Ames Housing – Property Value Prediction Report

EDA and Regression Models

Singh, Gurjeet

This report contains analysis done using Exploratory Data Analysis (EDA) methods and various regression models built for the home sale price on Ames housing data to understand and provide estimates of home values for typical homes in Ames, Iowa.

Table of Contents

Table of Figures

Introduction:

The purpose of this exercise is to build statistical models i.e. regression model to predict the value of a property or home. For our purposes, we will use Exploratory Data Analysis (EDA) methods to explore different aspects of the data, build two Simple Linear Regression Models using two predictor variables i.e. Total SQFT and Total Bathroom, build one Multiple Linear Regression Models using the same two variables, and build regression models for the transformed response variable log(SalePrice). There will be total of six (6) different models. We will perform the diagnostic test to assess the goodness-of-fit of each model.

For this assignment, we use the data set that contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2016 to 2010. The data set contains 2930 observations and 82 explanatory variables which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables, and 2 additional observation identifiers.

Now that we have some context for our analysis and dataset, let's look at the results in the next section.

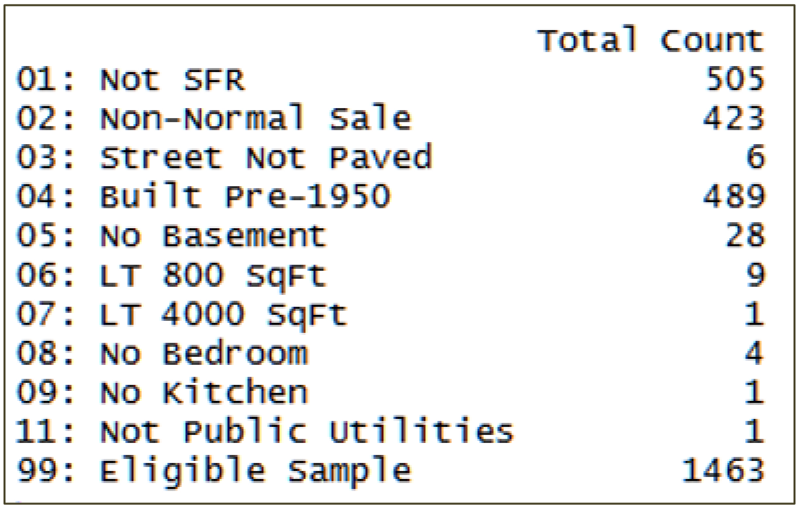
Results

Sample Definition:

For the purposes of the end goal, we have created a sample population. This eligible sample excludes the following kind of properties using the drop conditions. Figure 1 shows the drop conditions with the total count. Each of the drop conditions used to eliminate a property is explained below.

- Building type is not a single family.
- It not a normal sale
- Street are not paved
- Any house built prior to 1950
- There is no basement
- Living room square feet is below 800 and above 4000.
- There is no bedroom
- There is no kitchen
- There is no full bath
- There is no public utilities

Figure 1: Drop Conditions with count

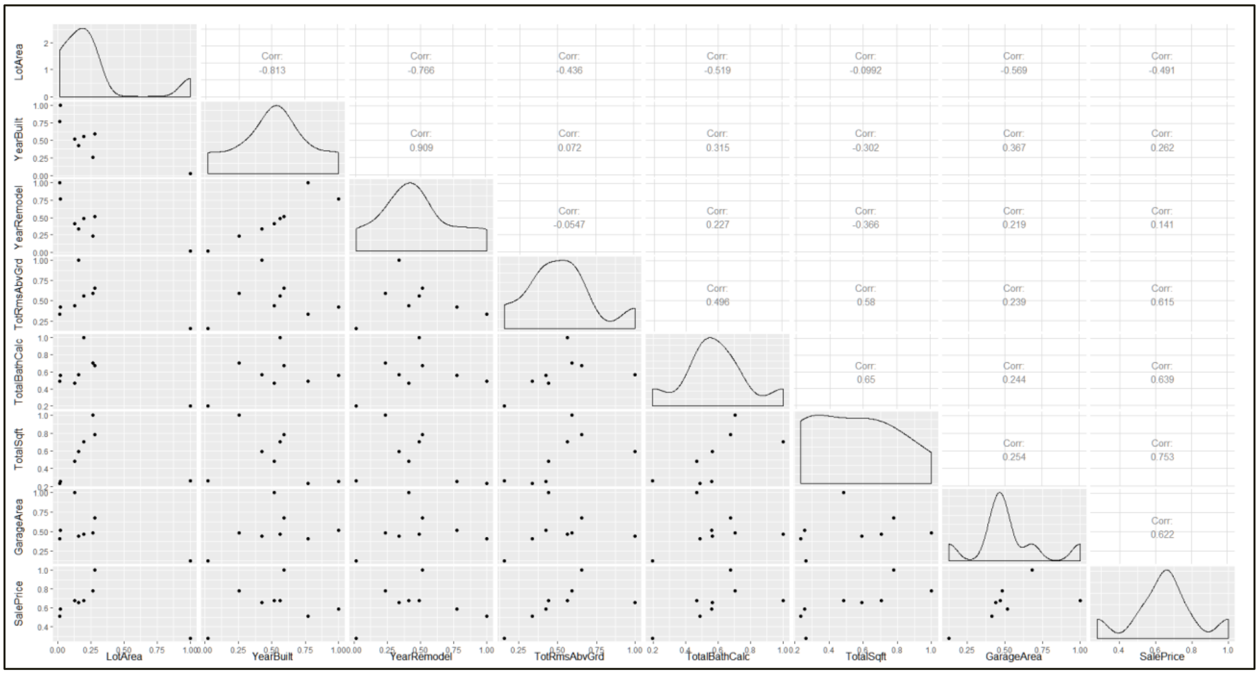


```
                          Total Count
01: Not SFR                       505
02: Non-Normal Sale               423
03: Street Not Paved                6
04: Built Pre-1950                489
05: No Basement                    28
06: LT 800 SqFt                     9
07: LT 4000 SqFt                    1
08: No Bedroom                      4
09: No Kitchen                      1
11: Not Public Utilities            1
99: Eligible Sample              1463
```

Exploratory Data Analysis

The first step towards our Exploratory Data Analysis (EDA) approach is to understand and analyze the data set to summarize the main characteristics of variables. For this purposes, we pick two (2) predictor variables i.e. TotalSQFT and TotalBathCalc to explore relationships between sale price and these predictor variables. The reason I picked these variables because both these variables have the highest correlation value between them and SalePrice.

Figure 2: Scatter plot matrix to select predictor variables



Continuous variables

Figure 3 shows the relationship between Total SQFT of the house and Sale Price. We see the positive strong relationship between Sale price and SQFT. As the SQFT increases, Sale Price of the house increases too. We do see that there are a few outliers on the top.

Figure 3: Scatter plot of Sale Price by SQFT

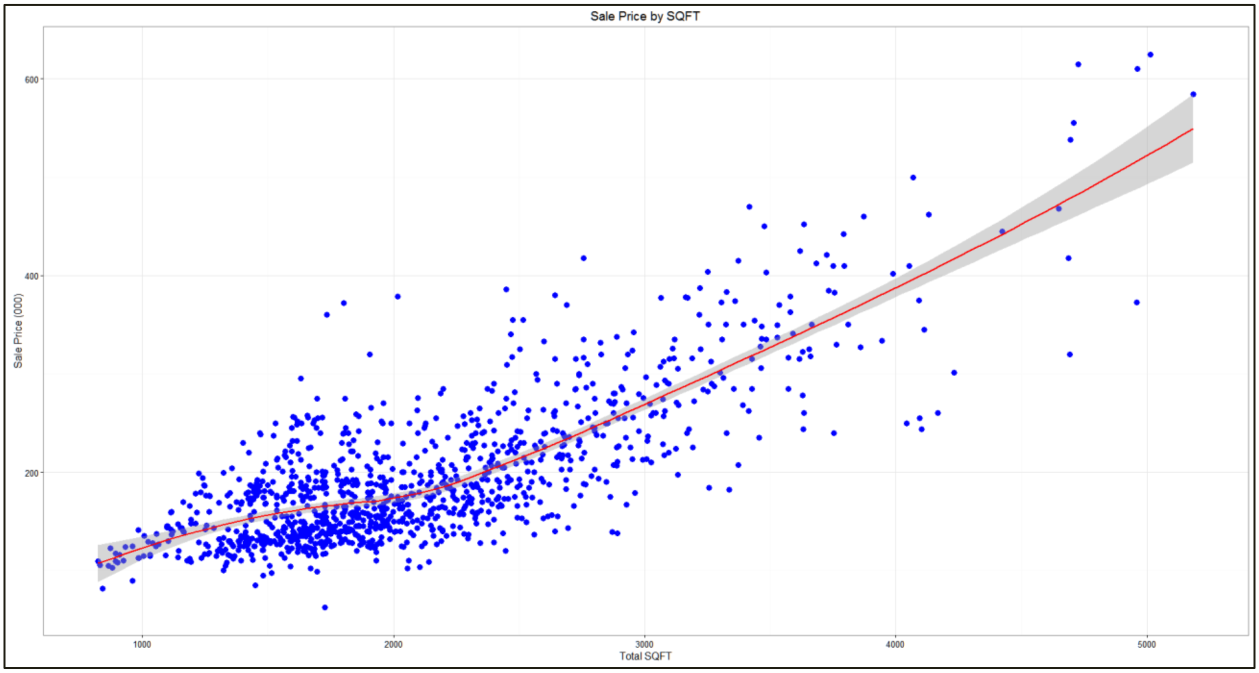
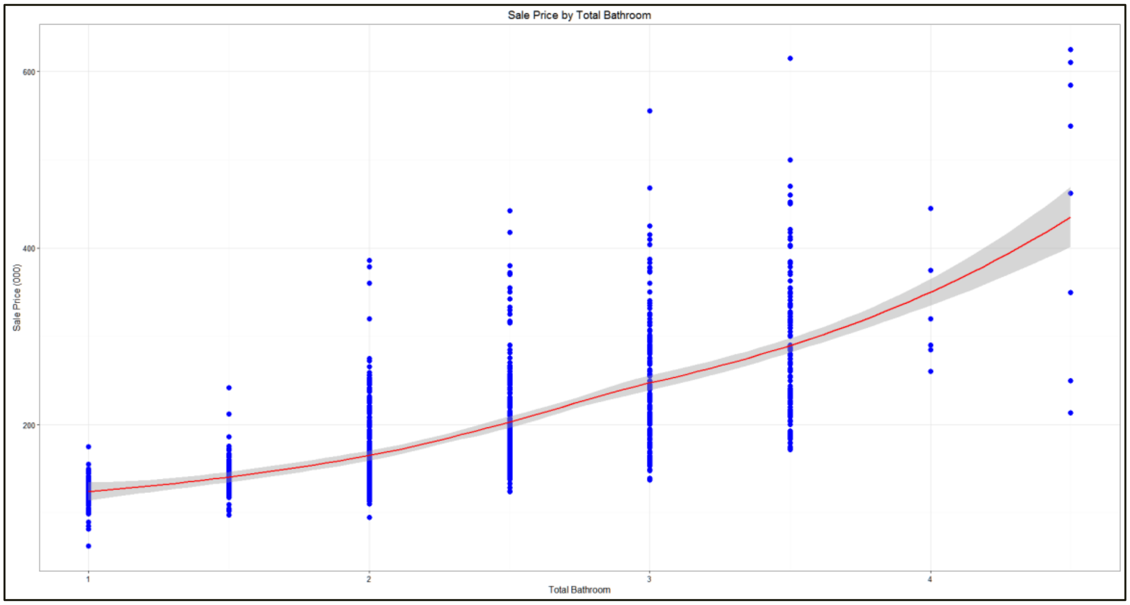


Figure 4 shows the relationship between Total Bathroom in the house and Sale Price. Again, we see the positive relationship between Sale price and Total Bathroom. As the number of bathrooms increases, Sale Price of the house increases too. Hence, Total Bathroom is also selected as the predictor variable.

Figure 4: Scatter plot of Sale Price by Total Bathroom



Simple Linear Regression Models

This section explains the two models that we have created based on their relationship with Sale Price.

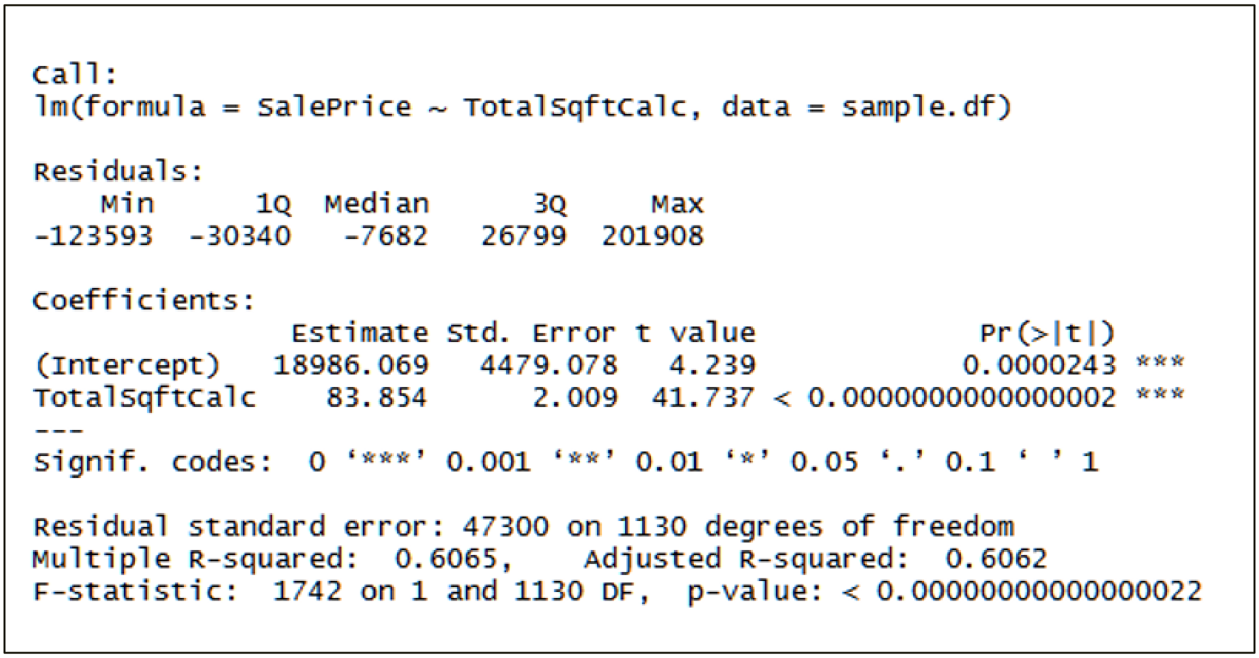
Model 1

The first simple linear regression model i.e. Model 1 is created using the predictor variable TotalSQFTCalc and response variable Sale Price.

Figure 5 shows the output of the model. Residuals in Figure 5 are essentially the difference between the actual observed response values and the response values that the model predicted. The Residuals section of the model output breaks it down into 5 summary points. When assessing how well the model fit the data, we should look for a symmetrical distribution across these points on the mean value zero (0). In our case, we can see that the distribution of the residuals do not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actual observed points. We further investigate this by plotting the residuals to see whether this normally distributed. Figure 6 shows the Q-Q plot of residuals and we see some evidence of non-normality.

*** significance stars in coefficients in Figure 5 shows high significane indicating that it's unlikely that no relationship exists between total square foot and sale price.

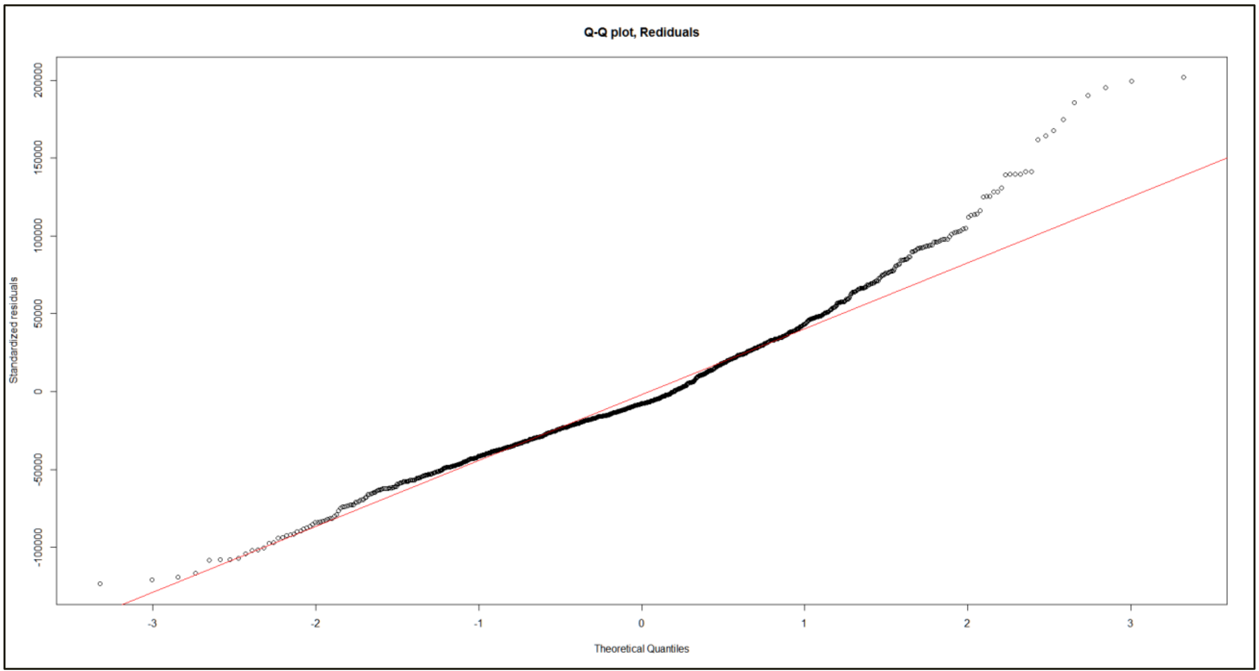
Figure 5: Model 1: Output



```
Call:
lm(formula = SalePrice ~ TotalSqftCalc, data = sample.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-123593  -30340   -7682   26799  201908 

Coefficients:
               Estimate Std. Error t value            Pr(>|t|)    
(Intercept)   18986.069   4479.078   4.239           0.0000243 ***
TotalSqftCalc    83.854      2.009  41.737 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47300 on 1130 degrees of freedom
Multiple R-squared:  0.6065,	Adjusted R-squared:  0.6062 
F-statistic:  1742 on 1 and 1130 DF,  p-value: < 0.00000000000000022
```

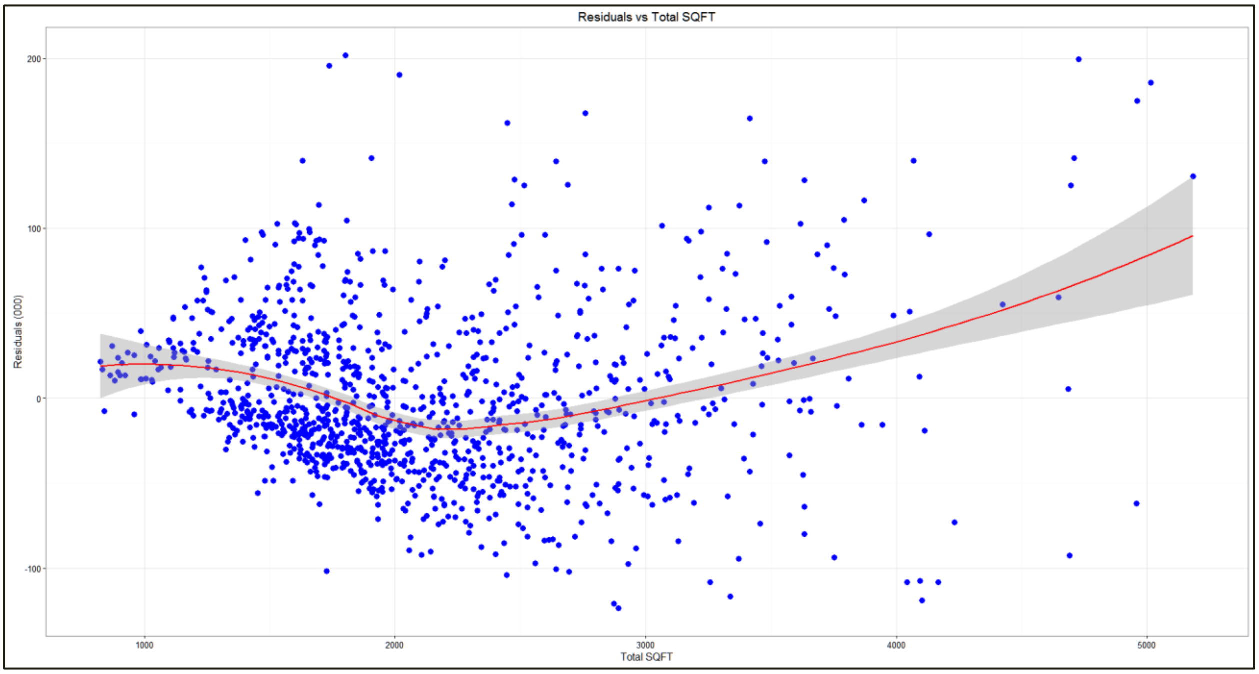
To assess the goodness-of-fit of this model, we produced a QQ-Plot, Figure 6, of the residuals to compare their distribution to the normal distribution. Figure 6 shows that model has some asymmetric distribution and evidence of non-normality.

Figure 6: Model 1: Q-Q plot for Residuals



Next we test for the second assumption to validate the homoscedasticity. Figure 7 shows the relationship between residual and Total SQFT. Again, we noticed that the distribution of the residuals do not appear to be strongly symmetrical. It is quite evident that why our model was predicting certain points that fall far away from the actual observed points in Figure 6. The structure in Figure 6 suggest that the model will need additional predictor variables.

Figure 7: Model 1: Scatterplot of Residuals vs Total SQFT



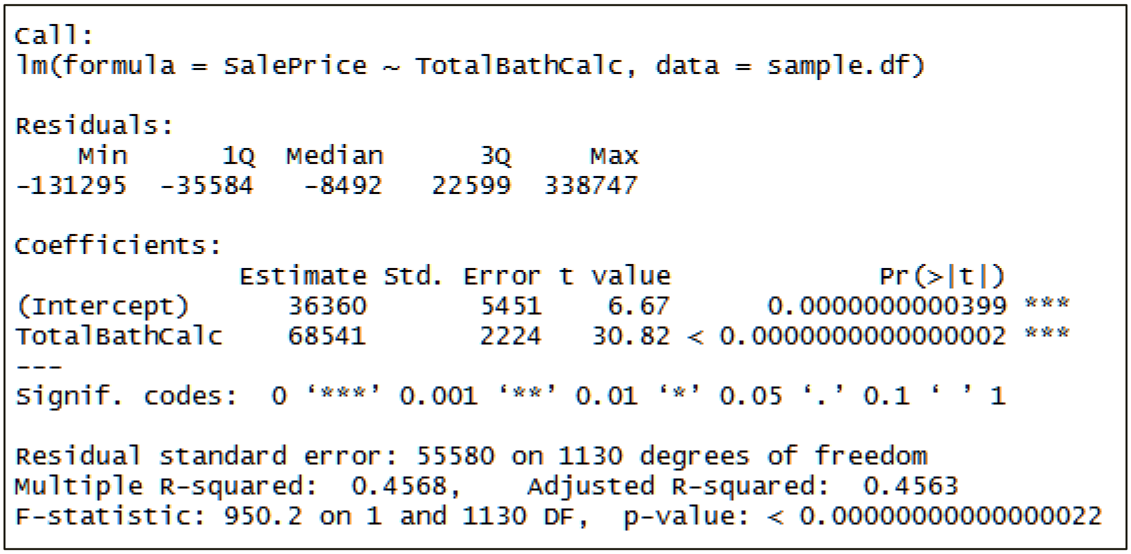
Next look at the R-Squared statistic to measure how well our model is fitting the actual data. The R-Squared measure the linear relationship between our predictor variable (TotalSQFTCalc) and our response variable (SalePrice). The R-Squared value in Figure 5 is **0.6065** which is roughly 60% (approx.) of the variance found in the response variable (SalePrice) can be explained by the predictor variable (TotalSQFTCalc). As we know that Total SQFT is one of the key important factor in price of house, therefore, we see a relatively strong R-Squared value. However, it hard to define what level of R-Squared is appropriate to claim that this model fits well. Therefore, we create another model with our next predictor variable.

Model 2

The second simple linear regression model i.e. Model 2 is created using the predictor variable TotalBathCalc and response variable Sale Price.

Figure 8 shows the output of the model. Residuals in Figure 8 again shows that the distribution of the residuals do not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actual observed points. We use Q-Q plot to further investigate by plotting the residuals to see whether this is normally distributed. Figure 9 shows the Q-Q plot of residuals and we again see some evidence of non-normality with this predictor variable as well.

Figure 8: Model 2: Output



```
Call:
lm(formula = SalePrice ~ TotalBathCalc, data = sample.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-131295  -35584   -8492   22599  338747 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept)      36360       5451    6.67      0.0000000000399 ***
TotalBathCalc    68541       2224   30.82 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55580 on 1130 degrees of freedom
Multiple R-squared:  0.4568,	Adjusted R-squared:  0.4563 
F-statistic: 950.2 on 1 and 1130 DF,  p-value: < 0.00000000000000022
```

To assess the goodness-of-fit of this model, we produced a QQ-Plot, Figure 9, of the residuals to compare their distribution to the normal distribution. Figure 9 shows that model has some asymmetric distribution and evidence of non-normality exists in this model as well.

Figure 9: Model 2: Q-Q plot for Residuals

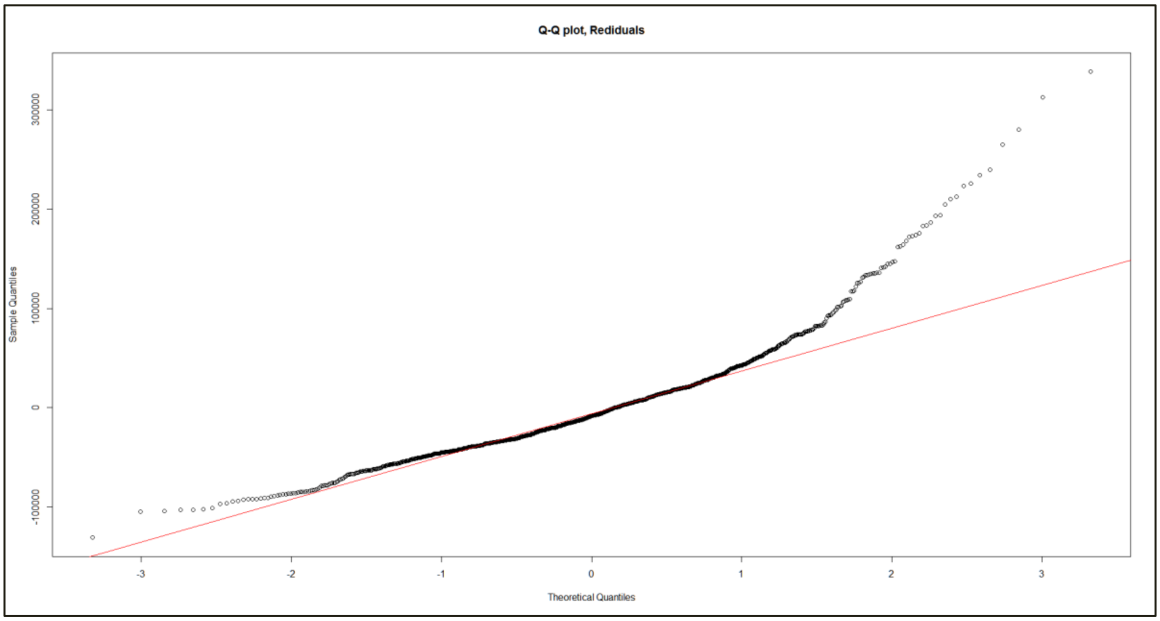
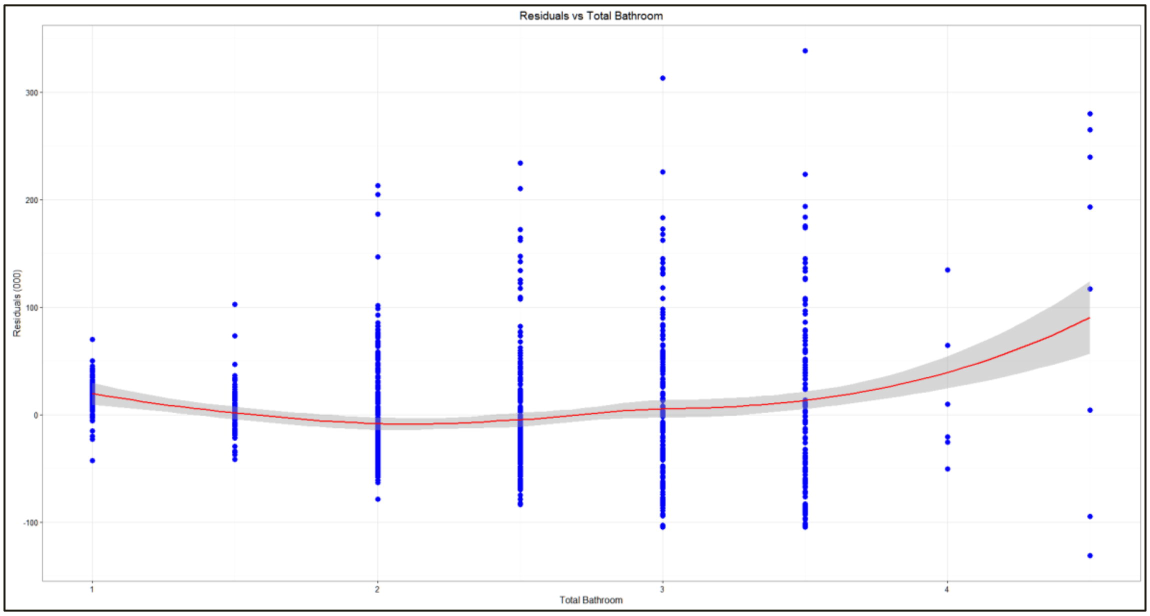


Figure 10 shows the relationship between residual and Total Bathroom in the house. Again, we noticed that the distribution of the residuals do not appear to be strongly symmetrical. It is quite evident that why our model was predicting certain points that fall far away from the actual observed points in Figure 8.

Figure 10: Model 2: Scatterplot of Residuals vs Total Bathroom



Lastly, we look at the R-Squared statistic to measure how well our model is fitting the actual data. The R-Squared value in Figure 8 is **0.4568** which is roughly 46% (approx.) of the variance found in the response variable (SalePrice) can be explained by the predictor variable (TotalBathCalc). We see that R-Squared value is not relatively strong as compared to Model 1. Hence, between Model 1 and Model 2, it seems like Model 1 fits better than Model 2.

Multiple Linear Regression Model

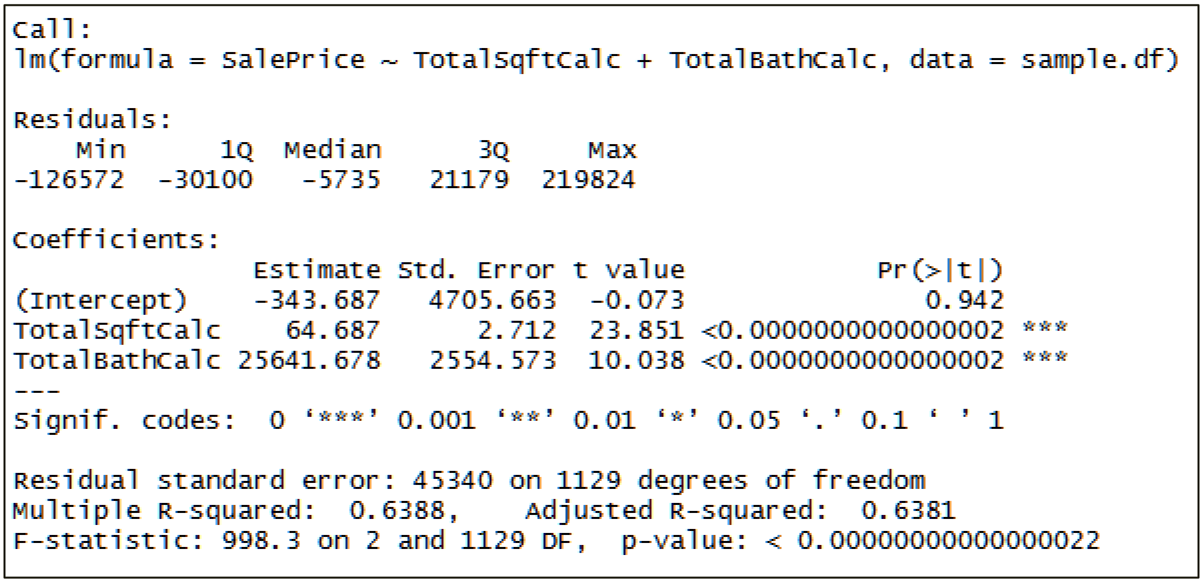
This section explains the Multiple Linear Regression Model using the two predictor variable i.e. TotalSQFTCalc and TotalBathCalc that we have selected based on their relationship with Sale Price.

Model 3

The third model is our multiple linear regression model i.e. Model 3 is created using the same predictor variables TotalSQFTCalc and TotalBathCalc used in Model 1 and Model 2 with the response variable SalePrice.

Figure 11 shows the output of the model. Residuals in Figure 11 are essentially the difference between the actual observed response values and the response values that the model predicted. In our case, we still see that the distribution of the residuals do not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actual observed points. We further investigate this by plotting the residuals to see whether this normally distributed. Figure 12 shows the Q-Q plot of residuals and we see some evidence of non-normality. However, there's has been a slight improvement in the R-Squared which we discussed more in detail while assessing the goodness-of-fit.

Figure 11: Model 3: Output



```
Call:
lm(formula = SalePrice ~ TotalSqftCalc + TotalBathCalc, data = sample.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-126572  -30100   -5735   21179  219824 

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)    
(Intercept)    -343.687   4705.663  -0.073                0.942    
TotalSqftCalc    64.687      2.712  23.851 <0.0000000000000002 ***
TotalBathCalc 25641.678   2554.573  10.038 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45340 on 1129 degrees of freedom
Multiple R-squared:  0.6388,	Adjusted R-squared:  0.6381 
F-statistic: 998.3 on 2 and 1129 DF,  p-value: < 0.00000000000000022
```

To assess the goodness-of-fit of this model, we produced a QQ-Plot, Figure 12, of the residuals to compare their distribution to the normal distribution. Figure 12 shows that model has some asymmetric distribution and evidence of non-normality exists in this model as well. However, there's definitely some improvement as compared to Model 1 and Model 2.

Figure 12: Model 3: Q-Q plot for Residuals

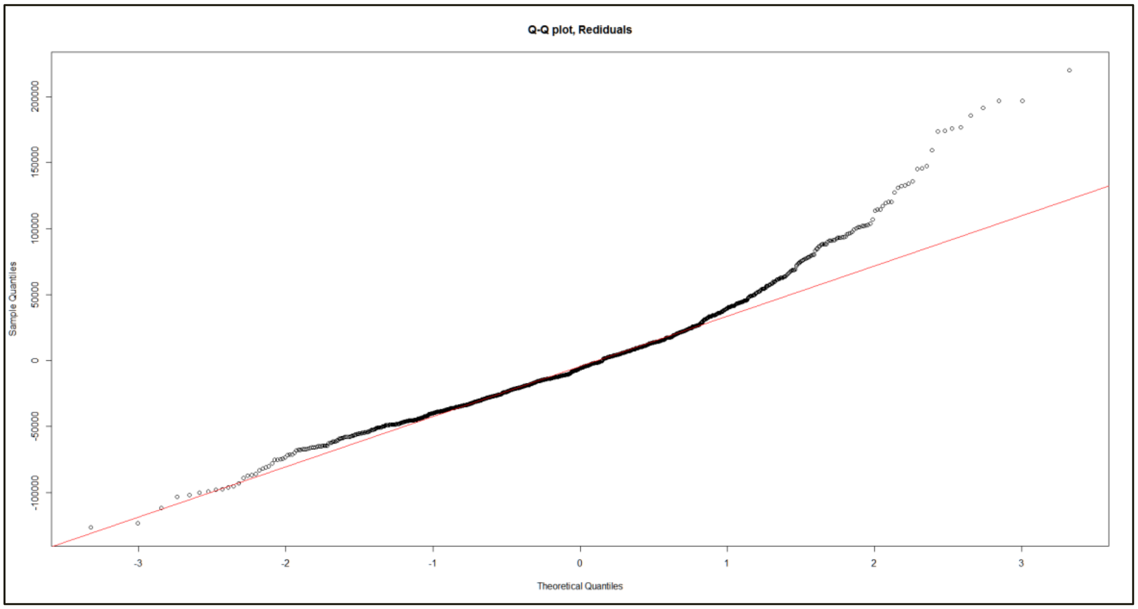
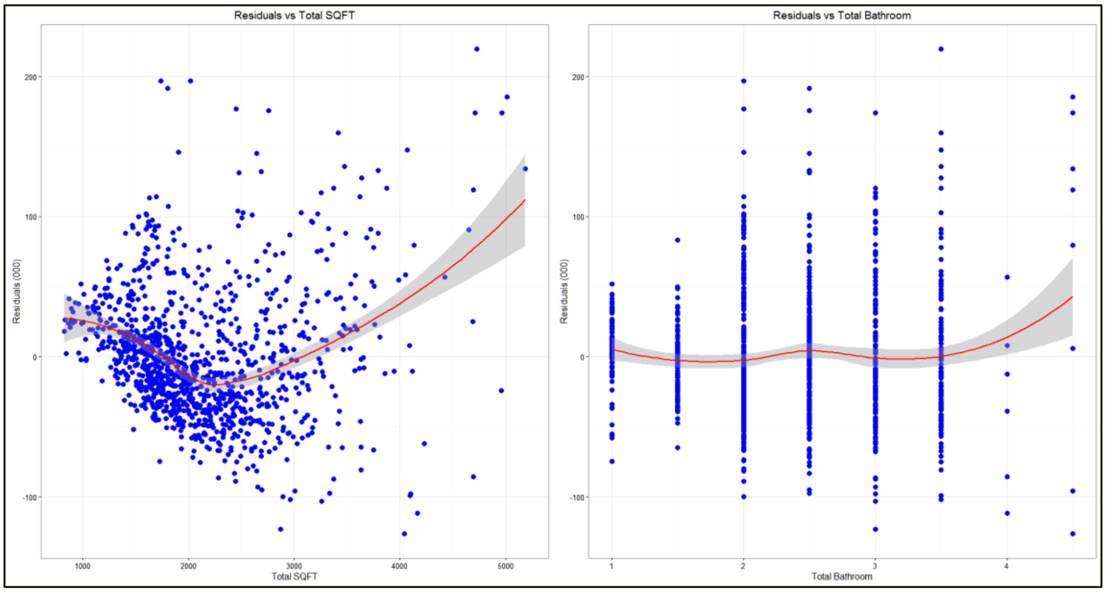


Figure 13 shows the side by side comparison of the relationship between Model 3 residual and Total SQFT and Total Bathroom. Again, we noticed that the distribution of the residuals do not appear to be strongly symmetrical. It is quite evident that why our model was predicting certain points that fall far away from the actual observed points in Figure 11. The structure in Figure 13 suggest that the model will need transformation or additional predictor variables.

Figure 13: Model 3: Side by Side - Scatterplot of Residuals vs Total SQFT and Total Bathroom



Lastly, we next look at the R-Squared statistic to measure how well our model is fitting the actual data. The R-Squared measure the linear relationship between our predictor variables (TotalSQFTCalc and TotalBathCalc) and our response variable (SalePrice). The R-Squared we got is **0.6388** which is roughly 64% (approx.) of the variance found in the response variable (SalePrice) can be explained by the predictor variable (TotalSQFTCalc and TotalBathCalc). R-Squared statistic has improved in the Multiple Linear Regression model and have a relatively strong R-Squared value. Therefore, based on R-Squared statistic and other conditions such as QQ plot and scatter plot, it seem like the multiple linear regression model fit better with these two predictior variables than each of the simple linear regression models.

Neighborhood Accuracy

Figure 14 shows the boxplot of the Multiple Linear Model i.e. Model 3's residuals by the neighborhood. Figure 14 shows that neighborhoods like Clear Creek, College Creek, Gilbert, Old Town, and Sawyer West better fit by the model. Figure 14 also shows that certain neighborhoods are consistently over predicted such as Northridge Heights, Northridge, Somerset, and Stone Brook. As well, there are neighborhoods that are under predicted such as Brookside, Iowa DOT and Rail Road, South & West of Iowa State University, and Veenker.

Figure 14: Boxplot of Residuals by Neighborhood

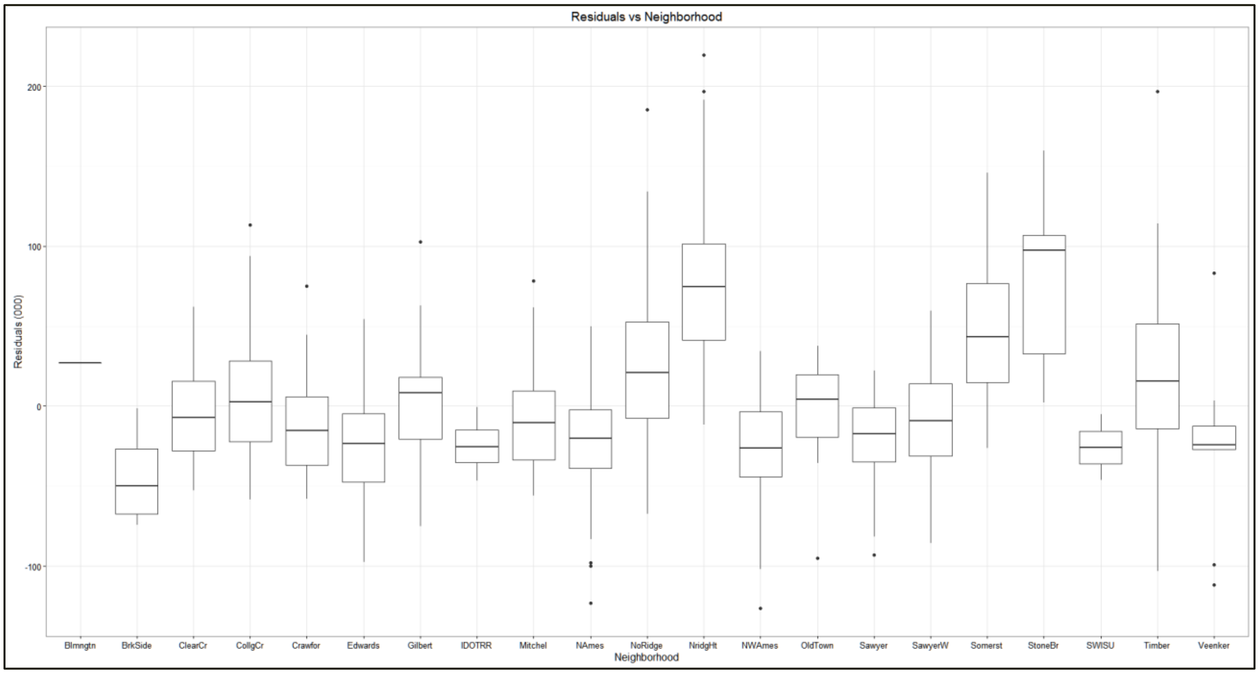
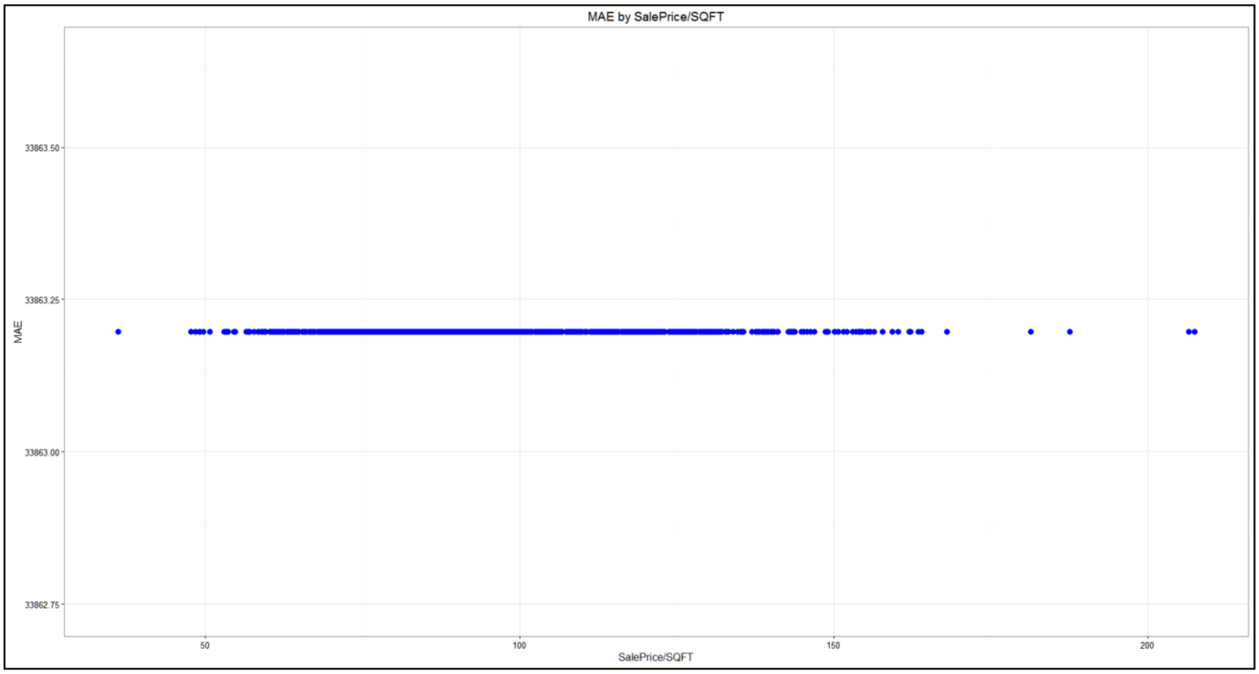


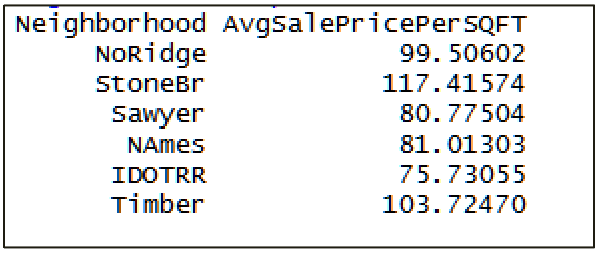
Figure 15 shows the scatter plot of Mean Absolute Error (MAE) calculated from Model 3 by sale price per SQFT. Figure 15 shows that there's no relationship between these two quantities.

Figure 15: Scatter plot of Sale Price per SQFT by MAE of Model 3



For the neighborhood accuracy, I have created 6 groups based on mean price per square foot of each neighborhood. These groups are shown in Figure 16. I used these groups to create indicator variables to include in the multiple regression model that we created in Model 3. For the base category, I have selected North Ames (NAmes) neighborhood.

Figure 16: Neighborhood Groups

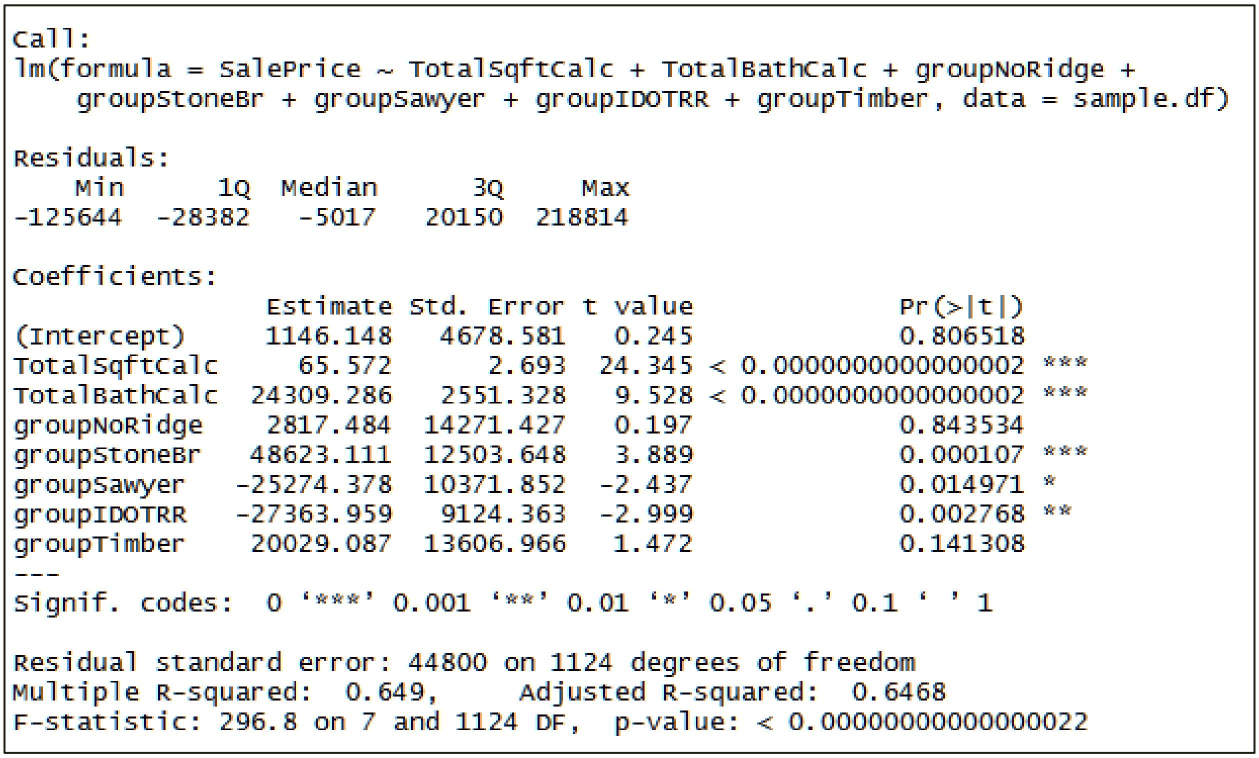


```
Neighborhood AvgSalePricePerSQFT
     NoRidge            99.50602
     StoneBr           117.41574
      Sawyer            80.77504
       NAmes            81.01303
      IDOTRR            75.73055
      Timber           103.72470
```

Model 4

Using these indicatory variables, we created a new multiple regression mode i.e. Model 4. The output is listed in Figure 17. We see that the R-Squared value is **0.649** which is roughly 65% (approx.) has increased a little indicating that indicator variables helped a little to better fit this model.

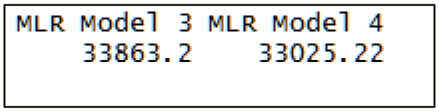
Figure 17: Model 4: Output



```
Call:
lm(formula = SalePrice ~ TotalSqftCalc + TotalBathCalc + groupNoRidge +
    groupStoneBr + groupSawyer + groupIDOTRR + groupTimber, data = sample.df)

Residuals:
    Min      1Q  Median      3Q     Max
-125644  -28382   -5017   20150  218814

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)     1146.148   4678.581   0.245             0.806518
TotalSqftCalc     65.572      2.693  24.345 < 0.0000000000000002 ***
TotalBathCalc  24309.286   2551.328   9.528 < 0.0000000000000002 ***
groupNoRidge    2817.484  14271.427   0.197             0.843534
groupStoneBr   48623.111  12503.648   3.889             0.000107 ***
groupSawyer   -25274.378  10371.852  -2.437             0.014971 *
groupIDOTRR   -27363.959   9124.363  -2.999             0.002768 **
groupTimber    20029.087  13606.966   1.472             0.141308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44800 on 1124 degrees of freedom
Multiple R-squared:  0.649,	Adjusted R-squared:  0.6468
F-statistic: 296.8 on 7 and 1124 DF,  p-value: < 0.00000000000000022
```

Figure 18 shows the Mean Absolute Error (MAE) comparison between Model 3 and Model 4. It is quite evident here that Model 4, which includes the indicator variables, better fit between the two.

Figure 18: MAE Results



```
MLR Model 3 MLR Model 4
    33863.2    33025.22
```

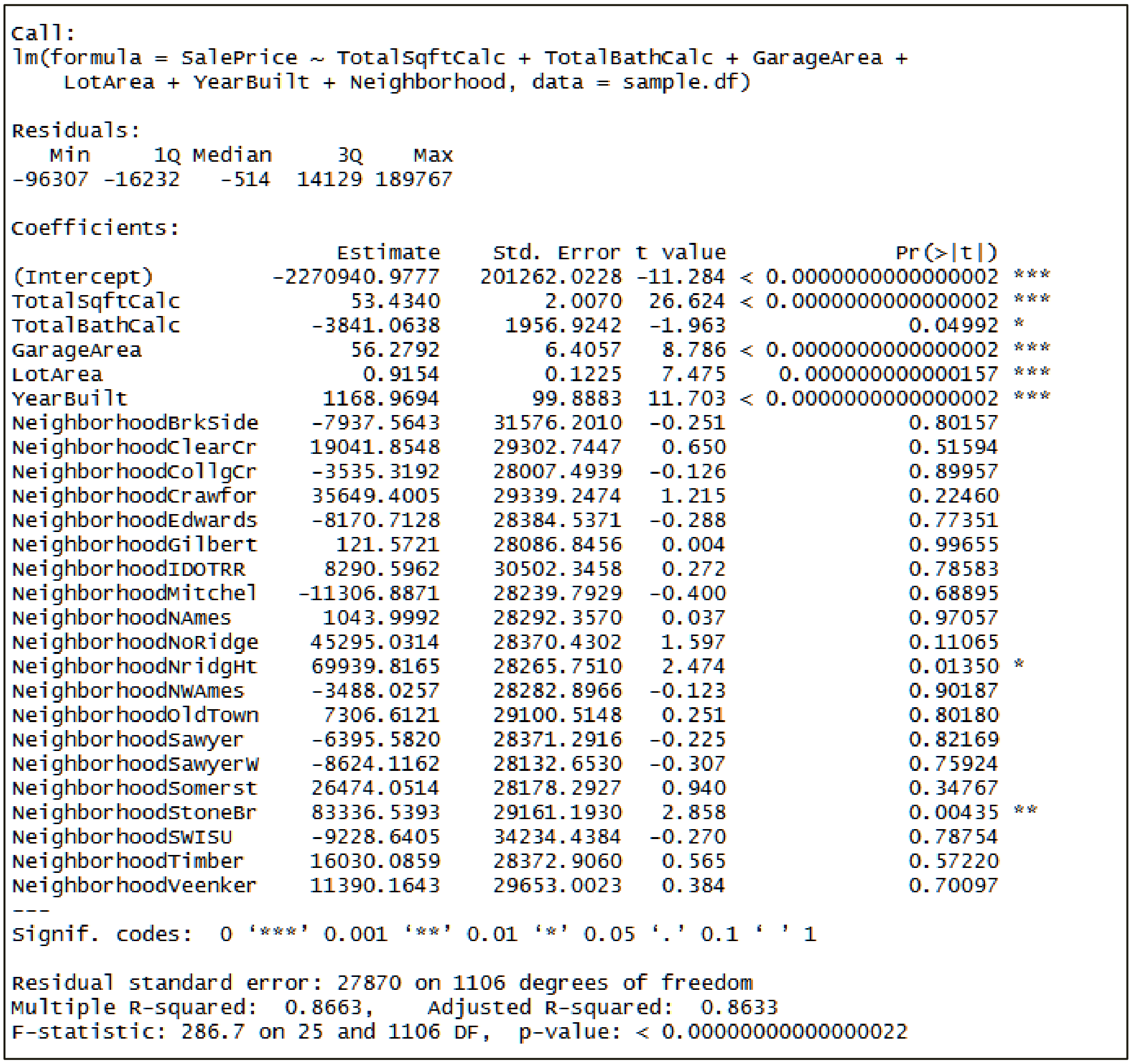
SalePrice versus Log SalePrice as the Response

This section explains the 2 models. One based on the SalePrice and the other one with the transformed response log(SalesPrice). These response variables are used against the four (4) continuous predictor variables, one (1) discrete variable, and one (1) nominal variable for each model. These predictor variables are selected based on their correlation values in Figure 2.

Model 5

Figure 19 shows the output of the Model 5. We notice that R-Squared value has tremendously increased. As it was suggested earlier that model may need a transformation or additional predictor variables, it seems like added more predictor variables definitely made this model better.

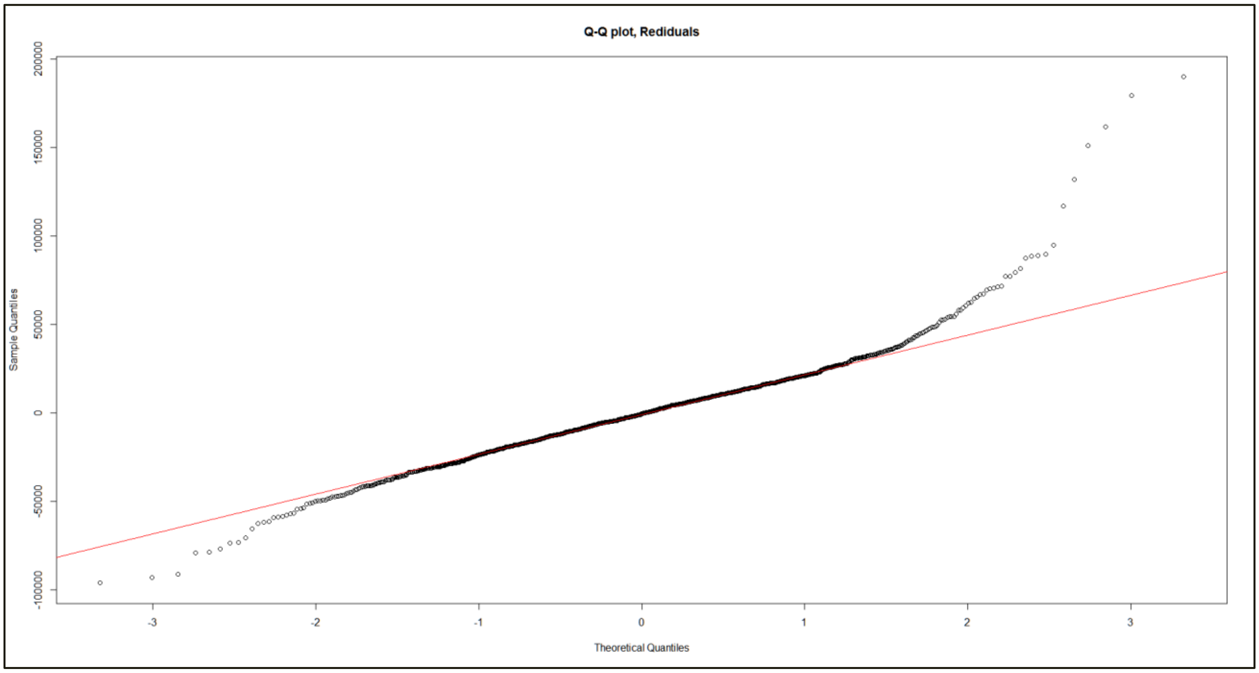
Figure 19: Model 5: Output



```
Call:
lm(formula = SalePrice ~ TotalSqftCalc + TotalBathCalc + GarageArea +
    LotArea + YearBuilt + Neighborhood, data = sample.df)

Residuals:
   Min     1Q Median     3Q    Max
-96307 -16232   -514  14129 189767

Coefficients:
                         Estimate    Std. Error t value              Pr(>|t|)
(Intercept)          -2270940.9777  201262.0228 -11.284 < 0.0000000000000002 ***
TotalSqftCalc              53.4340       2.0070  26.624 < 0.0000000000000002 ***
TotalBathCalc           -3841.0638    1956.9242  -1.963               0.04992 *
GarageArea                 56.2792       6.4057   8.786 < 0.0000000000000002 ***
LotArea                     0.9154       0.1225   7.475   0.000000000000157 ***
YearBuilt                1168.9694      99.8883  11.703 < 0.0000000000000002 ***
NeighborhoodBrkSide     -7937.5643   31576.2010  -0.251               0.80157
NeighborhoodClearCr     19041.8548   29302.7447   0.650               0.51594
NeighborhoodCollgCr     -3535.3192   28007.4939  -0.126               0.89957
NeighborhoodCrawfor     35649.4005   29339.2474   1.215               0.22460
NeighborhoodEdwards     -8170.7128   28384.5371  -0.288               0.77351
NeighborhoodGilbert       121.5721   28086.8456   0.004               0.99655
NeighborhoodIDOTRR       8290.5962   30502.3458   0.272               0.78583
NeighborhoodMitchel    -11306.8871   28239.7929  -0.400               0.68895
NeighborhoodNAmes        1043.9992   28292.3570   0.037               0.97057
NeighborhoodNoRidge     45295.0314   28370.4302   1.597               0.11065
NeighborhoodNridgHt     69939.8165   28265.7510   2.474               0.01350 *
NeighborhoodNWAmes      -3488.0257   28282.8966  -0.123               0.90187
NeighborhoodOldTown      7306.6121   29100.5148   0.251               0.80180
NeighborhoodSawyer      -6395.5820   28371.2916  -0.225               0.82169
NeighborhoodSawyerW     -8624.1162   28132.6530  -0.307               0.75924
NeighborhoodSomerst     26474.0514   28178.2927   0.940               0.34767
NeighborhoodStoneBr     83336.5393   29161.1930   2.858               0.00435 **
NeighborhoodSWISU       -9228.6405   34234.4384  -0.270               0.78754
NeighborhoodTimber      16030.0859   28372.9060   0.565               0.57220
NeighborhoodVeenker     11390.1643   29653.0023   0.384               0.70097
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27870 on 1106 degrees of freedom
Multiple R-squared:  0.8663,    Adjusted R-squared:  0.8633
F-statistic: 286.7 on 25 and 1106 DF,  p-value: < 0.00000000000000022
```

To assess the goodness-of-fit of this model, we next look at the R-Squared statistic to measure how well our model is fitting the actual data. The R-Squared we got is **0.8663** which is roughly 87% (approx.) of the variance found in the response variable (SalePrice) can be explained by the predictor variables. Even though the R-Squared value is high, our Q-Q plot suggest that residuals are not normally distributed and have some asymmetric distribution. Hence, evidence of non-normality exists in this model as well.

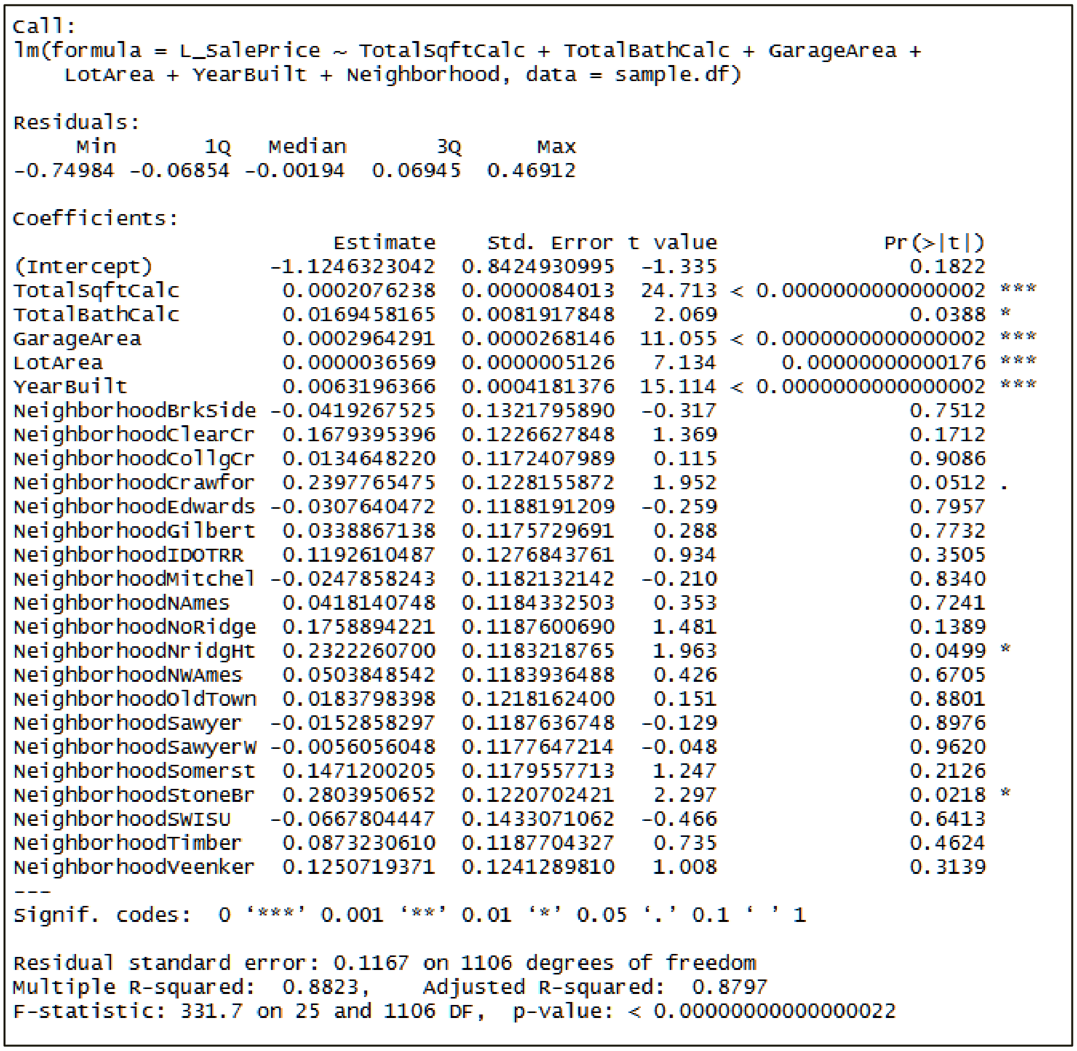
Figure 20: Model 5: Q-Q plot for Residuals



Model 6

Figure 21 shows the output of the model. Residuals in Figure 21 show that the distribution of the transformed residuals appear to be strongly symmetrical. That means that the model predicts certain points are not far away from the actual observed points. We further investigate this by plotting the residuals to see whether this normally distributed. Figure 22 shows the Q-Q plot of residuals and we see that points are now aligned on the line. This shows that transformation really helped in removing the non-normality.

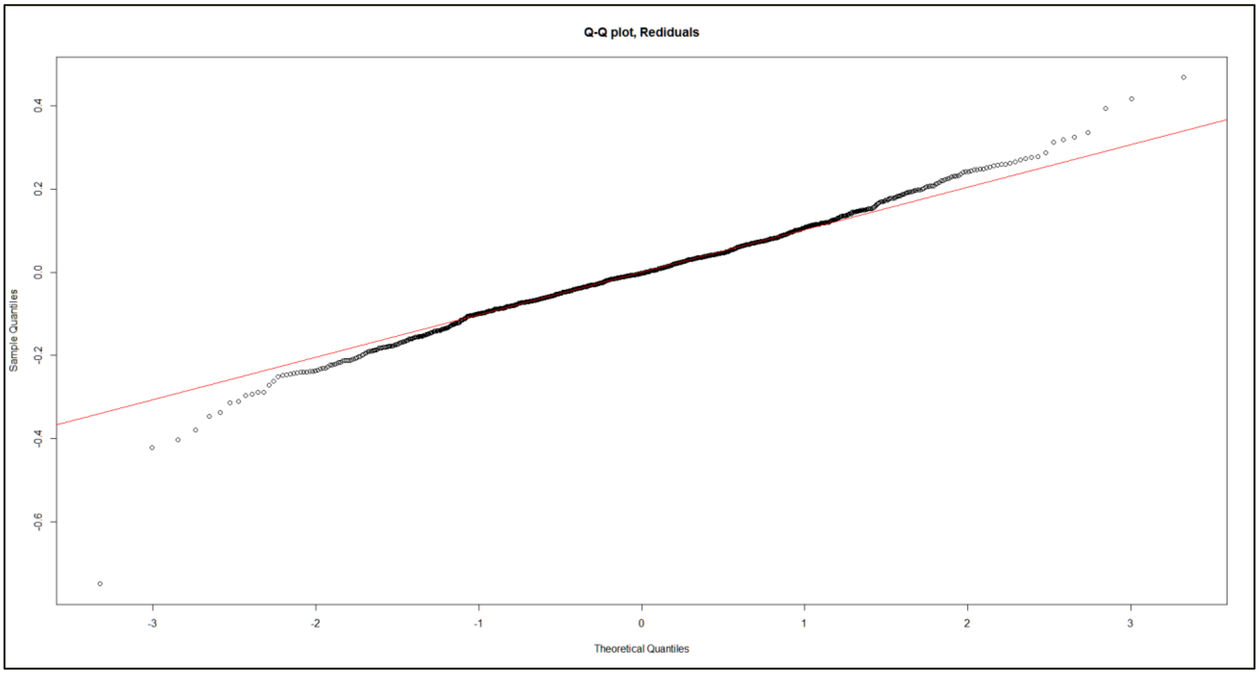
Figure 21: Model 5: Output



```
Call:
lm(formula = L_SalePrice ~ TotalSqftCalc + TotalBathCalc + GarageArea +
    LotArea + YearBuilt + Neighborhood, data = sample.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.74984 -0.06854 -0.00194  0.06945  0.46912 

Coefficients:
                        Estimate    Std. Error t value             Pr(>|t|)    
(Intercept)         -1.1246323042  0.8424930995  -1.335               0.1822    
TotalSqftCalc        0.0002076238  0.0000084013  24.713 < 0.0000000000000002 ***
TotalBathCalc        0.0169458165  0.0081917848   2.069               0.0388 *  
GarageArea           0.0002964291  0.0000268146  11.055 < 0.0000000000000002 ***
LotArea              0.0000036569  0.0000005126   7.134     0.00000000000176 ***
YearBuilt            0.0063196366  0.0004181376  15.114 < 0.0000000000000002 ***
NeighborhoodBrkSide -0.0419267525  0.1321795890  -0.317               0.7512    
NeighborhoodClearCr  0.1679395396  0.1226627848   1.369               0.1712    
NeighborhoodCollgCr  0.0134648220  0.1172407989   0.115               0.9086    
NeighborhoodCrawfor  0.2397765475  0.1228155872   1.952               0.0512 .  
NeighborhoodEdwards -0.0307640472  0.1188191209  -0.259               0.7957    
NeighborhoodGilbert  0.0338867138  0.1175729691   0.288               0.7732    
NeighborhoodIDOTRR   0.1192610487  0.1276843761   0.934               0.3505    
NeighborhoodMitchel -0.0247858243  0.1182132142  -0.210               0.8340    
NeighborhoodNAmes    0.0418140748  0.1184332503   0.353               0.7241    
NeighborhoodNoRidge  0.1758894221  0.1187600690   1.481               0.1389    
NeighborhoodNridgHt  0.2322260700  0.1183218765   1.963               0.0499 *  
NeighborhoodNWAmes   0.0503848542  0.1183936488   0.426               0.6705    
NeighborhoodOldTown  0.0183798398  0.1218162400   0.151               0.8801    
NeighborhoodSawyer  -0.0152858297  0.1187636748  -0.129               0.8976    
NeighborhoodSawyerW -0.0056056048  0.1177647214  -0.048               0.9620    
NeighborhoodSomerst  0.1471200205  0.1179557713   1.247               0.2126    
NeighborhoodStoneBr  0.2803950652  0.1220702421   2.297               0.0218 *  
NeighborhoodSWISU   -0.0667804447  0.1433071062  -0.466               0.6413    
NeighborhoodTimber   0.0873230610  0.1187704327   0.735               0.4624    
NeighborhoodVeenker  0.1250719371  0.1241289810   1.008               0.3139    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1167 on 1106 degrees of freedom
Multiple R-squared:  0.8823,	Adjusted R-squared:  0.8797 
F-statistic: 331.7 on 25 and 1106 DF,  p-value: < 0.00000000000000022
```

To assess the goodness-of-fit of this model, we next look at the R-Squared statistic to measure how well our model is fitting the actual data. The R-Squared we got is **0.8823** which is roughly 88% (approx.) of the variance found in the response variable (SalePrice) can be explained by the predictor variables. Even though our Q-Q plot in Figure 22 suggest that residuals are not normally distributed, we see that the R-Squared value is relatively strong as compared to Model 5.

Figure 22: Model 6: Q-Q plot for Residuals



Comparison of Model 5 and Model 6

Based on the results, it obvious that Model 6 i.e. with transformed log(SalePrice) is better compared to Model 5 i.e. without transformation. R-Squared value is higher for Model 6 as compared to Model 5 and both models satisfies the other goodness-of-fit (GOF) conditions. At this point, I do not believe that we need to consider transforming any predictor variable.

Figure 23 shows how transformation in Model 6 has distributed the residuals better as compared to Model 5. Hence, Model 6 better fits as compared to Model 5.

Figure 23: Model 5 vs Model 6: Q-Q plot for Residuals

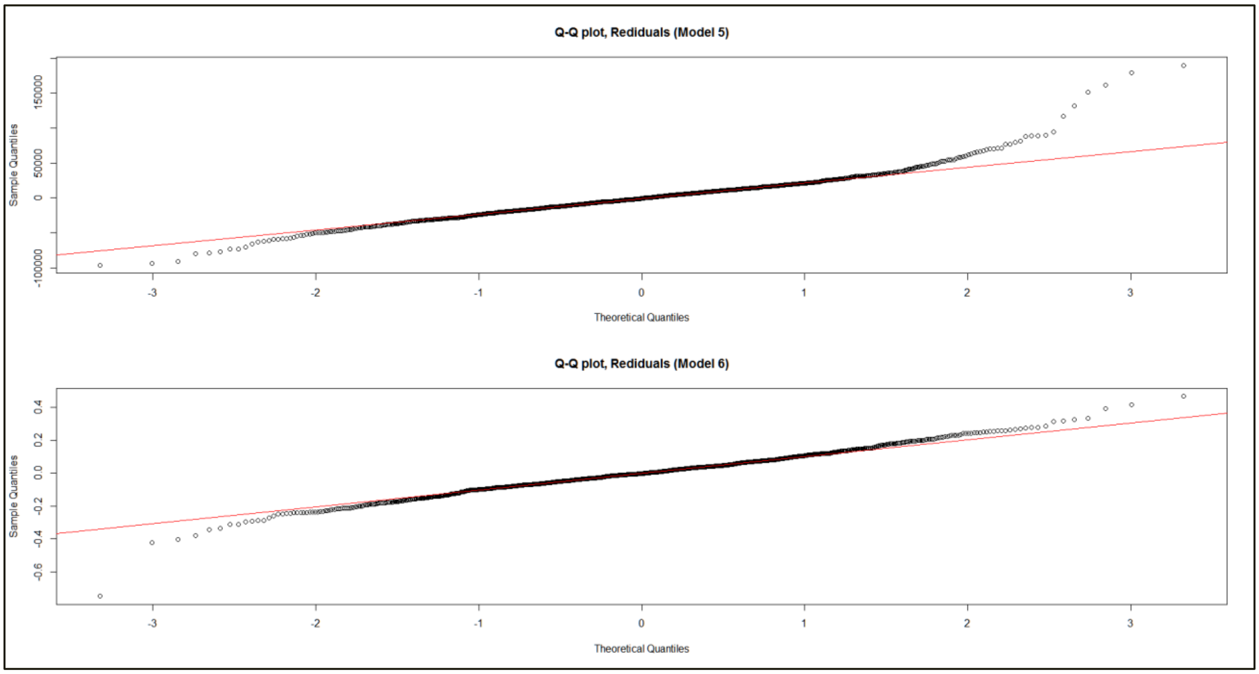
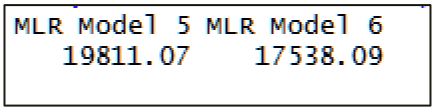


Figure 24 shows the Mean Absolute Error (MAE) comparison between Model 5 and Model 6. It is quite evident here that again Model 6 fit better between the two.

Figure 24: Model 5 vs Model 6: MAE



```
MLR Model 5 MLR Model 6
   19811.07    17538.09
```

Conclusion:

In conclusion, I would like to state that transformation of SalePrice to Log(SalePrice) as well as adding more predictor variables to the model really improved models. However, based on the QQ-plots normality assumption, the model still is not the best.