Regression Diagnostics with R

Ames Housing Project – OLS Approach

Kieu Van Dang Email: dang.v@northeastern.edu

LinkedIn: https://www.linkedin.com/in/kirudang/ Phone: +1 647 782 4558

Data science enthusiast with a great passion for data pre-processing and prediction with a strong background in business. Relevant skills include machine learning, statistics, problem-solving, programming (including SQL, Python, and R), and critical & creative thinking.

INDEX

A. Introduction

Data introduction: The set comprises data from the Ames Assessor's Office that was used to compute assessed values for individual residential properties sold in Ames, Iowa between 2006 and 2010. As it features 82 variables, including 23 nominal, 23 ordinal, 14 discrete, 20 continuous columns, and 2 extra observation identifiers. The data set has a varied range of predictors, which relates to the fact that one prediction of outcome is dependent on a large number of inputs. These explanatory variables describe almost every aspect of residential homes in Ames, Iowa intending to predict the selling price of each home, ranging across 2930 observations. Because of the large number of variables, this data set is a strong candidate for our goal in this study.

Project objective: Since the type of information contained in the data set is similar to what a typical home buyer would want to know before making a purchase and since most variables are straightforward and understandable, this report is to investigate several useful explanatory variables to predict the selling price using regression models. The study also compares the effectiveness of various modeling approaches for linear regression deploying methods learned in Module 1 of the course.

Incorporated Methods: Within the scope of this report, data scrubbing includes checking missing values, removing duplications, dropping unnecessary columns, converting to proper data types, and manipulating data for analysis and investigations. The manipulation also incorporates with calculating new data points, dealing with omission, and so on.

In terms of statistical modeling, the report employs diversified techniques to develop the most usable regression. The process starts with explanatory data analysis to grab a good sense of the set and then moves towards variable selection. After an initial line is fitted, it progresses to underlying assumptions for Ordinary Least Square regression to check the problems and detect outliers and unusual observations. Finally, corrective measures are applied to further improve the regression line.

B. Analysis

1. Initial data cleaning and selection

Prior to any analysis, it is always crucial to have data cleaning and manipulation to make them ready for effective investigation. Since the number of independent variables is huge (82 columns), retaining all of them for the model appears unnecessary. Furthermore, within the scope of this report which is to demonstrate several techniques learned for linear regression, only several independent variables are sufficient. My rationale to decide whether to keep a particular column or not and hence narrow down predictors is primarily based on my intuition and previous domain knowledge of real estate. Of course, in fact, house buyers will consider multiple factors as it is costly to buy a house, but the price of a house depends mostly on: Location, total area, age, condition of the structure, numbers of rooms, particularly bedrooms and bathrooms, and other miscellaneous values. Other methodologies to clean data are removing duplicating rows or eliminating columns with an insignificant proportion of missing values. Another important note is that categorical data with many values will impose a heavy burden on regression work by generating too many corresponding columns for dummy data. As a result, these categorical regardless of nominal or ordinal type should be removed from the set whenever possible.

Firstly, I examine the number of missing values to facilitate our coding later on in this report by generating a map to locate the available missing values by deploying the missmap() function from Amelia package. Then I have a double check by computing total missing values, using sum(is.na(x)) of sapply. As shown in Figure 1, there are total 6% of missing values in the data set, falling into some variables. Also referring to Figure 2, columns have a huge numbers of N/A count such as **Alley (Nominal)**: Type of alley access to property, **Pool QC** (Ordinal): Pool quality, **Fence (Ordinal)**: Fence quality or **Misc Feature (Nominal)**: Miscellaneous feature not covered in other categories. This result also reflects well the fact, in which, for example, not all properties have a pool, hence missing values for Pool are apparent. I drop these columns at the end of this step.

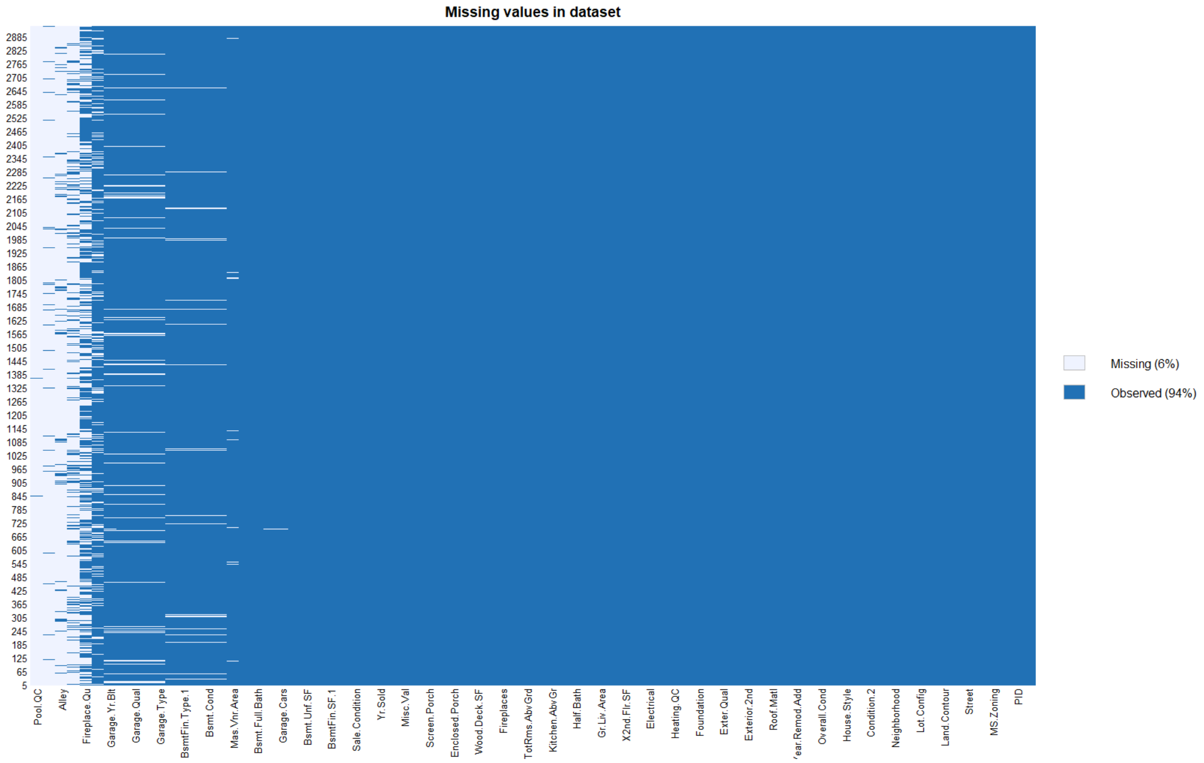
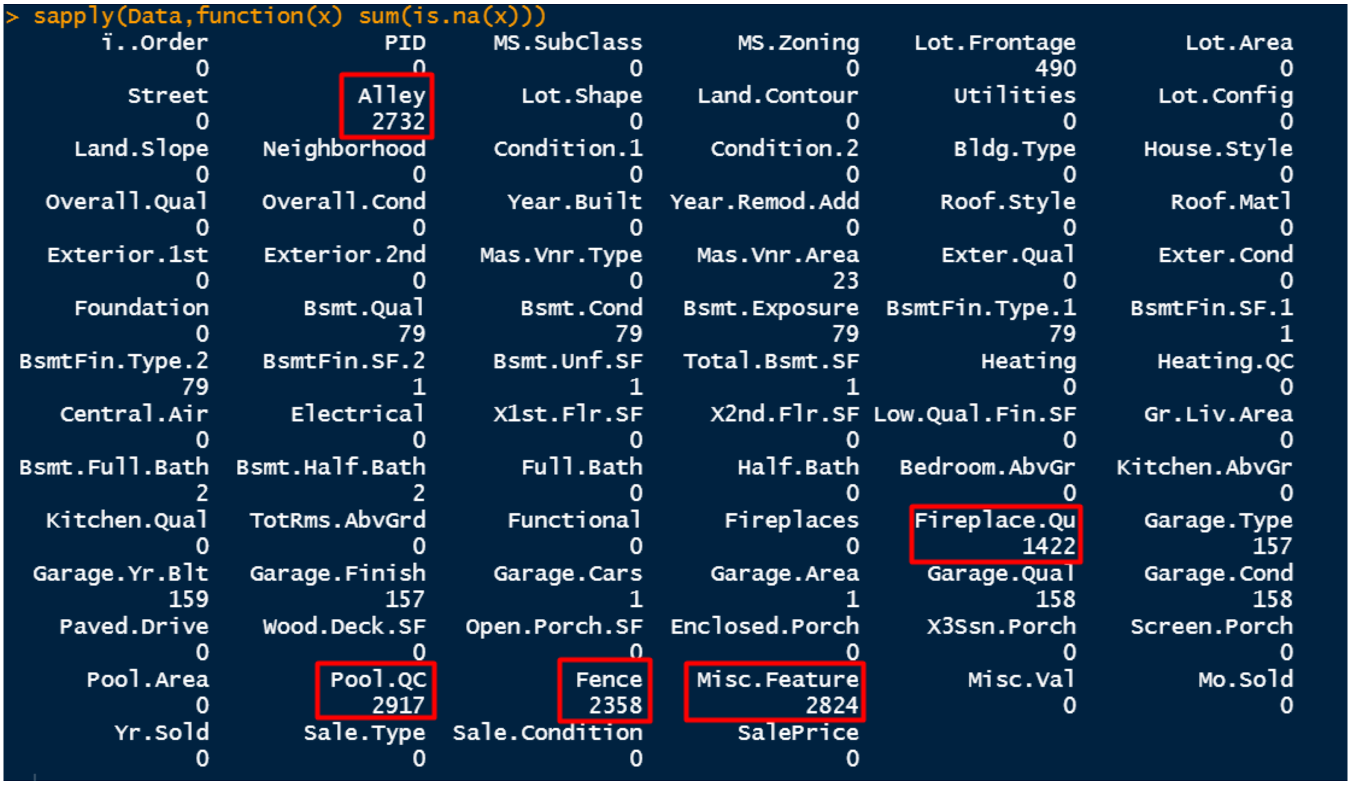


Figure 1: *Map of missing values*



```
> sapply(Data,function(x) sum(is.na(x)))
      ï..Order           PID    MS.SubClass      MS.Zoning   Lot.Frontage       Lot.Area
             0             0              0              0            490              0
        Street         Alley      Lot.Shape   Land.Contour      Utilities     Lot.Config
             0          2732              0              0              0              0
    Land.Slope  Neighborhood    Condition.1    Condition.2      Bldg.Type    House.Style
             0             0              0              0              0              0
  Overall.Qual  Overall.Cond     Year.Built Year.Remod.Add     Roof.Style      Roof.Matl
             0             0              0              0              0              0
  Exterior.1st  Exterior.2nd   Mas.Vnr.Type   Mas.Vnr.Area     Exter.Qual     Exter.Cond
             0             0              0             23              0              0
    Foundation     Bsmt.Qual      Bsmt.Cond  Bsmt.Exposure BsmtFin.Type.1   BsmtFin.SF.1
             0            79             79             79             79              1
BsmtFin.Type.2   BsmtFin.SF.2    Bsmt.Unf.SF  Total.Bsmt.SF        Heating     Heating.QC
            79             1              1              1              0              0
   Central.Air    Electrical     X1st.Flr.SF    X2nd.Flr.SF Low.Qual.Fin.SF    Gr.Liv.Area
             0             0              0              0              0              0
Bsmt.Full.Bath Bsmt.Half.Bath      Full.Bath      Half.Bath  Bedroom.AbvGr   Kitchen.AbvGr
             2             2              0              0              0              0
  Kitchen.Qual  TotRms.AbvGrd     Functional     Fireplaces   Fireplace.Qu    Garage.Type
             0             0              0              0           1422            157
 Garage.Yr.Blt  Garage.Finish    Garage.Cars    Garage.Area    Garage.Qual    Garage.Cond
           159            157              1              1            158            158
   Paved.Drive   Wood.Deck.SF  Open.Porch.SF Enclosed.Porch    X3Ssn.Porch   Screen.Porch
             0             0              0              0              0              0
     Pool.Area       Pool.QC          Fence   Misc.Feature       Misc.Val        Mo.Sold
             0          2917           2358           2824              0              0
       Yr.Sold     Sale.Type Sale.Condition      SalePrice
             0             0              0              0
```

Figure 2: *Sum of N.A value*

Next, I conduct a process of removing duplications as these observations will distort the view toward data. The set remains total of 2930 rows spanning across 77 variables up to this stage.

With respect to our inital assumption about imperative factors affecting house price, we take a look at the age of the house first. For time series data, we have three important variables: **Year Built (Discrete)**: Original construction date, **Year Remod/Add (Discrete)**: Remodel date (same as construction date if no remodeling or additions) and **Yr Sold (Discrete)**: Year Sold (YYYY). In this case, I calculate the age of the house by creating a new column called "**Age**" equal to the subtraction of the year sold for the year of remodel. It is more rational to have a subtrahend of this manipulation is **Year Remod/Add** instead of **Year Built.**

In terms of the total property area, here are the list of predictors with the most explanatory power: **Lot Frontage (Continuous):** Linear feet of street connected to property, **Lot Area (Continuous):** Lot size in square feet**, Total Bsmt SF (Continuous):** Total square feet of basement area, **Gr Liv Area (Continuous):** Above grade (ground) living area square feet, **Garage Area (Continuous)**: Size of garage in square feet.

As per the condition of structures, **Overall Qual (Ordinal):** Rates the overall material and finish of the house, **Overall Cond (Ordinal)**: Rates the overall condition of the house are two important ordinal variables. These are already decoded to numerical values ranging from 01: very poor to 10: excellent in the original set. I keep these two for the analysis without transforming it to an appropriate format to preserve the authenticity of data.

In addition, when it comes to the number of rooms that have a great influence on housing price, bedrooms and bathrooms seem to be significant among all. However, I still pick **Full Bath (Discrete)**: Full bathrooms above grade, **Bedroom (Discrete):** Bedrooms above grade (does NOT include basement bedrooms), **Kitchen (Discrete)**: Kitchens above grade, **TotRmsAbvGrd (Discrete):** Total rooms above grade (does not include bathrooms), **Fireplaces (Discrete):** Number of fireplaces, and let check if my intuitive prediction is correct or not in the final model.

Lastly, for the location, this is a very important factor yet challenging to define. It is noticeable that there are two main variables within the data set for the place of houses, but they compose of a high number of values.

Variable	Data type	Definition	Number of values	Sample values
MS Zoning	Nominal	Identifies the general zoning classification of the sale	7	**A**: Agriculture **C**: Commercial **FV**: Floating Village Residential **I**: Industrial …
Neighborhood	Nominal	Physical locations within Ames city limits	28	**Blmngtn**: Bloomington Heights **Blueste**: Bluestem **BrDale**: Briardale **BrkSide**: Brookside …

Table 1: *Location illustration*

I intended to put this selection at the last stage because this process is the most challenging one for me. For example, there were 28 neighborhoods listed in the "Neighborhood" variable, but it is unclear how to rank and transform them to numerical data. We could actually impute these neighborhoods based on their price per square foot and use this as a metric of location desirability because prime locations often have higher prices per area and vice versa. But this concept would naturally be correlated to house area and then cause multicollinearity issue in our regression model. Another simple solution to tackle this problem is to dummy the data values and bring them to our model. However, this method also leads to hard work as stated above because it imposes a heavy burden on manual data decoding; Furthermore, dummy data may also add multicollinearity to the model as well. A closer look to price distribution by location of boxplot as seen in Figure 3, it remains confused to the question how to transform this variable.

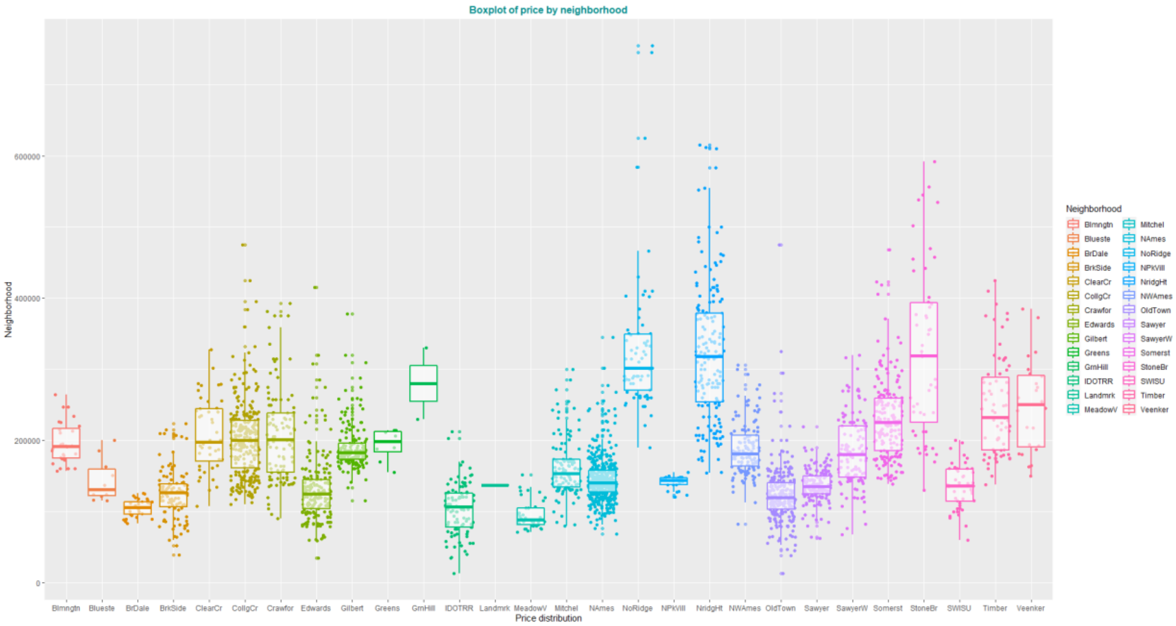


Figure 3: *Boxplots of Price by Neighborhood*

In this case, within the scope of report and due to time resource constraint, I choose to quickly convert **MS Zoning** into dummy variables by employing ‘fastDummies’ library with dummy_cols() function. But before doing so, I remove all punctuation symbols such as () and white space in the column using a default gsub function in R.

There are more procedures needed to manipulate data for further analysis, but for this initial cleaning part, the process ends here and below are our pre-final data set for next part of investigation.

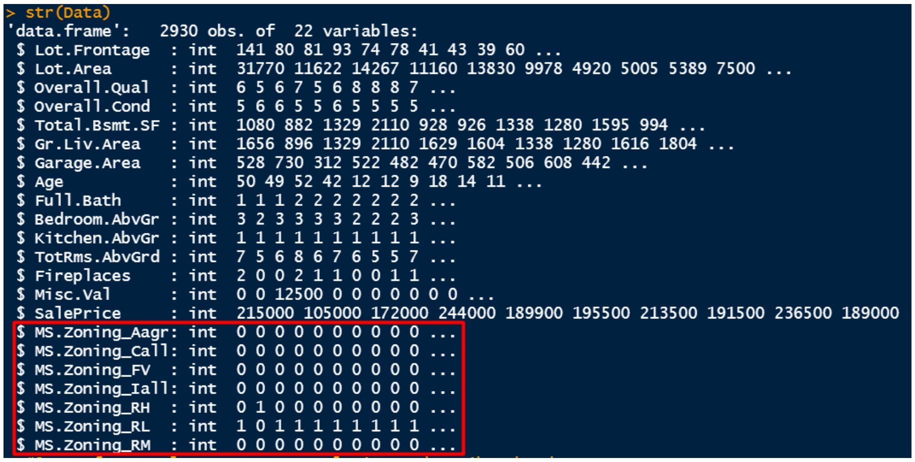

```
> str(Data)
'data.frame':	2930 obs. of  22 variables:
 $ Lot.Frontage  : int  141 80 81 93 74 78 41 43 39 60 ...
 $ Lot.Area      : int  31770 11622 14267 11160 13830 9978 4920 5005 5389 7500 ...
 $ Overall.Qual  : int  6 5 6 7 5 6 8 8 8 7 ...
 $ Overall.Cond  : int  5 6 6 5 5 6 5 5 5 5 ...
 $ Total.Bsmt.SF : int  1080 882 1329 2110 928 926 1338 1280 1595 994 ...
 $ Gr.Liv.Area   : int  1656 896 1329 2110 1629 1604 1338 1280 1616 1804 ...
 $ Garage.Area   : int  528 730 312 522 482 470 582 506 608 442 ...
 $ Age           : int  50 49 52 42 12 12 9 18 14 11 ...
 $ Full.Bath     : int  1 1 1 2 2 2 2 2 2 2 ...
 $ Bedroom.AbvGr : int  3 2 3 3 3 3 2 2 2 3 ...
 $ Kitchen.AbvGr : int  1 1 1 1 1 1 1 1 1 1 ...
 $ TotRms.AbvGrd : int  7 5 6 8 6 7 6 5 5 7 ...
 $ Fireplaces    : int  2 0 0 2 1 1 0 0 1 1 ...
 $ Misc.Val      : int  0 0 12500 0 0 0 0 0 0 0 ...
 $ SalePrice     : int  215000 105000 172000 244000 189900 195500 213500 191500 236500 189000
 $ MS.Zoning_Aagr: int  0 0 0 0 0 0 0 0 0 0 ...
 $ MS.Zoning_Call: int  0 0 0 0 0 0 0 0 0 0 ...
 $ MS.Zoning_FV  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ MS.Zoning_Iall: int  0 0 0 0 0 0 0 0 0 0 ...
 $ MS.Zoning_RH  : int  0 1 0 0 0 0 0 0 0 0 ...
 $ MS.Zoning_RL  : int  1 0 1 1 1 1 1 1 1 1 ...
 $ MS.Zoning_RM  : int  0 0 0 0 0 0 0 0 0 0 ...
```

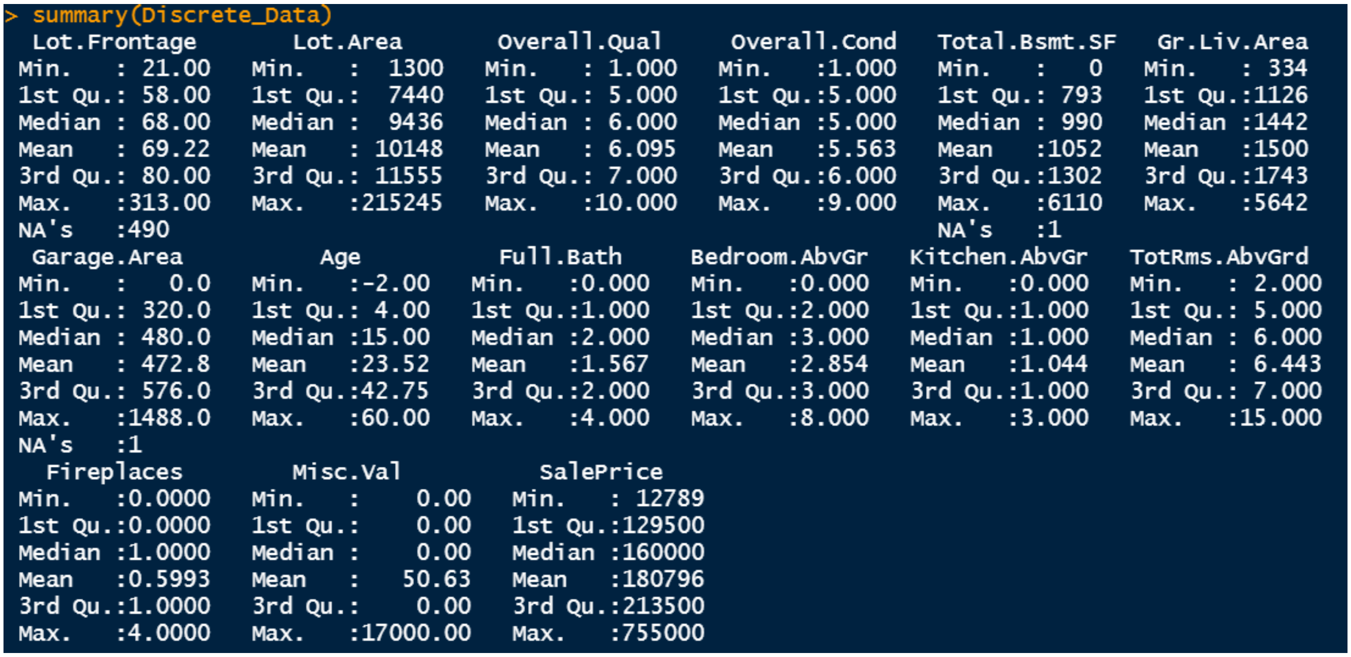
Figure 4: *Structure of pre-final data*

In conclusion, our data are now scaling down to 22 variables in which 6 of them are dummy data, and cross 2930 observations.

2. Explanatory Data Analysis

a. Data are still unclean

After conducting fundamental data cleaning and manipulation, we move forwards to descriptive statistics to describe data.



```
> summary(Discrete_Data)
  Lot.Frontage       Lot.Area       Overall.Qual     Overall.Cond   Total.Bsmt.SF   Gr.Liv.Area
 Min.   : 21.00   Min.   :  1300   Min.   : 1.000   Min.   :1.000   Min.   :   0   Min.   : 334
 1st Qu.: 58.00   1st Qu.:  7440   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.: 793   1st Qu.:1126
 Median : 68.00   Median :  9436   Median : 6.000   Median :5.000   Median : 990   Median :1442
 Mean   : 69.22   Mean   : 10148   Mean   : 6.095   Mean   :5.563   Mean   :1052   Mean   :1500
 3rd Qu.: 80.00   3rd Qu.: 11555   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:1302   3rd Qu.:1743
 Max.   :313.00   Max.   :215245   Max.   :10.000   Max.   :9.000   Max.   :6110   Max.   :5642
 NA's   :490                                                        NA's   :1
  Garage.Area          Age          Full.Bath      Bedroom.AbvGr   Kitchen.AbvGr   TotRms.AbvGrd
 Min.   :   0.0   Min.   :-2.00   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   : 2.000
 1st Qu.: 320.0   1st Qu.: 4.00   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 5.000
 Median : 480.0   Median :15.00   Median :2.000   Median :3.000   Median :1.000   Median : 6.000
 Mean   : 472.8   Mean   :23.52   Mean   :1.567   Mean   :2.854   Mean   :1.044   Mean   : 6.443
 3rd Qu.: 576.0   3rd Qu.:42.75   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.: 7.000
 Max.   :1488.0   Max.   :60.00   Max.   :4.000   Max.   :8.000   Max.   :3.000   Max.   :15.000
 NA's   :1
   Fireplaces        Misc.Val            SalePrice
 Min.   :0.0000   Min.   :    0.00   Min.   : 12789
 1st Qu.:0.0000   1st Qu.:    0.00   1st Qu.:129500
 Median :1.0000   Median :    0.00   Median :160000
 Mean   :0.5993   Mean   :   50.63   Mean   :180796
 3rd Qu.:1.0000   3rd Qu.:    0.00   3rd Qu.:213500
 Max.   :4.0000   Max.   :17000.00   Max.   :755000
```

Figure 5: *Summary of data*

While some variables appear to be clean, it could be seen that our data still contain missing values, in which: Lot.Frontage has 490 observations, Total.Bsmt.SF and Garage.Area each has one value. Since the proportions of missing values are different significantly and the nature of each column also differs, we need to have distinct approaches to handle this issue. Furthermore, the set supposedly contains many outliers and unusual observations as the mean and the media for lot area are quite close, identical at 9500 – 10000 square feet, while the maximum value is 215245 square feet, which is 21 times higher.

b. Descriptive analysis for variables: Lot frontage

Within this part, visualization techniques such as histogram, boxplot or summary descriptive analysis for variables will be limited to Lot frontage only. This is because we need to dive deep into this variable for the replacement of 490 missing values and the process of EDA would be the same for the other variables.

The pair variable I would like to use with Lot frontage is explanatory variable Ms.Zoning, since this predictor could explain well for the missing data of Lot frontage: a location zone is usually accompanied with an according frontage area.

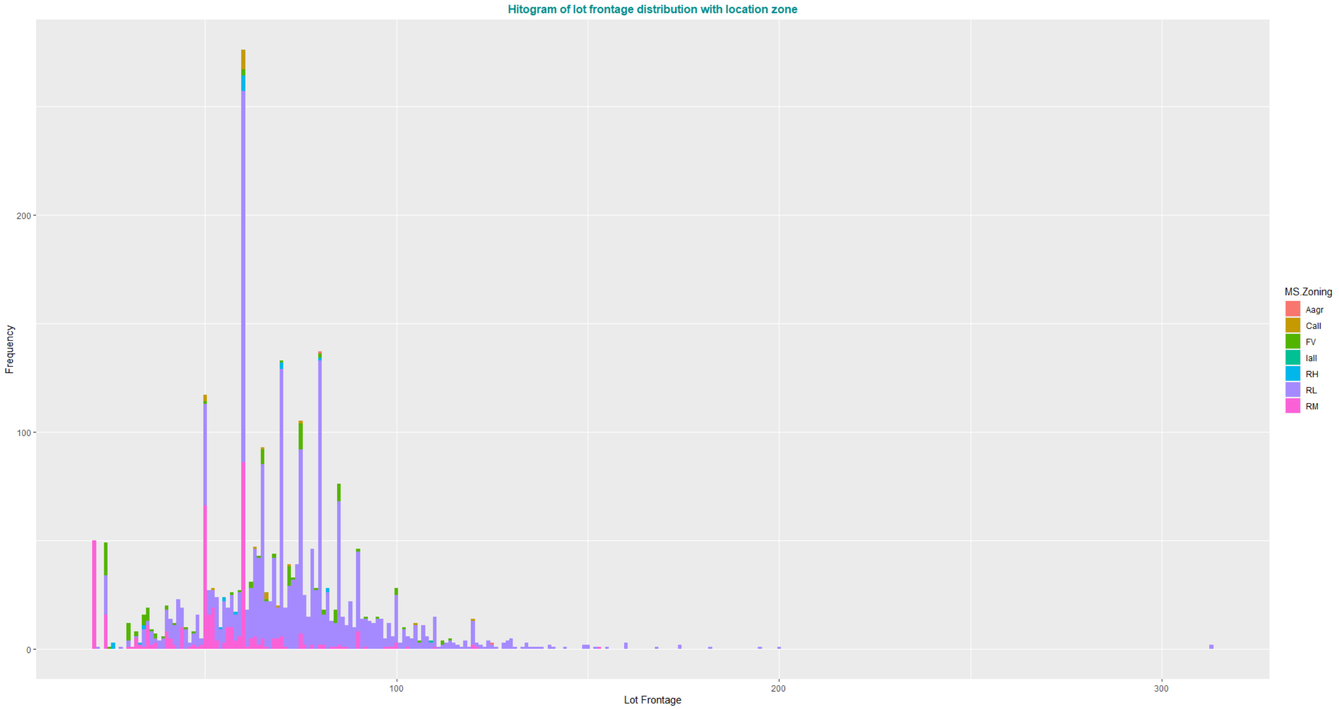
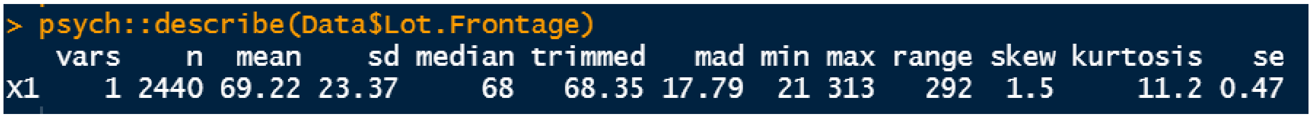


Figure 6: Histogram of lot frontage with stacking by location zone



```
> psych::describe(Data$Lot.Frontage)
   vars    n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 2440 69.22 23.37     68   68.35 17.79  21 313   292  1.5     11.2 0.47
```

Figure 7: Descriptive analysis of lot frontage

Both Figures 6 and 7 show us the data are highly right skewness at 1.5 with several outliers, in which maximum value is up to 313. This finding is also reinforced when the mean and median of lot frontage are quite similar, at around 68 square feet only. I will remove this unusual observation since it does not make sense to have such a value.

For the normality of data, since we combine all location zones together, the data seem not normal with the inflation of zone **RL**: Residential Low Density, where the frontage area would be significantly larger compared to the areas of the rest zones. To have a fair view of the normality of data, I split the histogram by each zone to check. As per Figure 8, only RL zone seems to have the most normal distribution among all.

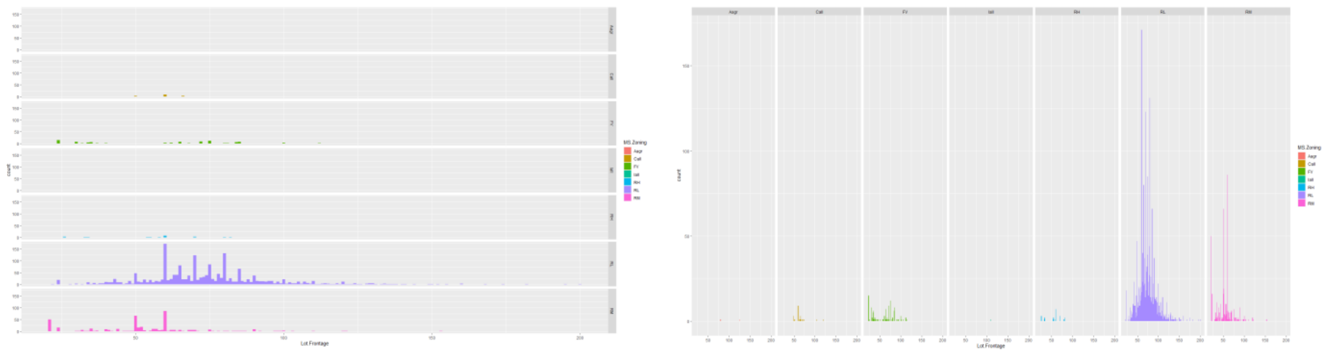
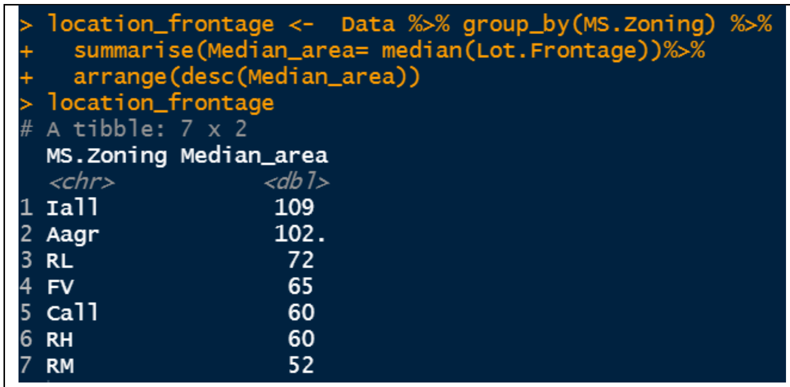


Figure 8: *Histogram of lot frontage by each location zone*

c. Dealing with missing data for modelling

There are several techniques to handle missing data, there is no such optimal solution for every case. The two most popular methods are deletion and imputation. If the proportion of missing values is insignificant, we can simply remove observations with those N/A values. Otherwise, more complex procedures of imputation should be done. Several examples are using the mean to substitute or employing regression to predict the missing values.

In our situation, since **Total Bsmt SF (Continuous):** Total square feet of basement area and **Garage Area (Continuous)**: Size of garage in square feet each have only one N/A value, I simply drop observation for this issue. And for **Lot frontage**, I replace the missing value by the median of each zone as the mean is influenced by some high leverage points in the above analysis. The replacement is by the median instead of the mean as shown in Figure 10 and the values for replacement can be seen in Table 2.



```
> location_frontage <-  Data %>% group_by(MS.Zoning) %>%
+   summarise(Median_area= median(Lot.Frontage))%>%
+   arrange(desc(Median_area))
> location_frontage
# A tibble: 7 x 2
  MS.Zoning Median_area
  <chr>           <dbl>
1 Iall            109
2 Aagr            102.
3 RL               72
4 FV               65
5 Call             60
6 RH               60
7 RM               52
```

***Table 2:** Values for replacement (by median)*

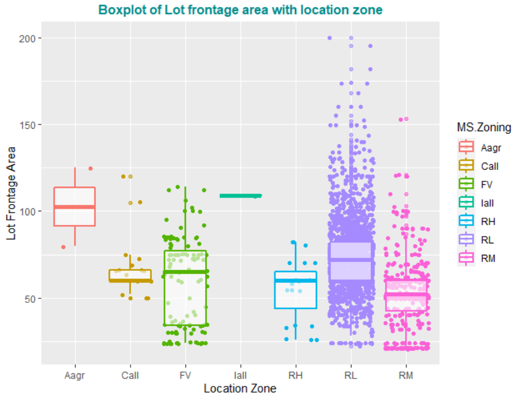
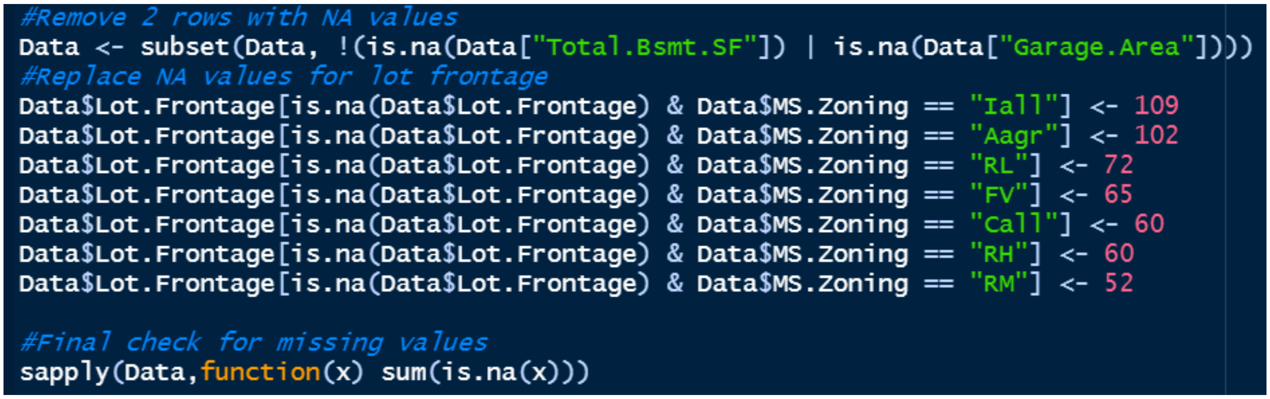
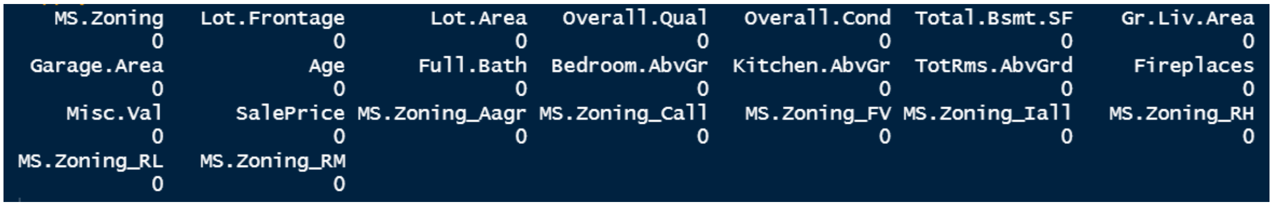


***Figure 10:** Boxplot to decide using the mean or median*

After managing all these missing values by both removing and imputation, here are the code and final check to see if any omission is there.



```
#Remove 2 rows with NA values
Data <- subset(Data, !(is.na(Data["Total.Bsmt.SF"]) | is.na(Data["Garage.Area"])))
#Replace NA values for lot frontage
Data$Lot.Frontage[is.na(Data$Lot.Frontage) & Data$MS.Zoning == "Iall"] <- 109
Data$Lot.Frontage[is.na(Data$Lot.Frontage) & Data$MS.Zoning == "Aagr"] <- 102
Data$Lot.Frontage[is.na(Data$Lot.Frontage) & Data$MS.Zoning == "RL"] <- 72
Data$Lot.Frontage[is.na(Data$Lot.Frontage) & Data$MS.Zoning == "FV"] <- 65
Data$Lot.Frontage[is.na(Data$Lot.Frontage) & Data$MS.Zoning == "Call"] <- 60
Data$Lot.Frontage[is.na(Data$Lot.Frontage) & Data$MS.Zoning == "RH"] <- 60
Data$Lot.Frontage[is.na(Data$Lot.Frontage) & Data$MS.Zoning == "RM"] <- 52

#Final check for missing values
sapply(Data,function(x) sum(is.na(x)))
```

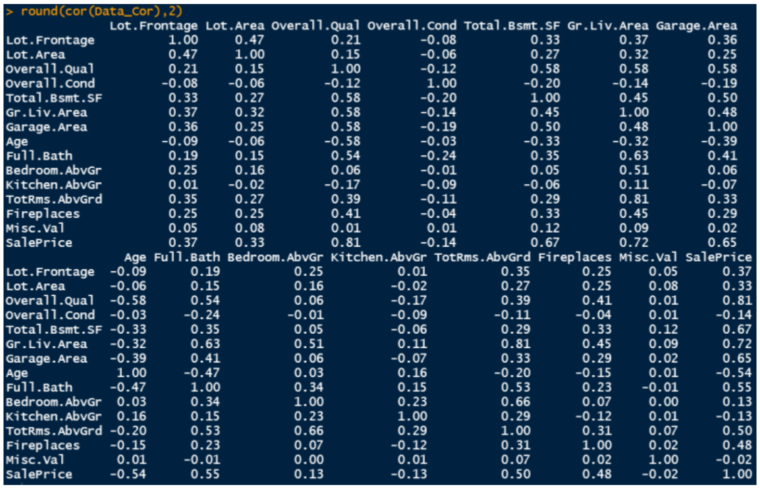


```
   MS.Zoning    Lot.Frontage        Lot.Area    Overall.Qual    Overall.Cond   Total.Bsmt.SF     Gr.Liv.Area
           0               0               0               0               0               0               0
 Garage.Area             Age       Full.Bath    Bedroom.AbvGr   Kitchen.AbvGr   TotRms.AbvGrd      Fireplaces
           0               0               0               0               0               0               0
    Misc.Val       SalePrice  MS.Zoning_Aagr  MS.Zoning_Call    MS.Zoning_FV  MS.Zoning_Iall    MS.Zoning_RH
           0               0               0               0               0               0               0
MS.Zoning_RL    MS.Zoning_RM
           0               0
```

***Figure 11:** Final check for data missing*

Up to this stage of processing, the data appear clean and ready to be employed.

d. Correlation matrix



```
> round(cor(Data_Cor),2)
              Lot.Frontage Lot.Area Overall.Qual Overall.Cond Total.Bsmt.SF Gr.Liv.Area Garage.Area
Lot.Frontage          1.00     0.47         0.21        -0.08          0.33        0.37        0.36
Lot.Area              0.47     1.00         0.15        -0.06          0.27        0.32        0.25
Overall.Qual          0.21     0.15         1.00        -0.12          0.58        0.58        0.58
Overall.Cond         -0.08    -0.06        -0.12         1.00         -0.20       -0.14       -0.19
Total.Bsmt.SF         0.33     0.27         0.58        -0.20          1.00        0.45        0.50
Gr.Liv.Area           0.37     0.32         0.58        -0.14          0.45        1.00        0.48
Garage.Area           0.36     0.25         0.58        -0.19          0.50        0.48        1.00
Age                  -0.09    -0.06        -0.58        -0.03         -0.33       -0.32       -0.39
Full.Bath             0.19     0.15         0.54        -0.24          0.35        0.63        0.41
Bedroom.AbvGr         0.25     0.16         0.06        -0.01          0.05        0.51        0.06
Kitchen.AbvGr         0.01    -0.02        -0.17        -0.09         -0.06        0.11       -0.07
TotRms.AbvGrd         0.35     0.27         0.39        -0.11          0.29        0.81        0.33
Fireplaces            0.25     0.25         0.41        -0.04          0.33        0.45        0.29
Misc.Val              0.05     0.08         0.01         0.01          0.12        0.09        0.02
SalePrice             0.37     0.33         0.81        -0.14          0.67        0.72        0.65
                Age Full.Bath Bedroom.AbvGr Kitchen.AbvGr TotRms.AbvGrd Fireplaces Misc.Val SalePrice
Lot.Frontage  -0.09      0.19          0.25          0.01          0.35       0.25     0.05      0.37
Lot.Area      -0.06      0.15          0.16         -0.02          0.27       0.25     0.08      0.33
Overall.Qual  -0.58      0.54          0.06         -0.17          0.39       0.41     0.01      0.81
Overall.Cond  -0.03     -0.24         -0.01         -0.09         -0.11      -0.04     0.01     -0.14
Total.Bsmt.SF -0.33      0.35          0.05         -0.06          0.29       0.33     0.12      0.67
Gr.Liv.Area   -0.32      0.63          0.51          0.11          0.81       0.45     0.09      0.72
Garage.Area   -0.39      0.41          0.06         -0.07          0.33       0.29     0.02      0.65
Age            1.00     -0.47          0.03          0.16         -0.20      -0.15     0.01     -0.54
Full.Bath     -0.47      1.00          0.34          0.15          0.53       0.23    -0.01      0.55
Bedroom.AbvGr  0.03      0.34          1.00          0.23          0.66       0.07     0.00      0.13
Kitchen.AbvGr  0.16      0.15          0.23          1.00          0.29      -0.12     0.01     -0.13
TotRms.AbvGrd -0.20      0.53          0.66          0.29          1.00       0.31     0.07      0.50
Fireplaces    -0.15      0.23          0.07         -0.12          0.31       1.00     0.02      0.48
Misc.Val       0.01     -0.01          0.00          0.01          0.07       0.02     1.00     -0.02
SalePrice     -0.54      0.55          0.13         -0.13          0.50       0.48    -0.02      1.00
```

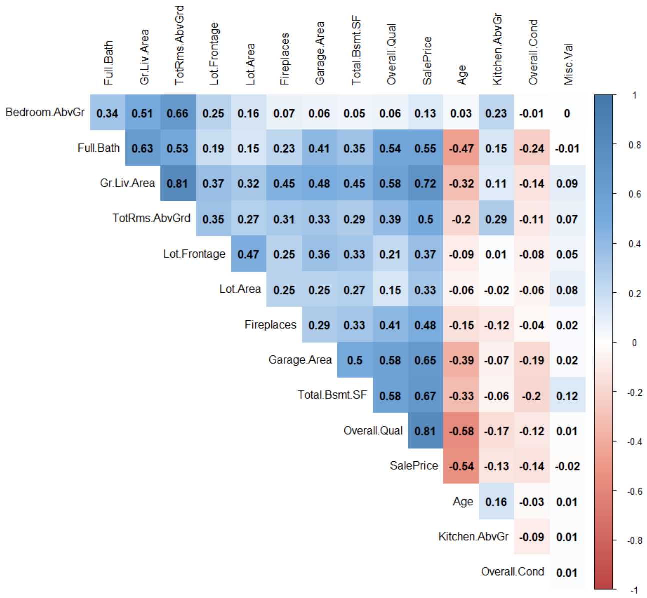
Figure 12: *Correlation table for variables*

Correlation is a statistical term describing the degree to which two variables move in coordination with one another. In this part, we validate the correlation between each pair of variables to detect their relationships as well as the problem of multicollinearity for regression. In R, to compute correlation for multiple variables, we could use cor() function to obtain the relationship quantifier for each pair of variables.

However, a table like that in Figure 12 would be hard to interpret even though I round it to 2 decimal points, hence we use Correlogram by R corrplot() function to visualize the correlation matrix.

In the matrix below, I use color to demonstrate how strong is the correlation: Blue for + 1 and Red for - 1. Associated with each color level is the correlation number inside the square box to let us quantify the correlation level. I also drop a redundant part of the bottom left haft and keep the top right area for a better visualization, gathering more audience focus as the chart is symmetric.

Finally, I calculate P-value for the significance of correlation coefficient (r) with $\alpha = 0.05$ (the null hypothesis is there is no relationship between two variables or r = 0) and remove non-relationship (by white blanks) as seen in Figure 8. This step is imperative to check the inter-relationship and because if r is too significant, there are multicollinear issues.



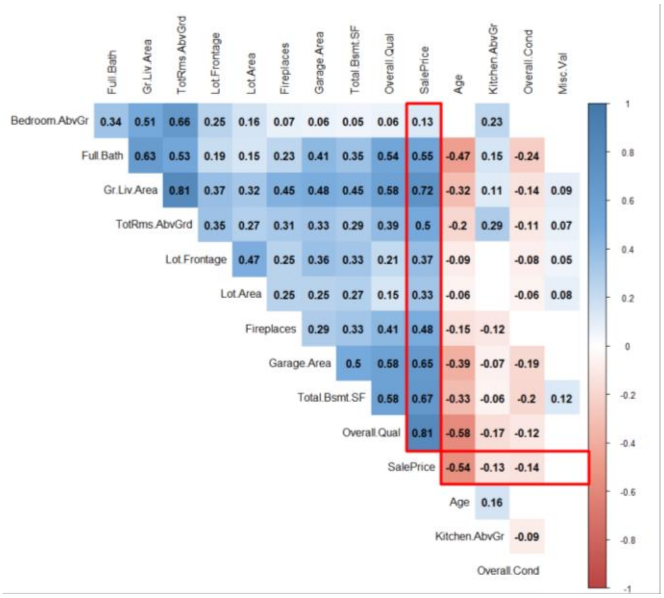


Figure 13: *Correlogram*

Figure 14: *Correlogram with a significant level of 0.05 to detect non-relationship pairs*

Taking a look at the relationship of explanatory predictors and **Sales prices** of houses, it is transparent to notice that most of our variables have positive relationships with outcome variables, in which Over quality of house is the variable with the highest correlation at +0.81, followed by **Gr Liv Area (Continuous):** Above grade (ground) living area square feet with the figure of +0.72. The minority of variables have negative relationships with the outcome; but surprisingly, **Overall Cond (Ordinal)**: Rates the overall condition of the house has a negative relationship with house price. However, this figure is quite marginal, at -0.14. And there is one variable that has a very low correlation with the outcome and cannot pass the hypothesis testing with $\alpha = 0.05$ is **Misc Val (Continuous**): Value of miscellaneous feature.

For the inter-relationship between groups of predictors, multicollinearity may happen for **Gr Liv Area (Continuous):** Above grade (ground) living area square feet with **TotRmsAbvGrd (Discrete):** Total rooms above grade (does not include bathrooms). This prediction has strong evidence in reality as the more rooms a house has, the higher its total area is, and therefore, we may drop one variable when developing our model.

e. Scatter plot to check the explanatory power

The insight we get from Anscombe's quartet is that always check for data visualization for a good grab of data set, and we just cannot rely solely on statistics outcome. Thus, after drawing a correlogram, we demonstrate the true explanatory power of 3 continuous variables to the dependent variable by correlation matrix.

Gr Liv Area (Continuous): Above grade (ground) living area square feet: r = +0.72

Misc Val (Continuous): Value of miscellaneous feature: r = -0.02

Lot Frontage (Continuous): Linear feet of street connected to property: r = +0.37 (the continuous variable with r value close the most to 0.5)

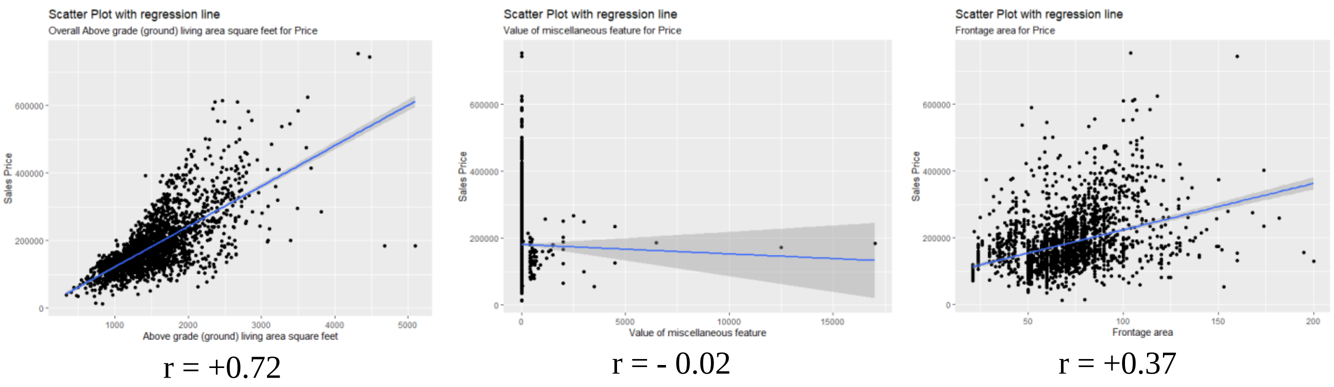


Figure 15: *Mixed Scatter plots for explanatory powers.*

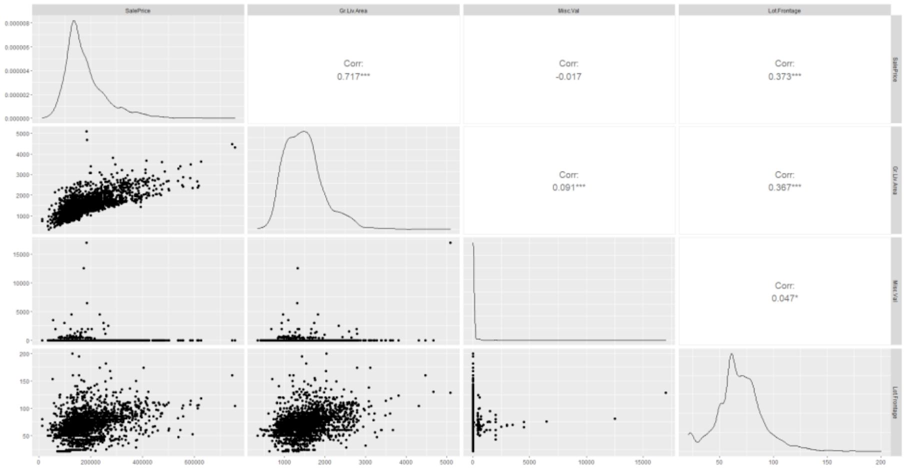


Figure 16: *Correlation matrix for explanatory powers.*

In general, both regression and correlation are most often used together since they have several similarities to talk about the relationship between variables. Evidently in Figures 15 and 16, the correlation could somewhat show how one variable affects another and therefore is used to predict the outcome. To elaborate, **Gr Liv Area** clusters close to the regression line and explains the sales price quite well. Even though there are high errors terms when area becomes large, we could further update model by scaling down the outcome, i.e., by logarithm transformation. Conversely, **Misc Val** with the majority of data points at 0 could not capture sufficiently the house price. The regression line between Misc Val for SalePrice is highly influenced by outliers. Lastly, as data of **Lot Frontage** shape rounded but not in line with regression, around the first half of frontage area are useful to explain the outcome variable, while the second half of data disperse largely beyond the line. It is also noticeable that four variables have right-skewed distributions, and the major errors come from the tails with influential points.

3. Fitting the model

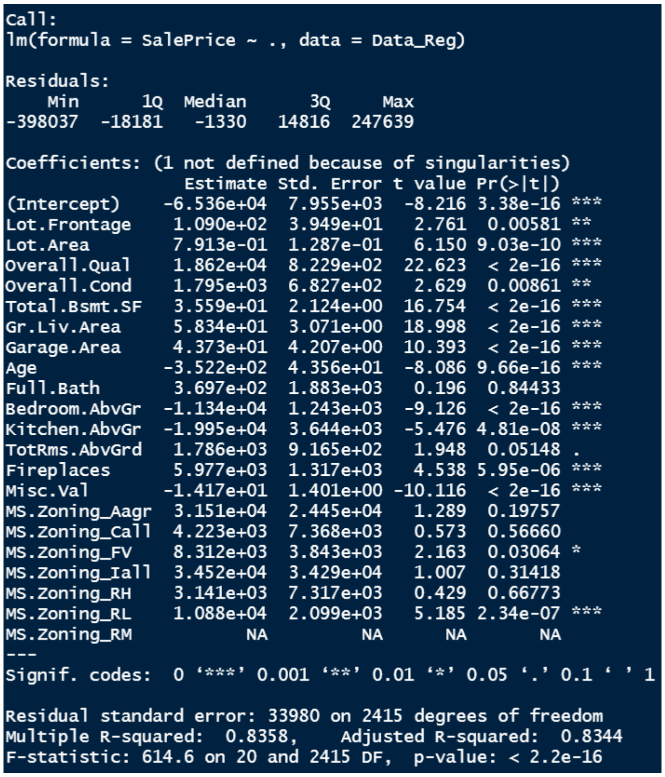
a. Full model (the first model)

The first step is to determine whether there is a strong connection between the outcome and the variables by fitting a multiple linear regression model using all of the predictors. The null hypothesis is that no relationship exists between any of the predictors and the outcome, which may be checked by obtaining F statistic.

Coefficients in estimate columns show us the causality of each predictor on the house price. Since this is our initial model without any correction, some explanatory variables have negative slopes such as Number of bedrooms above grade, which contradicts with the fact of buying a house that a house with more bedroom usually costs more. This could be either a concern or the norm due to the interaction of variables in regression, thus will be checked in the next part of this report.

In our model, a high F statistic combined with an extremely low p-value (2.2e-16) indicates that the null hypothesis can be rejected. As a result, there is a strong explanation of the predictors to the outcome.

The standard deviation of irreducible error is estimated using RSE (Residual Standard Error). A extremely large RSE of 33980 in our model indicates that our model deviates significantly from the real regression line, signifying that our model has several flaws needed to investigate.



```
Call:
lm(formula = SalePrice ~ ., data = Data_Reg)

Residuals:
    Min      1Q  Median      3Q     Max
-398037  -18181   -1330   14816  247639

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.536e+04  7.955e+03  -8.216 3.38e-16 ***
Lot.Frontage    1.090e+02  3.949e+01   2.761  0.00581 **
Lot.Area        7.913e-01  1.287e-01   6.150 9.03e-10 ***
Overall.Qual    1.862e+04  8.229e+02  22.623  < 2e-16 ***
Overall.Cond    1.795e+03  6.827e+02   2.629  0.00861 **
Total.Bsmt.SF   3.559e+01  2.124e+00  16.754  < 2e-16 ***
Gr.Liv.Area     5.834e+01  3.071e+00  18.998  < 2e-16 ***
Garage.Area     4.373e+01  4.207e+00  10.393  < 2e-16 ***
Age            -3.522e+02  4.356e+01  -8.086 9.66e-16 ***
Full.Bath       3.697e+02  1.883e+03   0.196  0.84433
Bedroom.AbvGr  -1.134e+04  1.243e+03  -9.126  < 2e-16 ***
Kitchen.AbvGr  -1.995e+04  3.644e+03  -5.476 4.81e-08 ***
TotRms.AbvGrd   1.786e+03  9.165e+02   1.948  0.05148 .
Fireplaces      5.977e+03  1.317e+03   4.538 5.95e-06 ***
Misc.Val       -1.417e+01  1.401e+00 -10.116  < 2e-16 ***
MS.Zoning_Aagr  3.151e+04  2.445e+04   1.289  0.19757
MS.Zoning_Call  4.223e+03  7.368e+03   0.573  0.56660
MS.Zoning_FV    8.312e+03  3.843e+03   2.163  0.03064 *
MS.Zoning_Iall  3.452e+04  3.429e+04   1.007  0.31418
MS.Zoning_RH    3.141e+03  7.317e+03   0.429  0.66773
MS.Zoning_RL    1.088e+04  2.099e+03   5.185 2.34e-07 ***
MS.Zoning_RM           NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33980 on 2415 degrees of freedom
Multiple R-squared:  0.8358,    Adjusted R-squared:  0.8344
F-statistic: 614.6 on 20 and 2415 DF,  p-value: < 2.2e-16
```

Figure 17: *Full predictor model*

The amount of variability in the result that can be explained by the model is measured by R-squared (R^2), and its real value constantly swings between 0 and 1. However, the increasement in number of predictors mostly results in a higher value of R^2 due to the inflation of R-squared for explanatory power. To prevent this impact, Adjusted R-squared modifies the value of R^2. The result shows that 83.44% of the variance in the data is being explained by the model.

The t value of a predictor indicates how far its predicted coefficient is from 0 in standard deviations. An extremely low p-value (smaller than 0.05) is desirable because we can reject the null hypothesis. Clearly, there are many predictors could not contribute to our line as p value is large. They are Full.Bath, TotRms.AbvGrd, and majority of dummy variables of house zones. It is clear that MS.Zoning_RM has NA as a coefficient in a regression, signifying that MS.Zoning_RM is linearly related to the other variables of dummy set and retaining all dummy variables would worsen the model.

b. Defining problems of model by assumption checking

To further improve and develop our model, we need perform regression analysis validation by testing the linear model's underlying assumptions of Ordinary Least Square in this part. First, let look at regression plot.

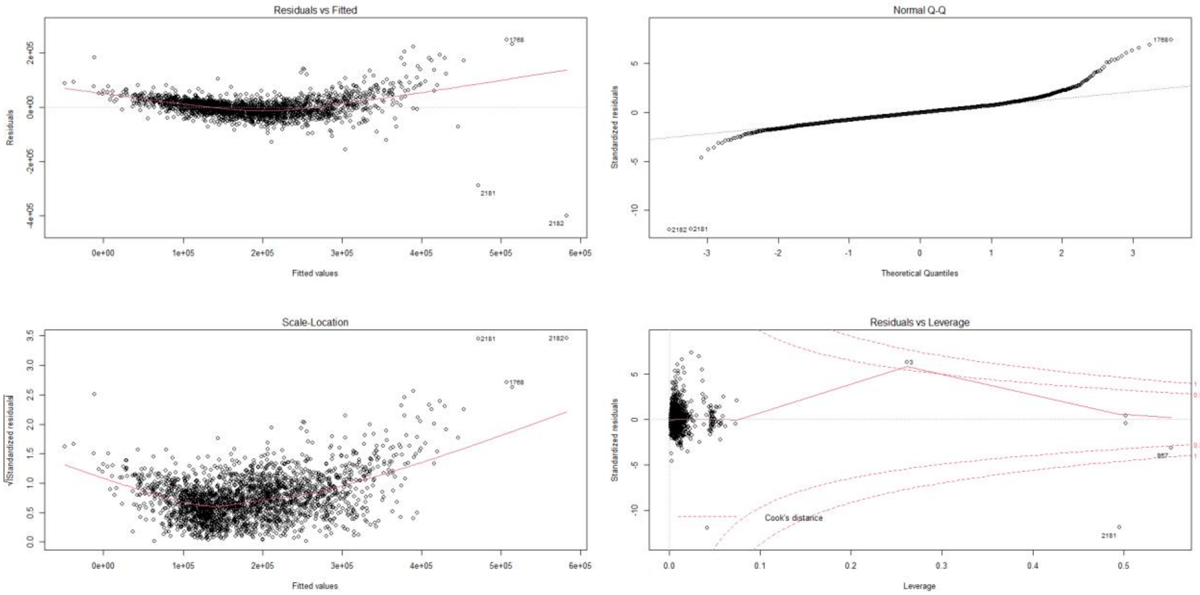
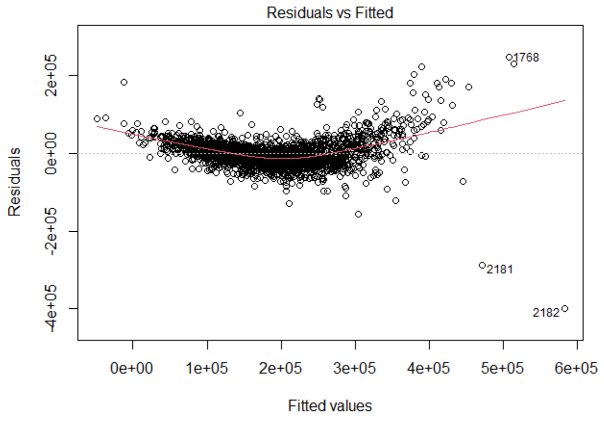


Figure 17: *Full predictor model*

- **Linearity:**

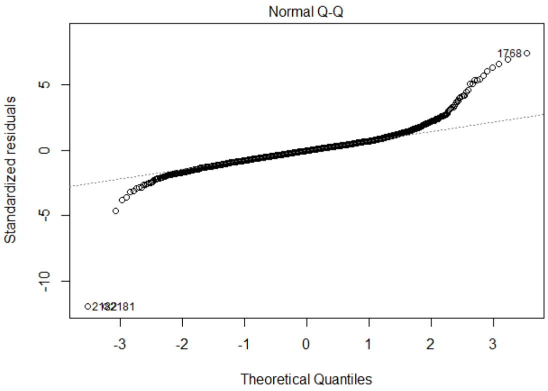
We assume that there is a linear connection between the predictors and the outcome when we use linear regression. If it is not linear, then the prediction would be invalid. The plot of fitted values versus residuals can be used to determine the model's non-linearity. The difference between the actual values and the fitted outcomes according to the model is the residual for any observation. If there is the presence of a pattern in the residual plot, this would indicate that there is an issue with the model's linear assumption.



It is quite straightforward to notice that there is a curve pattern in our residual plot as data do not randomly spread around the horizontal line. The curve in our case denotes slight non-linearity in our data and violates homoscedasticity of error terms. I suppose we can further improve this model by observing 3 outliers or modifying several variables.

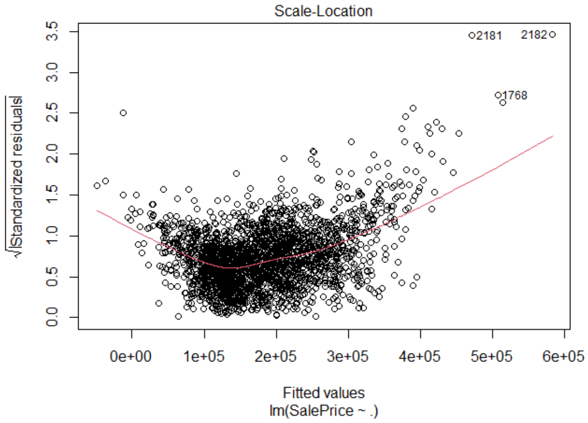
- **Normality:**

The normality is the assumption that residuals have a normal distribution to validate the hypothesis testing for each regression coefficient. In the normal Q-Q plot, any points that do not fall on the diagonal line are potential outliers, hence violating the normality assumption.



In our plot, we could see that 2 tails of observation have many potential outlier data points as they are located far from the line. The model is most effective when targeting those properties with prices in the inter-quantile range from 25% to 75%. This finding aligns with several previous notices about outliers in our previous EDA.

- **Homoscedasticity:**



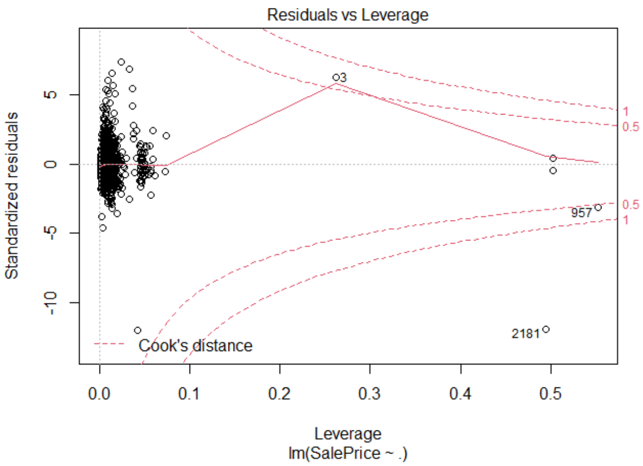
A scale-location plot shows the square root of the standardized residuals along the y-axis and the fitted values of a regression model along the x-axis. This graph is used to identify homoscedasticity and see if the residuals have any clear pattern.

Our plot appears to have the same view as the Residuals vs Fitted plot above when there is an apparent pattern and the red line just does not spread horizontally.

- **Independence of Errors and Observations**

While the other three major assumptions can be tested through the regression plot, this assumption verification depends largely on the knowledge of data set about how it was collected, the knowledge about sampling techniques were used or how observations were chosen. Let consider our data set can meet this assumption to facilitate our model.

Residuals vs. Leverage Plot Interpretation: A residuals vs. leverage plot is a form of diagnostic diagram that helps us to discover influential observations in a regression model. Leverage here refers to the amount to which the coefficients in the regression model would vary and change if a certain observation were removed from the data set. The observation of data with high leverage usually has a strong impact on the coefficients in the regression model; so if we eliminate these observations, the coefficients of the model would improve noticeably.



For our plot, there are 3 highly influential points that fall outside of the red dashed lines: the observations of 3rd, 957th and 2181st.

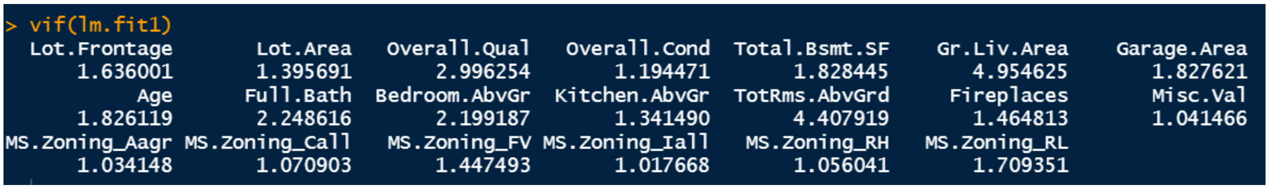
In conclusion, we may figure out that there are several problems with our model as it violates underlying assumptions and there is the presence of outliers, influenced points and unusual observations within the data set. Practical solutions to tackle these issues are:

- To transform data.
- To validate the observation if it is not an error or not.
- Attempt to fit another regression model.
- Or simply to remove the influential observations.

I prioritize to fit the model first as it could be intuitive to do so. Removing unusual observations should be taken into consideration carefully because we are given an available data set but did not collect ourselves.

c. Multicollinearity checking

Prior to conducting any corrective measures, we check some common issues of linear regression such as multicollinearity. Our initial model has a perfect linear relation between dummy variables that I drop out the column of MS.Zoning_RM, otherwise it always appears error whenever calculating the Variance Inflation Factor (VIF) in R.



```
> vif(lm.fit1)
  Lot.Frontage        Lot.Area    Overall.Qual    Overall.Cond   Total.Bsmt.SF     Gr.Liv.Area     Garage.Area
      1.636001        1.395691        2.996254        1.194471        1.828445        4.954625        1.827621
           Age       Full.Bath    Bedroom.AbvGr   Kitchen.AbvGr   TotRms.AbvGrd     Fireplaces        Misc.Val
      1.826119        2.248616        2.199187        1.341490        4.407919        1.464813        1.041466
MS.Zoning_Aagr  MS.Zoning_Call    MS.Zoning_FV  MS.Zoning_Iall    MS.Zoning_RH    MS.Zoning_RL
      1.034148        1.070903        1.447493        1.017668        1.056041        1.709351
```

Figure 18: *VIF checking for multicollinearity*

Referring to Figure 18, all values of VIF for the model is smaller than 5, hence the problem of multicollinearity could be satisfied.

There are several ways to mitigate multicollinearity if it exists in our model. The potential solutions include the following:

- To remove some of the highly correlated independent variables.

- To conduct a highly correlated variable study, such as principal components analysis or partial least squares regression.

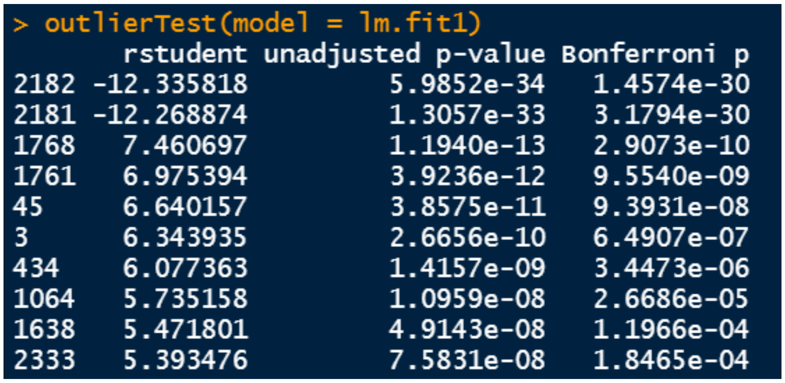
- To combine the independent linear variables together.

- LASSO and Ridge regression are sophisticated regression analysis techniques that can handle multicollinearity.

It is noticeable that each solution has pros and cons. The most practical and simplest way is to remove highly correlated predictors. However, if we are willing to accept fewer accurate coefficients or a regression model with a high R-squared but hardly any statistically significant variables, then keeping multicollinearity without any actions might be the best solution.

d. Outlier detection

Within the normality or linearity testing above, we already detected several outliers for the assumption validation. Here we further employ Bonferroni Outlier Test from car package to find more outliers in our set.



```
> outlierTest(model = lm.fit1)
       rstudent unadjusted p-value Bonferroni p
2182 -12.335818         5.9852e-34   1.4574e-30
2181 -12.268874         1.3057e-33   3.1794e-30
1768   7.460697         1.1940e-13   2.9073e-10
1761   6.975394         3.9236e-12   9.5540e-09
45     6.640157         3.8575e-11   9.3931e-08
3      6.343935         2.6656e-10   6.4907e-07
434    6.077363         1.4157e-09   3.4473e-06
1064   5.735158         1.0959e-08   2.6686e-05
1638   5.471801         4.9143e-08   1.1966e-04
2333   5.393476         7.5831e-08   1.8465e-04
```

Figure 19: *Outlier detection by Bonferroni method*

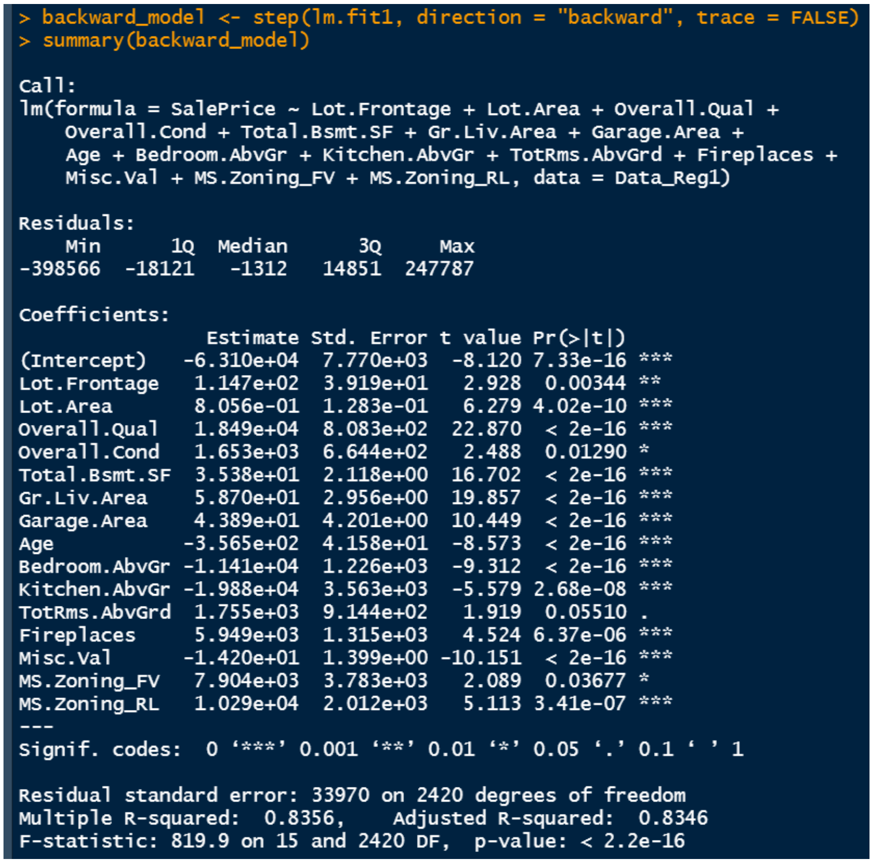
Figure 19 demonstrates there 10 observations with Studentized residuals with Bonferroni p < 0.05. The first column shows the index of rows that have unusual observations, and it ranks outliers in order starting from the row with the largest studentized residual.

Again, if there are outliers, do not exclude them just because they do not match our model well. Removing them is always justified by sound arguments and information. Instead, we should first update our model to determine whether it has improved, and then return to tackle outliers later. As in the analysis of Independence of Errors and Observations assumption above, within the scope of this report, I will leave the outliers and just try to fix the model.

4. Updating model

a. Further improvement: Stepwise selection (The second model)

As I started the regression line with a full model and then I dropped one variable that has perfect linearity issue, I will use the backward selection method to improve our model in the second attempt to find the best fit line.



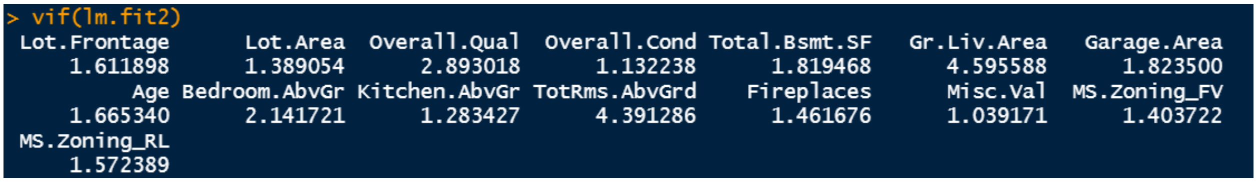
```
> backward_model <- step(lm.fit1, direction = "backward", trace = FALSE)
> summary(backward_model)

Call:
lm(formula = SalePrice ~ Lot.Frontage + Lot.Area + Overall.Qual + 
    Overall.Cond + Total.Bsmt.SF + Gr.Liv.Area + Garage.Area + 
    Age + Bedroom.AbvGr + Kitchen.AbvGr + TotRms.AbvGrd + Fireplaces + 
    Misc.Val + MS.Zoning_FV + MS.Zoning_RL, data = Data_Reg1)

Residuals:
    Min      1Q  Median      3Q     Max 
-398566  -18121   -1312   14851  247787 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -6.310e+04  7.770e+03  -8.120 7.33e-16 ***
Lot.Frontage   1.147e+02  3.919e+01   2.928  0.00344 ** 
Lot.Area       8.056e-01  1.283e-01   6.279 4.02e-10 ***
Overall.Qual   1.849e+04  8.083e+02  22.870  < 2e-16 ***
Overall.Cond   1.653e+03  6.644e+02   2.488  0.01290 *  
Total.Bsmt.SF  3.538e+01  2.118e+00  16.702  < 2e-16 ***
Gr.Liv.Area    5.870e+01  2.956e+00  19.857  < 2e-16 ***
Garage.Area    4.389e+01  4.201e+00  10.449  < 2e-16 ***
Age           -3.565e+02  4.158e+01  -8.573  < 2e-16 ***
Bedroom.AbvGr -1.141e+04  1.226e+03  -9.312  < 2e-16 ***
Kitchen.AbvGr -1.988e+04  3.563e+03  -5.579 2.68e-08 ***
TotRms.AbvGrd  1.755e+03  9.144e+02   1.919  0.05510 .  
Fireplaces     5.949e+03  1.315e+03   4.524 6.37e-06 ***
Misc.Val      -1.420e+01  1.399e+00 -10.151  < 2e-16 ***
MS.Zoning_FV   7.904e+03  3.783e+03   2.089  0.03677 *  
MS.Zoning_RL   1.029e+04  2.012e+03   5.113 3.41e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33970 on 2420 degrees of freedom
Multiple R-squared:  0.8356,	Adjusted R-squared:  0.8346 
F-statistic: 819.9 on 15 and 2420 DF,  p-value: < 2.2e-16
```

Figure 20: *Result of backward selection*

Our second model is:

SalePrice ~ Lot.Frontage + Lot.Area + Overall.Qual + Overall.Cond + Total.Bsmt.SF + Gr.Liv.Area + Garage.Area + Age + Bedroom.AbvGr + Kitchen.AbvGr + TotRms.AbvGrd + Fireplaces + Misc.Val + MS.Zoning_FV + MS.Zoning_RL

The most common concern when using stepwise selection for regression is multicollinearity, so I check it one more time to validate.



```
> vif(lm.fit2)
 Lot.Frontage      Lot.Area  Overall.Qual  Overall.Cond Total.Bsmt.SF   Gr.Liv.Area   Garage.Area 
     1.611898      1.389054      2.893018      1.132238      1.819468      4.595588      1.823500 
          Age Bedroom.AbvGr Kitchen.AbvGr TotRms.AbvGrd    Fireplaces      Misc.Val  MS.Zoning_FV 
     1.665340      2.141721      1.283427      4.391286      1.461676      1.039171      1.403722 
 MS.Zoning_RL 
     1.572389 
```

Figure 21: *VIF checking for the updated model*

Fortunately, there is no multicollinearity, but the improvement measured by Adjusted R Square just slightly increases from 83.44% to 83.46%, which is marginal. All the issues with the assumption remain unchanged with the updated model by the method of backward selection as seen in Figure 22. We continue to the second phase of corrective measures: transformation.

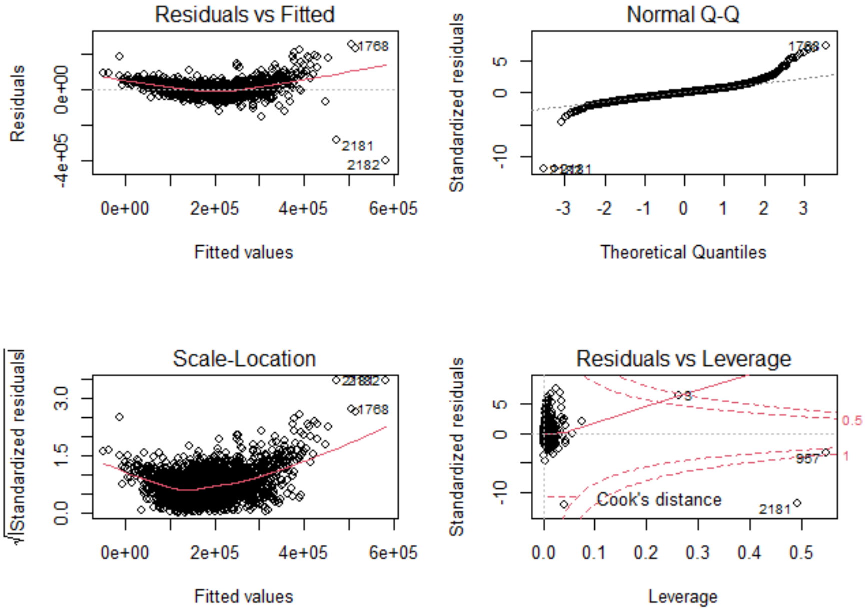
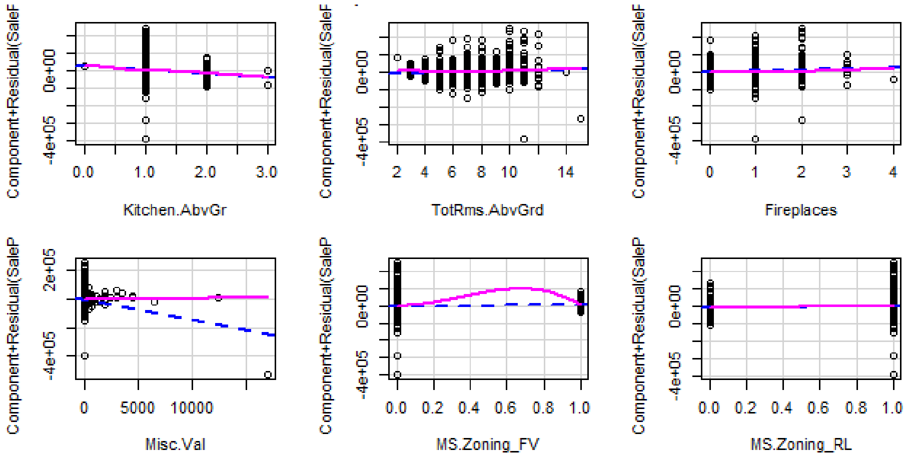


Figure 22: Regression plot for updated model

b. Further improvement: Data transformation (the third model)

To make a meaningful transformation and best guess to improve our model, I come up with Component Residual Plots to detect which variable is needed to modify.



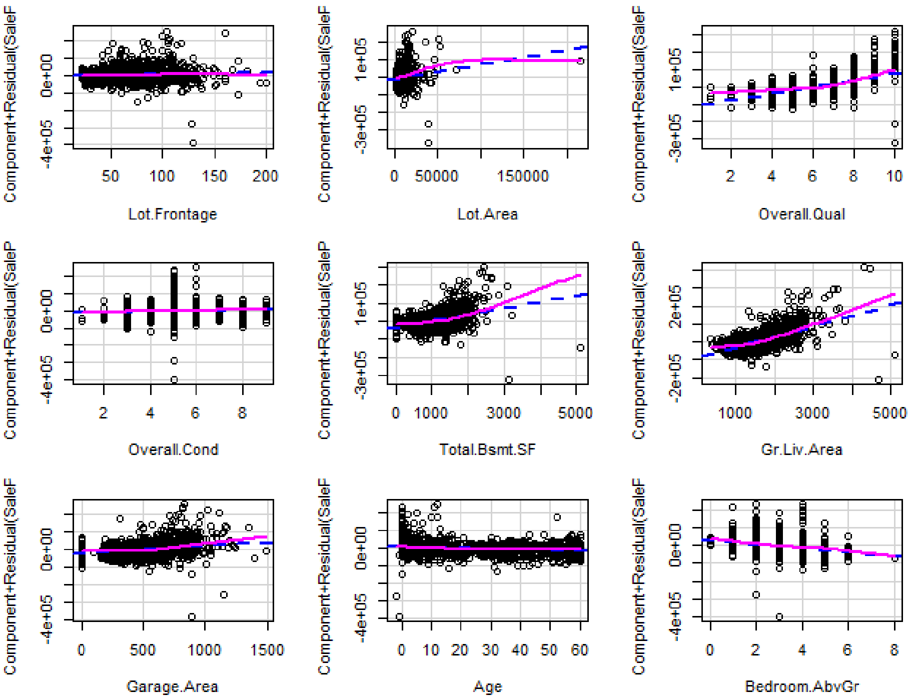


Figure 23: *Component Residual Plots for updated model*

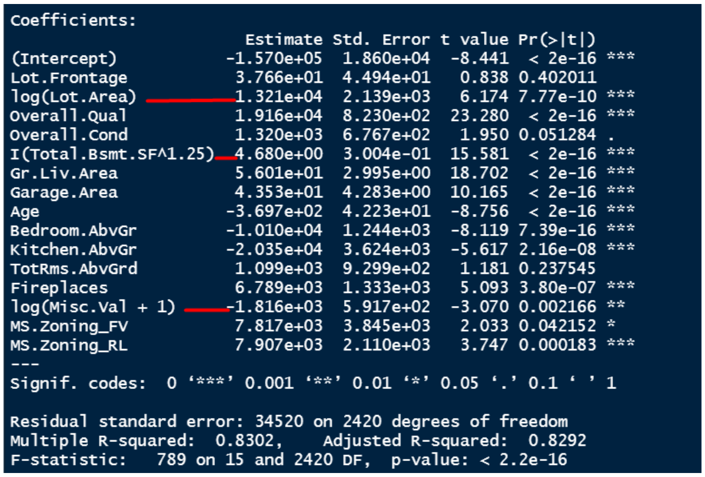
For the relationship between the predictor and the residuals, the pink line (residual line) is simulated. The blue dashed line (component line) represents the best fit line. A considerable discrepancy between the two lines for a predictor indicates that there is no linear link between the predictor and the outcome.

From the Figure 23, there are 3 variables that have considerable difference between the two lines: Lot.Area, Total.Bsmt.SF, and Misc.Val. For Lot.Area and Misc.Val, as the fit lines are below the residual line, I try to smoosh them up by logarithm transformation. Conversely, the opposite direction is applied for the plot of Total.Bsmt.SF, I increase the values by an exponential of 1.25.

Three new variables will have modified data as:

Original data	Transform data	Note
Lot.Area	Log(Lot.Area)	
Misc.Val	Log(Misc.Val + 1)	Since Miscellaneous values have many 0 data, the logarithm of 0 would return undefined values. Thus, I come up with the simplest solution: adding 1 to the sum.
Total.Bsmt.SF	Total.Bsmt.SF^1.25	

While all three transformed predictors contribute greatly for the outcome as the p-values are insignificant. The overall fit of new transformed model reduces to 82.92%, indicating that this third model is even worse than our initial full-predictor model.



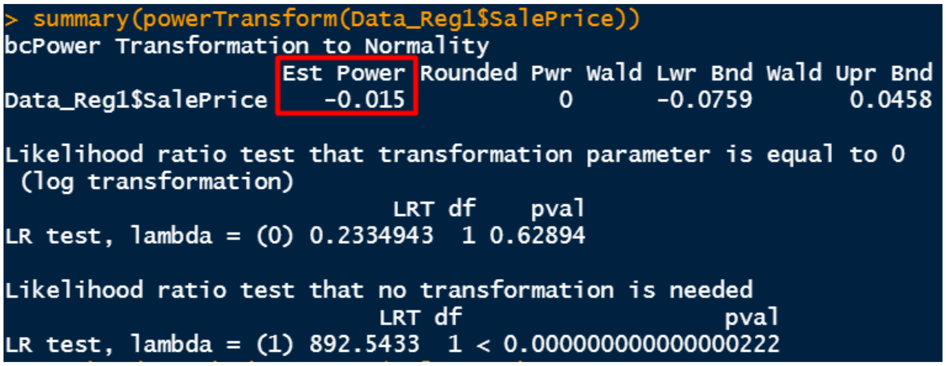
```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)            -1.570e+05  1.860e+04  -8.441  < 2e-16 ***
Lot.Frontage            3.766e+01  4.494e+01   0.838 0.402011    
log(Lot.Area)           1.321e+04  2.139e+03   6.174 7.77e-10 ***
Overall.Qual            1.916e+04  8.230e+02  23.280  < 2e-16 ***
Overall.Cond            1.320e+03  6.767e+02   1.950 0.051284 .  
I(Total.Bsmt.SF^1.25)   4.680e+00  3.004e-01  15.581  < 2e-16 ***
Gr.Liv.Area             5.601e+01  2.995e+00  18.702  < 2e-16 ***
Garage.Area             4.353e+01  4.283e+00  10.165  < 2e-16 ***
Age                    -3.697e+02  4.223e+01  -8.756  < 2e-16 ***
Bedroom.AbvGr          -1.010e+04  1.244e+03  -8.119 7.39e-16 ***
Kitchen.AbvGr          -2.035e+04  3.624e+03  -5.617 2.16e-08 ***
TotRms.AbvGrd           1.099e+03  9.299e+02   1.181 0.237545    
Fireplaces              6.789e+03  1.333e+03   5.093 3.80e-07 ***
log(Misc.Val + 1)      -1.816e+03  5.917e+02  -3.070 0.002166 ** 
MS.Zoning_FV            7.817e+03  3.845e+03   2.033 0.042152 *  
MS.Zoning_RL            7.907e+03  2.110e+03   3.747 0.000183 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34520 on 2420 degrees of freedom
Multiple R-squared:  0.8302,	Adjusted R-squared:  0.8292 
F-statistic:   789 on 15 and 2420 DF,  p-value: < 2.2e-16
```

Figure 24: *transformed model regression result*

c. Further improvement: Box-Cox Transformation (the fourth model)

The distribution of original sales price is highly positively skewed with skewness of 1.78 due to several outliers.

For the further corrective transformation, I also employ powerTransform () function, which uses the maximum likelihood-like approach of Box and Cox (1964) to select a transformation of a response for normality. The Figure 25 detects the appropriate power to apply the transformation.



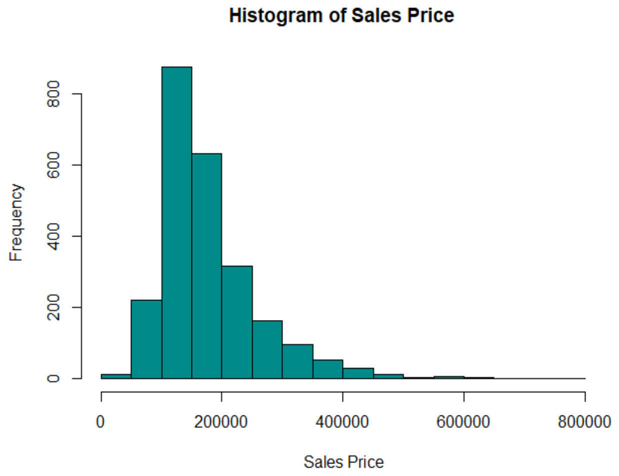
```
> summary(powerTransform(Data_Reg1$SalePrice))
bcPower Transformation to Normality 
                   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Data_Reg1$SalePrice   -0.015           0      -0.0759       0.0458

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df    pval
LR test, lambda = (0) 0.2334943  1 0.62894

Likelihood ratio test that no transformation is needed
                          LRT df                  pval
LR test, lambda = (1) 892.5433  1 < 0.000000000000000222
```

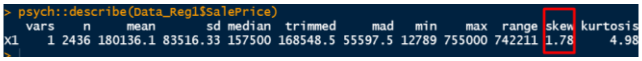
Figure 25: *Power transform test to approach normality.*

My transformed value for the outcome of sales price is as below calculation:

Original data	Transform data	Note
SalePrice	SalePrice_New = SalePrice^(-0.015)	

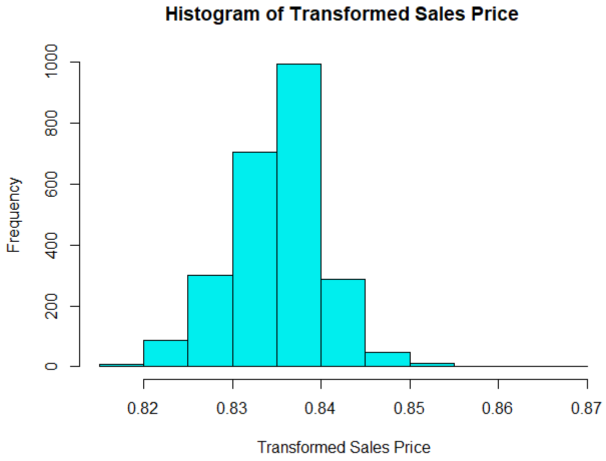
Please note that the remaining independent variables retain the same values without any transformation in this model.

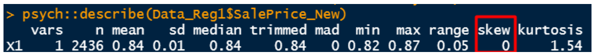




```
> psych::describe(Data_Reg1$SalePrice)
   vars    n     mean       sd median  trimmed     mad   min    max  range skew kurtosis
X1    1 2436 180136.1 83516.33 157500 168548.5 55597.5 12789 755000 742211 1.78     4.98
>
```

Figure 26: Distribution of original sales price

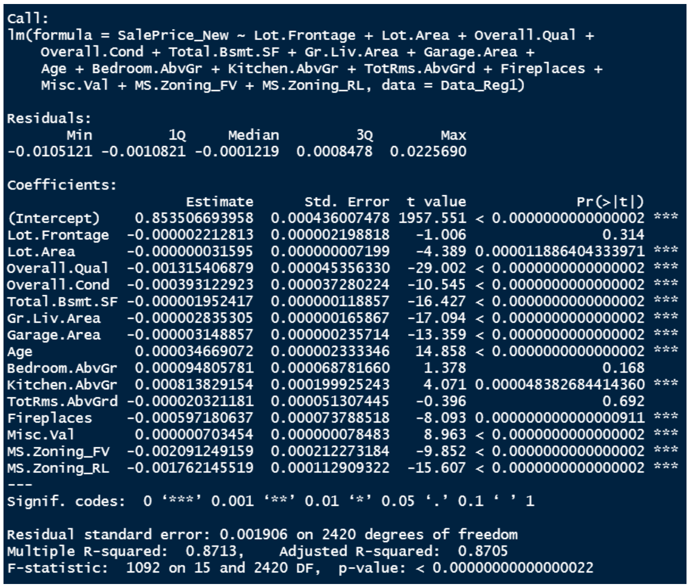




```
> psych::describe(Data_Reg1$SalePrice_New)
   vars    n mean   sd median trimmed mad  min  max range skew kurtosis
X1    1 2436 0.84 0.01   0.84    0.84   0 0.82 0.87  0.05    0     1.54
```

Figure 27: Distribution of transformed sales price

Applying the new sales price to the regression line, the result of our fourth model becomes better with the R adjusted square increases to 87.05%. The residual plot also improves significantly as shown in Figure 28.



```
Call:
lm(formula = SalePrice_New ~ Lot.Frontage + Lot.Area + Overall.Qual +
    Overall.Cond + Total.Bsmt.SF + Gr.Liv.Area + Garage.Area +
    Age + Bedroom.AbvGr + Kitchen.AbvGr + TotRms.AbvGrd + Fireplaces +
    Misc.Val + MS.Zoning_FV + MS.Zoning_RL, data = Data_Reg1)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0105121 -0.0010821 -0.0001219  0.0008478  0.0225690 

Coefficients:
                   Estimate      Std. Error  t value              Pr(>|t|)    
(Intercept)    0.853506693958  0.000436007478 1957.551 < 0.0000000000000002 ***
Lot.Frontage  -0.000002212813  0.000002198818   -1.006                0.314    
Lot.Area      -0.000000031595  0.000000007199   -4.389 0.0000118864043339 71 ***
Overall.Qual  -0.001315406879  0.000045356330  -29.002 < 0.0000000000000002 ***
Overall.Cond  -0.000393122923  0.000037280224  -10.545 < 0.0000000000000002 ***
Total.Bsmt.SF -0.000001952417  0.000000118857  -16.427 < 0.0000000000000002 ***
Gr.Liv.Area   -0.000002835305  0.000000165867  -17.094 < 0.0000000000000002 ***
Garage.Area   -0.000003148857  0.000000235714  -13.359 < 0.0000000000000002 ***
Age            0.000034669072  0.000002333346   14.858 < 0.0000000000000002 ***
Bedroom.AbvGr  0.000094805781  0.000068781660    1.378                0.168    
Kitchen.AbvGr  0.000813829154  0.000199925243    4.071 0.0000483826844143 60 ***
TotRms.AbvGrd -0.000020321181  0.000051307445   -0.396                0.692    
Fireplaces    -0.000597180637  0.000073788518   -8.093 0.000000000000000911 ***
Misc.Val       0.000000703454  0.000000078483    8.963 < 0.0000000000000002 ***
MS.Zoning_FV  -0.002091249159  0.000212273184   -9.852 < 0.0000000000000002 ***
MS.Zoning_RL  -0.001762145519  0.000112909322  -15.607 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001906 on 2420 degrees of freedom
Multiple R-squared:  0.8713,    Adjusted R-squared:  0.8705 
F-statistic:  1092 on 15 and 2420 DF,  p-value: < 0.00000000000000022
```

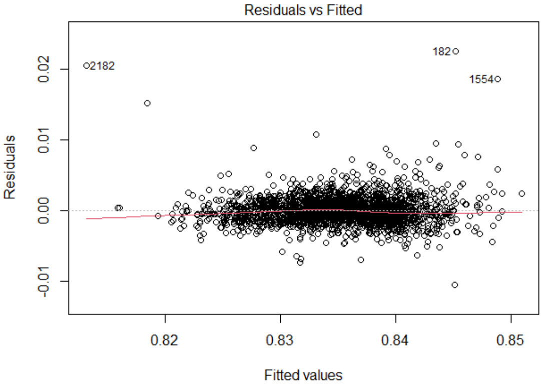


Figure 28: *Regression result of the modified model*

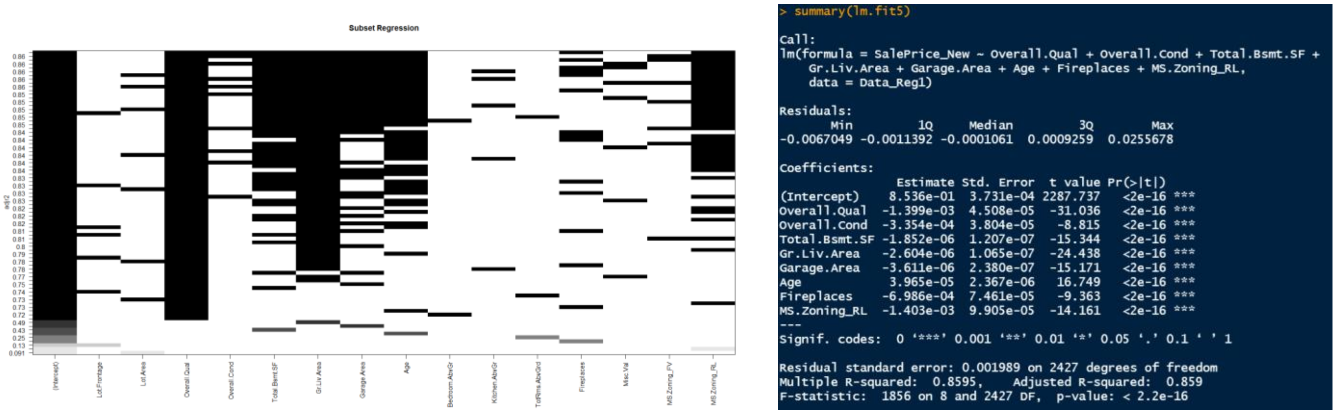
Our final model of transformation is:

SalePrice_New ~ Lot.Frontage + Lot.Area + Overall.Qual + Overall.Cond + Total.Bsmt.SF + Gr.Liv.Area + Garage.Area + Age + Bedroom.AbvGr + Kitchen.AbvGr + TotRms.AbvGrd + Fireplaces + Misc.Val + MS.Zoning_FV + MS.Zoning_RL

Important note: Even though the model improves considerably, the result still contains several variables with insignificant contribution as p-value is large. Furthermore, as the slopes for each predictor are scaled down to fit the values of outcome transformation (becomes smaller) , it gets tougher to interpret the effect of each explanatory variable to the outcome and grab the sense of the model.

d. Subset regression (the fifth model)

For our most updated model above, I continue to employ the subset regression to choose the best model in scenario of a different number of predictors. Our best subset model is the top one model with the highest Adjusted R Square of the plot in Figure 29. Besides the intercept, there are total 8 best explanatory variables to predict the outcome, they are: **SalePrice_New ~ Overall.Qual + Overall.Cond + Total.Bsmt.SF + Gr.Liv.Area + Garage.Area + Age + Fireplaces + MS.Zoning_RL**. This model has all significant coefficients with very low p-values and the overall fit is 85.9%.



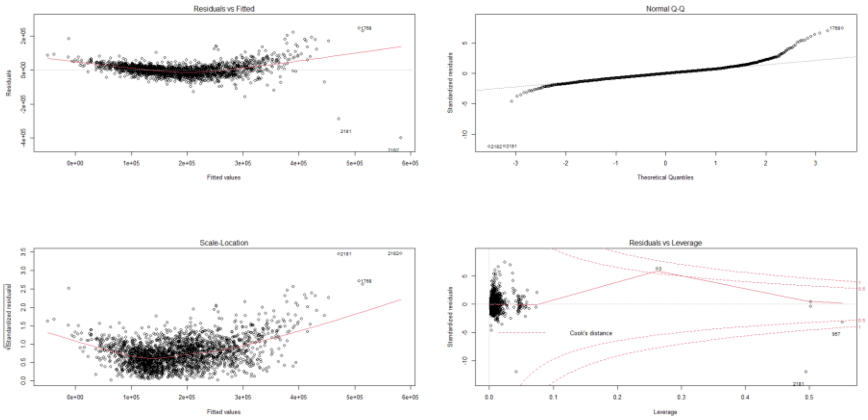
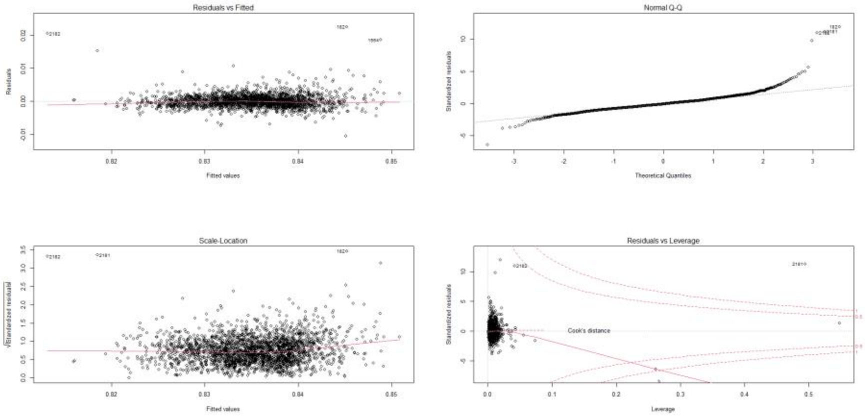
```
> summary(lm.fit5)

Call:
lm(formula = SalePrice_New ~ Overall.Qual + Overall.Cond + Total.Bsmt.SF +
    Gr.Liv.Area + Garage.Area + Age + Fireplaces + MS.Zoning_RL,
    data = Data_Reg1)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0067049 -0.0011392 -0.0001061  0.0009259  0.0255678

Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)    8.536e-01  3.731e-04 2287.737   <2e-16 ***
Overall.Qual  -1.399e-03  4.508e-05  -31.036   <2e-16 ***
Overall.Cond  -3.354e-04  3.804e-05   -8.815   <2e-16 ***
Total.Bsmt.SF -1.852e-06  1.207e-07  -15.344   <2e-16 ***
Gr.Liv.Area   -2.604e-06  1.065e-07  -24.438   <2e-16 ***
Garage.Area   -3.611e-06  2.380e-07  -15.171   <2e-16 ***
Age            3.965e-05  2.367e-06   16.749   <2e-16 ***
Fireplaces    -6.986e-04  7.461e-05   -9.363   <2e-16 ***
MS.Zoning_RL  -1.403e-03  9.905e-05  -14.161   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001989 on 2427 degrees of freedom
Multiple R-squared:  0.8595,	Adjusted R-squared:  0.859
F-statistic:  1856 on 8 and 2427 DF,  p-value: < 2.2e-16
```

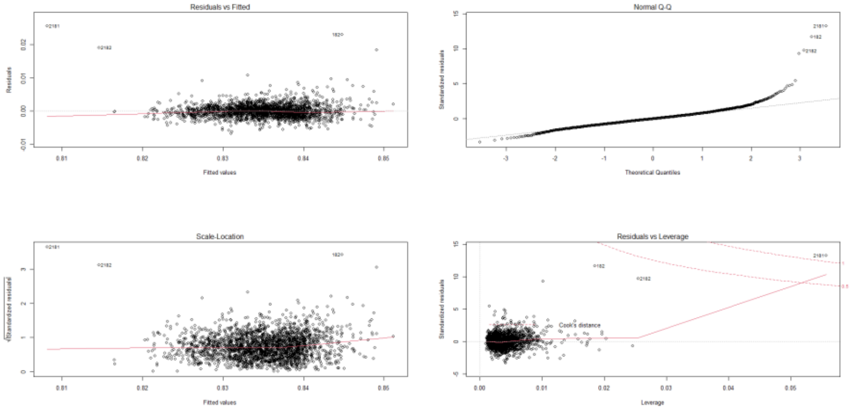
Figure 29: *Regression result of the modified model*

C. Conclusion

1. Model comparison

Between three models: The original model with full predictors (1st), the best model by transforming outcome to normality (4th), and the best subset model (5th), below are the comparison table with fundamental statistics and visualization.

Model	Number of predictors	Overall fit	Regression plot
The original model	21	83.44%	
The transformed model	15	87.05%	

The best subset model	8	85.9%	

Among the three, I would go with the best subset model, which is the upgraded version of the best transformed model of outcome normality. This is because this model removes 7 predictors from its former version, which still preserves several coefficients with significant p-value. Although the overall fit reduces slightly, collecting less data for a fewer predictor model would save money and resources. Especially, if we acquire less data, there is a higher chance that data will be more correct and thus eliminating the effect of outliers, which cause a tremendous harm to our work to define the best fit line.

2. Key takeaways

Our model is built through the process starting with explanatory data analysis to grab a good sense of the set at the first stage, then fitting the model and finally upgrading it to find the most useable one.

Firstly, it is recognizable that too many predictors could cause confusion when selecting the most powerful variables for the prediction, so the previous domain knowledge would help to narrow down the scope and mitigate the problem.

The second finding is that the data set naturally contains a lot of junk with missing values, outliers, unusual observations, and other issues. These problems distort the view and impose a heavy burden on the effort to update the model later on.

Fitting any full model is a primary step to define problems and thus for further improvement. Validating the underlying assumption and checking common issues of Ordinary Least Square are practical. However, it turns out that there are many concerns since the assumptions are violated by the presence of outliers and influenced points. Solutions are available to tackle these concerns, but each has its own advantage and disadvantage and there is no such an optimal solution for every case. Furthermore, we do not exclude outliers just because they do not match our model well. Removing them is always justified by sound arguments and information. Instead, we should first update our model to determine whether it has improved. In this report, I focus on updating the model only and leave outliers intact in the data set because it should be precautious to eliminate any observations.

The fourth advancement is that using automatic machine learning methods of variable selection such as AIC, BIC, Stepwise or Subset would be helpful but upgrading model by variable transformation would take time and be very challenging because it is about which variables should we transform, how to transform to fit regression but still avoid the overall fitting issue, how is the result of transformation, or transforming data could change the nature of the regression. I came up with at least 10 transformed models manually to see if they are improved and selected only 3 models in this report.

Finally, when comparing models, it is not merely about the overall fit, it should also take consideration of various aspects such as which model have less data but still be useful, the cost of acquire each variable and so on.

Bibliography

Kabacoff, R. I. (2015). *R in Action*.

Ames Housing.

Alvin T. Tan, P. D. (Nov 13, 2019). Cracking the Ames Housing Dataset with Linear Regression. https://towardsdatascience.com/wrangling-through-dataland-modeling-house-prices-in-ames-iowa-75b9b4086c96

Cotton, R. (Apr 24, 2012). *Remove all special characters from a string in R?* https://stackoverflow.com/questions/10294284/remove-all-special-characters-from-a-string-in-r

DataDaft. (Nov 4, 2019). *How to Make a Scatter Plot Matrix in R*. https://www.youtube.com/watch?v=AY9PYzJtCNA

Forst, J. *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Globe, S. (Sep 30, 2020). *R Replace Values in Data Frame Conditionally (4 Examples) | Exchange Value in Column & Entire Matrix*. https://www.youtube.com/watch?v=9c4tVegFWk0

Guinness, J. (Feb 22, 2021). *Multiple Linear Regression in R - Ames Housing Data*. https://www.youtube.com/watch?v=WtwMj9PakF8

Kaplan, J. (2020-11-29). *Making dummy variables with dummy_cols()*. https://cran.r-project.org/web/packages/fastDummies/vignettes/making-dummy-variables.html

Leon Adams, D. S. (March 5, 2017). *Ames housing prediction*. http://rstudio-pubs-static.s3.amazonaws.com/256459_5a62c0ca6d5849af92607011bb6c3e1d.html

Malhotra, K. R. (Sep 27, 2018). *Linear regression: Modeling and Assumptions*. https://towardsdatascience.com/linear-regression-modeling-and-assumptions-dcd7a201502a

O'Leary, M. (Apr 3 '12). *Why would R return NA as a lm() coefficient?* https://stats.stackexchange.com/questions/25804/why-would-r-return-na-as-a-lm-coefficient

Rajdev, D. (Jul 6 2017). *Outlier detection using outlierTest function*. https://stats.stackexchange.com/questions/288910/outlier-detection-using-outliertest-function

RPubs. https://www.rpubs.com/prakharprasad/511734

RPubs. https://rpubs.com/charliesangel/AmesHousing

University, N.). *ALY6015_M01_Regression Diagnostics and Feature Selection_Lab*. https://northeastern.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5bab984f-1923-49d0-8d0d-ac94003dfc7c&start=2460.885521

ZACH. (NOVEMBER 25, 2020). *How to Interpret a Scale-Location Plot (With Examples)*. How to Interpret a Scale-Location Plot (With Examples)

Niemann-Ross, M. (June 12th 2018). *Code Clinic: R*. https://www.linkedin.com/learning/code-clinic-r-2018/welcome?u=74653650

rstudio-pubs-static. *Regression Analysis of IMDB 5000 Movies Datasets*. https://rstudio-pubs-static.s3.amazonaws.com/281788_ba06442931084c42aeafe0cce17785c5.html

Hadley Wickham, G. G. *R for Data Science - 15 Factors*. https://r4ds.had.co.nz/factors.html

Prabhakaran, S. Top 50 ggplot2 Visualizations - The Master List (With Full R Code). http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

Moon, K.-W. (Jan 26 2020). *Plot for distribution of common statistics and p-value*. https://cran.r-project.org/web/packages/webr/vignettes/plot-htest.html

Changyong FENG, H. W., Naiji LU, Tian CHEN, Hua HE, Ying LU,Xin M. TU. (Apr 26 2014). Log-transformation and its implications for data analysis. 105–109. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/

datamentor.io. *R Box Plot*. https://www.datamentor.io/r-programming/box-plot/

Rdocumentation.org. *hist: Histograms*. https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/hist

Singh, A. (Feb 14, 2021). *Density plot with ggplot2*. https://abhishekstats.com/2021/02/14/density-plot-with-ggplot2/

Bun, B. (Apr 23, 2020). Variable selection Part 2/3: Regression with Stepwise Procedure using R. https://www.youtube.com/watch?v=4s_3L9ssO0w&t=1s

Caughlin, D. (May 3, 2020). *Multiple Linear Regression in R*. https://www.youtube.com/watch?v=zyEZop-5K9Q&t=1359s

DataExplained. (Jun 21, 2020). *Multiple Linear Regression using R (All about it)*. https://www.youtube.com/watch?v=_ymR-FFG44c&t=1236s

Dataslice. (Jun 14, 2020). *Interpreting Linear Regression Output in R*. https://www.youtube.com/watch?v=7WPfuHLCn_k

Introspective-Mode. (May 13, 2016). *Data Assumption: Multicollinearity*. https://www.introspective-mode.org/assumption-multicollinearity/

jmp. *Regression Model Assumptions*. https://www.jmp.com/en_ca/statistics-knowledge-portal/what-is-regression/simple-linear-regression-assumptions.html

Keith, T. Z. (January 25, 2019). *Multiple Regression and Beyond*.

Kubrick, R. (Apr 7 2012). *How to deal with multicollinearity when performing variable selection?* https://stats.stackexchange.com/questions/25611/how-to-deal-with-multicollinearity-when-performing-variable-selection

mts. (August 09 2015). *ggplot2: Plotting regression lines with different intercepts but with same slope*. https://stackoverflow.com/questions/31903606/ggplot2-plotting-regression-lines-with-different-intercepts-but-with-same-slope

Rai, D. B. (May 2, 2015). Multiple Linear Regression in R. https://www.youtube.com/watch?v=S-zKhFr91Tg researchconsultation.com. *Identifying Multicollinearity in Multiple Regression* http://www.researchconsultation.com/multicollinearity-regression-spss-collinearity-diagnostics-vif.asp

r-tutor. *Residual Plot*. https://www.r-tutor.com/elementary-statistics/simple-linear-regression/residual-plot

Statistics, M.-R. P. (Jan 6, 2021). *3.6 Collinearity in R: Checking For Collinearity In R*. https://www.youtube.com/watch?v=6TUXW-p_bI4

ZACH. (MAY 25, 2021). *How to Interpret Pr(>|t|) in Regression Model Output in R*. https://www.statology.org/interpret-prt-regression-output-r/

Dataslice. (Jun 14, 2020). *Interpreting Linear Regression Output in R*. https://www.youtube.com/watch?v=7WPfuHLCn_k

jmp. *Regression Model Assumptions*. https://www.jmp.com/en_ca/statistics-knowledge-portal/what-is-regression/simple-linear-regression-assumptions.html

r-tutor. *Residual Plot*. https://www.r-tutor.com/elementary-statistics/simple-linear-regression/residual-plot