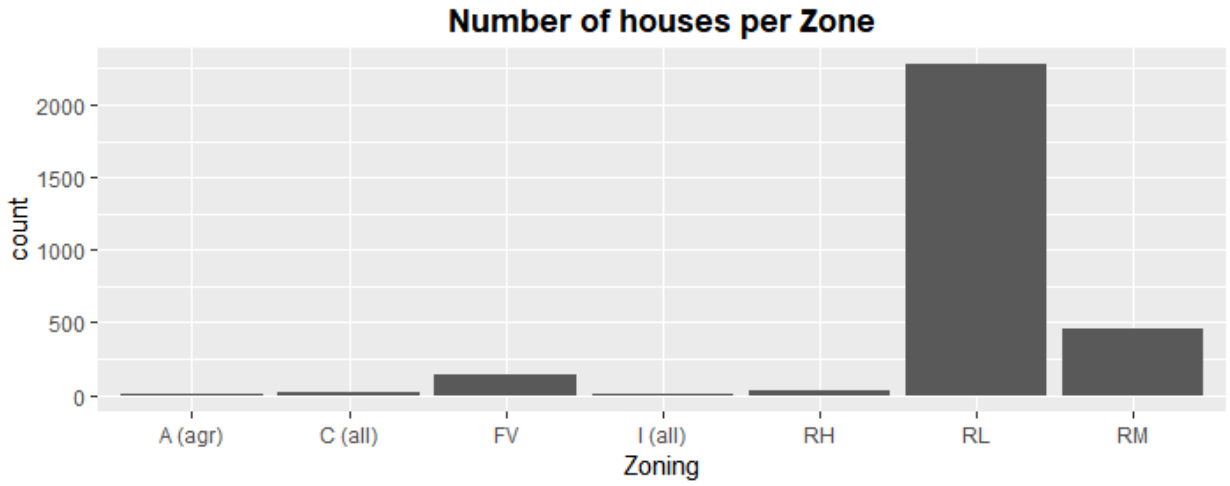
Assignment #3

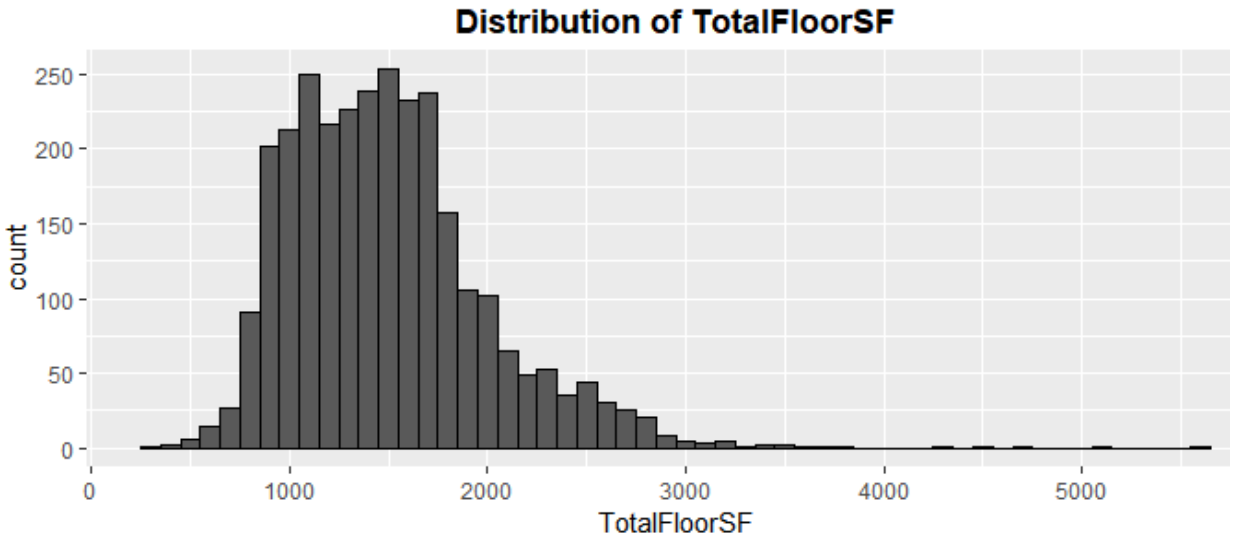
Syamala Srinivasan

Introduction:

The objective of this task is to provide estimates of home values for typical homes in Ames. A "typical" home in Ames would logically only include homes. Essentially this definition of "home" excludes lots identified as Agriculture, Commercial or Industrial lots. The removal of these populations ensures that future trends and analysis are comparing similar sample populations.



There are 29 lots in the data set that are marked as Agriculture, Commercial or Industrial. These lots will be removed to ensure that only "typical" homes are analyzed.



Lastly, only five homes have a total square footage of greater than 4,000. Therefore, these outliers will be removed as to reflect more accurately what a "typical" home in Ames is. Since

these values are well over 2,500 square feet greater than the mean these are viewed as not typical.

Variable Name	Populations Dropping	Drop Count	Cumulative Drop
Zoning	A, C, I	29	29
Total Square Feet	>4000	5	34

Categorical variables can be extremely large and can prove to be difficult to work with. This process can be enhanced by creating dummy variables to simplify the dataset and allow simpler comparisons in relationships to be drawn.

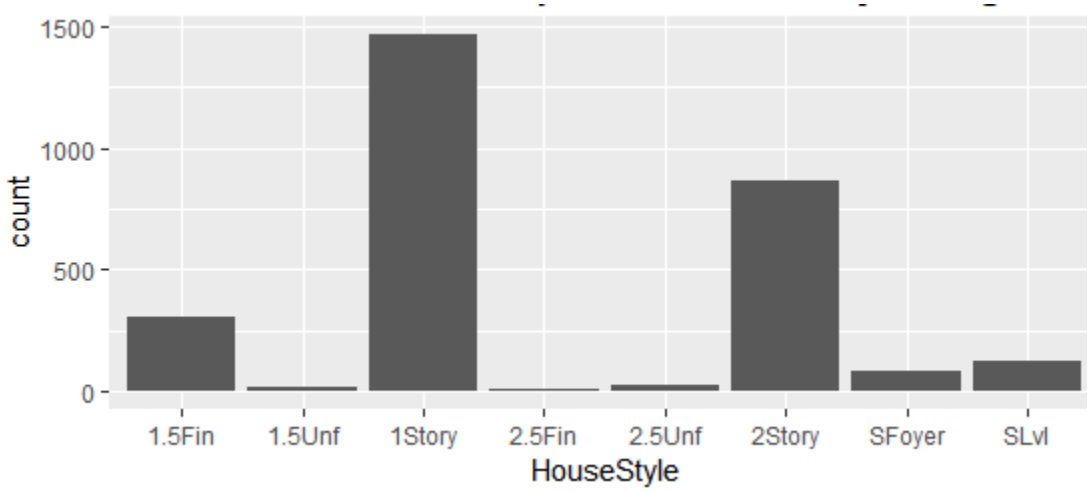
This assignment begins the fitting and validation of multiple linear regression models. The models in this assignment take into consideration discrete and categorical variables to analyze and determine which variables add the most benefit to the model and isolate those variables if necessary.

Results:

Task 1

In the previous analysis conducted, only numerical values were considered. However, categorical variables play just as an important role as predictors. In the next model, the variable House Style will be assessed.

There are 8 types of homes in our data set that indicate primarily the number of floors in each home. Most homes are either a 1 story or 2 story home but there are a few that do not fit those categories and have fallen into other such as 1.5 or 2.5. Additionally, the house style indicates if the home is finished or not.



Below are the individual sale price means for each house style and the predicted value after a simple linear regression model was created utilizing only the house style variable.

House Style	Mean Sale Price	Predicted Sale Price
1.5 Finished	$138,587.40	$138,587.40
1.5 Unfinished	$112,172.20	$112,172.20
1 Story	$179,786.50	$179,786.50
2.5 Finished	$220,000	$220,000
2.5 Unfinished	$182,495.50	$182,495.50
2 Story	$206,604.60	$206,604.60
SFoyer	$143,472.70	$143,472.70
SLvl	$165,527.40	$165,527.40

It is easily noticeable that the predicted sale price for each house style is exactly the same as the mean sale price. This indicates that the predicted sale price heavy takes into consideration the mean of the value.

Task 2

The 8 categories for house style can be summarized into 3 different categories, 1 story, 2 story and split. Therefore, dummy variables will be created for this categorical variable based on the number of floors and the last category will be used for interpretation.

Next a multiple regression model will be created using the dummy coded variables for the number of stories. For the purpose of the model the first category will be used for interpretation. The R squared of our model is .04 which indicates that this is a terrible predictor of sale price as only 4% of our data can be explained using this model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-159256  -47045  -19045   29556  442955

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   172045       1805  95.326  < 2e-16 ***
Stylegrp2      34024       3128  10.878  < 2e-16 ***
Stylegrp3     -15193       5559  -2.733  0.00632 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76380 on 2893 degrees of freedom
Multiple R-squared:  0.04676,   Adjusted R-squared:  0.0461
F-statistic: 70.96 on 2 and 2893 DF,  p-value: < 2.2e-16
```

House Style	Mean Sale Price	Predicted Sale Price
Group 1	$172,044	$172,044
Group 2	$206,068	$206,068
Group 3	$156,851	$156,851

The mean for each group is again identical to the predicted sale prices.

Task 3:

Two hypothesis tests will be conducted on each of the betas. These tests include:

- Reduced Model
    - Null: Betas corresponding to predictors are 0
    - Alternative: At least one inequality
    - The removal of two predictors that were identified to be potentially insignificant (1.5 Unfinished and Split Foyer) obtained a small F statistic of 18, compared to the 30 in the full model. This leads us to reject the null hypothesis.
- Individual Significance
    - Null: Beta =0
    - Alternative: At least one inequality
    - The t values obtained for our model include 95 as the intercept, 10 and -2. Additionally we can observe extremely small p values that allow us a over 99% confidence that we can reject the null hypothesis.
- Individual Significance (B1)
    - Null: The variable does not have a relationship with the sale price of a home
    - Alternative: The variable has a strong relationship with the home
    - The significantly low p value indicates that this variable has a strong significance and thus we can reject the null hypothesis
- Individual Significance (B2)
    - Null: The variable does not have a relationship with the sale price of a home
    - Alternative: The variable has a strong relationship with the home
    - The significantly low p value indicates that this variable has a strong significance and thus we can reject the null hypothesis
- Individual Significance (B3)
    - Null: The variable does not have a relationship with the sale price of a home
    - Alternative: The variable has a strong relationship with the home
    - The significantly low p value indicates that this variable has a strong significance and thus we can reject the null hypothesis

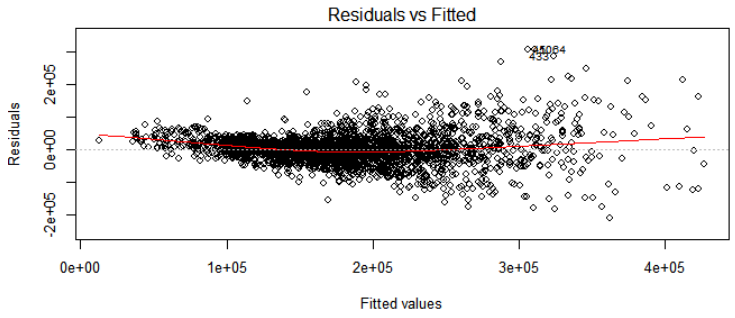
Task 4:

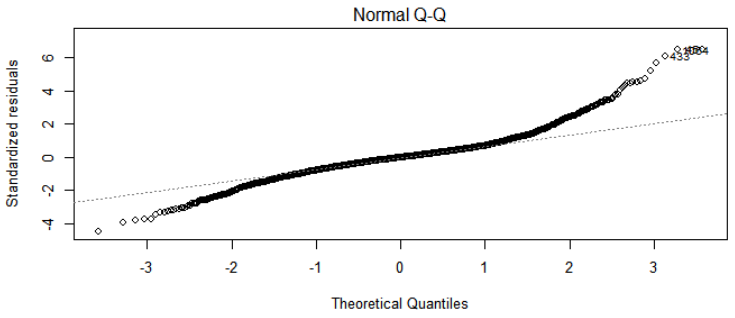
Model 1 will be creating utilizing the total livable floor space above ground (GrLivArea) and the size of the garage (GarageArea).

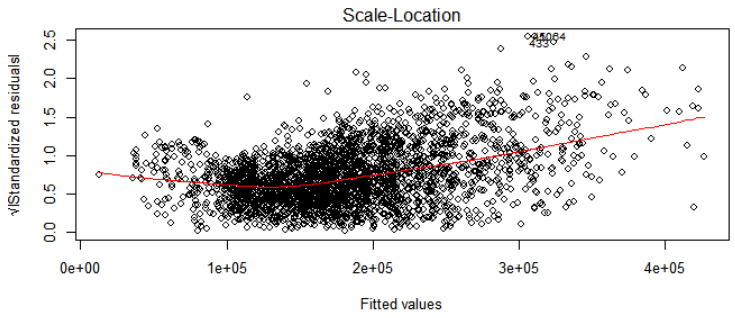
```
Residuals:
    Min      1Q  Median      3Q     Max
-211878  -23904    -787   20403  305716

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15542.340   2934.705  -5.296 1.27e-07 ***
GrLivArea       85.374      2.052  41.602  < 2e-16 ***
GarageArea     145.926      4.699  31.054  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47060 on 2892 degrees of freedom
Multiple R-squared:  0.6383,    Adjusted R-squared:  0.638
F-statistic:  2551 on 2 and 2892 DF,  p-value: < 2.2e-16
```

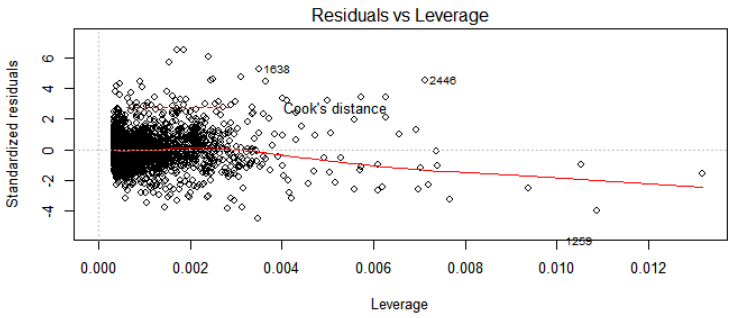
The equation for this model is Y= -15542 + 85X1 + 145X2. This indicates that for every square foot of livable floor space, if the garage size is held constant, adds $85 to the sale price. While each additional square foot of space in the garage, if the home size is held constant, adds $145 to the sale price.

When looking at the residuals vs fitted, it is noticeable that there is some randomization as the points follow a tight horizonal line. The QQ plot does not look good either, this graph is far from normalized as many points on the ends are far from the line and there are many outliers. The scale location graph is a bit more randomized than the fitted graph but still seems to have a bit of a funnel shape and a sharp positive increase on the right.



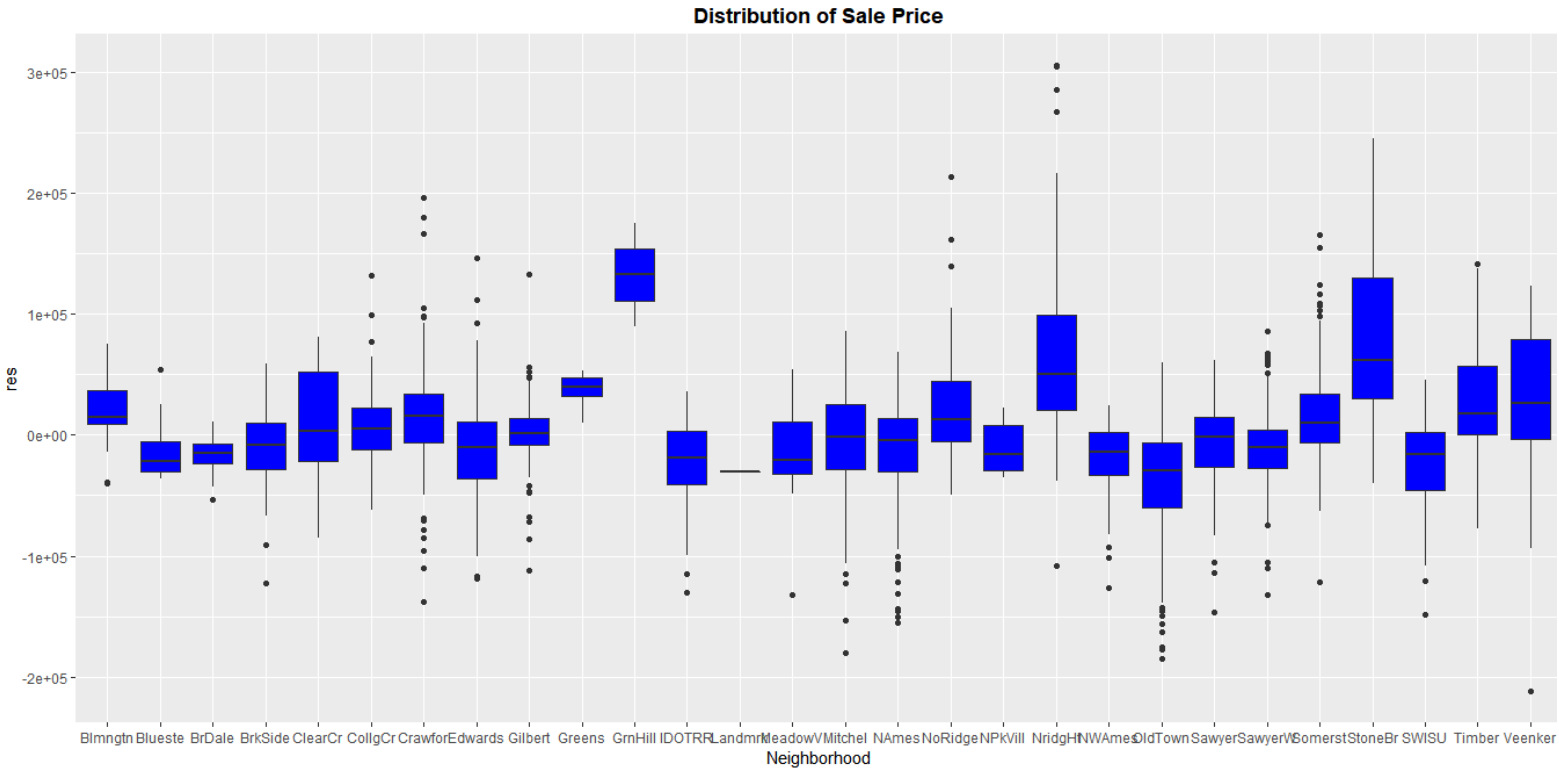






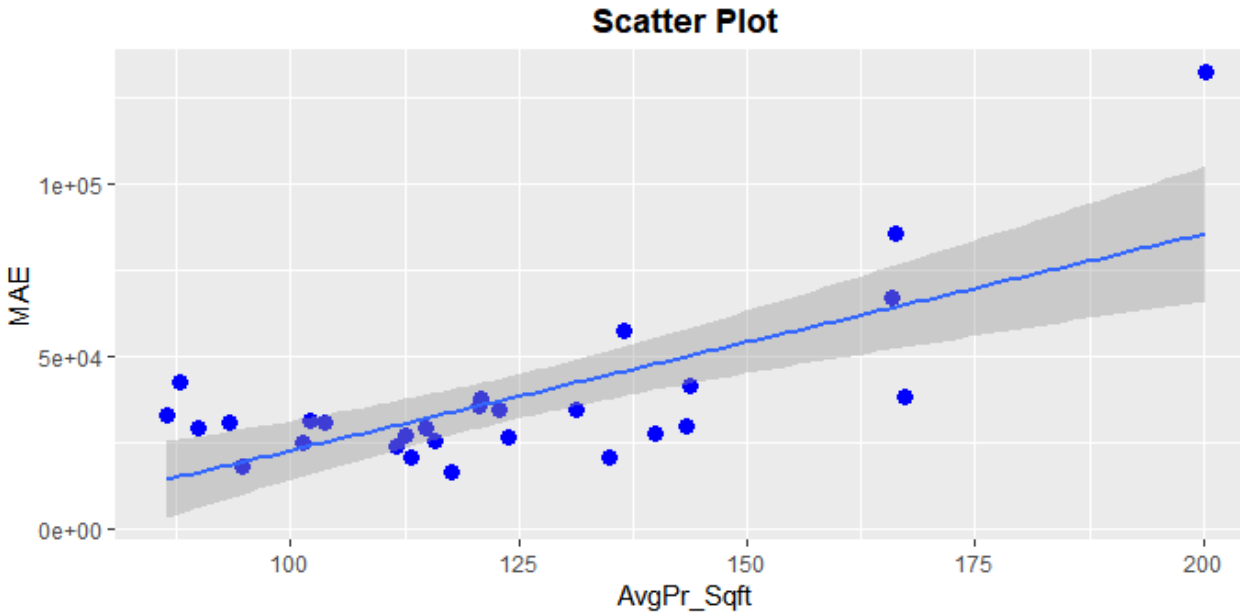
Adding additional variables will not always increase the accuracy of the model as some variables are not as influential to the sale price as others, some in fact may actually muddy the data and negatively affect the model.

Task 5:



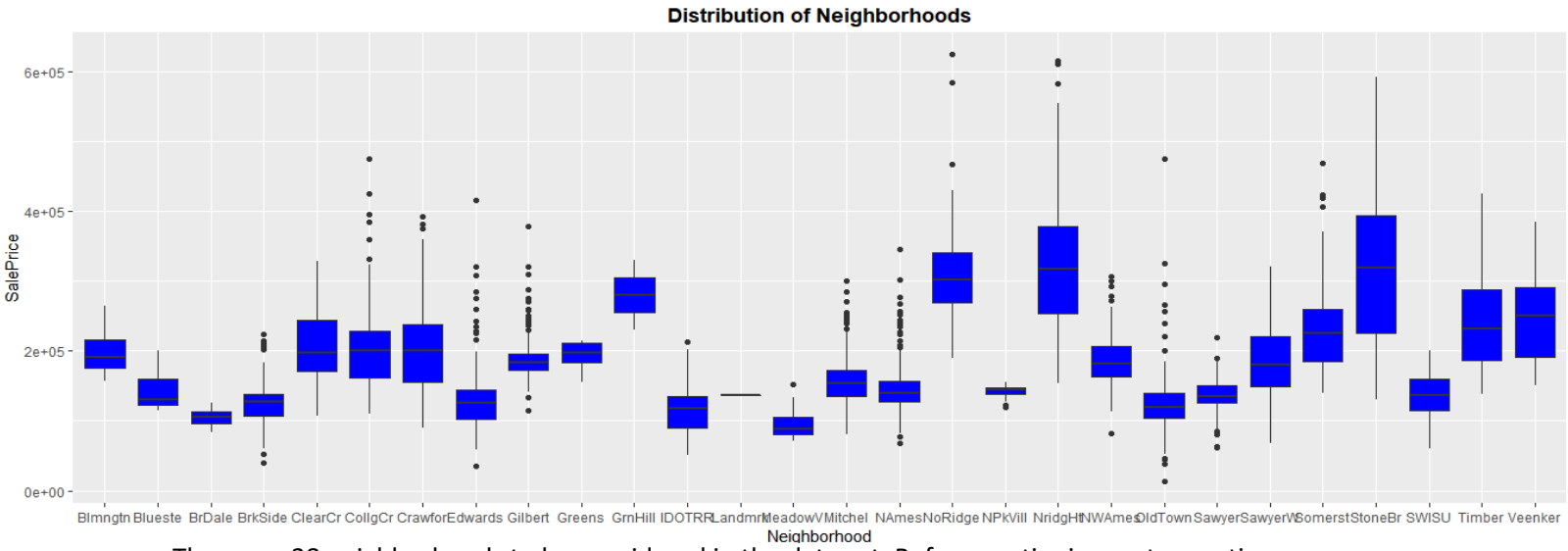
The residual of each neighborhood is the difference between the actual and the predicted prices. When viewing the above graph it becomes apparent that some neighborhoods have wider ranges of residuals than others. There are three neighborhoods that stand out as overpredicted, these are GrnHill, NridgHt and StoneBr. The neighborhood that has the least amount of fluctuation is Landmrk but this can be ignored because there is only one home in Landmrk. GrnHill is also an exception because there are only 2 houses in GrnHill.

Mean Absolute Error (MAE) measures the errors in a set of predictors. In Model 1 our predictors include the livable space above ground and the area of the garage. Therefore, the mean absolute error will indicate the average difference between the predicted and observed values. The MAE for Model 1 is 32,438.51. Therefore, Model 1 has an average miscalculation of prediction of $32,438.51.



The above graph indicates a significant positive increase between the mean absolute error and the average price per square foot. Meaning, the larger the home is, the larger the error in prediction is. One explanation for this could be that there are lesser large homes than there are smaller homes, providing more datapoints for small homes.

It is commonly said that the location of a home ("Location! Location! Location!") is one of the most important aspects to determine the price. When looking at the price distribution by neighborhood, this becomes apparent. There are some neighborhoods that struggle to sell homes over $100,000 while others have never sold a home lower than $250,000



There are 28 neighborhoods to be considered in the data set. Before continuing on to creating a multiple linear regression model, dummy variables must be created to group the 28 neighborhoods into more usable groupings. The three groups include homes less than $100 per square foot, between $100

and $120, between $120 and $140, and greater than $140. To complete the multiple regression model, the first group will be used for the basis of interpretation.

The next step of the analysis is to include the new neighborhood variables into the original multiple regression model of livable floor space and garage size.

```
Residuals:
    Min      1Q  Median      3Q     Max
-158141  -13339      -5   11120  256197

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.691e+04  2.182e+03   21.50   <2e-16 ***
GrLivArea    1.174e+02  1.378e+00   85.26   <2e-16 ***
GarageArea   3.770e+01  3.351e+00   11.25   <2e-16 ***
NbhdGrp1    -1.171e+05  1.777e+03  -65.88   <2e-16 ***
NbhdGrp2    -7.488e+04  1.634e+03  -45.82   <2e-16 ***
NbhdGrp3    -4.832e+04  1.571e+03  -30.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29420 on 2889 degrees of freedom
Multiple R-squared:  0.8587,    Adjusted R-squared:  0.8585
F-statistic:  3512 on 5 and 2889 DF,  p-value: < 2.2e-16
```

This multiple linear regression model is significantly improved by adding the neighborhood values. The R squared value has improved 35%, increasing from .63 to .85. The significance of this model is further evident by the very small p values for each of the predictor variables indicating that each is highly significant. The new mean absolute error of the model is 18,826.76 which is $13,612 less than the original calculation further indicating that Model 2 is a more accurate prediction model than Model 1.

Task 6:

The next two models created will utilize four continuous predictor variables and a discrete predictor variable. The four continuous predictor variables will include living area above ground, garage area, lot area, and the age of the home and the discrete variable will be the number of bedrooms above ground. Both of the models created in this section will utilize the same predictor variables, the only difference is that Model 3 will utilize the sale price of the home as the response variable while Model 4 will utilize the log transformation of the sale price as it’s response variable.

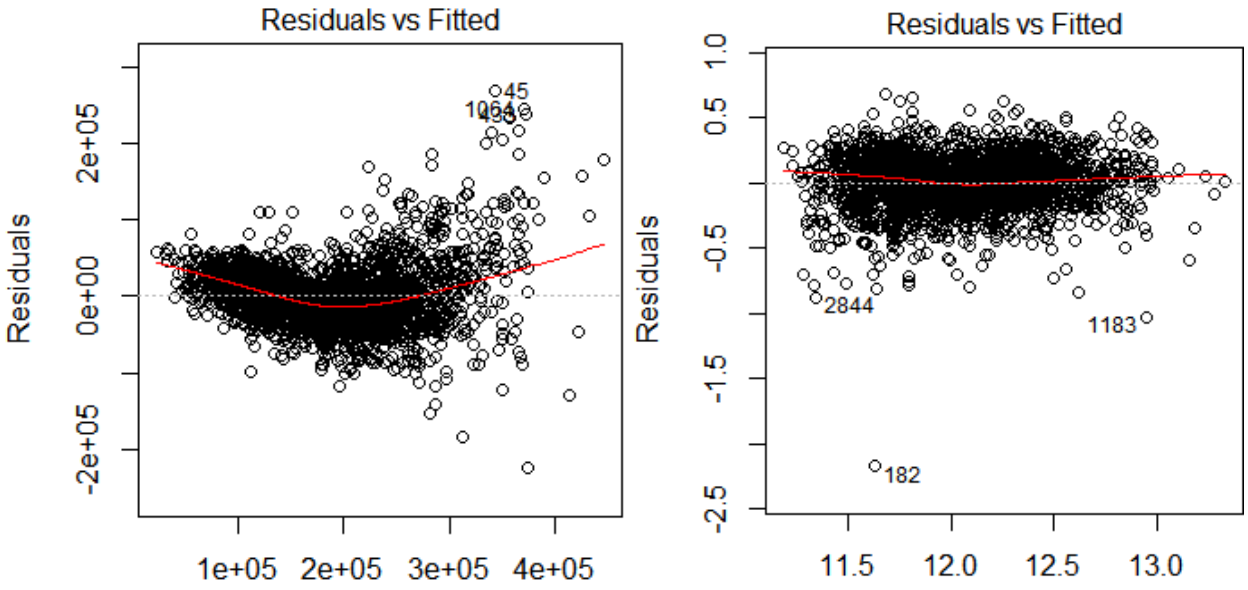
Summary of Sale Price Regression

```
Residuals:
    Min      1Q  Median      3Q     Max
-224782  -22955   -1961   19354  267471

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.859e+04  3.408e+03  20.129   <2e-16 ***
GrLivArea     1.042e+02  2.085e+00  50.005   <2e-16 ***
GarageArea    7.207e+01  4.365e+00  16.511   <2e-16 ***
LotArea       9.320e-01  9.702e-02   9.607   <2e-16 ***
HouseAge     -7.768e+02  2.769e+01 -28.049   <2e-16 ***
BedroomAbvGr -2.059e+04  1.071e+03 -19.222   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38750 on 2889 degrees of freedom
Multiple R-squared:  0.755,     Adjusted R-squared:  0.7546
F-statistic:  1781 on 5 and 2889 DF,  p-value: < 2.2e-16
```

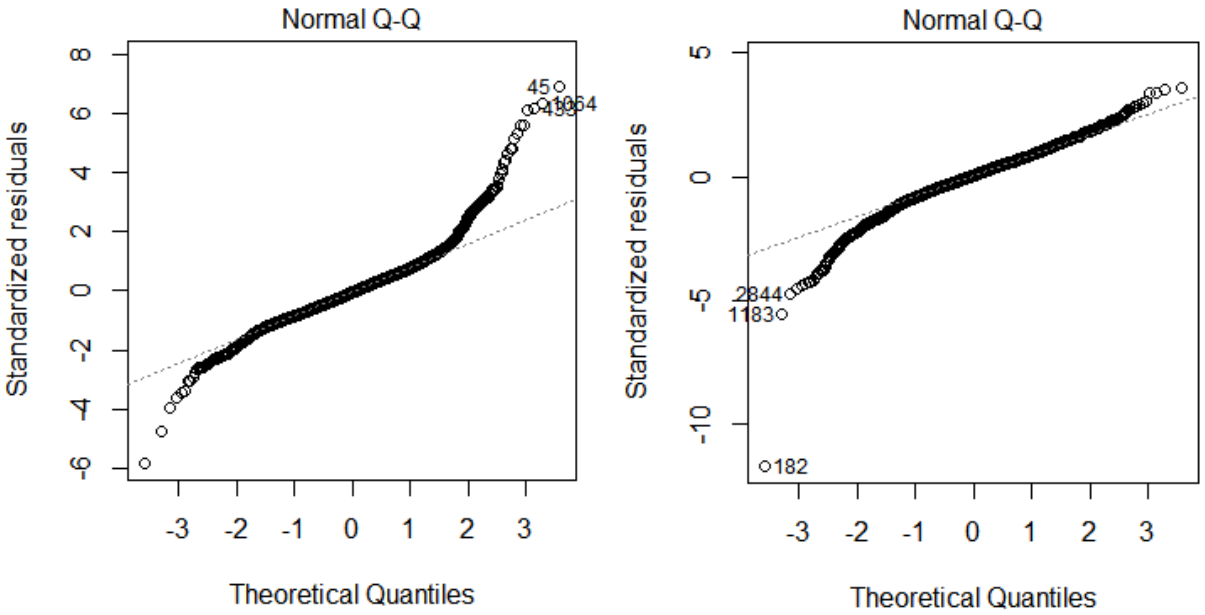
Summary of log Sale Price Regression

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.17374 -0.09779  0.00902  0.11235  0.67235

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.147e+01  1.643e-02 697.858   <2e-16 ***
GrLivArea     4.815e-04  1.005e-05  47.901   <2e-16 ***
GarageArea    3.652e-04  2.105e-05  17.352   <2e-16 ***
LotArea       4.531e-06  4.678e-07   9.686   <2e-16 ***
HouseAge     -4.872e-03  1.335e-04 -36.485   <2e-16 ***
BedroomAbvGr -7.076e-02  5.165e-03 -13.700   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1868 on 2889 degrees of freedom
Multiple R-squared:  0.7774,    Adjusted R-squared:  0.777
F-statistic:  2018 on 5 and 2889 DF,  p-value: < 2.2e-16
```

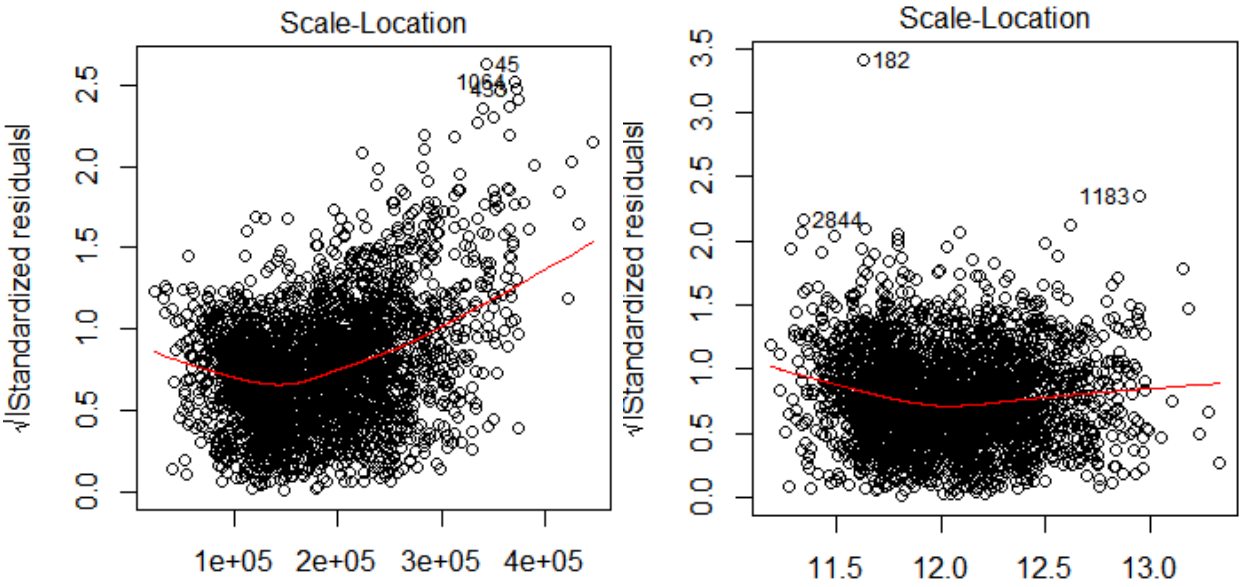
The two models produced using the new predictor variables both produce significant models. To determine while model is a more accurate predictor of sale price we will look at the R squared. R squared for the log sale price is .02 higher than the not transformed price, indicating that it is a slightly better model.



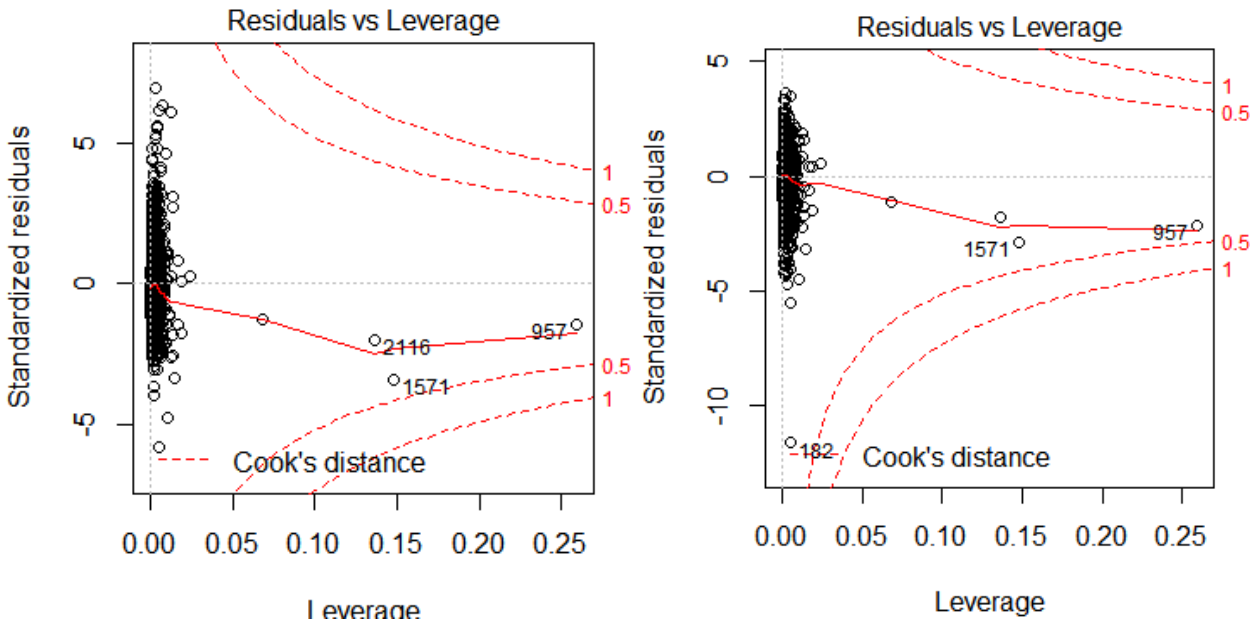
The fitted graph on the left illustrates the not transformed sale price while the graph on the right is of the log transformed sale price. There is a large difference between the two graphs in that the not transformed graph has a bit of a “U” shape which is not preferred as this graph should not follow any trend. The transformed graph has a much more randomized appearance and follows an almost horizonal line.



The fitted graph on the left illustrates the not transformed sale price while the graph on the right is of the log transformed sale price. As expected with a log transformed variable, the distribution along the QQ line is significantly more normalized for the transformed variable. Although, there are still many points far from the line towards the left of the graph indicating some outliers that were not present in the original.

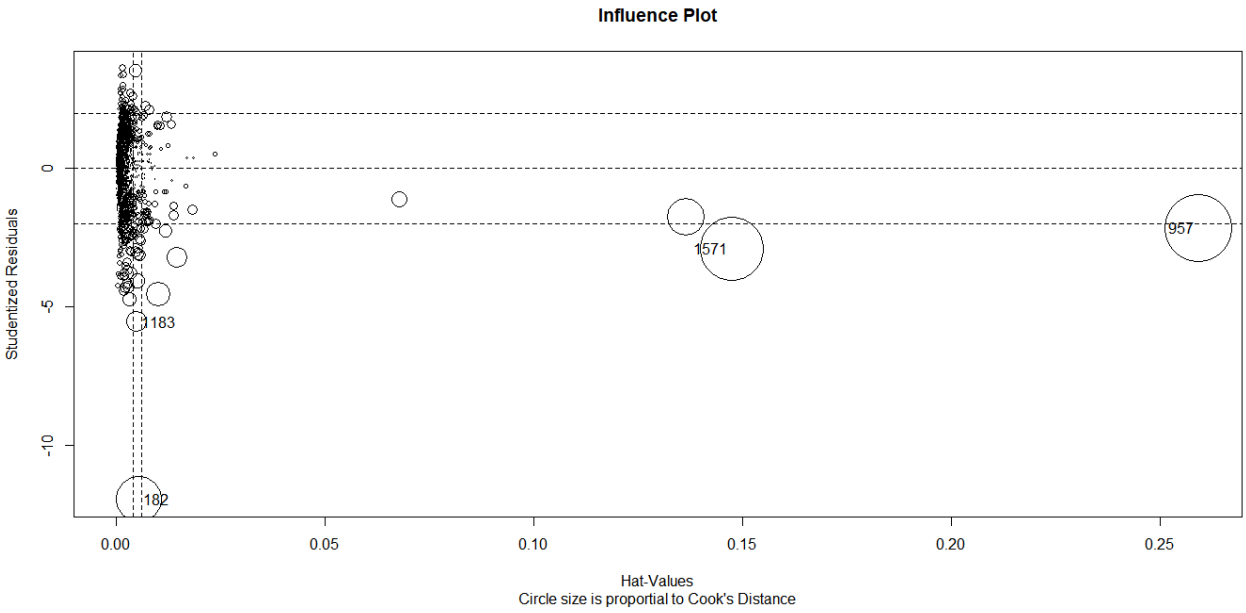


The fitted graph on the left illustrates the not transformed sale price while the graph on the right is of the log transformed sale price. The major difference noticed in the two graphs is that the graph with the log transformed price is a much tighter distribution of data than the not transformed price. This is good as it shows less of a funnel shaped trend and a more randomized set of data.

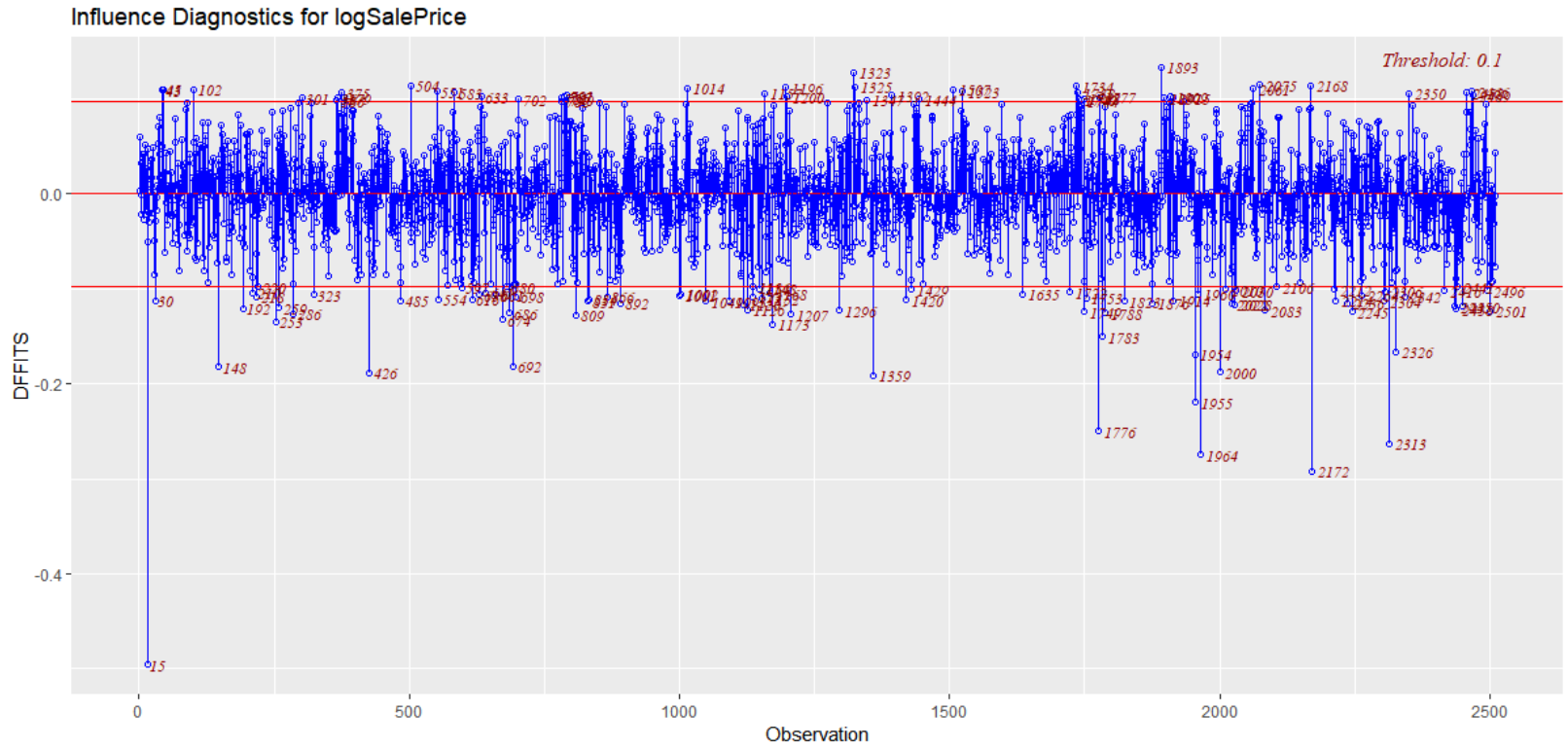


The fitted graph on the left illustrates the not transformed sale price while the graph on the right is of the log transformed sale price. Both of the graph illustrate three points that are highly influential. Two of the points are the same on both graphs, 1571 and 957. These points are more highlighted on a significance graph in task 7.

Task 7:



As you can see above, there are many points that are considered outliers that have a large degree of influence over the model. To determine which points are the largest influencers and remove those points, DFITS must be calculated. The n of our dataset is 2896 and the p of the dataset is 12. This results in a DFITS of .134.



When this value is set as the threshold, 74 data points are identified as highly influential and are subsequently removed. The above graph indicates the threshold line and the points observations that are outlies past the lines. Post this adjustment, the adjusted R squared value becomes .81 which is .4 higher than the previous R squared indicating that the removal of these points has resulted in a moderately more reliable model.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.80068 -0.09623  0.00392  0.10263  0.63304

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.144e+01  1.511e-02  757.04   <2e-16 ***
GrLivArea     4.717e-04  9.262e-06   50.93   <2e-16 ***
GarageArea    3.753e-04  1.960e-05   19.15   <2e-16 ***
LotArea       9.977e-06  7.193e-07   13.87   <2e-16 ***
HouseAge     -4.564e-03  1.221e-04  -37.37   <2e-16 ***
BedroomAbvGr -7.632e-02  4.711e-03  -16.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1628 on 2815 degrees of freedom
Multiple R-squared:  0.8165,    Adjusted R-squared:  0.8162
F-statistic:  2505 on 5 and 2815 DF,  p-value: < 2.2e-16
```

Task 8:

A model must be validated to ensure that the predictions and conclusions presented are accurate. Without appropriate validation of models, there would be a substantial lack of confidence in any of the conclusions presented by the model. These validations include observing the individual values that appear to be highly influential on the data or are considered outliers. These data points can skew the data towards results that may not be accurate for the entire data set, thus removing or transforming the data can present a more accurate story. It is important to consider the points that are being transformed or removed and to carefully examine the impact that adjusting these variables presents. Removing variables should be done with extreme caution to prevent artificially crafting a conclusion that is not correct.

Some next steps to continue this modeling process could be to examine the affects of other various predictor variables in the model as well as utilizing the model to create predictions.