Introduction:

In Assignment #3, we will build off of Assignment #2 and develop predictive models. We will study models using categorical variables, continuous variables and a combination of both. We will also look at the effect of dummy variables and log transformations.

Section 1: Sample Definition

The Ames Housing dataset is comprised of 82 columns including 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. There are also 2 columns (SID and PID) used as observation identifiers. These are not relevant for the purposes of this assignment and were removed. There were also 5 additional variables created. This gives a total of 85 variables used for 2930 residential properties.

It was observed that multiple building types were included in the original data set. It does not make sense to compare one building type to another as each type has a different perception of value in the eyes of a buyer and this is reflected in the different zoning types associated with each building.

The original dataset consists of 5 building types and 8 zoning classifications :

	A (agr)	C (all)	FV	I (all)	RH	RL	RM
1Fam	2	22	77	2	12	2009	301
2fmCon	0	3	0	0	4	30	25
Duplex	0	0	0	0	4	92	13
Twnhs	0	0	19	0	1	28	53
TwnhsE	0	0	43	0	6	114	70

As the purpose of this assignment is to predict the sale price of a “typical” home, it does not make sense to include buildings falling under the zoning classifications industrial (I), agriculture (A) or commercial (C). Residential zoning types including residential high density (RH), residential low density (RL), residential low density park (RP), residential medium density (RM) and floating village residential will be used for further analysis. For building type, we are looking to analyze houses that represent the typical “house” that a family would buy. Duplexes (Duplx) and townhouses (TwnhsE and TwnhsI) are not good representations and are removed leaving single family detached (1Fam) and two family conversion houses (2FmCon) left for analysis.

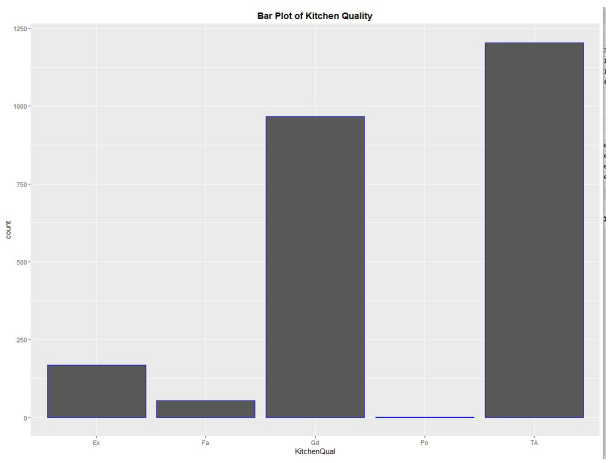
After these conditions are met, there are 2399 observations remaining. Here is a table showing the distribution of homes in the four zoning classifications by building type.

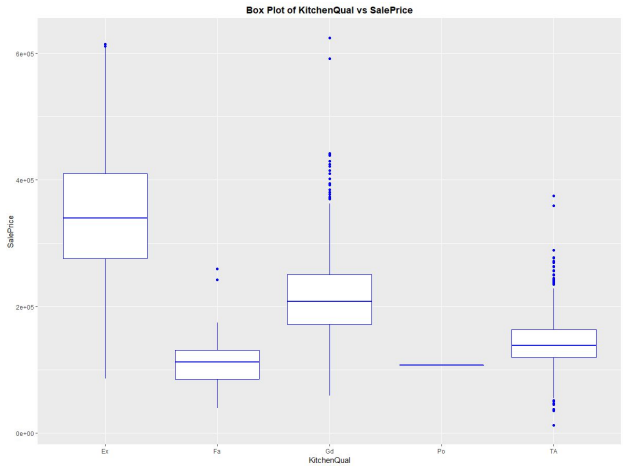
	A (agr)	C (all)	FV	I (all)	RH	RL	RM
1Fam	0	0	77	0	12	2004	301
2fmCon	0	0	0	0	0	0	0
Duplex	0	0	0	0	0	0	0
Twnhs	0	0	0	0	0	0	0
TwnhsE	0	0	0	0	0	0	0

It is evident that none of the properties fall under the two family conversion building type so this can be removed from consideration. While investigating the data further, it was also observed that the majority of properties ranged from 250 total sq ft. to 4000 sq ft. There were 5 properties with over 4000 total sq ft. which were identified as outliers and removed from the sample. This leaves us with a dataset of 85 variables and 2395 properties.

Section 2: Regression with Categorical Variable

From Assignments #1 and #2, it was apparent that KitchenQual had a strong relationship with SalePrice. We will use it to build a regression model. Below is a visualization of the distribution of the 5 possible categories for KitchenQual: Excellent, Good, Typical/Average, Fair and Poor. The other visualization is a box plot illustrating the relationship between the categories of KitchenQual and the response variable SalePrice.





Before we build the regression model using KitchenQual as the only predictor variable, we will first need to calculate the mean sale prices for each category. Below are the results:

```
  KitchenQual Mean_Sale_Price
1          Ex        344899.8
2          Fa        110702.8
3          Gd        215140.0
4          Po        107500.0
5          TA        142907.2
```

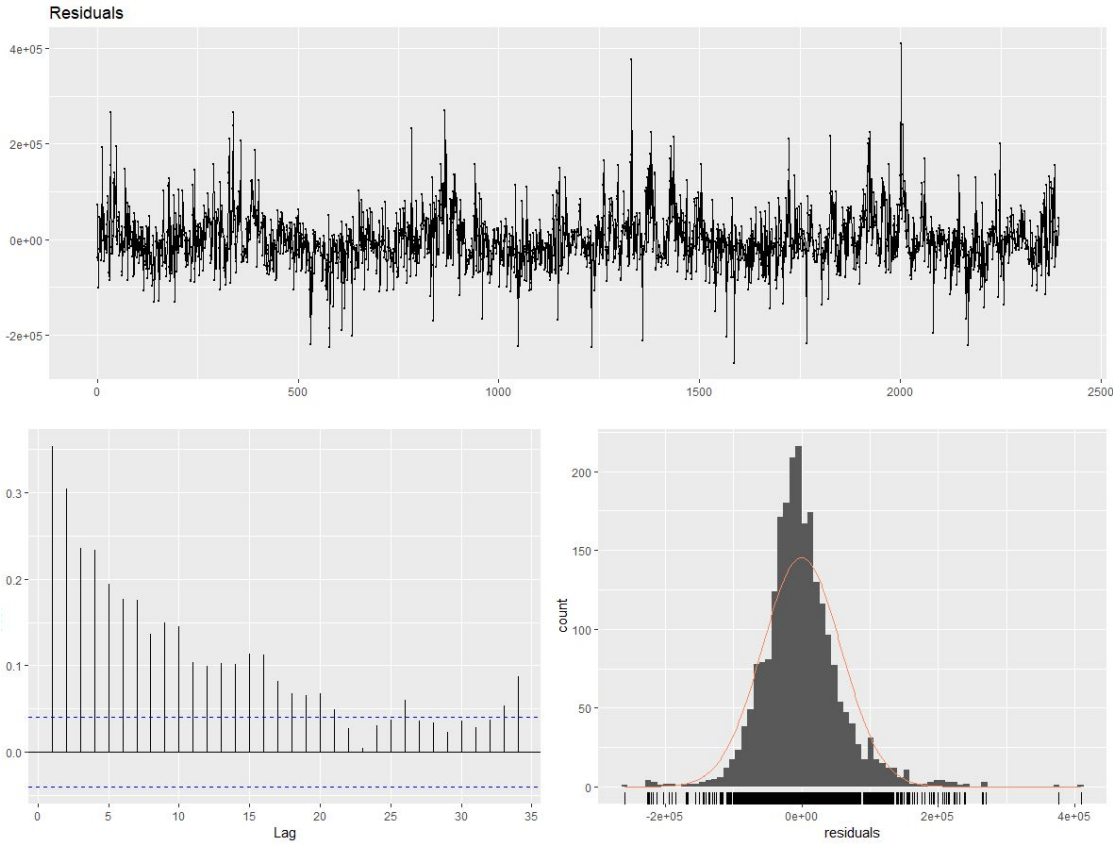
This means that the average price of a house with an “Excellent” rated kitchen is $344,899 and so on for the other categories. Now that we know the mean prices for each category, we will build the regression model and analyze the results. One of the main questions to keep in mind is if the predicted model goes through the mean of SalePrice in each category.

```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq F value    Pr(>F)
KitchenQual    4 7.6351e+12 1.9088e+12  569.93 < 2.2e-16 ***
Residuals   2389 8.0011e+12 3.3491e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = SalePrice ~ KitchenQual, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-258900  -31839   -4907   27093  409860

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     344900       4452   77.48  < 2e-16 ***
KitchenQualFa  -234197       9046  -25.89  < 2e-16 ***
KitchenQualGd  -129760       4825  -26.89  < 2e-16 ***
KitchenQualPo  -237400      58043   -4.09 4.45e-05 ***
KitchenQualTA  -201993       4754  -42.49  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57870 on 2389 degrees of freedom
Multiple R-squared:  0.4883,    Adjusted R-squared:  0.4874
F-statistic: 569.9 on 4 and 2389 DF,  p-value: < 2.2e-16
```

Based on the resulting ANOVA, the table shows that the model is statistically significant because it has a very high F-value (569.93). This means that we can reject the null hypothesis that there is no significant difference between house values based on kitchen quality. A large F-value also indicates a small P-value. Similar to a a large F-value, a small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis. From the R calculations of the coefficient estimates, KitchenQual category "Excellent" is chosen as the baseline to calculate the other categories. Below are the resulting equations for each category:

SalePrice = $344,900 for "Excellent"
SalePrice = $344,900 - $129,760 for "Good"
SalePrice = $344,900 - $201,993 for "Typical/Average"
SalePrice = $344,900 - $234,197 for "Fair"
SalePrice = $344,900 - $237,400 for "Poor"

Comparing these results to the computed sale price averages of each category, we can see that the values are in fact the same. The resulting R^2 value .4883 means approximately 49% of the variation in SalePrice is accounted for by the differences in means of each KitchenQual category. The residual standard error 57870 means that sale price estimates for each category may be off by as much as $57,870. This is an error range between 16.8% for "Excellent" to 28.6% for "Poor". As the sole predictor, these results indicate that KitchenQual may be useful to implement into other multiple regression models depending on the other predictor variables used. To further confirm the validity of this model, the residuals are plotted using the checkresiduals() function from the Forecast package in R.



Based on the first plot, the residuals don't seem to follow any trend and instead resemble white noise. The horizontal-band pattern suggests that the variance of the residuals is constant. The histogram of the residuals can be used to check whether the variance is normally distributed. A symmetric bell-shaped histogram that is evenly distributed around zero indicates that the normality assumption is likely to be true. The residuals appear to be normally distributed with a mean around zero.

Section 3: Dummy Variable Regression (MLR) using Categorical Variable

Categorical variables are coded as dummy variables and included as multiple predictor variables for fitting a regression model. Here, we have 5 categories in KitchenQual so 4 dummy variables can be created with 1's and 0's for each observation. The regression equation can be written as:

SalePrice = a + b1*KitchenQualExcellent + b2*KitchenQualGood +
b3*KitchenQualTypical + b4*KitchenQualFair + b5*KitchenQualPoor

We will now build a regression model using this equation. Below are the results from the ANOVA table and summary() function.

```
Analysis of Variance Table

Response: SalePrice
                     Df     Sum Sq    Mean Sq  F value    Pr(>F)
KitchenQualGood1      1 1.4159e+12 1.4159e+12  422.754 < 2.2e-16 ***
KitchenQualTypical1   1 3.9421e+12 3.9421e+12 1177.052 < 2.2e-16 ***
KitchenQualFair1      1 2.2211e+12 2.2211e+12  663.176 < 2.2e-16 ***
KitchenQualPoor1      1 5.6027e+10 5.6027e+10   16.729 4.455e-05 ***
Residuals          2389 8.0011e+12 3.3491e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
lm(formula = SalePrice ~ KitchenQualGood1 + KitchenQualTypical1 +
    KitchenQualFair1 + KitchenQualPoor1, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-258900  -31839   -4907   27093  409860

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         344900       4452   77.48  < 2e-16 ***
KitchenQualGood1   -129760       4825  -26.89  < 2e-16 ***
KitchenQualTypical1 -201993       4754  -42.49  < 2e-16 ***
KitchenQualFair1   -234197       9046  -25.89  < 2e-16 ***
KitchenQualPoor1   -237400      58043   -4.09 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57870 on 2389 degrees of freedom
Multiple R-squared:  0.4883,    Adjusted R-squared:  0.4874
F-statistic: 569.9 on 4 and 2389 DF,  p-value: < 2.2e-16
```

The large F-values indicate that the model is statistically significant and there is also a significant difference in sale price based on these categories. From these results, we can also build regression equations for each category using “Excellent” as the baseline. This is the estimated Intercept in the results above.

“Excellent”:

SalePrice = a = $344,900

This is the same value that was calculated as the average of homes with “Excellent” kitchens.

“Good”:

SalePrice = $344,900 - $129,760 = $215,140

“Typical”:

SalePrice = $344,900 - $201,993 = $142,907

“Fair”:

SalePrice = $344,900 - $234,197 = $110,703

“Poor”:

SalePrice = $344,900 - $237,400 = $107,500

Once again, these values are the same as the mean values in each kitchen type so the model does go through the mean of SalePrice in each category. The baseline does not have to be “Excellent”, it can be any category, however, whichever category is chosen as the baseline will still output the same results just with different equations.

Further analysis of the results shows that the R^2 value (.4883) is the same as the previous model and means approximately 49% of the variation in SalePrice is accounted for by KitchenQual. This is the same for residual standard error which measures how close the fit is to the actual data points. In this case it is 57870 or $57870

in possible error. The t-tests and p-values for each category indicate that the mean values are significantly different than the mean value of the baseline (Excellent). This is a good indicator that kitchen quality is a good predictor of value.

Section 4: Hypothesis Tests

Hypothesis test for complete model:

H0: b1=0 (there is no difference in mean home sale prices based on KitchenQual categories)

H1: b1 !=0 (there is at least 1 difference in mean sale price based on KitchenQual categories)

From the overall F-test, we can see there is a significant difference at least four mean house prices. This means we reject the null hypothesis and accept the alternative hypothesis. We can conclude that there is significant difference in mean sale prices based on KitchenQual.

Hypothesis tests for each of the coefficients:

When we build a regression model using a categorical variable, the beta coefficients can be interpreted by using the baseline value. This value does not have any slope and can be tested against the null hypothesis (mean house price = $0). Each coefficient represents how much the mean value in sale price changes from the baseline category to the corresponding category. If we are able to reject the null hypothesis for the baseline value, each subsequent category can be tested against this baseline value.

To interpret hypothesis tests, it is important to understand t-tests and p-values associated with these scores. The t score is a ratio between the difference between two groups and the difference within the groups. The larger the t score, the more difference there is between groups. To see if a t-score is “big enough”, it is important to look at the associated p-value. A p-value is the probability that the results from the data occurred

by chance. P-values are scored from 0% to 100% and use the absolute value of t. Low p-values are good and indicate the results did not occur by chance.

Baseline value (Excellent): SalePrice = a (intercept)
H0: a = 0 (Mean sale price = $0)
H1: a!=0 (Mean sale price != $0)

Based on the t-test (77.48) and low p-value (2e-16) associated with the test, we can conclude that the baseline mean house price is significantly different from zero. We can reject the null hypothesis.

Beta coefficient #1 (Good): SalePrice = a + b1*1
H0: b1 = 0 (the mean price is not different)
H1: b1!=0 (the mean price is different)

Based on the t-test (-26.89) and low p-value (2e-16) associated with the test, we can conclude that the mean house price for “Good” kitchens is significantly different from the baseline. We can reject the null hypothesis.

Beta coefficient #2 (Typical): SalePrice = a+b2*1
H0: b2 = 0 (the mean price is not different)
H1: b2!=0 (the mean price is different)

Based on the t-test (-42.49) and low p-value (2e-16) associated with the test, we can conclude that the mean house price for “Typical” kitchens is significantly different from the baseline. We can reject the null hypothesis.

Beta coefficient #3 (Fair): SalePrice = a+b3*1

H0: b3 = 0 (the mean price is not different)

H1: b3!=0 (the mean price is different)

Based on the t-test (-25.89) and low p-value (2e-16) associated with the test, we can conclude that the mean house price for “Fair” kitchens is significantly different from the baseline. We can reject the null hypothesis.

Beta coefficient #4 (Poor): SalePrice = a+b4*1

H0: b4 = 0 (the mean price is not different)

H1: b4!=0 (the mean price is different)

Based on the t-test (-4.09) and low p-value (4.45e-05) associated with the test, we can conclude that the mean house price for “Poor” kitchens is significantly different from the baseline. We can reject the null hypothesis.

Section 4: Multiple Linear Regression Model:

In this section, we will pick two or more continuous predictors and fit a multiple linear regression model. This will be called Model 1. We will evaluate this model by plotting relevant diagnostic plots and assessing the goodness of fit. For Model 1, we will use TotalFloorSF (FirstFlrSF + SeconfFlrSF) and QualityIndex (OverallQual + OverallCond) as the two continuous predictor variables to build the model.

The general multiple linear regression model can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1,2, \ldots n.$$

For this model, the regression equation can be written as:

$$SalePrice = \beta_0 + \beta_1 TotalFloorSF + \beta_2 QualityIndex$$

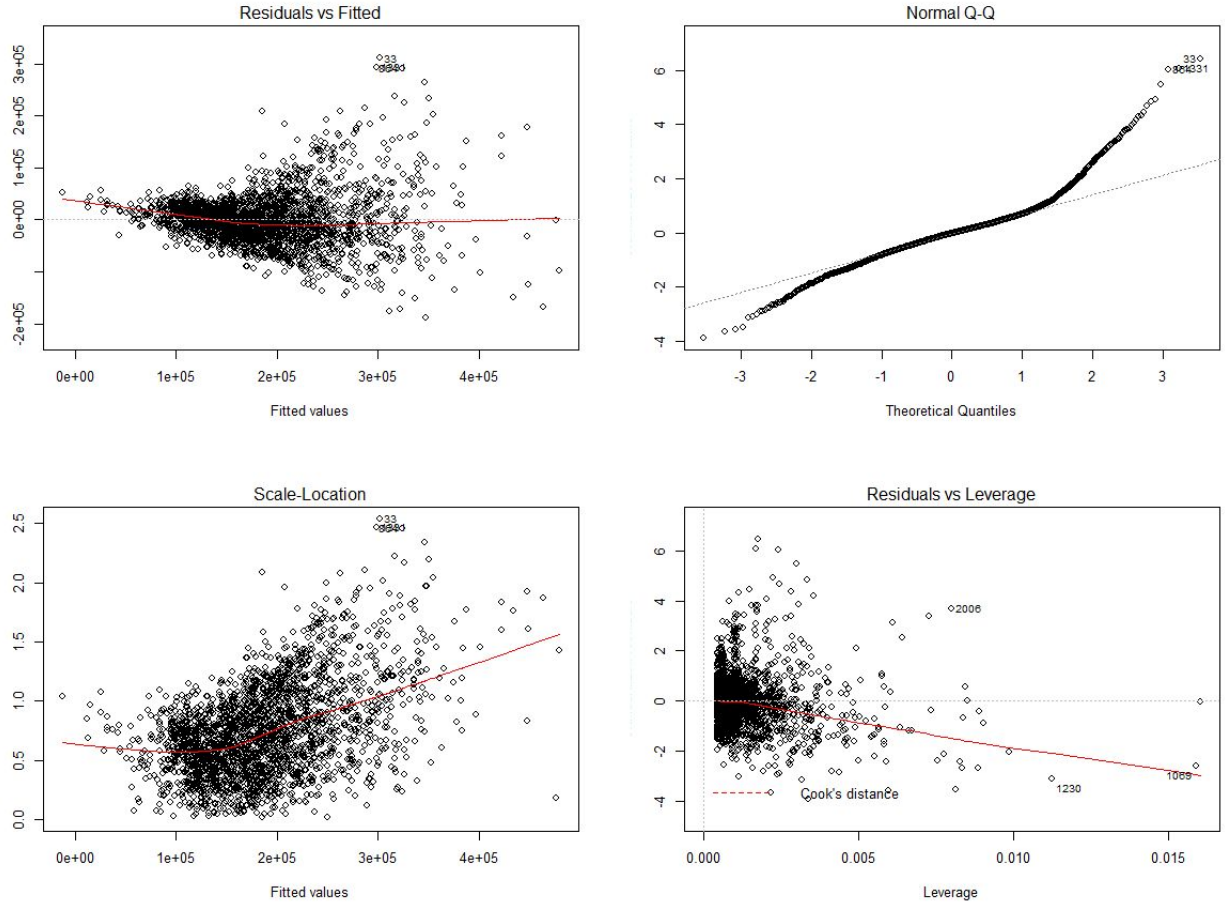
Now we will build the model and report the results:

```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq F value    Pr(>F)
TotalFloorSF    1 9.1912e+12 9.1912e+12 3942.60 < 2.2e-16 ***
QualityIndex    1 8.7092e+11 8.7092e+11  373.58 < 2.2e-16 ***
Residuals    2391 5.5740e+12 2.3313e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

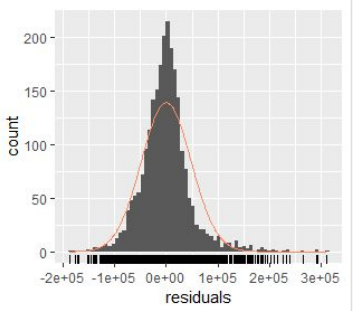
```
Call:
lm(formula = SalePrice ~ TotalFloorSF + QualityIndex, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-187347  -25966    -738   20949  310401

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -54739.71    4146.46  -13.20   <2e-16 ***
TotalFloorSF    106.82       2.19   48.76   <2e-16 ***
QualityIndex   2299.59     118.97   19.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48280 on 2391 degrees of freedom
Multiple R-squared:  0.6435,    Adjusted R-squared:  0.6432
F-statistic:  2158 on 2 and 2391 DF,  p-value: < 2.2e-16
```

The overall F-statistic (2158) and low p-value (2.2e-16) indicate that this model is significant in predicting house prices. The adjusted R^2 value (.6432) estimates that the model accounts for approximately 64% of the variation found in the SalePrice. The residual standard error approximates that the estimated sale price from this model could be off by an average of $48,280 in either direction. Assuming that the predictor variables are independent, the estimate for the TotalFloorSF beta coefficient (106.82) means that for every sq. ft increase to the first floor or second floor, there is a $106.82 increase in sale price (holding QualityIndex constant). Similarly, the estimate for the QualityIndex beta coefficient (2299.59) means that for every unit increase in QualityIndex, there is a $2299.59 increase in sale price (holding TotalFloorSF constant). Both of these predictors are statistically significant based on the t-tests and associated p-values.

To further explore the goodness of fit of this model, it is important to plot the residuals:



Based on the Residuals vs. Fitted plot, the residuals do not appear to be random as they seem to create a fan shape. Similarly, the qq-plot shows deviation at both the upper and lower quantiles. In particular, there appears to be a “heavy tail” (larger values) than the normal distribution This model appears to violate the properties of normality and constant variation for residuals. Below is a histogram of the residuals. The steepness of the distribution confirms that the model violates normality and in particular has more larger values than the normal distribution.

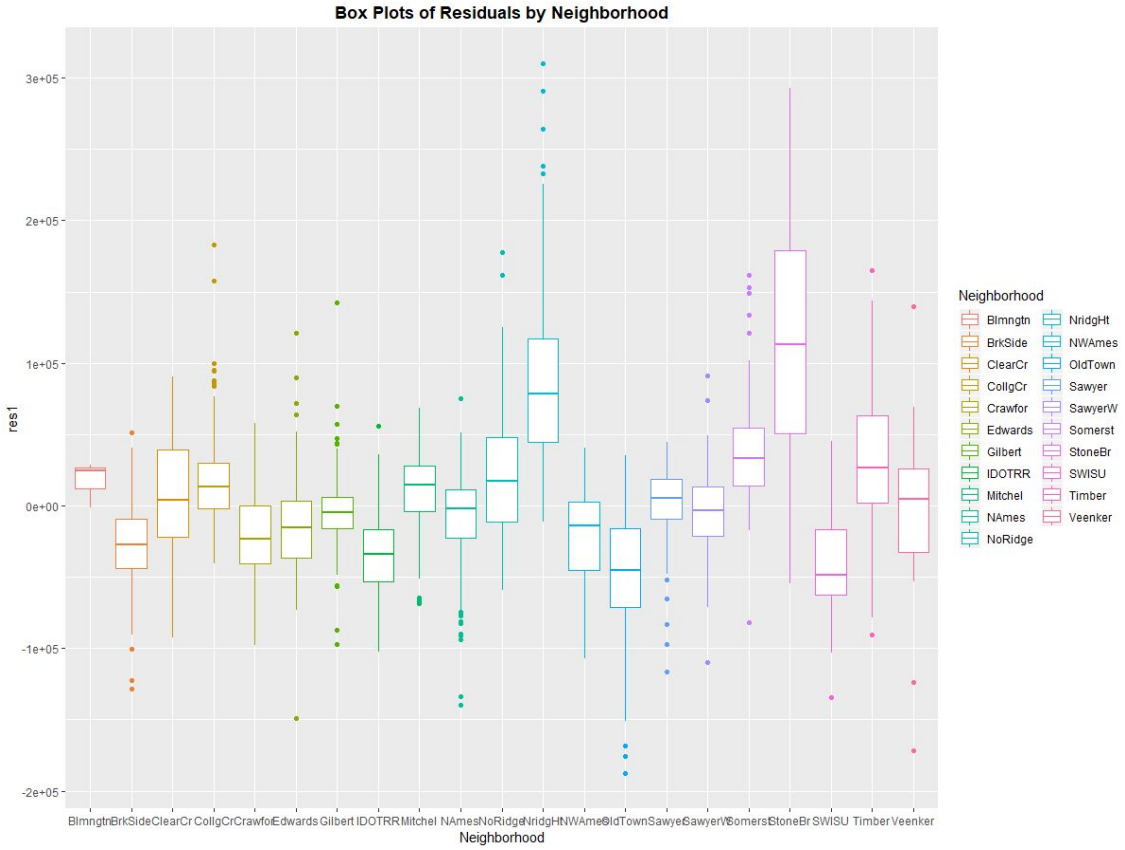


While this model in itself will probably not produce accurate predictions, it is still a better fit than the previous simple linear regression models. This is based on the RMSE and R^2 produced for each model. However, as we see from this model, more variables can produce other problems such as violations of the underlying assumptions of the model. Therefore, the addition of more variables does not always equal a better model. As more variables are added, multicollinearity can also become a problem and further complicates model building. To address these problems, different variables may need to be selected or transformations to the variables may need to be implemented

Section 5: Neighborhood Accuracy

One way to address the problems that come with using multiple variables for linear regression is to make changes to the model by introducing a categorical or indicator variable. This new variable can act as a repairman for the model by dividing the patterns found in the residuals into “normal” portions. In this section, we will study the neighborhood variable and see if can help improve Model 1.

We will begin by creating a boxplot of the residuals, sorted by neighborhood:

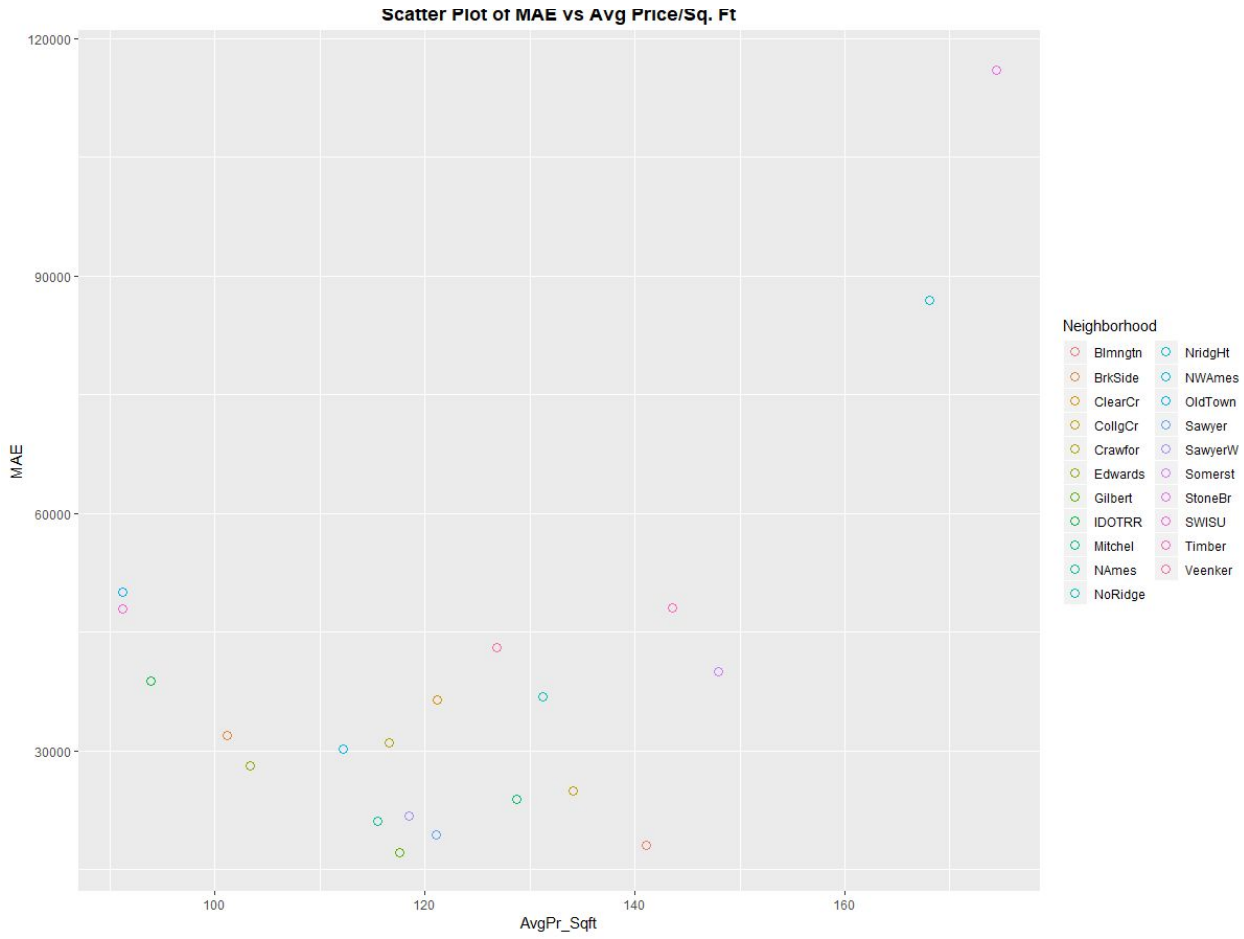


Based on this boxplot, it appears that neighborhoods NridgeHT and StoneBr are significantly underpredicted by the model. On the other hand, OldTown and SWISU seem to be especially overpredicted. Model1 seems to predict ClearCr, NAmes, SawyerW and Veenker well.

Here is a table showing the residual means of Model 1 sorted by Neighborhood:

	Neighborhood	Res_mean
1	Blmngtn	17425.363
2	BrkSide	-27063.476
3	ClearCr	2864.862
4	CollgCr	16195.772
5	Crawfor	-21374.005
6	Edwards	-13897.817
7	Gilbert	-4168.905
8	IDOTRR	-33066.979
9	Mitchel	11091.500
10	NAmes	-7419.146
11	NoRidge	20035.037
12	NridgHt	86319.052
13	NWAmes	-22602.956
14	OldTown	-46267.876
15	Sawyer	1707.760
16	SawyerW	-3978.261
17	Somerst	36924.931
18	StoneBr	112224.033
19	SWISU	-40867.079
20	Timber	32252.414
21	Veenker	-5548.572

Next, we will calculate the mean MAE (based on Model 1) and the mean price per square foot for each neighborhood. We will then plot them against each other and study the relationship.



Based on this plot, the overall picture shows sort of a check shape. However, there appears to be four different subgroups. One group (AvgPr_Sqft < $110) shows an almost negative, linear relationship with MAE. As price per square foot increases, the MAE decreases linearly. Another group (AvgPr_Sqft between $110 and $140) sort of shows 2 sets of points that increase in MAE as price per sq ft increase. The third group (AvgPr_Sqft < 140) does not show much of a pattern. The fourth group seems to contain 2 outliers.

We will create these 4 new subgroups for each neighborhood. These groups will be included in a new variable called Nghbrhd_subgroups. A new model, Model 2, will be built using the additional variable as a predictor.

Here is the equation for Model 2:

$$SalePrice = \beta_0 + \beta_1 TotalFloorSF + \beta_2 QualityIndex + \beta_3 Nghbrhd\_subgroups$$

We will now build this model and interpret the results:

```
Analysis of Variance Table

Response: SalePrice
                   Df     Sum Sq    Mean Sq   F value Pr(>F)
TotalFloorSF        1 9.1912e+12 9.1912e+12 3939.2398 <2e-16 ***
QualityIndex        1 8.7092e+11 8.7092e+11  373.2654 <2e-16 ***
Nghbrhd_subgroups   3 2.2504e+09 7.5013e+08    0.3215 0.8098
Residuals        2388 5.5718e+12 2.3332e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Call:
lm(formula = SalePrice ~ TotalFloorSF + QualityIndex + Nghbrhd_subgroups,
    data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-188063  -25694    -876   20947  313015

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -55019.512   4487.828 -12.260   <2e-16 ***
TotalFloorSF              106.925      2.195  48.720   <2e-16 ***
QualityIndex             2296.403    119.149  19.273   <2e-16 ***
Nghbrhd_subgroupsGroup2   897.619   2441.905   0.368    0.713
Nghbrhd_subgroupsGroup3   -80.012   3304.076  -0.024    0.981
Nghbrhd_subgroupsGroup4 -2447.285   3786.550  -0.646    0.518
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48300 on 2388 degrees of freedom
Multiple R-squared:  0.6437,    Adjusted R-squared:  0.6429
F-statistic: 862.7 on 5 and 2388 DF,  p-value: < 2.2e-16
```

The overall F-statistic is 862.7 indicates that the model is significant (or at least 1 predictor variable is).

The baseline equation for Group 1 neighborhoods can be estimated as:
SalePrice = $-55019.512 + $106.925*TotalFloorSF + $2296.403*QualityIndex

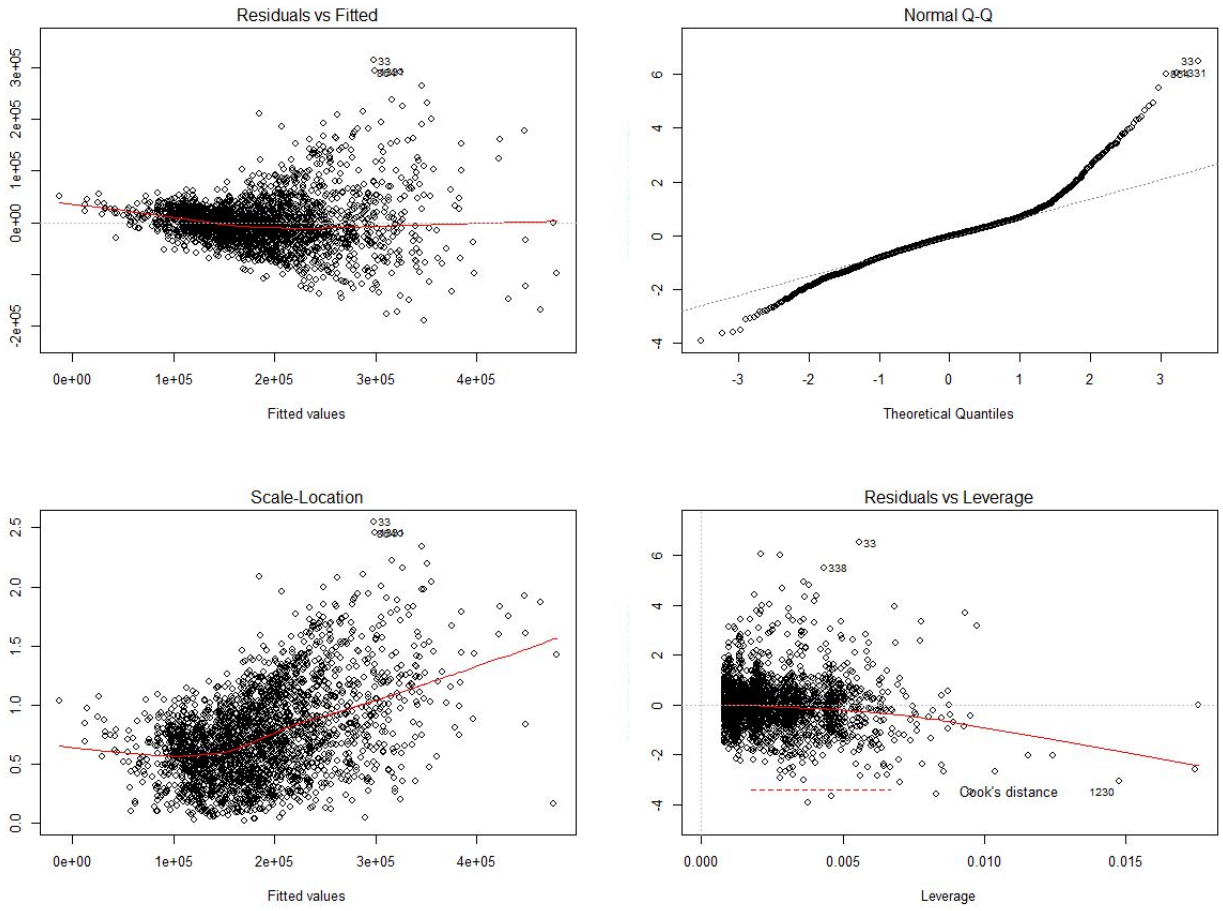
For Group 2 neighborhoods, the equation can be estimated as:
SalePrice = $-55019.512 + $106.925*TotalFloorSF + $2296.403*QualityIndex + $897.619*1

For Group 3 neighborhoods, the equation can be estimated as:
SalePrice = $-55019.512 + $106.925*TotalFloorSF + $2296.403*QualityIndex + -$80.012

For Group 4 neighborhoods, the equation can be estimated as:
SalePrice = $-55019.512 + $106.925*TotalFloorSF + $2296.403*QualityIndex + -$2447.285

Below is a calculation of the Mean Absolute Error from both Model and Model 2. As we can see, there is a slight improvement in Model 2 so it fits the data better. We will also plot the residuals from Model 2.

```
> c(MAE1,MAE2)
[1] 33492.86 33476.87
```



Section 6: Model Comparision of Y vs. log(Y)

In this section we will fit two models using the same set of predictor variables, but the response variables will be SalePrice and log(SalePrice). These models must include at least four continuous predictor variables and any discrete variables. We will call these models Model 3 and Model 4.

Model 3 will be comprised of the following variables:

Model3 <- lm(SalePrice ~ QualityIndex + HouseArea + LotArea + GarageArea + BsmtQual

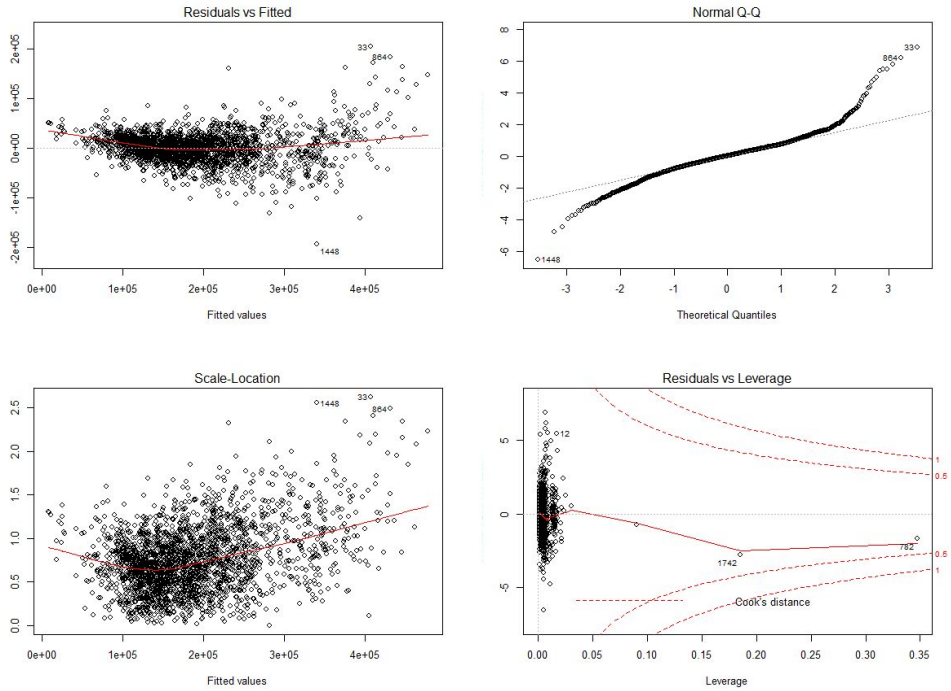
Model 4 will be comprised of the following variables:

Model4 <- lm(logSalePrice ~ QualityIndex + HouseArea + LotArea + GarageArea + BsmtQual

Here are the results from Model 3:

```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq F value    Pr(>F)
QualityIndex   1 4.2760e+12 4.2760e+12 4815.42 < 2.2e-16 ***
HouseArea      1 7.5383e+12 7.5383e+12 8489.36 < 2.2e-16 ***
LotArea        1 3.8804e+10 3.8804e+10   43.70 4.726e-11 ***
GarageArea     1 4.5463e+11 4.5463e+11  511.99 < 2.2e-16 ***
BsmtQual       3 9.0162e+11 3.0054e+11  338.46 < 2.2e-16 ***
Residuals   2339 2.0770e+12 8.8797e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
lm(formula = SalePrice ~ QualityIndex + HouseArea + LotArea +
    GarageArea + BsmtQual, data = newdata)

Residuals:
    Min      1Q  Median      3Q     Max
-193523  -14987     613   15313  204047

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.507e+04  4.837e+03   5.184 2.36e-07 ***
QualityIndex  1.680e+03  7.509e+01  22.377  < 2e-16 ***
HouseArea     5.483e+01  1.188e+00  46.151  < 2e-16 ***
LotArea       6.514e-01  8.763e-02   7.433 1.48e-13 ***
GarageArea    5.015e+01  3.811e+00  13.159  < 2e-16 ***
BsmtQualFa   -9.215e+04  4.422e+03 -20.839  < 2e-16 ***
BsmtQualGd   -6.313e+04  2.490e+03 -25.355  < 2e-16 ***
BsmtQualTA   -8.778e+04  2.775e+03 -31.632  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29800 on 2339 degrees of freedom
  (47 observations deleted due to missingness)
Multiple R-squared:  0.8641,    Adjusted R-squared:  0.8637
F-statistic:  2125 on 7 and 2339 DF,  p-value: < 2.2e-16
```

The large F-value of 2125 and p-value less than 2.2e-16 indicates that the model is statistically significant. Overall, the model accounts for 86.37% of the variation in SalePrice. For each variable, the t-tests and p-values indicate that the variables are significant predictors of value.



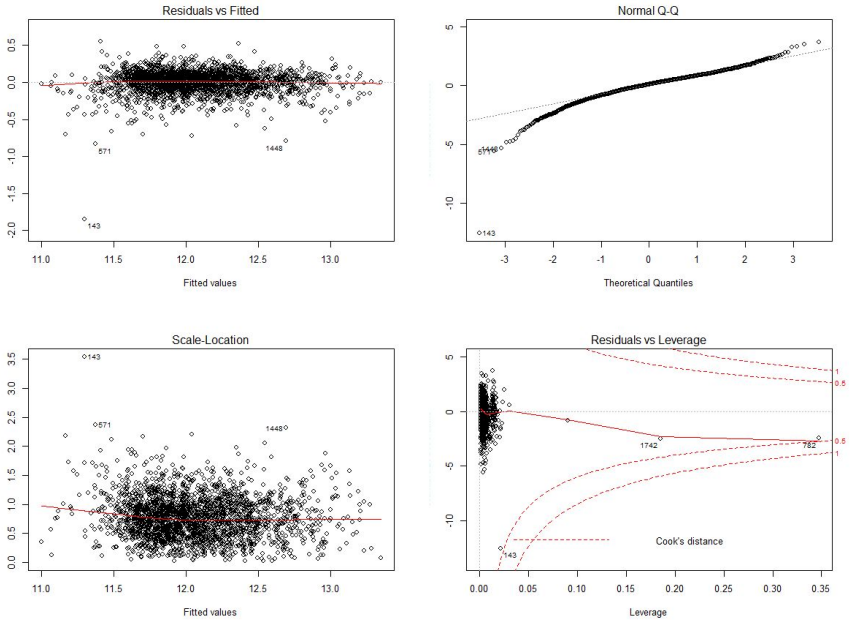
However, by looking at the Residuals vs Fitted plot, we can see that there is an increase in variation of the residuals and so they are not constant. Similarly, the QQ plot shows that the residuals deviate from normality at the upper and lower quantiles. We will see if the log transformation of SalePrice solves these problems in Model 4.

Here are the results for Model 4:

```
Analysis of Variance Table

Response: logSalePrice
              Df  Sum Sq Mean Sq  F value    Pr(>F)
QualityIndex   1 119.092 119.092 5400.721 < 2.2e-16 ***
HouseArea      1 164.506 164.506 7460.202 < 2.2e-16 ***
LotArea        1   0.970   0.970   43.987 4.092e-11 ***
GarageArea     1  12.468  12.468  565.433 < 2.2e-16 ***
BsmtQual       3  18.638   6.213  281.744 < 2.2e-16 ***
Residuals   2339  51.577   0.022
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Call:
lm(formula = logSalePrice ~ QualityIndex + HouseArea + LotArea +
    GarageArea + BsmtQual, data = newdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84333 -0.07330  0.01530  0.09004  0.54708

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.107e+01  2.411e-02 459.177  < 2e-16 ***
QualityIndex  1.073e-02  3.742e-04  28.669  < 2e-16 ***
HouseArea     2.533e-04  5.921e-06  42.788  < 2e-16 ***
LotArea       2.830e-06  4.367e-07   6.481 1.11e-10 ***
GarageArea    2.858e-04  1.899e-05  15.052  < 2e-16 ***
BsmtQualFa   -4.447e-01  2.204e-02 -20.178  < 2e-16 ***
BsmtQualGd   -1.263e-01  1.241e-02 -10.181  < 2e-16 ***
BsmtQualTA   -3.016e-01  1.383e-02 -21.811  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1485 on 2339 degrees of freedom
  (47 observations deleted due to missingness)
Multiple R-squared:  0.8596,    Adjusted R-squared:  0.8591
F-statistic:  2045 on 7 and 2339 DF,  p-value: < 2.2e-16
```

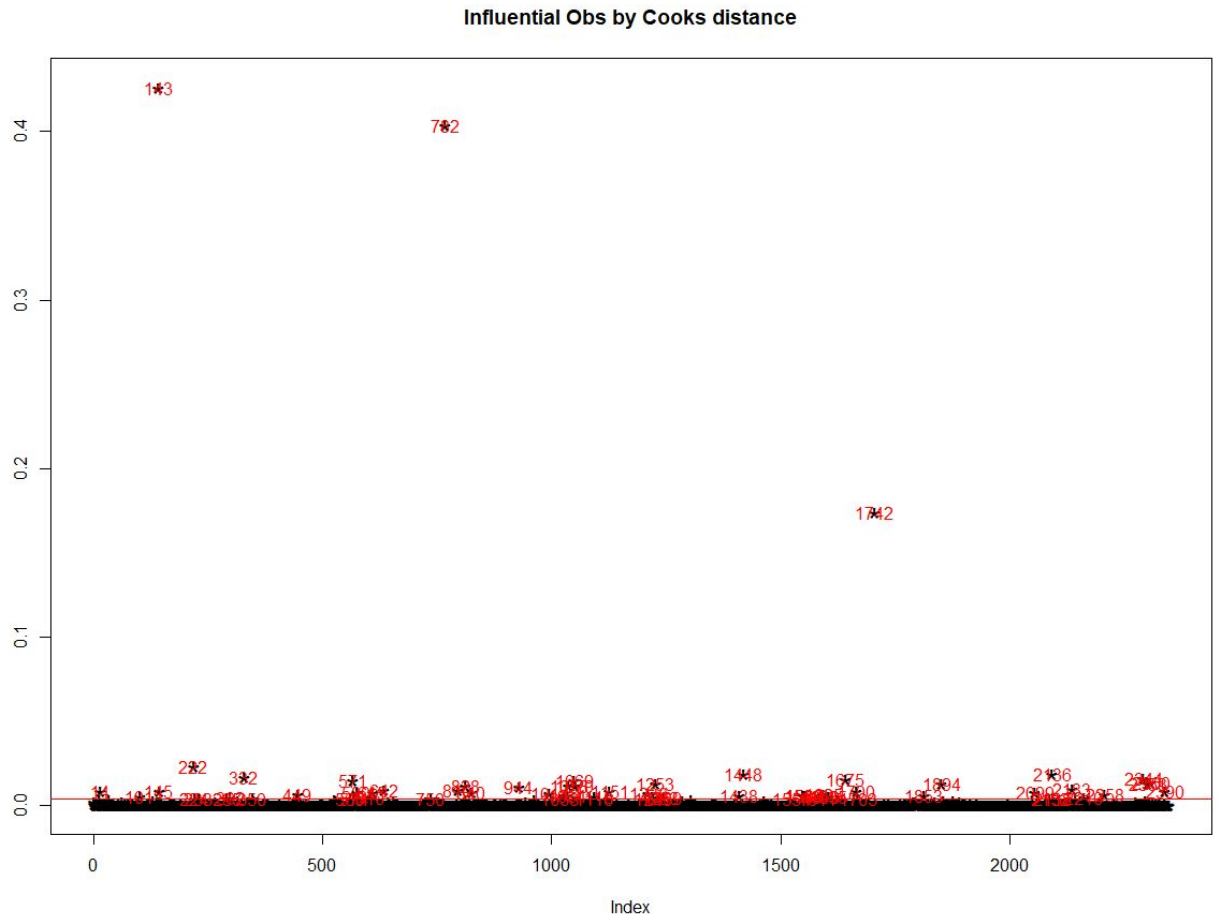
The large F-value of 2045 and p-value less than 2.2e-16 indicates that the model is statistically significant. Overall, the model accounts for 85.91% of the variation in SalePrice. For each variable, the t-tests and p-values indicate that the variables are significant predictors of value.



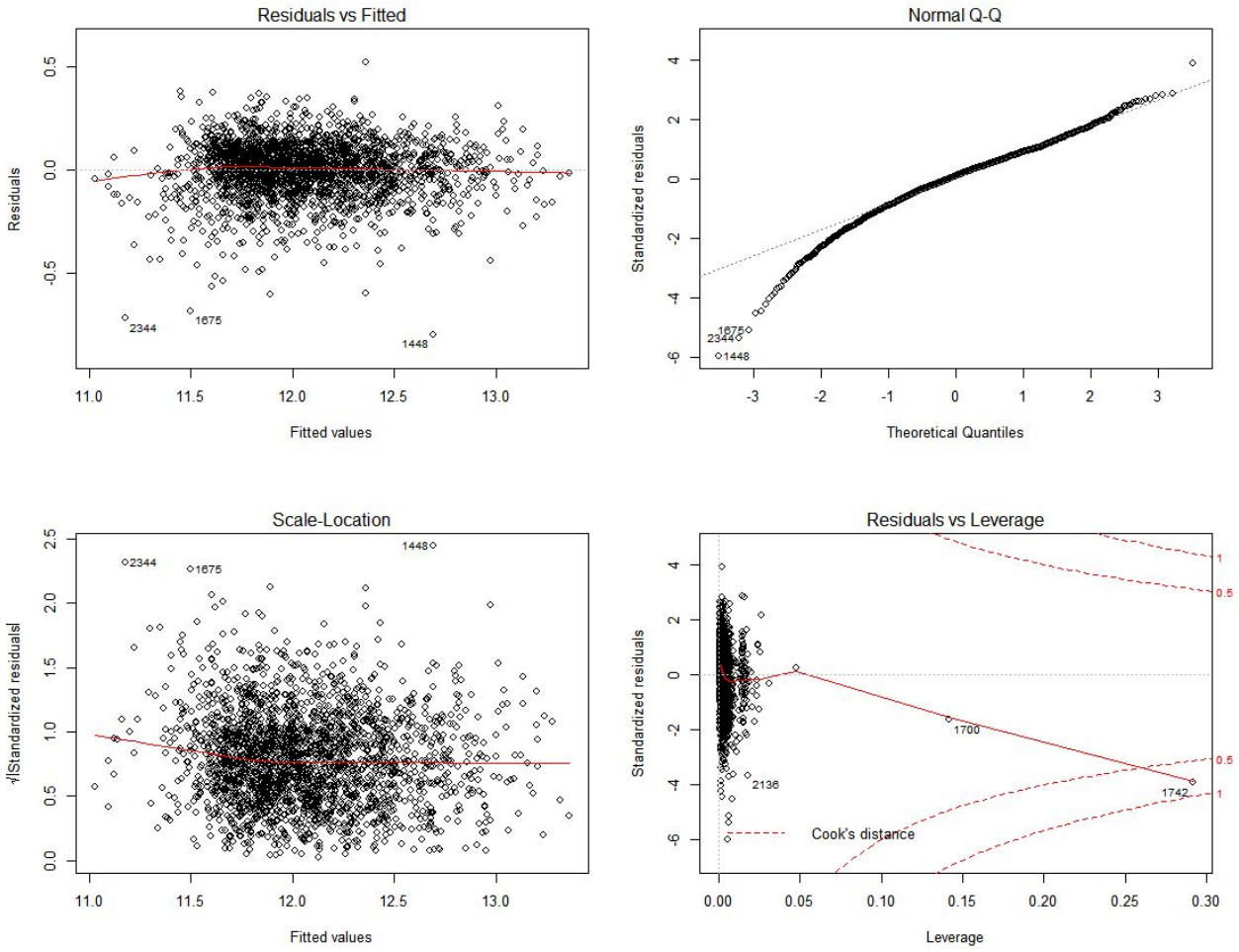
By looking at the Residual vs. Fitted plot, we can see that there is no trend in the residuals and is fairly horizontal. Also, the residuals appear to be distributed fairly even around 0. The QQ plot is also close to normal except for the lower quantiles. However, Model 4 is much improved from Model 3 in terms of satisfying the underlying assumptions of the model.

Section 7: Model Based Outliers

In this section, we will use Model 4 to identify “influential points” that may be having a disproportionate influence on the model coefficients. We will use the threshold value that is given in the textbook on Page 112 and refit the model after removing them.



```
Analysis of Variance Table

Response: logSalePrice
                  Df Sum Sq Mean Sq  F value    Pr(>F)
QualityIndex       1 279500  279500 12719434 < 2.2e-16 ***
HouseArea          1 386089  386089 17570050 < 2.2e-16 ***
LotArea            1   2277    2277   103604 < 2.2e-16 ***
GarageArea         1  29264   29264  1331737 < 2.2e-16 ***
BsmtQual           3  43742   14581   663527 < 2.2e-16 ***
Residuals    5508337 121042       0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = logSalePrice ~ QualityIndex + HouseArea + LotArea +
    GarageArea + BsmtQual, data = model4_screen)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84334 -0.07335  0.01530  0.09025  0.54707

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.107e+01  4.967e-04 22283.8   <2e-16 ***
QualityIndex  1.073e-02  7.711e-06  1391.3   <2e-16 ***
HouseArea     2.533e-04  1.220e-07  2076.5   <2e-16 ***
LotArea       2.830e-06  8.999e-09   314.5   <2e-16 ***
GarageArea    2.858e-04  3.913e-07   730.5   <2e-16 ***
BsmtQualFa   -4.446e-01  4.541e-04  -979.2   <2e-16 ***
BsmtQualGd   -1.263e-01  2.557e-04  -494.1   <2e-16 ***
BsmtQualTA   -3.016e-01  2.850e-04 -1058.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1482 on 5508337 degrees of freedom
  (110308 observations deleted due to missingness)
Multiple R-squared:  0.8596,    Adjusted R-squared:  0.8596
F-statistic: 4.816e+06 on 7 and 5508337 DF,  p-value: < 2.2e-16
```



From these Residuals vs Fitted plot, we can see that the data hugs the fitted line much tighter. Also, the RMSE score appears to slightly lower for this model which is an improvement.

Section 8: Summary/Conclusions

- Model building is not a one step process. Even if you build a model that seems to fit well, it still may be violating the underlying assumptions of the model. Correcting these violations can take many forms including choosing new variables, adding variables and performing variable transformations
- Variable transformation can address models that violate assumptions of normality and constant variance (when the residuals show different patterns such as funnel or fan shape instead of looking like “white noise”)
- Similarly, treatment of outliers and influential points can lead to increased normality and improve the model fit. However, there is a tradeoff between removing outliers and the representativeness of the data. EDA is required to ensure that there are no substantive effects to model performance after treatment.

- There are many models that can be built from this data and we were only scratching the surface with the ones we built in this assignment. It is important that all possible and relevant models be compared and analyzed.
- For models with more than 3 predictor variables, it is important to study collinearity and the interactions between these variables. High levels of multicollinearity can have a substantial impact on model performance. This another next step.
- Since we are building predictive models, we want to study model performance on out-of-sample data. This may require getting new data or splitting the current data into training/test splits and putting the models into a cross-validation process.