared. This mean that 80.26% of variance in the dependent variable can be explained by model without outliers.

3. Number 9 Deliverable, Interpret the four graphs of plot()

   Normality (Residuals vs fitted & Normal Q-Q): Checking residual normality of each variable. Linearity (Residuals vs fitted): Checking linear relationship, we usually do this with scatter plot or other plots. Homoscedasticity (Scale-Location): Checking same variance of errors. Identifying outliers with (Residuals vs Leverage). Independence of errors (Durbin-Watson Test): Cehcking autocorrelation and distribution of errors is not correlated to the errors in prior observations (Paro, 2023)

4. Number 10 Deliverable, what steps would you take to correct multicollinearity if it exists?

   First, remove some of the highly correlated independent variables. Second, linearly combine the independent variables, such as adding them together. Third, perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression (Jim, 2022).

5. Number 11 Deliverable, should these observations be removed from the model (outliers)?

   This may or may not be removed. First, when I find outliers, I need to look at the

data. As an Analyst, if I'm not familiar with the meaning of the data, I should consult with an expert of dataset.

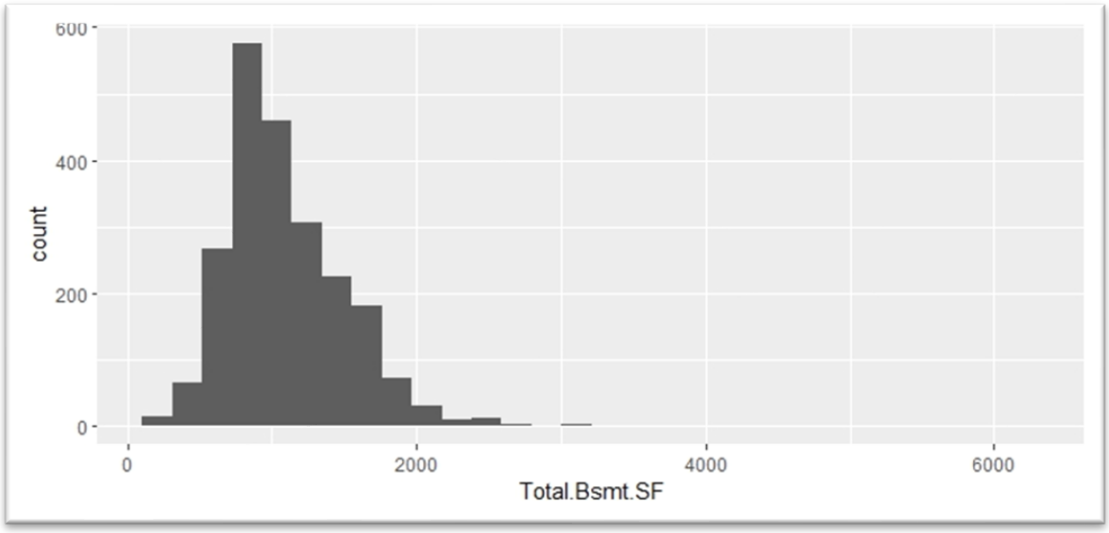
6. Number 12 Deliverable, did my changes improve the model? Why?
   In this case, the model is improved by removing the outliers. After deleting outliers in dataset, R-squared getting higher as 0.8033 from 0.755. Adjusted R-squared getting better as 0.0007 difference from 0.0009. The model with the lowest AIC offers the best fit (ZACH, 2021). AIC and BIC getting smaller as much as the difference of 486 after deleting outliers.
   By deleting the outliers among the values used to create the model, the residuals decreased and the ability to explain the dependent variable, SalePrice, increased.
7. Number 14 Deliverable, compare the preferred model from step 13 with my model from step 12. How do they differ? Which model do you prefer and why?
   I preferred the model of mine. The difference between two model is steps which make them. I start making model with understanding variables. However, all subsets regression method only uses the numbers. if there are many variables in dataset, it could be better to use all subsets regression model, because it informs the size and effect of the model considering all possibilities of various variables.
   If I have a lot of time, I will analyze the data by following the analysis learned in this module step by step. First, understanding variables. Second, checking numbers with EDA. Third, understanding variables with correlation matrix and scatter plots. Forth, testing assumption. Fifth, making model with feature selection. Sixth, checking model with AIC & BIC. Seventh, checking multicollinearity. Eighth, Checking outliers. Nineth, thinking about method to improve model.

## CONCLUSION

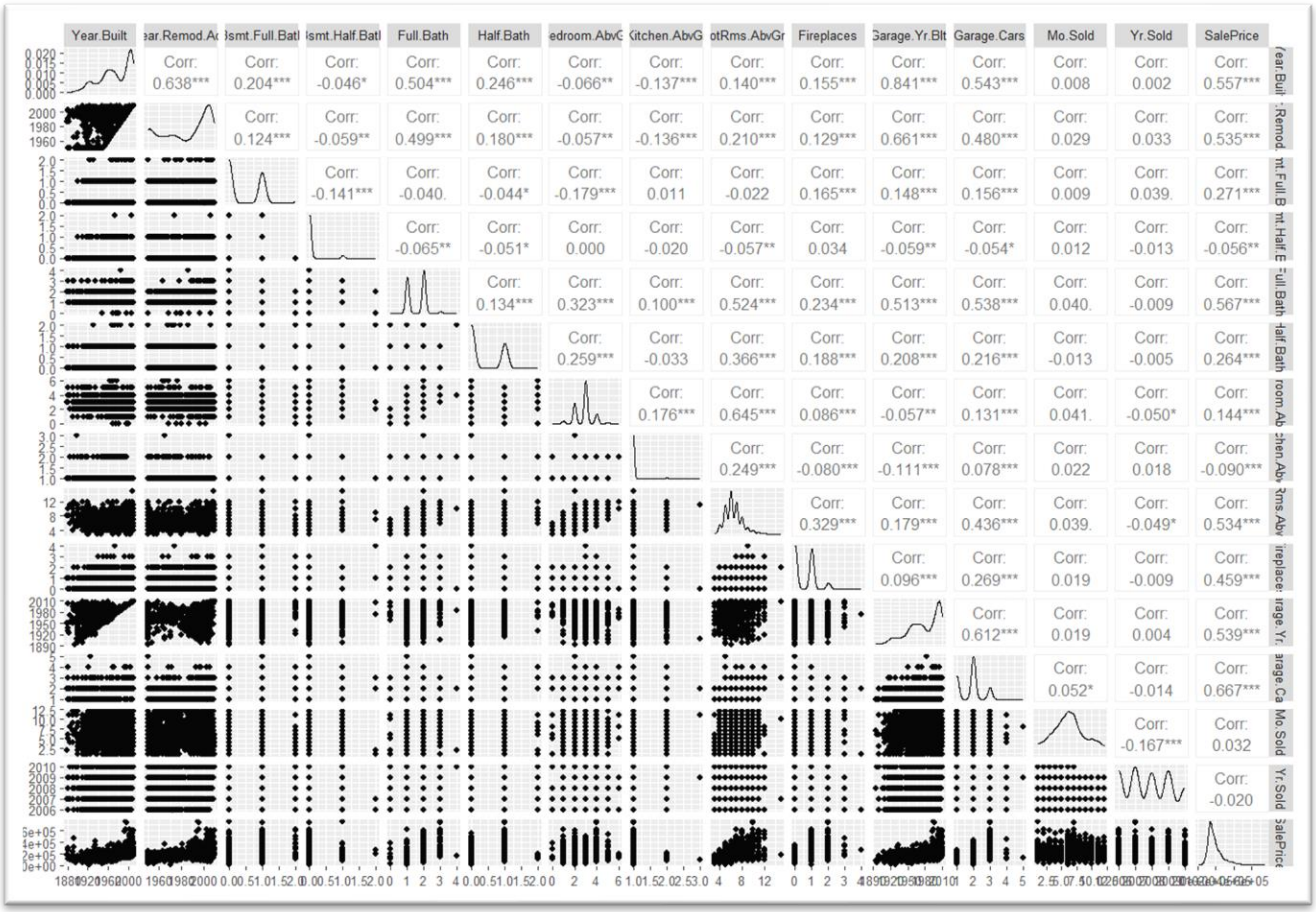
In this project, I focused on checking the assumptions and evaluating the efficiency of model with AIC & BIC. I realize making proper steps to analyze is important dealing with real world problem. I will improve my own steps and feedback process. I also learn all subsets regression method which can be used for specific situation. When time is short and there are many variables, I think using all subsets regression method is better and it can also be used for cross-checking models.

# APPENDIX

## Histogram of Gr.Liv.Area

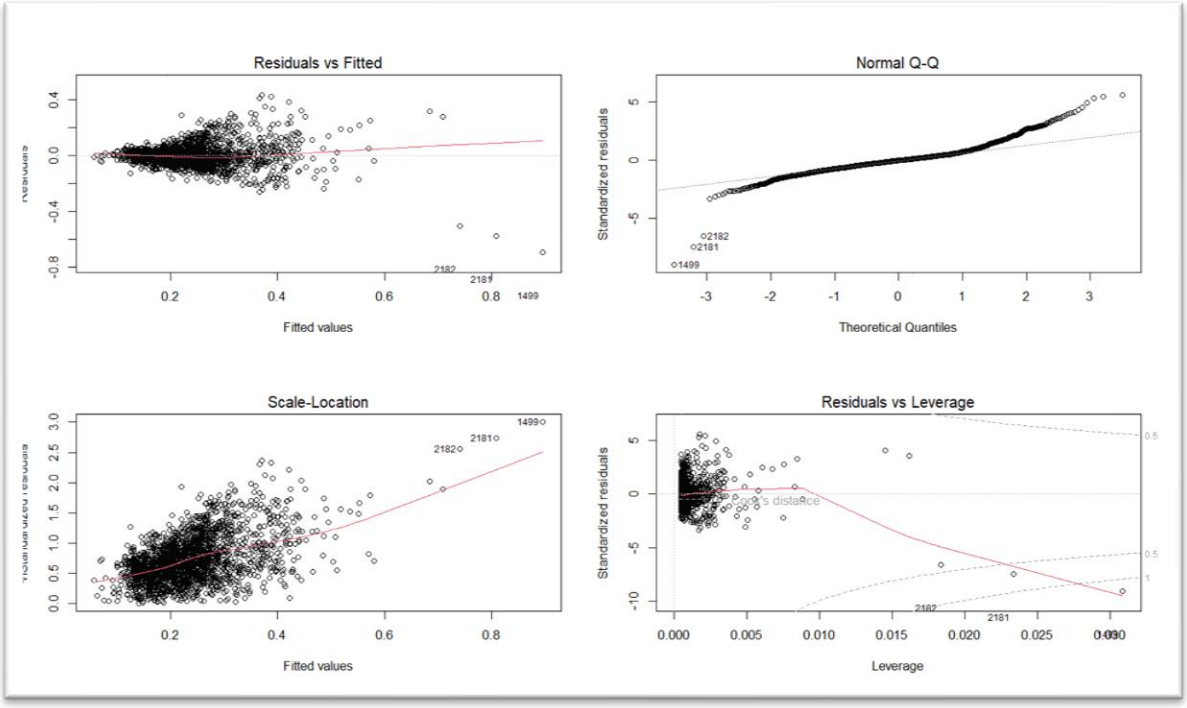



## Plot of Correlation matrix of Discrete Variables

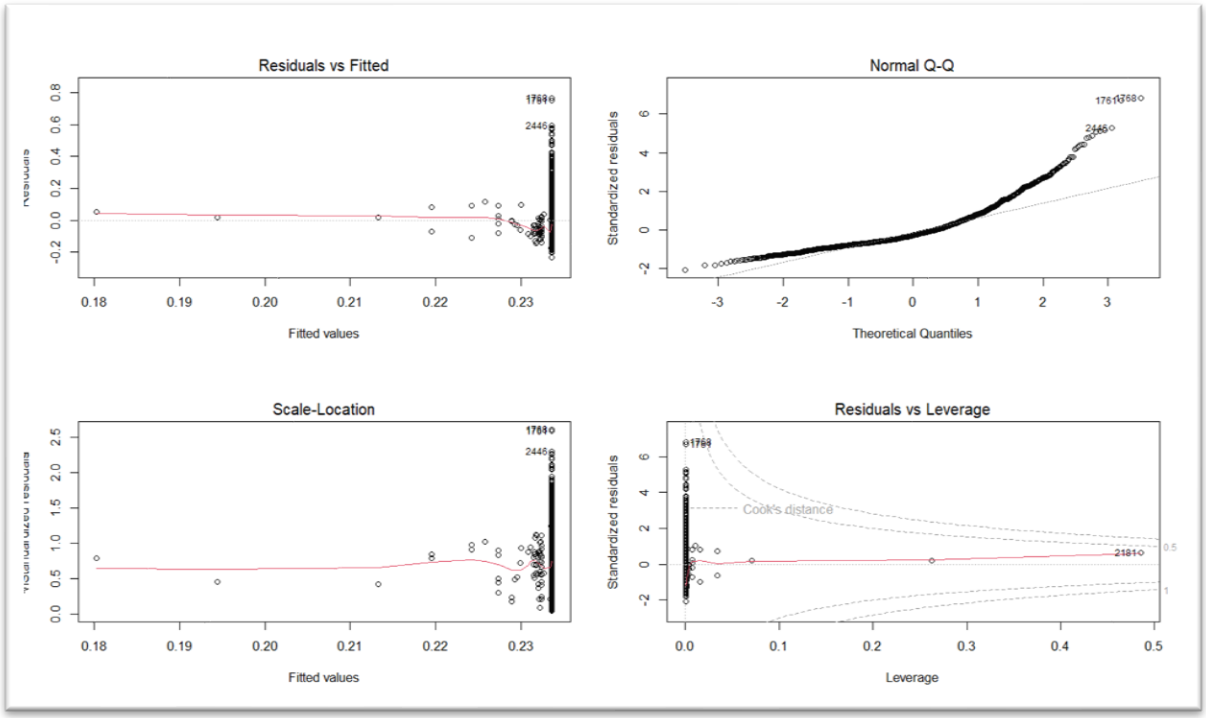


## 4 Plot of each independent variables

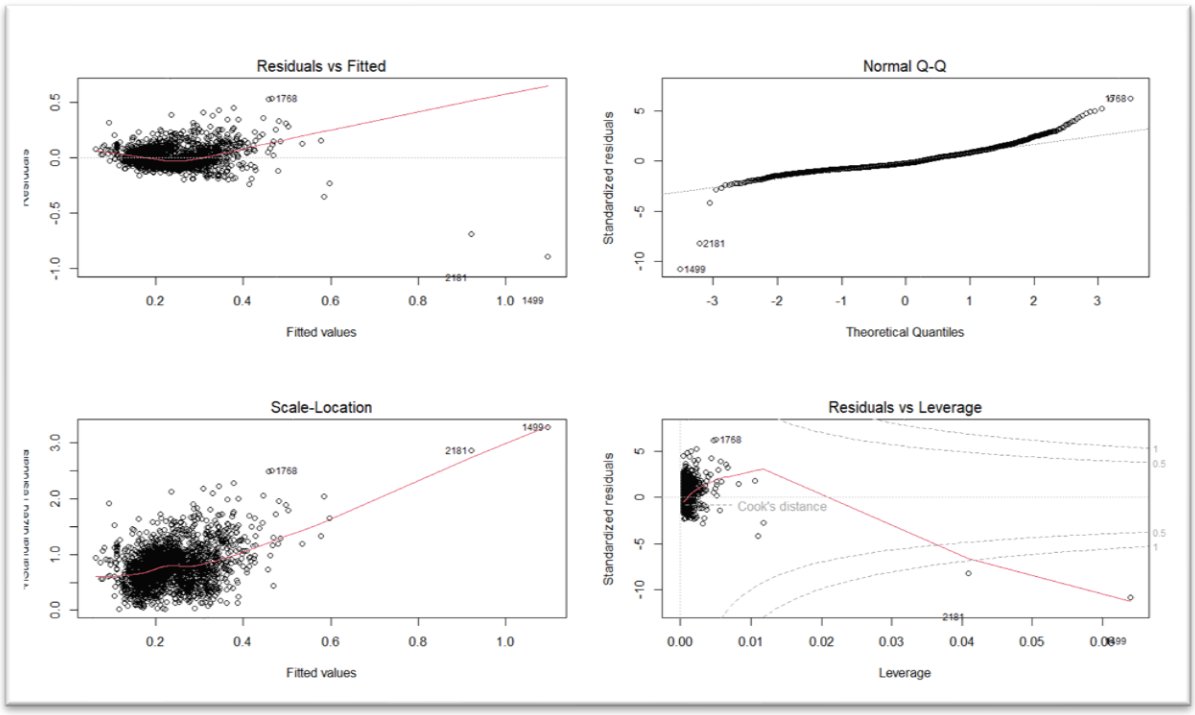
### Gr.Liv.Area



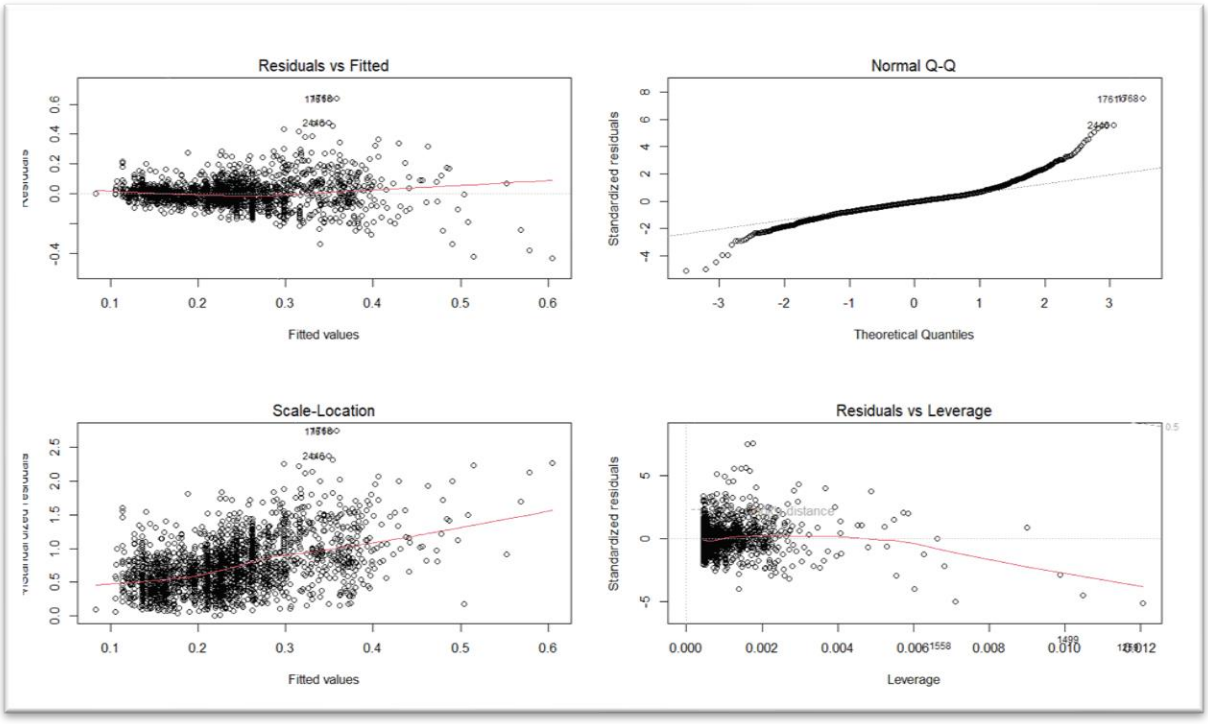

### Misc.Val



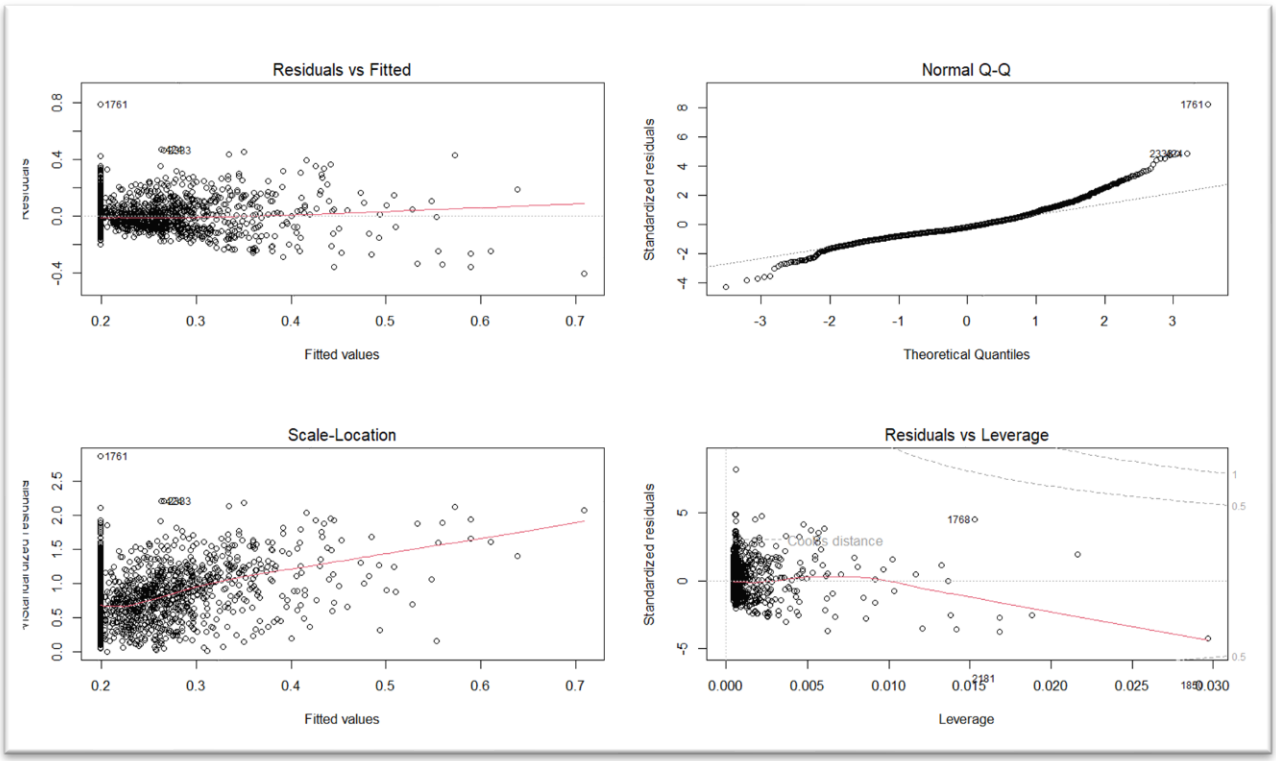
## Total.Bsmt.SF



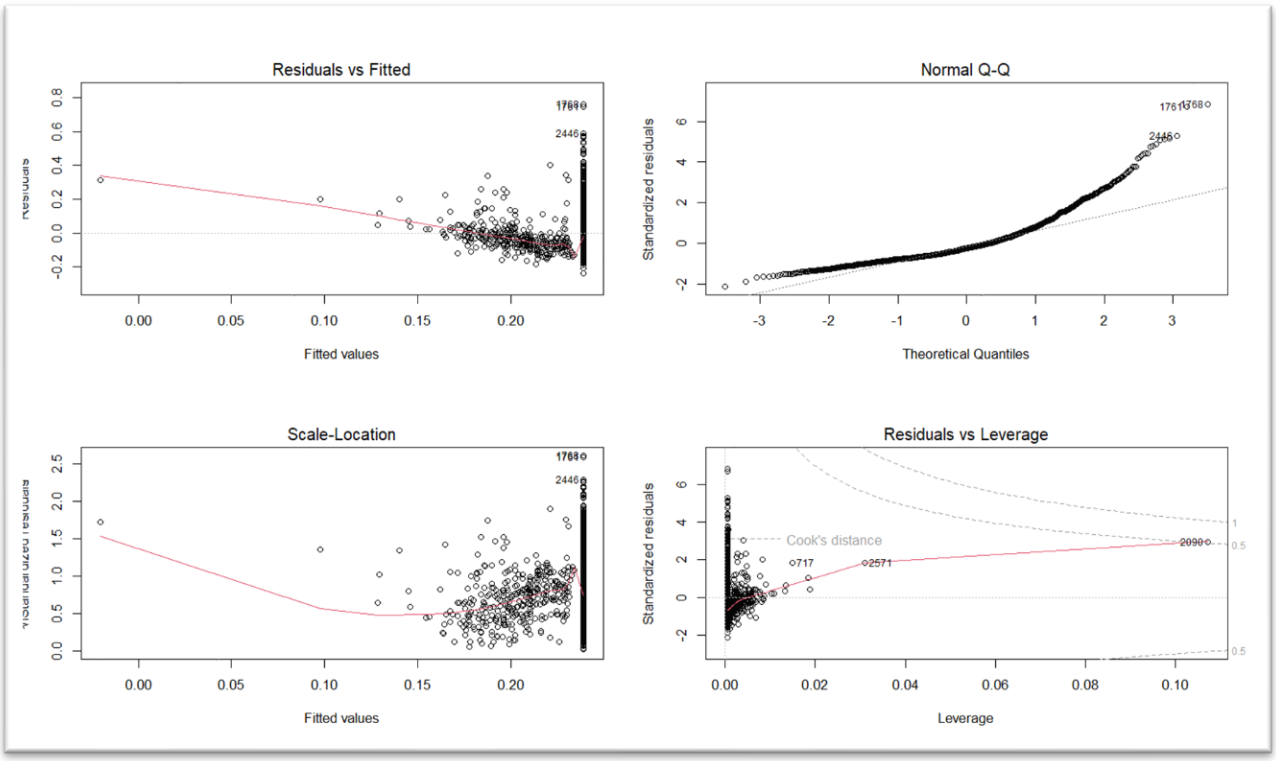

## Garage.Area



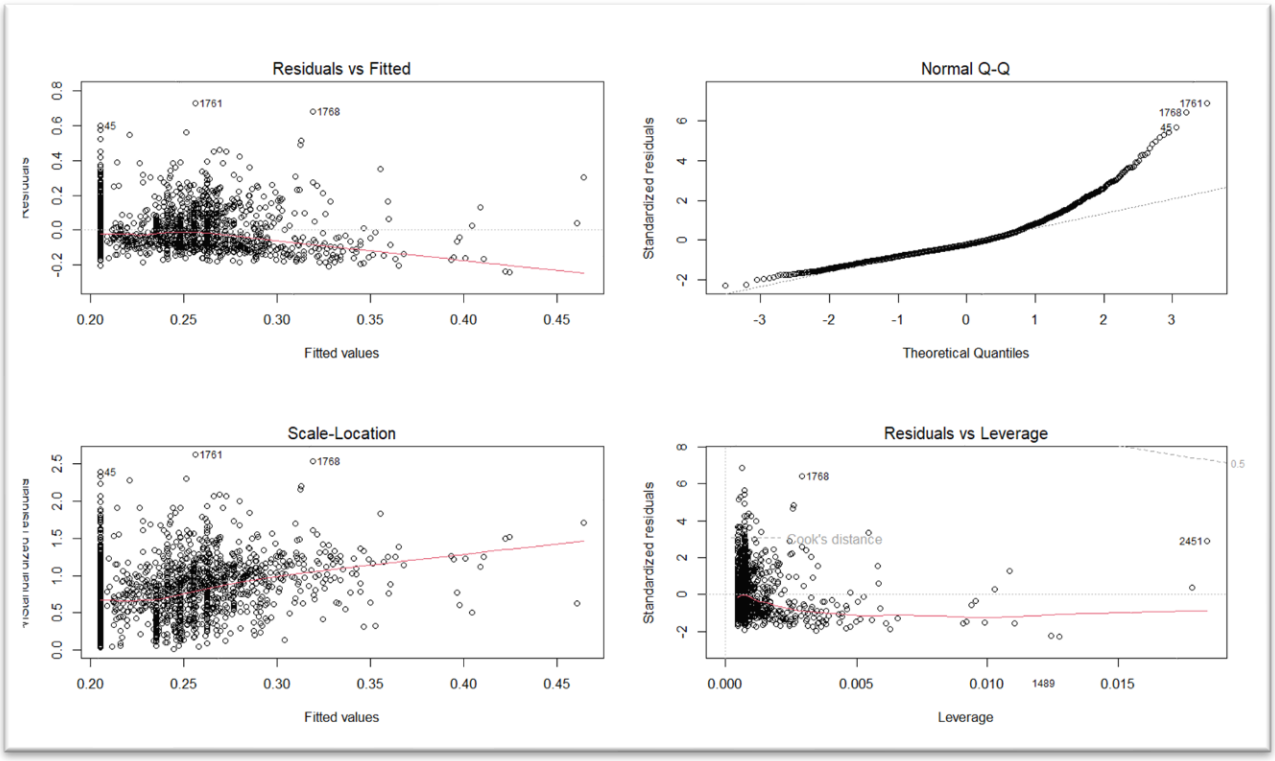
## Mas.Vnr.Area



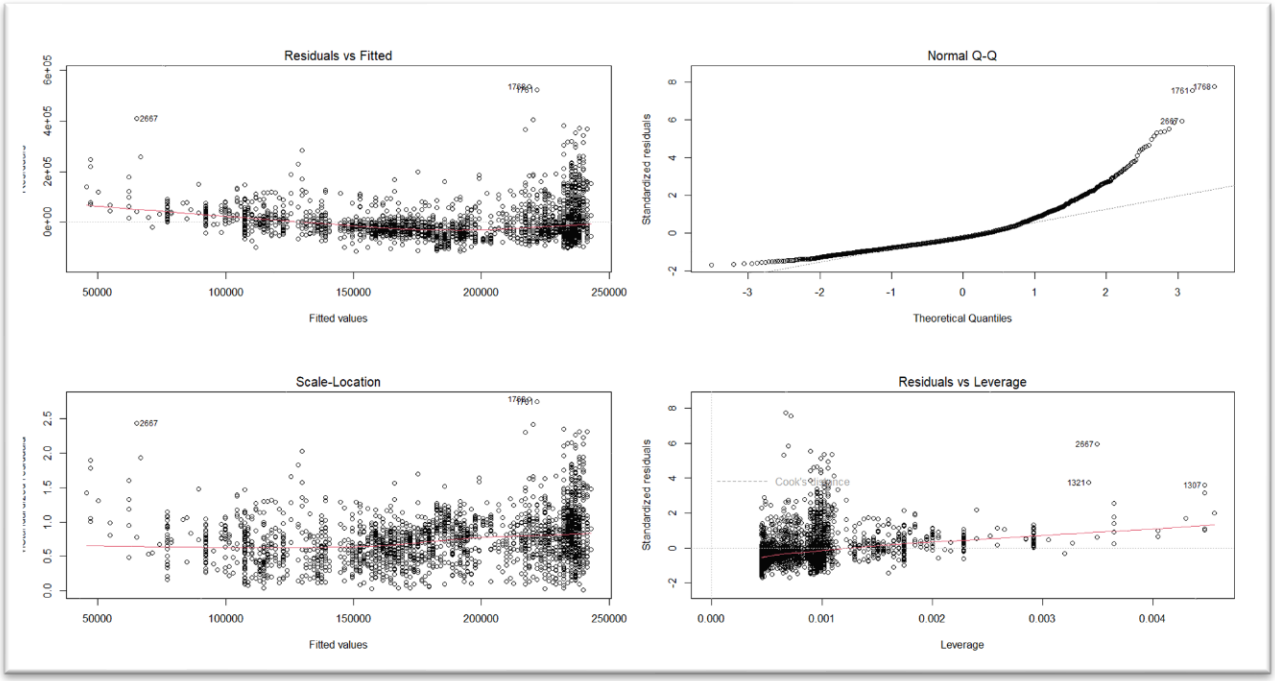

## Enclosed.Porch



## Wood.Deck.SF




## Year.Built



# REFERENCES


Kabacoff, Robert I. (2015). R in Action 2nd Edition. Manning Publisher.

Kicklighter, Clois E., Ronald J. Baird, and Joan C. Kicklighter. (1995). Architecture: Residential Drawing and Design. South Holland.

Harshita_Dudhe. (2015, August 15). How to count the missing value in R. Analytics Vidhya. Retrieved from https://discuss.analyticsvidhya.com/t/how-to-count-the-missing-value-in-r/2949

ggplot2. (n.d). Histograms and frequency polygons. Retrieved from https://ggplot2.tidyverse.org/reference/geom_histogram.html

STHDA. (n.d.). ggplot2 Quick correlation matrix heatmap - R software and data visualization. Retrieved from http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

ZACH. (2021, July 30). How to plot multiple histograms in R (with examples). STATOLOGY. Retrieved from https://www.statology.org/multiple-histograms-r/

danas.zuokas. (2012, May 30). How to write a linear model formula with 100 variables in R. Cross Validated. Retrieved from https://stats.stackexchange.com/questions/29477/how-to-write-a-linear-model-formula-with-100-variables-in-r

Peter. (2012, February 11). How to deal with an error such as "Coefficients: 14 not defined because of singularities" in R? Cross Validated. Retrieved from https://stats.stackexchange.com/questions/13465/how-to-deal-with-an-error-such-as-coefficients-14-not-defined-because-of-singu

R Documentation. (n.d.). All-subsets regression. Retrieved from https://search.r-project.org/CRAN/refmans/lmSubsets/html/lmSubsets.html

Coding Prof. (2022, May 21) 5 Ways to Check the Normality of Residuals in R [Examples]. Retrieved from https://search.r-project.org/CRAN/refmans/lmSubsets/html/lmSubsets.html

ZACH. (2020, April 2). How to Perform a Durbin-Watson Test in R. STATOLOGY. Retrieved from https://www.statology.org/durbin-watson-test-r/

Paromita Guha. (2023). ALY6015 - Intermediate Analytics. Power Point.

ZACH. (2021, May 20). What is Considered a Good AIC Value?. STATOLOGY. Retrieved from https://www.statology.org/what-is-a-good-aic-value/

Safa, Mulani. (2022, August 3). How to Normalize data in R [3 easy methods]. DigitalOcean. Retrieved from https://www.digitalocean.com/community/tutorials/normalize-data-in-r

ZACH. (2019, May 9). How to calculate variance inflation factor (VIF) in R. STATOLOGY. Retrieved from https://www.statology.org/variance-inflation-factor-r/

CFI Team. (2022, November 24). R-Squared. Retrieved from https://corporatefinanceinstitute.com/resources/data-science/r-squared/

CFI Team. (2022, December 5). Adjusted R-squared. Retrieved from https://corporatefinanceinstitute.com/resources/data-science/adjusted-r-squared/

Jim Frost. (2022). Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. Retrieved from https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

# R-Codes

```
#Install.Packages
library("psych")
library("ggplot2")
library("reshape2")
library("dplyr")
library("ggpubr")
library(tidyverse)
library(car)
library(GGally)
library(insight)
library(ggiraphExtra)
library(caret)
library(Hmisc)

library(stats)
install.packages("lmSubsets")
library(lmSubsets)


#Setwd in file direction
setwd("C:\\Users\\14083\\Desktop\\0. Winter course in 2023 Northeastern\\ALY
6015\\Module1\\dataset")

#1. Import dataset using read.csv()
h <- read.csv("AmesHousing.csv", stringsAsFactors = T,
                header=T)

#Checking dataset & data structure
headtail(h,5)
str(h)
summary(h)

#check NA value
table(h$Alley, exclude=NULL)
table(h$Fireplace.Qu, exclude=NULL)
table(h$Pool.QC, exclude=NULL)
table(h$Fence, exclude=NULL)
table(h$Misc.Feature, exclude=NULL)

#Cleaning Dataset without NA value
h <- subset(h, select = -c(Alley,Fireplace.Qu,Fence,Pool.QC,Misc.Feature))
h <- na.omit(h)
str(h)
```

```
#######
#EDA#
#######

#describe
describe(cordata3)
describe(h$Year.Built)

#SalePrice histogram
ggplot(h, aes(SalePrice)) +
  geom_histogram()

#SalePrice histogram by Year.Built
ggplot(h, aes(SalePrice, after_stat(density), colour = Paved.Drive)) +
  geom_freqpoly()

#Gr.Liv.Area histogram
ggplot(h, aes(Gr.Liv.Area)) +
  geom_histogram()

#4. cor()
#Correlation heatmap Continuous
cordata3 <- subset(h, select = c(Lot.Frontage, Lot.Area, Mas.Vnr.Area, BsmtFin.SF.1,
BsmtFin.SF.2, 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, Screen.Porch,
Pool.Area, Misc.Val, SalePrice))
cordata3
str(cordata3)

aR <- round(cor(cordata3), 2)

cor(cordata3)
aR <- round(cor(cordata3), 2)

get_upper_tri<-function(aR){
  aR[lower.tri(aR)] <- NA
  return(aR)
}

upper_tri <- get_upper_tri(aR)
upper_tri
melted_cormat <- melt(upper_tri, na.rm = TRUE)
melted_cormat

reorder_cormat <- function(aR){
  # Use correlation between variables as distance
```

```
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  aR <-aR[hc$order, hc$order]
}

aR <- reorder_cormat(aR)
aR
upper_tri <- get_upper_tri(aR)

melted_cormat <- melt(upper_tri, na.rm = TRUE)

#NEW
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Rating\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()

ggheatmap +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))

#Correlation heatmap Discrete
cordata4 <- subset(h, select = c(Year.Built, Year.Remod.Add, Bsmt.Full.Bath, Bsmt.Half.Bath,
Full.Bath, Half.Bath, Bedroom.AbvGr, Kitchen.AbvGr, TotRms.AbvGrd, Fireplaces,
Garage.Yr.Blt, Garage.Cars, Mo.Sold, Yr.Sold, SalePrice))
cordata4
str(cordata4)

aR <- round(cor(cordata4), 2)
```

```
cor(cordata4)
aR <- round(cor(cordata4), 2)

get_upper_tri<-function(aR){
  aR[lower.tri(aR)] <- NA
  return(aR)
}

upper_tri <- get_upper_tri(aR)
upper_tri
melted_cormat <- melt(upper_tri, na.rm = TRUE)
melted_cormat

reorder_cormat <- function(aR){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  aR <-aR[hc$order, hc$order]
}

aR <- reorder_cormat(aR)
aR
upper_tri <- get_upper_tri(aR)

melted_cormat <- melt(upper_tri, na.rm = TRUE)

#NEW
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Rating\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed()

ggheatmap +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
```

```
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))

#5. correlation matrix()
cordata1 <- subset(h, select = c(Lot.Frontage, Lot.Area, Mas.Vnr.Area, BsmtFin.SF.1,
BsmtFin.SF.2, Bsmt.Unf.SF, Total.Bsmt.SF, X1st.Flr.SF, X2nd.Flr.SF, SalePrice))

cordata2 <- subset(h, select = c(Low.Qual.Fin.SF, Gr.Liv.Area, Garage.Area, Wood.Deck.SF,
Open.Porch.SF, Enclosed.Porch, Screen.Porch, Pool.Area, Misc.Val, SalePrice))

ggpairs(cordata1)
ggpairs(cordata2)
ggpairs(cordata4)

#6. Scatterplot highest/lowest/closest to 0.5
#scatter plot
ggscatter(h, x = "Gr.Liv.Area", y = "SalePrice",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Above ground living area (ft.)", ylab = "SalePrice")

ggscatter(h, x = "Enclosed.Porch", y = "SalePrice",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Enclosed porch area (ft.)", ylab = "SalePrice")

ggscatter(h, x = "Mas.Vnr.Area", y = "SalePrice",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Masonry veneer area (ft.)", ylab = "SalePrice")

# 7. making model
# 7-1. normalization
# normalization
summary(cordata3)
process <- preProcess(as.data.frame(cordata3), method=c("range"))
summary(process)

norm_scale <- predict(process, as.data.frame(cordata3))
norm_scale

#model method=
reg_mod1 <- lm(SalePrice ~., data = norm_scale)
```

```
summary(reg_mod1)

#check alias
alias(reg_mod1)

#fix singularity
norm_scale2 <- subset(norm_scale, select = -
c(X1st.Flr.SF,X2nd.Flr.SF,Low.Qual.Fin.SF,BsmtFin.SF.1,BsmtFin.SF.2,Bsmt.Unf.SF))
reg_mod2 <- lm(SalePrice ~., data = norm_scale2)
summary(reg_mod2)

#Model for checking OLS
#use plot()
par(mfrow=c(2,2))

for_mod1 <- lm(SalePrice ~ Gr.Liv.Area, data = norm_scale2)
plot(for_mod1)

for_mod2 <- lm(SalePrice ~ Misc.Val, data = norm_scale2)
plot(for_mod2)

for_mod3 <- lm(SalePrice ~ Total.Bsmt.SF, data = norm_scale2)
plot(for_mod3)

for_mod4 <- lm(SalePrice ~ Garage.Area, data = norm_scale2)
plot(for_mod4)

for_mod5 <- lm(SalePrice ~ Mas.Vnr.Area, data = norm_scale2)
plot(for_mod5)

for_mod6 <- lm(SalePrice ~ Enclosed.Porch , data = norm_scale2)
plot(for_mod6)

for_mod7 <- lm(SalePrice ~ Wood.Deck.SF, data = norm_scale2)
plot(for_mod7)

for_mod8 <- lm(SalePrice ~ Screen.Porch, data = norm_scale2)
plot(for_mod8)

for_mod9 <- lm(SalePrice ~ Year.Built, data = h2)
plot(for_mod9)

#Durbin-Watson Test
durbinWatsonTest(for_mod1)

#use plot()
```

```
par(mfrow=c(2,2))
plot(for7)
dev.off()

#my model
fit_mod <-
lm(SalePrice~Gr.Liv.Area+Misc.Val+Total.Bsmt.SF+Garage.Area+Mas.Vnr.Area+Enclosed.Porc
h+Wood.Deck.SF+Screen.Porch, data=h)
summary(fit_mod)

AIC(fit_mod)
BIC(fit_mod)

fit_mod2 <-
lm(SalePrice~Gr.Liv.Area+Misc.Val+Total.Bsmt.SF+Garage.Area+Mas.Vnr.Area+Enclosed.Porc
h+Wood.Deck.SF+Screen.Porch+Year.Built, data=h)
summary(fit_mod2)

AIC(fit_mod2)
BIC(fit_mod2)

fit_mod3 <-
lm(SalePrice~Gr.Liv.Area+Misc.Val+Total.Bsmt.SF+Garage.Area+Mas.Vnr.Area+Wood.Deck.SF
+Screen.Porch+Year.Built, data=h)
summary(fit_mod3)

AIC(fit_mod3)
BIC(fit_mod3)

#plot of 4
par(mfrow=c(2,2))
plot(fit_mod3)
dev.off()

##Checking VIF
summary(fit_mod3)
vif_fit3 <- vif(fit_mod3)
vif_fit3

##Checking multicolinearity
#create horizontal bar chart to display each VIF value
opar <- par(no.readonly = TRUE)
par(fig=c(0.2, 1, 0, 1))
barplot(vif_fit3, main = "VIF Values", horiz = TRUE, las=2, cex.names=0.8, xlim=c(0,5), col =
"steelblue")
```

```
#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)

##Checking VIF again
vif_all <- vif(reg_mod1)
vif_all

##checking alias . What if ?
alias(reg_mod1)

#create horizontal bar chart to display each VIF value
opar <- par(no.readonly = TRUE)
par(fig=c(0.2, 1, 0, 1))
barplot(vif_all, main = "VIF Values", horiz = TRUE, las=2, cex.names=0.8, xlim=c(0,5), col =
"steelblue")

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)

#checking outliers with plot
plot(fit_mod3)

#Deleting outliers
par(mfrow=c(2,2))
plot(fit_mod3)
str(h)

## checking outliers
check1 <- subset(h, Order == 3,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check1

check2 <- subset(h, Order == 1499,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check2

check3 <- subset(h, Order == 2181,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check3

check4 <- subset(h, Order == 2182,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
```

```
check4

#comparing with others
check5 <- subset(h, Order == 4,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check5

check6 <- subset(h, Order == 5,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check6

check7 <- subset(h, Order == 6,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check7

check8 <- subset(h, Order == 7,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check8

check9 <- subset(h, Order == 8,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check9

check10 <- subset(h, Order == 1000,
                 select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check10

#checking tables
table(h$Mas.Vnr.Area, exclude=NULL)
table(h$Wood.Deck.SF, exclude=NULL)

#checking outliers status
check11 <- subset(h, Mas.Vnr.Area != 0 & Mas.Vnr.Area != 0,
                  select = c("Gr.Liv.Area", "Misc.Val", "Total.Bsmt.SF", "Garage.Area", "Mas.Vnr.Area",
"Wood.Deck.SF", "Screen.Porch", "Year.Built"))
check11

#Deleting outliers again2
str(h)
h2 <- h[ !(h$Order %in% c(3, 1499, 2181, 2182)), ]
```

```
str(h2)

#Making model without outliers
fit_mod4 <-
lm(SalePrice~Gr.Liv.Area+Misc.Val+Total.Bsmt.SF+Garage.Area+Mas.Vnr.Area+Wood.Deck.SF
+Screen.Porch+Year.Built, data=h2)
summary(fit_mod4)

AIC(fit_mod4)
BIC(fit_mod4)


par(mfrow=c(2,2))
plot(fit_mod4)
dev.off()

#all subset regression
lm_all <- lmSubsets(SalePrice ~ ., data=norm_scale)
lm_all
summary(lm_all)

#making model with SIZE 8
lm_all8 <-
lm(SalePrice~Gr.Liv.Area+Total.Bsmt.SF+Garage.Area+Mas.Vnr.Area+Misc.Val+BsmtFin.SF.1+
Wood.Deck.SF,data=h2)
summary(lm_all8)

#Deleting Misc.Val
lm_all9 <-
lm(SalePrice~Gr.Liv.Area+Total.Bsmt.SF+Garage.Area+Mas.Vnr.Area+BsmtFin.SF.1+Wood.De
ck.SF,data=h2)
summary(lm_all9)

##Checking VIF with lmSUb
vif_all2 <- vif(lm_all9)
vif_all2

#create horizontal bar chart to display each VIF value
opar <- par(no.readonly = TRUE)
par(fig=c(0.2, 1, 0, 1))
barplot(vif_all2, main = "VIF Values", horiz = TRUE, las=2, cex.names=0.8, xlim=c(0,5), col =
"steelblue")

#add vertical line at 5
abline(v = 5, lwd = 3, lty = 2)
```

```
#adding Year.Built
lm_all10 <-
lm(SalePrice~Gr.Liv.Area+Total.Bsmt.SF+Garage.Area+Mas.Vnr.Area+BsmtFin.SF.1+Wood.De
ck.SF+Year.Built,data=h2)
summary(lm_all10)

fit_mod3 <-
lm(SalePrice~Gr.Liv.Area+Misc.Val+Total.Bsmt.SF+Garage.Area+Mas.Vnr.Area+Wood.Deck.SF
+Screen.Porch+Year.Built, data=h)
summary(fit_mod3)

##APPENDIX
#Total.Bsmt.SF histogram
ggplot(h, aes(Total.Bsmt.SF)) +
  geom_histogram()
```