Julia Rodd Assignment 3

Contents

Introduction 1

Results 1

Section 1 2

Section 2 4

Section 3 7

Section 4 9

Model 1: TotalFloorSF and HouseAge 9

Section 5 12

Model 2: TotalFloorSF, HouseAge, and NbhdGrp 15

Section 6 16

Model 3: SalePrice Model 16

Model 4: logSalePrice Model 18

Model 3 and 4 Comparison 21

Summary and Conclusions 25

Introduction

This assignment utilizes housing data from Ames, IA from 2006 - 2010 and expands upon the analyses in Assignments 1 and 2. The overall goal of this assignment is to define and incorporate categorical variables into a regression model and compare the fits of several types of models using various goodness of fits metrics and techniques. Some consideration is given to identifying and removing influential points. Ultimately, simple and multiple linear regression models are created and commentary is provided on their model interpretation and goodness of fit. The R ggplot and dplyr packages are the two primary packages used in this assignment.

Results

Before we get started, it is worth noting that for this assignment we utilize the same drop conditions defined in Assignment 2.

That is, we will fit various models for typical Ames houses.

Typical Ames houses are defined as:

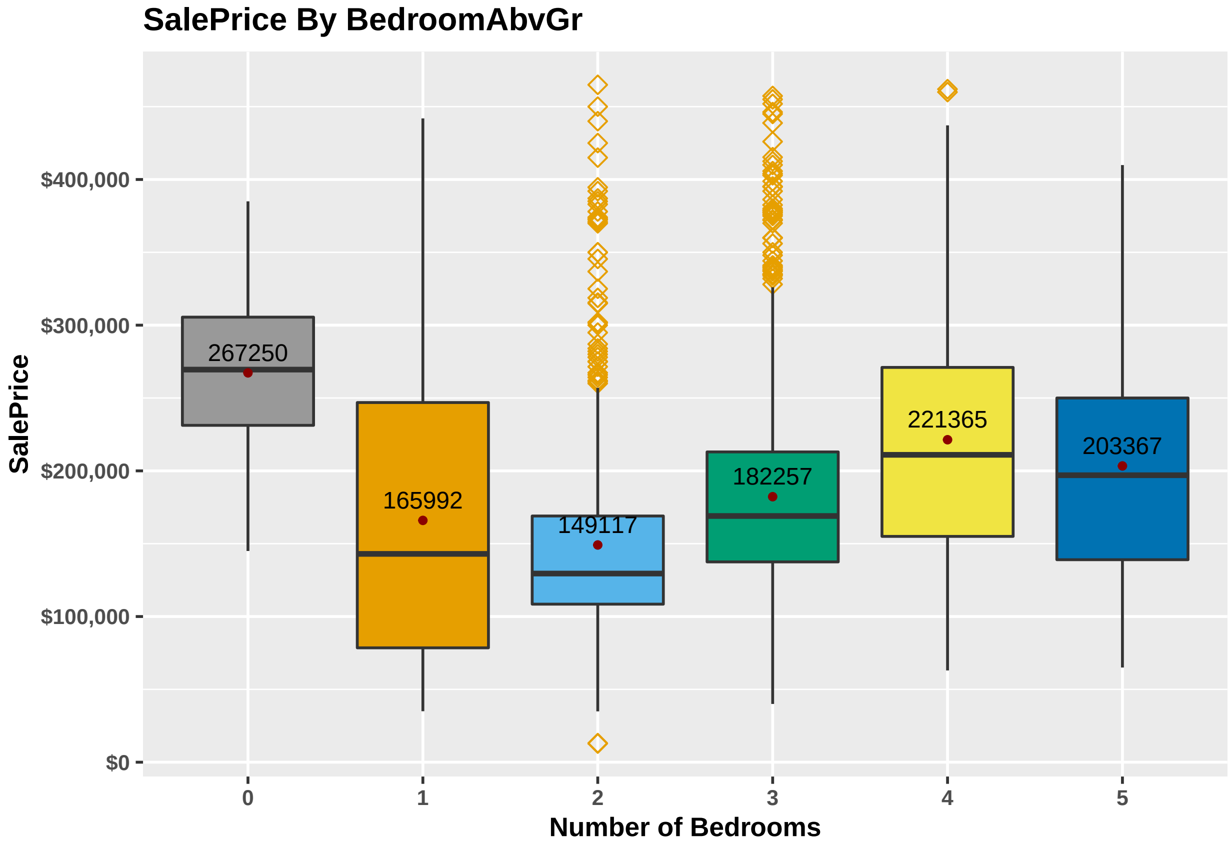
1. A single family home (BldgType == "1Fam");
2. Houses with a total square feet < 3,600 (TotalFloorSF < 3600); and,
3. Houses a sale price < 465,500 (SalePrice < 465500)

In defining our data this way, we remove 536 homes from our data set and are left with 2394 homes for modeling. The rationale in limiting our scope of observations is to improve model performance by removing variability.

Section 1

In this section, we perform some analysis using the BedroomAbvGr variable. For this analysis, we will be treating BedroomAbvGr as a factor (categorical) variable.

We begin by analyzing its spread and report the mean SalePrice for each category in the below visual.



We can see that the mean sale prices go up and down based on the number of bedrooms. This pattern is very interesting, as we would expect the sale price to increase as number of bedrooms increase. Since the means are not constant across each category in BedroomAbvGr, this makes BedroomAbvGr a worthwhile variable to explore further.

We now create a simple linear regression model using BedroomAbvGr as the predictor for SalePrice.

```
## Analysis of Variance Table
##
## Response: SalePrice
##                         Df         Sum Sq      Mean Sq F value
## as.factor(BedroomAbvGr)    5  1092840670850 218568134170  44.235
## Residuals               2388 11799264811961   4941065667
##                                      Pr(>F)
## as.factor(BedroomAbvGr) < 0.00000000000000022 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can quickly validate from the ANOVA results that our model is significant.

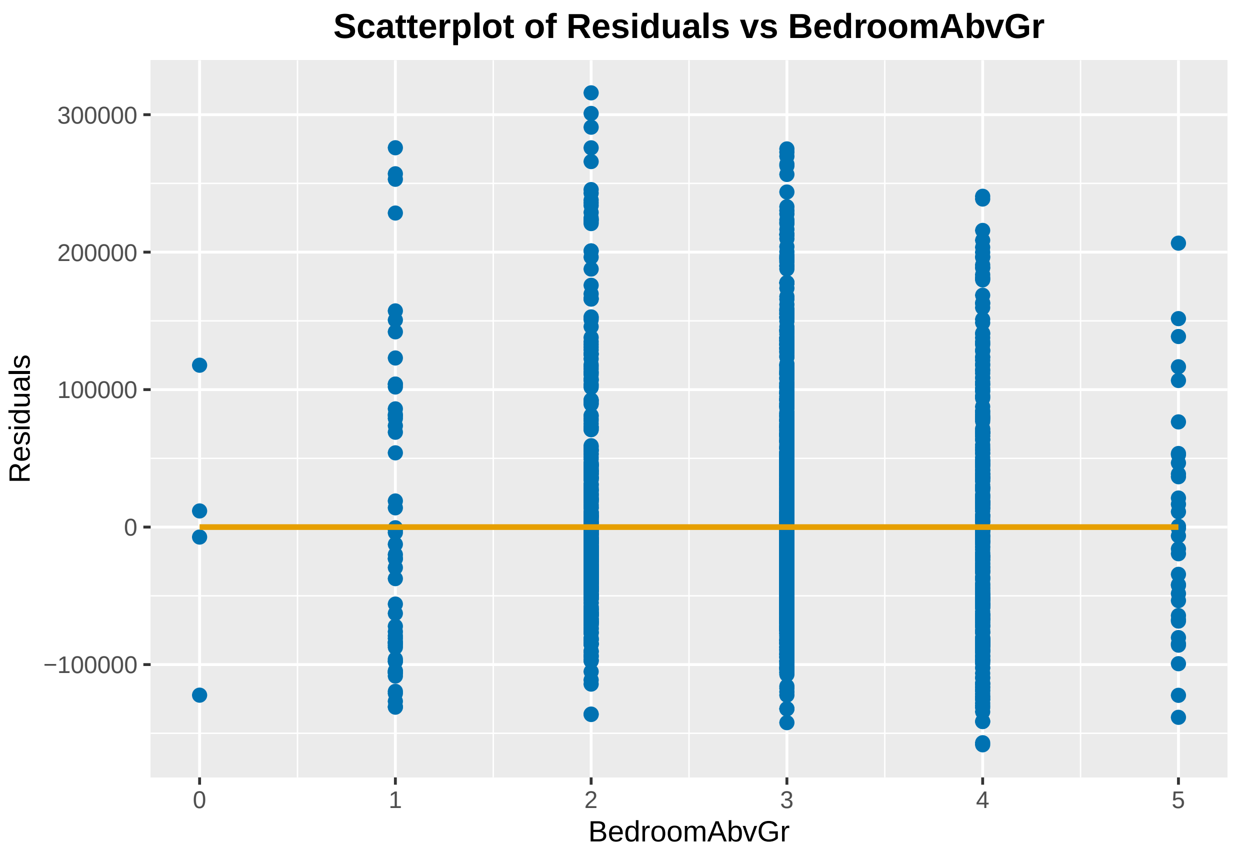
```
##                             Estimate Std. Error   t value
## (Intercept)                267250.00   35146.36  7.603918
## as.factor(BedroomAbvGr)1 -101258.49   36448.51 -2.778124
## as.factor(BedroomAbvGr)2 -118132.69   35287.22 -3.347747
## as.factor(BedroomAbvGr)3  -84993.08   35193.72 -2.415007
## as.factor(BedroomAbvGr)4  -45885.11   35363.31 -1.297534
## as.factor(BedroomAbvGr)5  -63883.33   37215.53 -1.716577
##                                       Pr(>|t|)
## (Intercept)              0.0000000000000410277
## as.factor(BedroomAbvGr)1 0.0055101175889978289
## as.factor(BedroomAbvGr)2 0.0008273497403602111
## as.factor(BedroomAbvGr)3 0.0158096080269842613
## as.factor(BedroomAbvGr)4 0.1945727568937927199
## as.factor(BedroomAbvGr)5 0.0861861187236832721
```

In interpreting the model, we can see that R has chosen BedroomAbvGr = 0 (0 bedrooms) as the baseline cateogry. We did not create dummy variables to alter this interpretation.

In calculating the average SalePrice for each category in BedroomAbvGr, we can see that the values are the mean SalePrices from the boxplot above. The interpretation of these coefficients is provided below:

- 0 bedrooms: has a mean SalePrice of 267250 dollars (the intercept)
- 1 bedroom: has a mean SalePrice that is -101258 dollars lower than homes with 0 bedrooms (mean = 165992)
- 2 bedrooms: has a mean SalePrice that is -118133 dollars lower than homes with 0 bedrooms (mean = 149117)
- 3 bedrooms: has a mean SalePrice that is -84993 dollars lower than homes with 0 bedrooms (mean = 182257)
- 4 bedrooms: has a mean SalePrice that is -45885 dollars lower than homes with 0 bedrooms (mean = 221365)
- 5 bedrooms: has a mean SalePrice that is -63883 dollars lower than homes with 0 bedrooms (mean = 203367)

Next, we examine a plot of the residuals versus our predictor, BedroomAbvGr.



From this plot of residuals vs BedroomAbvGr, we can see a slight pattern. We notice that the 0 line does not go through the average of each bedroom. If we were to create a line that goes through the average of each bar (representing the residuals for each number of bedrooms) then we would not get a completely straight line. This information tells us that our linearity assumption is violated.

Additionally, we can calculate the mean absolute error from this model, which is 52778. The mean error is quite large, which indicates that the fit of this model could be improved.

Section 2

In this section, we explicitly define dummy variables for BedroomAbvGr and use these new variables to fit a model predicting SalePrice.

We will use homes with 5 bedrooms as our baseline category.

```
##              Estimate Std. Error   t value
## (Intercept) 203366.67   12236.39 16.619824
## bed0         63883.33   37215.53  1.716577
## bed1        -37375.16   15587.08 -2.397830
## bed2        -54249.36   12635.31 -4.293473
## bed3        -21109.74   12371.79 -1.706281
## bed4         17998.22   12846.27  1.401047
##                                                                      Pr(>|t|)
## (Intercept) 0.00000000000000000000000000000000000000000000000000000000008832018
## bed0        0.08618611872362802461822184341144748032093048095703125000000000000
## bed1        0.01656886088971543102843497763387858867645263671875000000000000000
## bed2        0.00001829258340771867404409012358712516288505867123603820800781250
## bed3        0.08808577207180245949302843655459582805633544921875000000000000000
## bed4        0.16133006063115973294230798273929394781589508056640625000000000000
```

In interpreting the model, we can see that our coefficients have changed from above with 5 bedrooms as the baseline.

In calculating the average SalePrice for each category in BedroomAbvGr, we can see that the values are the mean SalePrices from the boxplot above. The interpretation of these coefficients is provided below:

- 5 bedrooms: has a mean SalePrice of 203367 dollars (the intercept)
- 4 bedrooms: has a mean SalePrice that is 63883 dollars higher than homes with 5 bedrooms (mean = 267250)
- 3 bedrooms: has a mean SalePrice that is -37375 dollars lower than homes with 5 bedrooms (mean = 165992)
- 2 bedrooms: has a mean SalePrice that is -54249 dollars lower than homes with 5 bedrooms (mean = 149117)
- 1 bedroom: has a mean SalePrice that is -21110 dollars lower than homes with 5 bedrooms (mean = 182257)
- 0 bedrooms: has a mean SalePrice that is 17998 dollars higher than homes with 5 bedrooms (mean = 221365)

We can see that while the coefficients remained the same, defining our own dummy variables altered the interpretation of the model. We were able to utilize a different baseline category and have control over the model interpretation. Similar to above, we can see that the predicted model goes through the mean of each bedroom category. Interpreting the coefficients helps to demonstrate this point.

Furthermore, we can add our newly defined dummy variables into a regression model with TotalFloorSF. We will perform this exercise to demonstrate how to interpret a model with both continuous and categorical variables. Since we are using our dummy variables instead of BedroomAbvGr, 5 bedrooms becoms our baseline category.

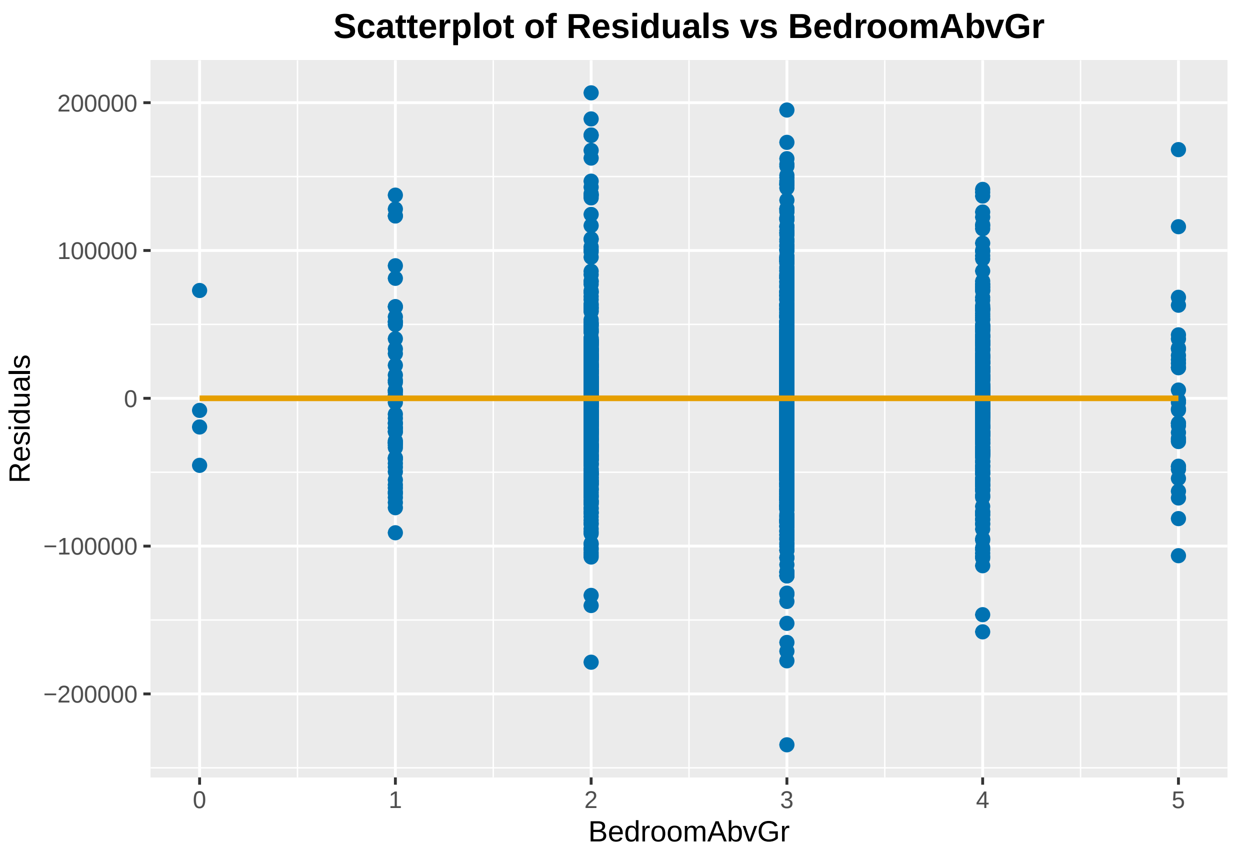
```
## Analysis of Variance Table
##
## Response: SalePrice
##                Df        Sum Sq       Mean Sq  F value
## TotalFloorSF    1 7470557209491 7470557209491 3709.236
## bed0            1   27994688495   27994688495   13.900
## bed1            1   42298972083   42298972083   21.002
## bed2            1   54177343977   54177343977   26.900
## bed3            1  431114564329  431114564329  214.054
## bed4            1   58444490268   58444490268   29.018
## Residuals    2387 4807518214166    2014041983
##                             Pr(>F)
## TotalFloorSF < 0.00000000000000022 ***
## bed0                     0.0001973 ***
## bed1                 0.00000482432 ***
## bed2                 0.00000023230 ***
## bed3         < 0.00000000000000022 ***
## bed4                 0.00000007871 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that in this multiple linear regression model, all coefficients are significant.

Below is the output of the coefficients. We will use this information to interpret our model.

```
##                  Estimate  Std. Error    t value
## (Intercept)  -106037.7153  9413.17483 -11.264819
## TotalFloorSF     137.9758     2.34177  58.919436
## bed0          163943.9855 23820.70876   6.882414
## bed1          116330.0819 10287.75338  11.307627
## bed2           95591.8663  8458.33737  11.301496
## bed3           84259.8355  8098.63682  10.404200
## bed4           44246.4348  8213.73627   5.386883
##                                           Pr(>|t|)
## (Intercept)  0.00000000000000000000000000010241926
## TotalFloorSF 0.00000000000000000000000000000000000
## bed0         0.00000000000749331740239979346819912
## bed1         0.00000000000000000000000000006451307
## bed2         0.00000000000000000000000000006893503
## bed3         0.00000000000000000000000079830395485
## bed4         0.00000007871223245883801702699883451
```

- 5 bedrooms: for each additional increase in 1 square foot, the average SalePrice increases by -106038 dollars
- 4 bedrooms: has a mean SalePrice that is 138 dollars higher than homes with 5 bedrooms when TotalFloorSF is held constant
- 3 bedrooms: has a mean SalePrice that is 163944 dollars higher than homes with 5 bedrooms when TotalFloorSF is held constant
- 2 bedrooms: has a mean SalePrice that is 116330 dollars higher than homes with 5 bedrooms when TotalFloorSF is held constant
- 1 bedroom: has a mean SalePrice that is 95592 dollars higher than homes with 5 bedrooms when TotalFloorSF is held constant
- 0 bedrooms: has a mean SalePrice that is 84260 dollars higher than homes with 5 bedrooms when TotalFloorSF is held constant



In comparing the first model with just BedroomAbvGr to now the multiple linear regression model with TotalFloorSF and BedroomAbvGr dummy variables, we can see that the residuals now go through the middle or average of the residuals for each bedroom.

Moerover, our mean absolute error is much improved. In this model, it is 32635.

Overall, we can see that adding a continuous variable has improved the fit of our model and enables us to meet the linearity assumption.

Section 3

In this section, we perform hypothesis testing on each of the coefficients in our regression model with BedroomAbvGr (dummy variables) as the sole predictor.

The hypothesis tests can be written as follows. We will perform an hypothesis testing on each coefficient to determine if it is significant or not. Therefore, the t test will be used.

bed0

- H0: bed0 = 0
- H1: bed0 is not equal to 0

bed1

- H0: bed1 = 0
- H1: bed1 is not equal to 0

bed2

- H0: bed2 = 0
- H1: bed2 is not equal to 0

bed3

- H0: bed3 = 0
- H1: bed3 is not equal to 0

bed4

- H0: bed4 = 0
- H1: bed4 is not equal to 0

Below is the output from our regression model. We will use this information to interpret the significance of our coefficients.

```
##
## Call:
## lm(formula = SalePrice ~ bed0 + bed1 + bed2 + bed3 + bed4, data = ames_subdat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158365  -46471  -15257   32743  315883
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   203367      12236  16.620 < 0.0000000000000002 ***
## bed0           63883      37216   1.717               0.0862 .
## bed1          -37375      15587  -2.398               0.0166 *
## bed2          -54249      12635  -4.293            0.0000183 ***
## bed3          -21110      12372  -1.706               0.0881 .
## bed4           17998      12846   1.401               0.1613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70290 on 2388 degrees of freedom
## Multiple R-squared:  0.08477,    Adjusted R-squared:  0.08285
## F-statistic: 44.24 on 5 and 2388 DF,  p-value: < 0.00000000000000022
```

We can see that bed1 and bed2 are significant predictors at the .05 level, since they have higher t-values and lower p-values. Therefore, we reject the null hypotheses for bed1 and bed2 and conclude that they are different than 0. However, bed0, bed3, and bed4 are not significant at the .05 level so we fail to reject the null hypothesis for each of these predictors. It is worth noting that bed0 and bed3 may be considered significant at the .10 level. For bed4, with such a high p-value, we fail to reject the null hypothesis and conclude that bed4 is not different than 0.

A potential next step might be to perform partial hypothesis testing to compare a reduced model just with bed1 and bed2 to this full model with all bedroom types, but that exercise is beyond the scope of the assignment.

Section 4

In this section, we will create a multiple linear regression model using the predictors of TotalFloorSF and HouseAge. We choose TotalFloorSF, since we know it has a strong positive linear relationship with SalePrice. We choose HouseAge, since it has a fairly strong negative linear relationship with SalePrice. In addition, we have not incorporated HouseAge into a regression model in any assignment thus far and are curious to see how it performs.

Model 1: TotalFloorSF and HouseAge

From the ANOVA results below, we can see that this model is significant.

```
##                Df        Sum Sq       Mean Sq F value              Pr(>F)
## TotalFloorSF    1 7470557209491 7470557209491    5185 <0.0000000000000002
## HouseAge        1 1976760791937 1976760791937    1372 <0.0000000000000002
## Residuals    2391 3444787481382    1440730858
##
## TotalFloorSF ***
## HouseAge     ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

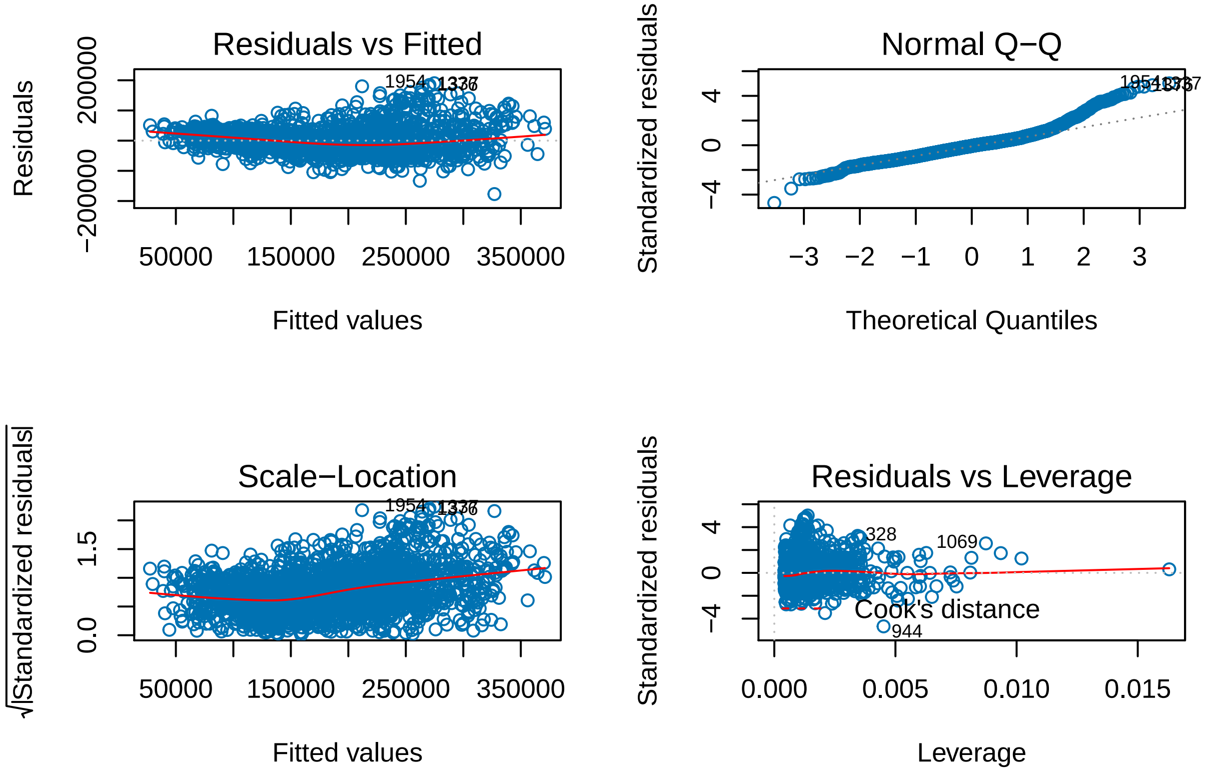
We will now examine some metrics to assess goodness of fit.

We can see that for Model 1 our R squared value is: 73.3%. In contrast, the R squared value for our simple linear regression model with BedroomAbvGr as the predictor is: 8.5%. As another data point, we can see that adding TotalFloorSF to the BedroomAbvGr model increased the R squared value to: 62.7%. Therefore, as more predictors are added to a model, the R squared value increases.

However, R squared is not a metric to be used when comparing models. R squared tells us how well a given model explains the variability of our response variable, but adjusted R squared takes into account the number of variables used in a model since adding more variables to a model is not always advantageous. Adjusted R squared can be used to compare models against one another.

For instance, the adjusted R squared for Model 1 is: 73.3%, while the adjusted R squared for our simple linear regression model is: 8.3%. Moreover, the adjusted R squared for the model with BedroomAbvGr and TotalFloorSF is: 62.6%. From the adjusted R squared value, Model 1 is the superior model. Further exploration is needed however to determine how well Model 1 meets our linear regression assumptions.

We will begin by examining diagnostic plots of our residuals.



We gain a lot of valuable information about the behavior of this model by examining the residuals.

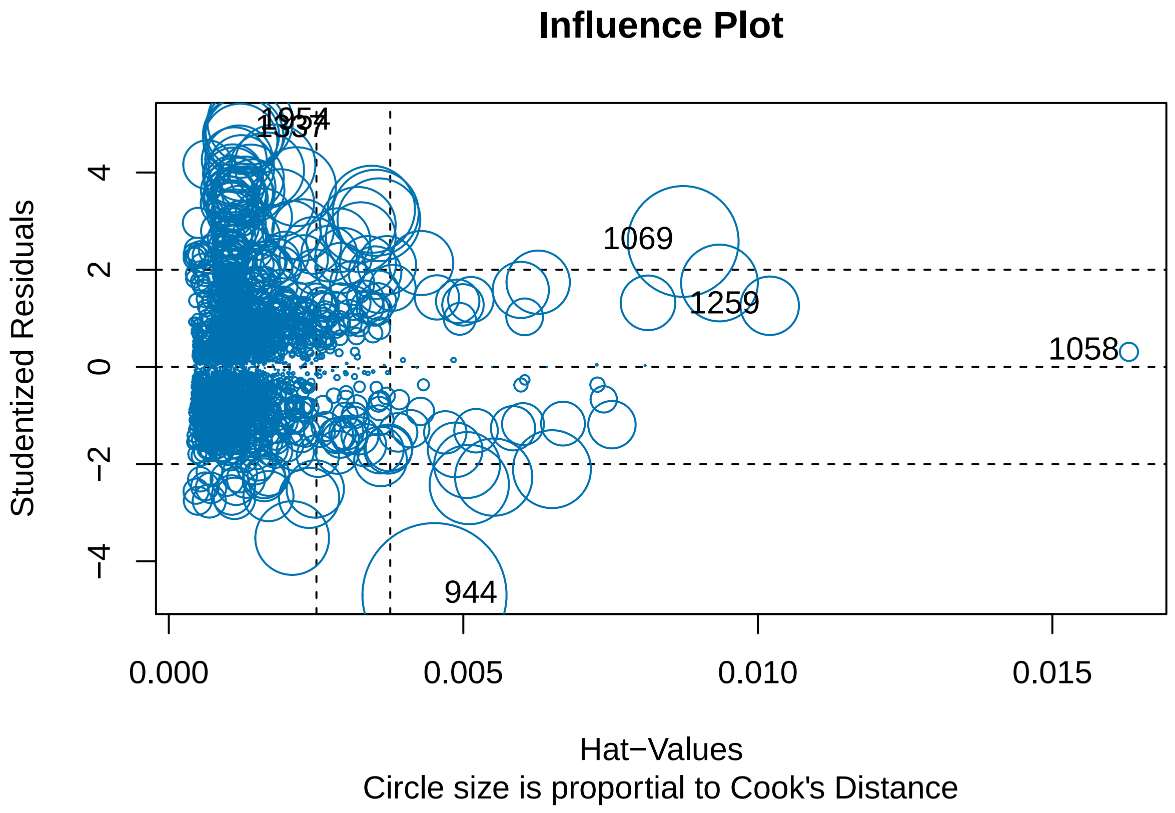
First, we can see that the residuals vs fitted values demostrate a pattern. They have a slight funnel shape, which is indiciative of heteroscedasticity. A transformation of the response variable would alleviate this behavior. Also, we can see that the red line is not straight and has a slight U-shape, which tells us that a linear model is not an appropriate fit here.

In examining the normal QQ-plot, we can see that the residuals do not fall perfectly on the line. We see a departure from the normal line especially in the right tail. Therefore, our normality assumption is violated. We can also see some outliers (numbered points) in the right tail.

The standardized residuals vs fitted values echo the conclusions drawn from the residuals vs fitted values plot. We can see increasing variation and a non straight red line, which tells us that this linear model is not an appropriate fit and constant variance is not met.

Lastly, from the residuals vs leverage plot, we can see that we have outliers in the x direction and some outliers in the y direction. The worst possible case would be outliers in both the x- and y- direction, which we do not have. Utilizing our drop conditions may have helped to remove some potential influential points. Further study of these outliers is needed to determine if they should be included in the model or not.

We now examine an influence plot of this multiple linear regression model.



```
##          StudRes         Hat        CookD
## 944  -4.6956212 0.004510628 0.0330110013
## 1058  0.3092368 0.016299986 0.0005283843
## 1069  2.5810519 0.008732912 0.0195170154
## 1259  1.2578461 0.010201753 0.0054344568
## 1337  4.8862990 0.001302563 0.0102817770
## 1954  5.0405660 0.001380021 0.0115854271
```

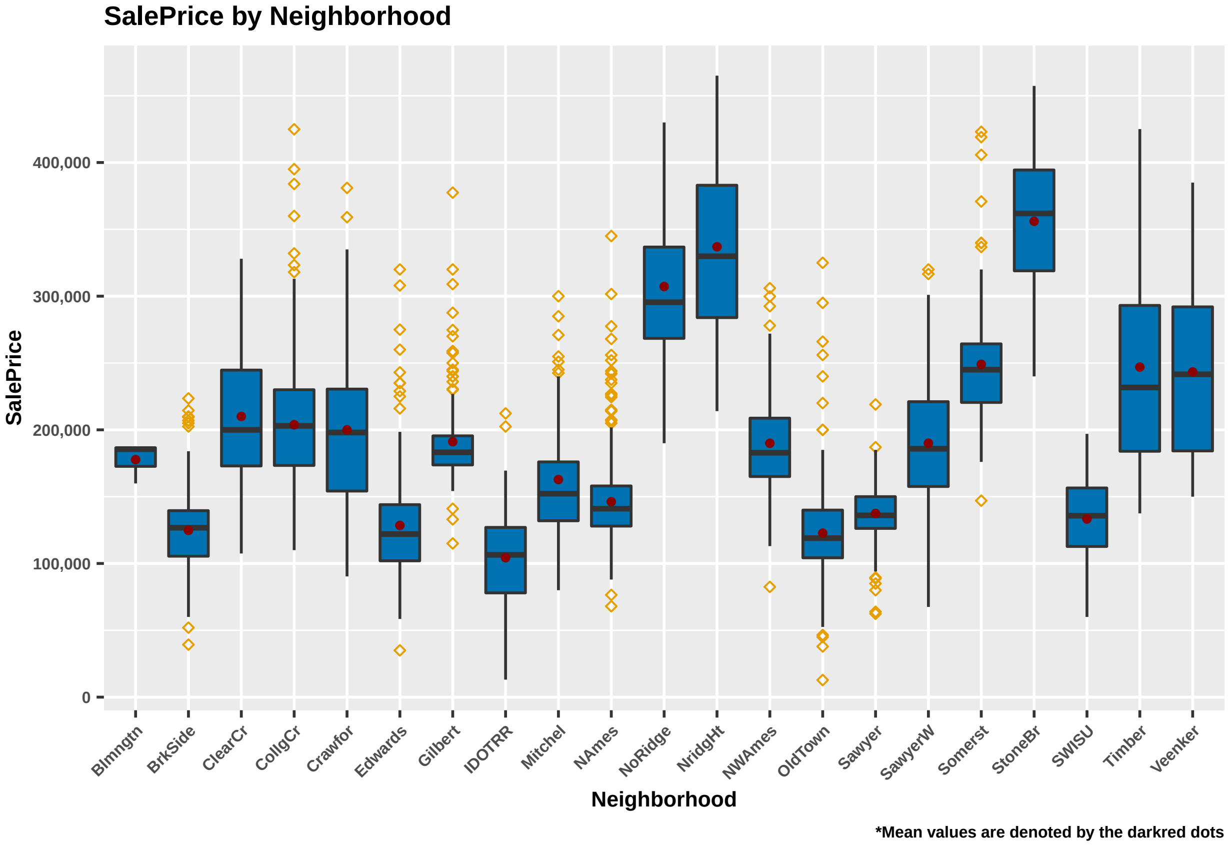
This plot presents a slightly different perspective to the residuals vs leverage plot shown above. Here, the circles are proportional to Cook's distance. We see that there are several larger circles, which means that these points have higher hat values and are thus influential points (e.g., circle near .005 on the x-axis, circle near (.08,2)). There are statistical methods to identify and aid in the removal of these points, but we will not go into that detail right now.

Finally, multicollinearity is an issue that should be evaluated as part of model adequacy checking process. We can assess this by calculating the Variance Inflation Factor (VIF). For Model 1, the VIF is 1.1209529, 1.1209529. We want the VIF to be close to 1, which means our predictors are not correlated. In this case, since our VIF is just above 1, we do not have multicollinearity issues to address.

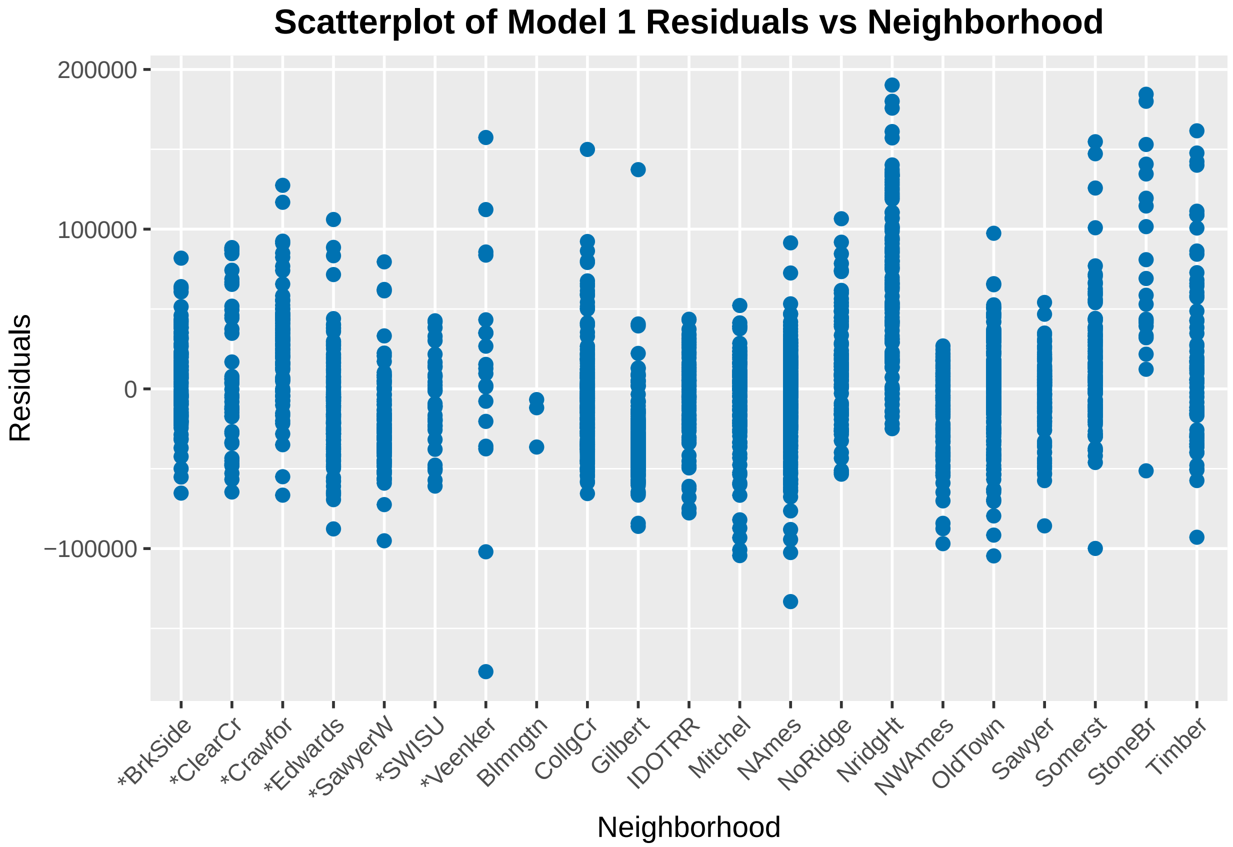
Overall, this multiple linear regression model does not fit the data well, since it violates several linear model assumptions. It is important to note that when assessing goodness of fit, both metrics, such as R squared, adjusted R squared, residual standard error, VIF, and graphical displays of residuals are needed to derive appropriate conclusions. When comparing a reduced model, such a simple linear regression model, to a multiple linear regression model, the model with more predictors has a higher R squared value. Goodness of fit is not just about if the model at hand explains variability in the response variable: it is also about validating that the model meets core assumptions of linearity, normality, and constant variance. If these assumptions were not validated, then inappropriate conclusions would be drawn. Goodness of fit is also a step to help identify any outliers, influential points, and multicollinearity that could be negatively impacting model fit/performance.

Section 5

In this section, we explore how the Neighborhood variable behaves. As we saw from Assignment 1 with a boxplot of SalePrice by Neighborhood, we could see variability in the SalePrice across neighborhoods. This boxplot is shown below.



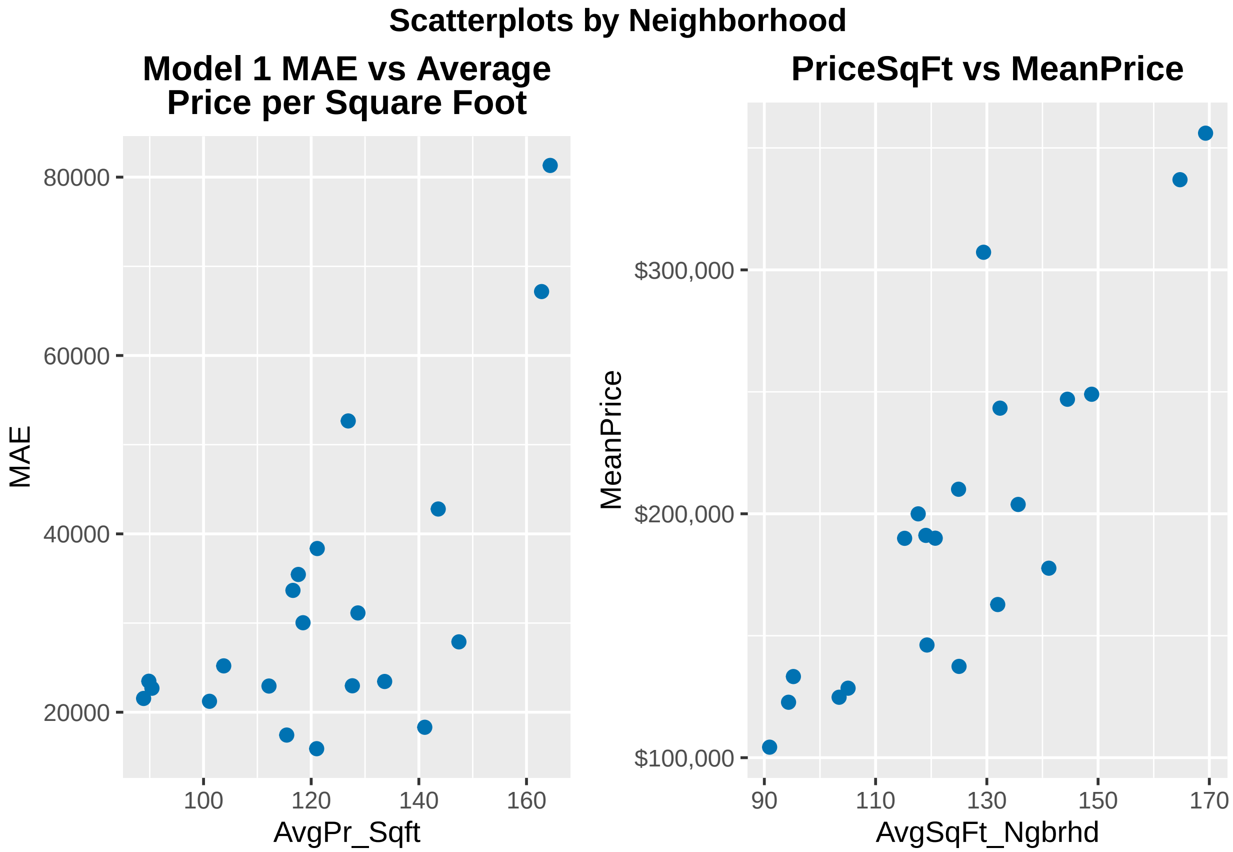
We now evaluate boxplots of the residuals from Model 1 by Neighborhood.



From the residual plot above, we can see that Model 1 fits certain neighborhoods better than others. The neighborhoods with an asterisk (on the left side of the plot) indicate that Model 1 fits these neighborhoods better, since 0 goes through the average of those points. We can see that for neighborhoods without an asterisk, the 0 line for the residuals does not go through the middle of their points.

Moreover, we can visually see that some neighborhoods have more points (or homes). When modeling using a categorical variable, it is important to ensure that the categories have an equal numer of homes, as that helps to increase the accuracy of the model.

Next, we will compare the mean absolute error from Model 1 to the price per square foot for each neighborhood to determine if there are any trends. We also plot mean SalePrice and price per square foot as another comparison.



From the plot of MAE vs AvgPr_Sqft above, we can see that the generally MAE and AvgPr_Sqft have a positive linear relationship. We can also start to notice that trends are different by neighborhood. For example, we can see there are 3 neighborhoods in the bottom left of the plot and 2 neighborhoods in the top right of the plot.

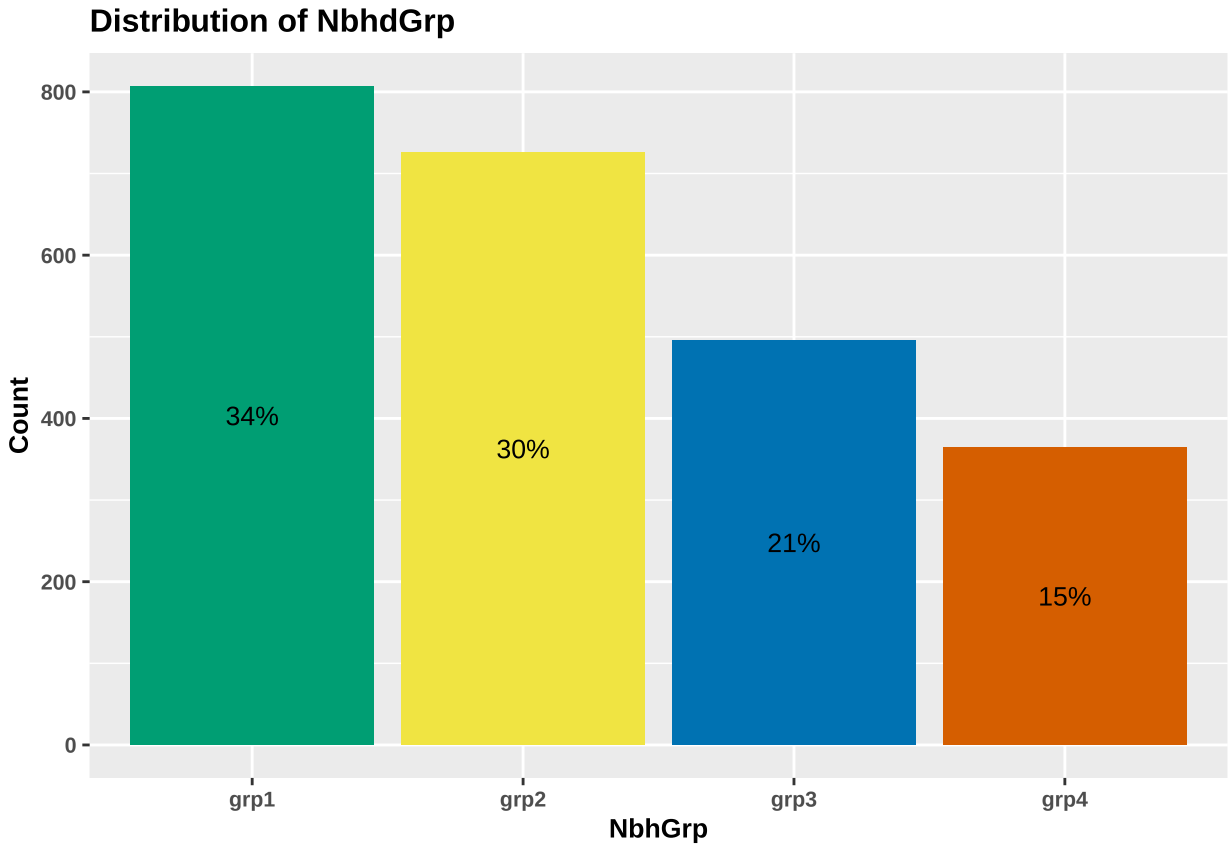
Similarly, in the plot on the right, we can see that MeanPrice and AvgSqFt_Ngbrhd also have a positive linear relationship.

From this information, we can separate our neighborhood variable into neighborhood groups by AvgPr_Sqft. Our groups will be:

- <= 110 AvgPr_Sqft (5 neighborhoods)
- 111 - 130 AvgPr_Sqft (8 neighborhoods)
- 131 - 150 AvgPr_Sqft (6 neighborhoods)
- 151+ AvgPr_Sqft (2 neighborhoods)

We will utilize group 4 or neighborhoods with AvgPr_Sqft 151+ as our baseline group.

We perform a quick check to see the number of homes in each neighborhood group.



From the bar chart above, we can see that there is a different number of homes in each group. Ideally these groups would have an equal number of neighborhoods as well as an equal number of houses, but we will move forward with these groups until further analysis tells us otherwise. One consideration might be to exclude neighborhoods with AvgPr_Sqft > 160 or to break our groups into even smaller ones.

Now, we will incorporate these new dummy variables into our multiple linear regression model.

Model 2: TotalFloorSF, HouseAge, and NbhdGrp

The coefficients of this new multiple regression model are shown below. We can readily see that all the NbhdGrp variables have negative coefficients, since group 4 is the baseline category, and it is based on higher price_sqft than the other groups.

```
##                  Estimate  Std. Error   t value
## (Intercept)    86756.9947 2233.503791  38.84345
## TotalFloorSF     115.8634    1.127332 102.77662
## HouseAge        -300.2721   20.507530 -14.64204
## NbhdGrp1     -106768.2734 1740.445351 -61.34538
## NbhdGrp2      -71082.0244 1540.933533 -46.12920
## NbhdGrp3      -47265.8439 1626.365548 -29.06225
##
## (Intercept)  0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## TotalFloorSF 0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## HouseAge     0.0000000000000000000000000000000000000000000001484068381215963067145530152046717375924 3
## NbhdGrp1     0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## NbhdGrp2     0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## NbhdGrp3     0.0000000000000000000000000000000000000000000000000000000000000000000000000000000000000
```

If we compare the mean absolute errors, Model 1 has a MAE of 27,300, while Model 2 has a MAE of 15,917. We can see that on the basis of MAE alone, Model 2 is the better model since it has a lower MAE.

Section 6

In this section, we will compare the results to two multiple linear regression models. These models will have the same predictors but the response variable will be SalePrice in one model and logSalePrice in another model.

The models will use the following predictors:

- TotalFloorSF (continuous)
- HouseAge (continuous)
- QualityIndex (continuous)
- Price per square feet (price_sqft; continuous)
- Neighborhood Group variable (categorical)

The above variables were chosen due to their relationships with SalePrice, as they are either correlated with SalePrice and/or demonstrate patterns in SalePrice based on different groups. This analysis was performed in Assignments 1 and 2.

Model 3: SalePrice Model

From the ANOVA results below, we can see that all predictors, except NbhdGrp3, are significant. If we were to select this model, then we might compare this model without NbhdGrp3, but for now, we will move forward with this variable included.

```
##                 Df        Sum Sq       Mean Sq   F value
## TotalFloorSF    1 7470557209491 7470557209491 41911.086
## HouseAge        1 1976760791937 1976760791937 11089.961
## QualityIndex    1  724977281136  724977281136  4067.245
## price_sqft      1 2283150894784 2283150894784 12808.862
## NbhdGrp1        1    3592039621    3592039621    20.152
## NbhdGrp2        1    7730826577    7730826577    43.371
## NbhdGrp3        1      37282097      37282097     0.209
## Residuals    2386  425299157167     178247761          
##                            Pr(>F)    
## TotalFloorSF < 0.0000000000000002 ***
## HouseAge     < 0.0000000000000002 ***
## QualityIndex < 0.0000000000000002 ***
## price_sqft   < 0.0000000000000002 ***
## NbhdGrp1         0.0000074927659 ***
## NbhdGrp2         0.0000000000555 ***
## NbhdGrp3                   0.647    
## Residuals                            
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can interpret the model using the coefficient table below.

```
##                  Estimate   Std. Error     t value         Pr(>|t|)
## (Intercept)  -208030.7830 4416.5548440 -47.1025019 0.00000000000000
## TotalFloorSF     121.6855    0.7681883 158.4057760 0.00000000000000
## HouseAge         -21.5434   13.0078109  -1.6561894 0.09781495718750
## QualityIndex    -142.6349   38.1463278  -3.7391526 0.00018898624712
## price_sqft      1703.0938   25.4705178  66.8652997 0.00000000000000
## NbhdGrp1       10893.4507 1950.6223918   5.5846025 0.00000002608048
## NbhdGrp2        6389.5852 1415.8687265   4.5128373 0.00000670702138
## NbhdGrp3         528.3768 1155.3288949   0.4573389 0.64746912425907
```

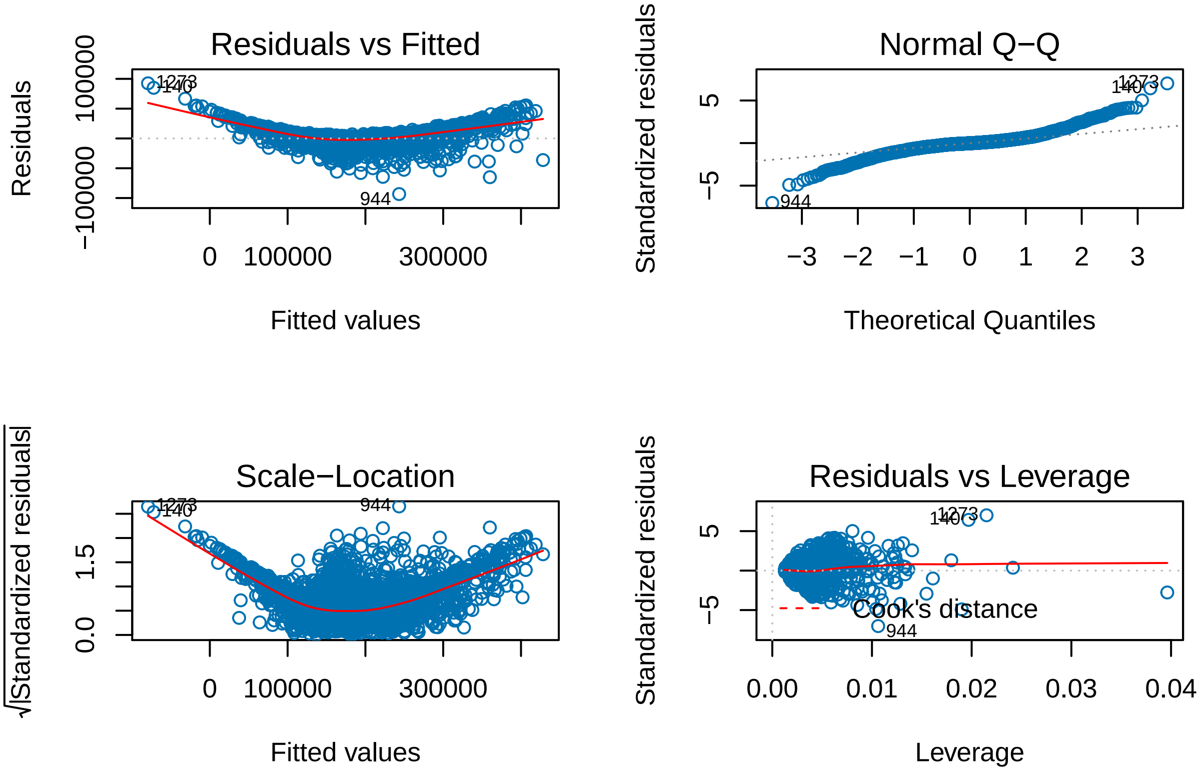
Here is a summary of the coefficient interpretation. Please note that y is SalePrice.

- NbhdGrp1
  - The model is: y = -208031 + 122 x TotalFloorSF - 22 x HouseAge - 143 x QualityIndex + 1703 x price_sqft + 10893
  - For each additional 1 square foot, the average SalePrice of a home goes up by 122 dollars when all other variables are held constant.
  - For each additional increase in house age, the average SalePrice of a home decreases by 22 dollars when all other variables are held constant.
  - For each additional increase in quality index, the average SalePrice of a home decreases by 143 dollars when all other variables are held constant.
  - For each additional increase in price per square foot, the average SalePrice of a home goes up by 1703 dollars when all other variables are held constant.
  - Compared to NbhdGrp4, NbhdGrp1 has an average SalePrice of a home that is 10893 dollars higher than NbhdGrp4.
- NbhdGrp2
  - The model is: y = -208031 + 122 x TotalFloorSF - 22 x HouseAge - 143 x QualityIndex + 1703 x price_sqft + 6390
  - For each additional 1 square foot, the average SalePrice of a home goes up by 122 dollars when all other variables are held constant.
  - For each additional increase in house age, the average SalePrice of a home decreases by 22 dollars when all other variables are held constant.
  - For each additional increase in quality index, the average SalePrice of a home decreases by 143 dollars when all other variables are held constant.
  - For each additional increase in price per square foot, the average SalePrice of a home goes up by 1703 dollars when all other variables are held constant.
  - Compared to NbhdGrp4, NbhdGrp2 has an average SalePrice of a home that is 6390 dollars higher than NbhdGrp4.
- NbhdGrp3
  - The model is: y = -208031 + 122 x TotalFloorSF - 22 x HouseAge - 143 x QualityIndex + 1703 x price_sqft +528
  - For each additional 1 square foot, the average SalePrice of a home goes up by 122 dollars when all other variables are held constant.
  - For each additional increase in house age, the average SalePrice of a home decreases by 22 dollars when all other variables are held constant.
  - For each additional increase in quality index, the average SalePrice of a home decreases by 143 dollars when all other variables are held constant.
  - For each additional increase in price per square foot, the average SalePrice of a home goes up by 1703 dollars when all other variables are held constant.
  - Compared to NbhdGrp4, NbhdGrp3 has an average SalePrice of a home that is 528 dollars higher than NbhdGrp4. Because this amount is so small, we can see why this variable was not significant.

- NbhdGrp4 (baseline group)
  - The model is: y = -208031 + 122 x TotalFloorSF - 22 x HouseAge - 143 x QualityIndex + 1703 x price_sqft
  - For each additional 1 square foot, the average SalePrice of a home goes up by 122 dollars when all other variables are held constant.
  - For each additional increase in house age, the average SalePrice of a home decreases by 22 dollars when all other variables are held constant.
  - For each additional increase in quality index, the average SalePrice of a home decreases by 143 dollars when all other variables are held constant.
  - For each additional increase in price per square foot, the average SalePrice of a home goes up by 1703 dollars when all other variables are held constant.

Additionally, in all cases, the NbhdGrp variables for each scenario above could have been combined with the intercept. For ease of interpretation, we kept it separate. Therefore, we can see that the intercept is different for each Neighborhood but the slopes of all the remaining variables are the same.

We continue our analysis of Model 3 by examining diagnostic plots of its residuals to assess goodness of fit.



We can see that Model 3 violates the assumptions of linearity, constant variance, and normality from the residual plots above. We can also see some outliers called out in different plots. There do not appear to be highly influential points, as the outliers appear to either be in the x- or y- direction. Even still, the residuals have a pattern, which is easier to see by the curve of the red line, and the residuals depart from the normal QQ-line in the tails.

Model 4: logSalePrice Model

From the ANOVA results below, we can see that all predictors are significant.

```
##                Df Sum Sq Mean Sq  F value               Pr(>F)    
## TotalFloorSF    1 215.23  215.23 35324.32 < 0.0000000000000002 ***
## HouseAge        1  70.66   70.66 11596.07 < 0.0000000000000002 ***
## QualityIndex    1  31.57   31.57  5180.74 < 0.0000000000000002 ***
## price_sqft      1  47.25   47.25  7755.30 < 0.0000000000000002 ***
## NbhdGrp1        1   0.07    0.07    11.56             0.000685 ***
## NbhdGrp2        1   0.92    0.92   151.15 < 0.0000000000000002 ***
## NbhdGrp3        1   0.91    0.91   149.75 < 0.0000000000000002 ***
## Residuals    2386  14.54    0.01                                  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now move forward with the model interpretation, as this interpretation is different from Model 3 above since the response variable is now a log variable.

```
##                   Estimate     Std. Error   t value
## (Intercept)   9.8716537539 0.025821937798 382.29717
## TotalFloorSF  0.0005975153 0.000004491309 133.03812
## HouseAge     -0.0012081286 0.000076051786 -15.88560
## QualityIndex  0.0031557918 0.000223027255  14.14980
## price_sqft    0.0089621371 0.000148916554  60.18228
## NbhdGrp1      0.1555170923 0.011404556685  13.63640
## NbhdGrp2      0.1428145689 0.008278052798  17.25219
## NbhdGrp3      0.0826605310 0.006754774233  12.23735
##                                                                              Pr(>|t|)
## (Intercept)  0.0000000000000000000000000000000000000000000000000000000000000000000000
## TotalFloorSF 0.0000000000000000000000000000000000000000000000000000000000000000000000
## HouseAge     0.0000000000000000000000000000000000000000000000000000000428365255703033
## QualityIndex 0.0000000000000000000000000000000000000000000010424302301572124005270628
## price_sqft   0.0000000000000000000000000000000000000000000000000000000000000000000000
## NbhdGrp1     0.0000000000000000000000000000000000000000007916863294648264744515775426
## NbhdGrp2     0.0000000000000000000000000000000000000000000000000000000000000000605729
## NbhdGrp3     0.0000000000000000000000000000000000193142554426314039840717473595077536
```

Here is a summary of the coefficient interpretation. Please note that y is logSalePrice and there may be slight differences in the coefficients due to rounding.

- NbhdGrp1
  - The model is: y = 9.872 + 0.001 x TotalFloorSF - 0.001 x HouseAge + 0.003 x QualityIndex + 0.009 x price_sqft + 0.156
  - Each additional 1 square foot results in a 0.06% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in house age results in a -0.121% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in quality index results in a 0.316% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in price per square foot results in a 0.896% percentage change in SalePrice when all other variables are held constant.
  - Compared to NbhdGrp4, NbhdGrp1 has an average SalePrice that is 15.552% higher than NbhdGrp4.
- NbhdGrp2

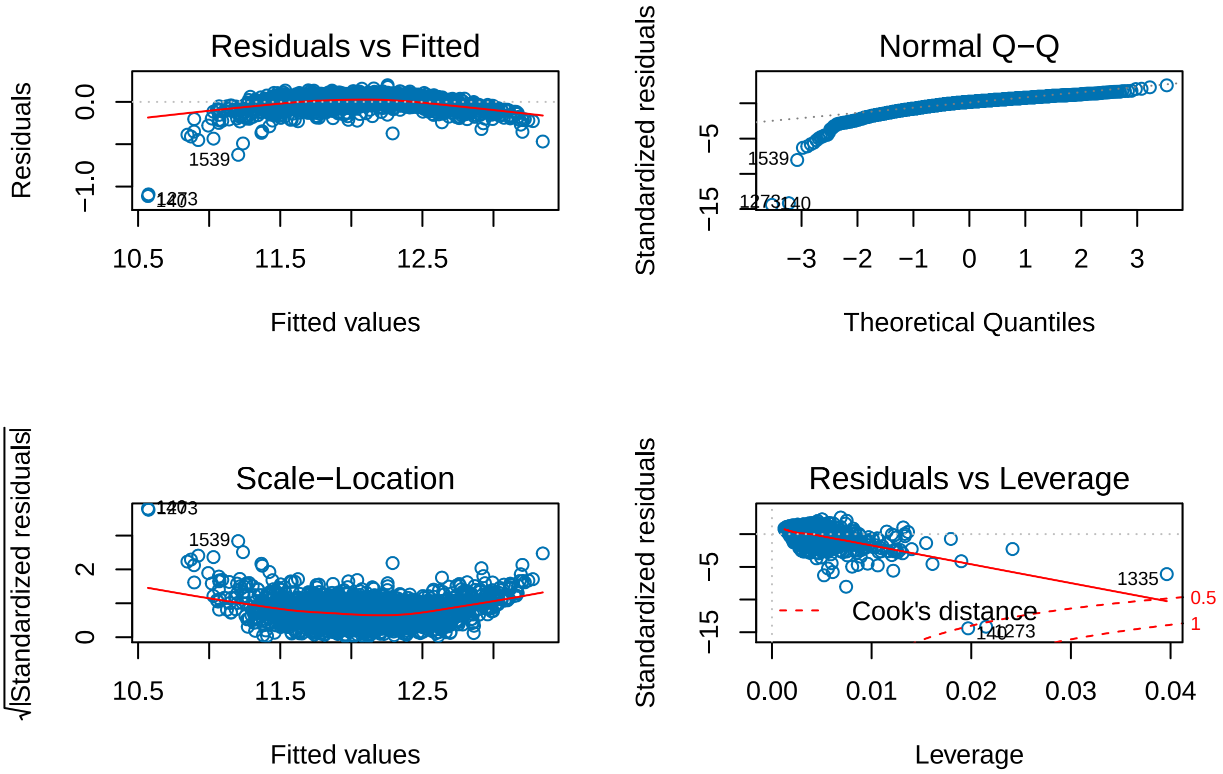
  - The model is: y = 9.872 + 0.001 x TotalFloorSF - 0.001 x HouseAge + 0.003 x QualityIndex + 0.009 x price_sqft + 0.143
  - Each additional 1 square foot results in a 0.06% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in house age results in a -0.121% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in quality index results in a 0.316% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in price per square foot results in a 0.896% percentage change in SalePrice when all other variables are held constant.
  - Compared to NbhdGrp4, NbhdGrp2 has an average SalePrice that is 14.281% higher than NbhdGrp4.

- NbhdGrp3
  - The model is: y = 9.872 + 0.001 x TotalFloorSF - 0.001 x HouseAge + 0.003 x QualityIndex + 0.009 x price_sqft + 0.083
  - Each additional 1 square foot results in a 0.06% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in house age results in a -0.121% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in quality index results in a 0.316% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in price per square foot results in a 0.896% percentage change in SalePrice when all other variables are held constant.
  - Compared to NbhdGrp4, NbhdGrp3 has an average SalePrice that is 8.266% higher than NbhdGrp4.

- NbhdGrp4 (baseline group)
  - The model is: y = 9.872 + 0.001 x TotalFloorSF - 0.001 x HouseAge + 0.003 x QualityIndex + 0.009 x price_sqft
  - Each additional 1 square foot results in a 0.06% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in house age results in a -0.121% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in quality index results in a 0.316% percentage change in SalePrice when all other variables are held constant.
  - Each additional increase in price per square foot results in a 0.896% percentage change in SalePrice when all other variables are held constant.

Similar to Model 3, for Model 4, the NbhdGrp variables could have been combined with the intercept. For ease of interpretation, we kept it separate. Therefore, we can see that the intercept is different for each Neighborhood but the slopes of all the remaining variables are the same. The log transformation of SalePrice does not change this model behavior.

Next, we examine diagnostic plots to determine goodness of fit.



Although the behavior of some of the residuals has improved relative to Model 3, we can see that the residuals vs fitted values now have an inverted U-pattern and the residuals depart from the normal QQ-line in the left tail. Additionally, there appear to be some influential points in the lower right of the plot. This means that further explortaion and decisioning is needed in how to handle these points. All in all, this model does not fit the data best as core assumptions of linearity, constant variance, and normality are violated.

Model 3 and 4 Comparison

In comparing Model 3 and Model 4, there is no clear winner.

First, if we compare the adjusted R squared values, Model 3 with SalePrice has an adjusted R squared value of 97% and Model 4 with logSalePrice has an adjusted R squared value of 96%. Although these adjusted R squared values are comparable, the slight edge is given to Model 3.

Next, if we compare their diagnostic plots, Model 4 has the slight advantage. Although both models do not meet assumptions of linear regression, Model 4 residuals have a less pronounced pattern than Model 3. Model 4 residuals also align better to the normal QQ-line, even though there is a depature in the left tail. This pattern supports the rationale for transformation: to help a model meet the assumptions of linear regression. Both models have outliers, but with Model 4, there appear to be influential points. It would be interesting to remove these influential points in Model 4 to see if the fit is improved.

Furthermore, in assessing multicollinearity, price_sqft and NbhdGrp1 have very high VIFs (price_sqft = 7.4, NbhdGrp1 = 11.4). These values are cause for concern and show that there are multicollinearity issues with these models. Therefore, one of these variables should be removed (e.g., price_sqft) and the model reassesed for multicollinearity.

Given all of this information, it is clear that the modeling process needs to continue before an appropriate model can be selected.

Lastly, in comparing the residual standard errors, Model 3 has a residual standard error of 13,350 and Model 4 has a residual standard error of .078. A metric like this or even MAE, however, is not as revealing since the log transformation is typically performed to address linearity issues and improve conformity to constant variance and normality. Log transformation is also helpful in situations where we are dealing with a skewed distribution. From assignment 2, we saw that before the transformation, SalePrice skewness was 1.13 and after the log transformation, it is -0.28. Comparing residual standard errors and MAE would be more revealing on two models with logSalePrice, for example.

In this specific instance, the desire to transform SalePrice was due the the wedge-shaped pattern with other predictors and the violation of constant variance in examining the residuals. Transformation was not due to theoretical considerations or the fact that the problem at hand involved a binomal or Poisson distribution. The evidence to transform SalePrice came from the patterns displayed in the diagnostic plots.

Deciding if a predictor requires a transformation is a more involved process. First, plotting the response variable, SalePrice, and each predictor in the model should be performed. Power transformations can be utilized to see which transformation yields a linear relationships between these two variables (if there is evidence of a lack of linear relationship). If the response variable is transformed first, and there are still concerns about linearity, then a predictor variable could be transformed. In our specific instance, log transformation of the predictor variables does not seem warranted since logSalePrice has a linear relationship with the continuous predictor variables. The next logical step, rather, appears to be exploration and potential removal of influential points.

##Section 7

In this section, we will explore the fit of Model 4 after identifying and removing influential points.

First, let's begin by calculating the DFFITS values for Model 4. DFFITS is one statistical approach to systematically identify and remove influential points.

For Model 4, we have n = 2386 and p = 7. Therefore, our DFFITs cuttoff value is: .116.

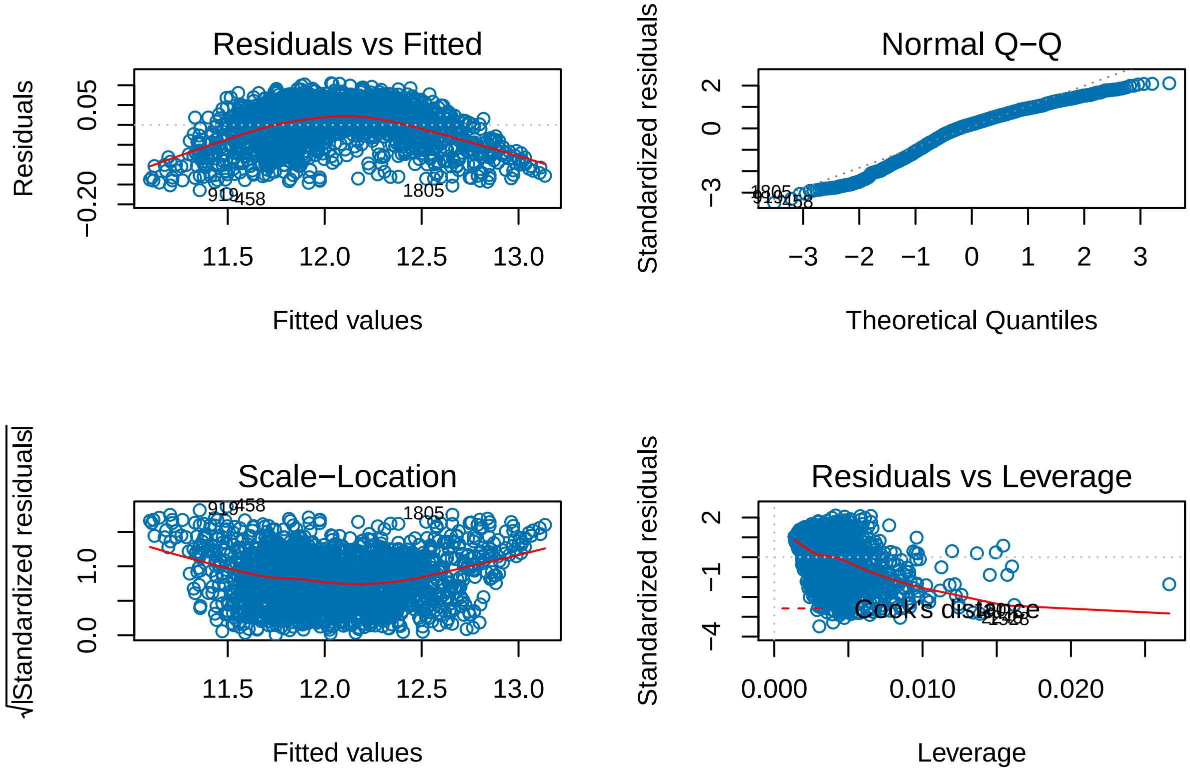
In calculating DFFITs, 139 points are identified to be influential points, which seems like a high number.

After removing the 139 influential points, we will now refit the model and examine the goodness of fit.

First, from the ANOVA results below, we can see that all predictors are significant in this new model. We note that the removal of these influential points did not alter significance.

```
##                Df Sum Sq Mean Sq  F value              Pr(>F)    
## TotalFloorSF    1 160.30  160.30 63928.82 < 0.0000000000000002 ***
## HouseAge        1  54.88   54.88 21886.84 < 0.0000000000000002 ***
## QualityIndex    1  19.00   19.00  7575.41 < 0.0000000000000002 ***
## price_sqft      1  34.89   34.89 13913.59 < 0.0000000000000002 ***
## NbhdGrp1        1   0.15    0.15    58.95   0.0000000000000239 ***
## NbhdGrp2        1   0.37    0.37   146.61 < 0.0000000000000002 ***
## NbhdGrp3        1   0.35    0.35   139.40 < 0.0000000000000002 ***
## Residuals    2247   5.63    0.00                                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

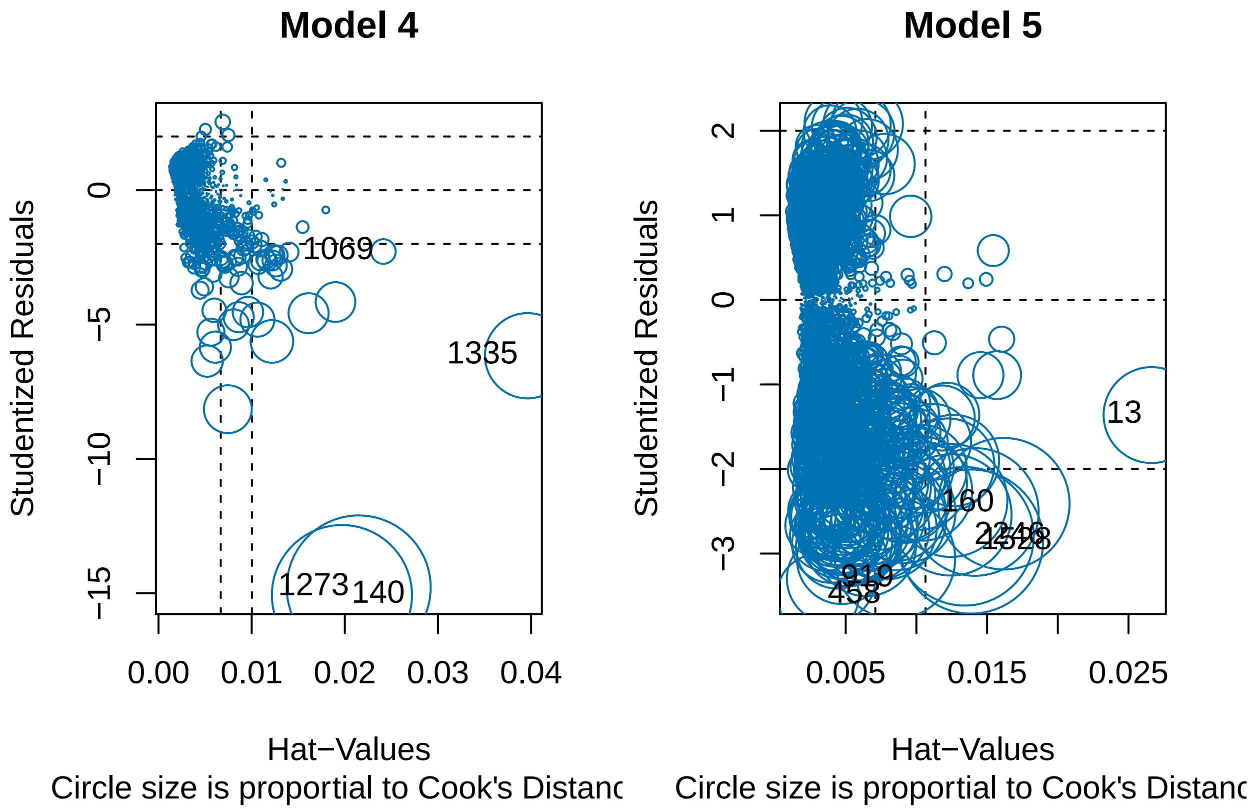
Next, we examine the diagnostic plots, which are shown below.



This result is very surprising. We can see that the residuals vs fitted values and scale-location visuals show that the residuals clearly follow a U-shape or quadratic pattern. Adding a quadratic term may be an appropriate next step in the modeling process. Moreover, we can see that the residuals more closely follow the normal QQ-line. We can see that the removal of the 139 influential points has corrected the skewness in the left tail. Additionally, we can see in the residuals vs leverage visual, we no longer have outliers in both the x- and y-direction.

A better visual might be to compare the influential plots before and after the removal of these points. Model 5 is Model 4 without the influential points.

```
##          StudRes        Hat      CookD
## 140  -15.071359 0.01968700 0.52083793
## 1069  -2.280552 0.02414079 0.01605426
## 1273 -14.787996 0.02149120 0.55018234
## 1335  -6.162378 0.03961878 0.19283495
```



```
##         StudRes          Hat       CookD
## 13   -1.362084 0.026631691 0.006342702
## 160  -2.410600 0.016182312 0.011922224
## 458  -3.491260 0.003039821 0.004622613
## 919  -3.298417 0.003967089 0.005392798
## 1528 -2.857587 0.013869241 0.014310158
## 2246 -2.795471 0.013407467 0.013234687
```

In comparing these influence plots, we can see that in Model 5 we have 'zoomed in' on the majority of points in the data set. However, in Model 5, there still appear to be some influential points as there are circles with large sizes.

Finally, we will compare the MAE values before and after the removal of these influential points.

- Before (Model 4): 180701
- After (Model 5): 178355

The MAE after the removal of these influential points is only marginally better. The MAE values for both models is very high, which suggests that there is a better model out there that will have a lower MAE.

All in all, the goodness of fit in this new model is hardly improved. It appears that the removal of these points has revealed a stronger pattern of the residuals following a U-shape. This pattern suggests that the addition of a quadratic term to the model is needed.

Summary and Conclusions

In this assignment, we explored various goodness of fit techniques on a host of models.

We began with a subset of Ames data and defined dummy variables for BedroomAbvGr. This step prepared us for sections later in the assignment.

From a modeling perspective, we created a multiple linear regression model with TotalFloorSF and HouseAge, since these two variables had strong linear relationships with SalePrice as well as no linear relationship with one another. Although this model (Model 1) had strong goodness of fit metrics, its residuals showed that this model violated regression assumption.

Additionally, we noticed some behavior with the Neighborhood Group variable when analyzing price per square feet and the fit of this model against neighborhoods. Therefore, we incorporated our newly formed Neighborhood Group variable into a multiple linear regression model (Model 2) and saw that on the basis of MAE, this model performed better than the model without this Neighborhood Group variable.

Overall, we saw that all models struggled to meet the core assumptions of linearity, constant variance, and normality when examining their residual plots. Therefore, we created a model using logSalePrice and although the fit was better, it was not perfect. In examining the diagnostic plots, we observed several influential points, calculated their DFFITS values, removed these 139 influential points, and refit the model. To our surprise, the residuals showed an even more distinct pattern. This entire process helped to demonstrate that creating a 'good' model is not a linear process.

In reflecting on this assignment, we saw that variable transformation and outlier deletion impact the modeling process. For one, variable transformation improves constant variance and thereby also improves normality as well as the linear relationship between variables. Although this observation is a plus, transforming a dependent variable makes the interpretation of the model more difficult. When trying to explain this model to someone without a statistical background, challenges may arise. Moreover, it is unclear right now if predictor variables would benefit from log transformation. We transformed SalePrice becasue of the wedge-shaped pattern it demonstrated, but we did not explore power transformations to determine if different transformations would help or if predictor variables in our model should be transformed as well.

Furthermore, the presence of outliers and influential points pose challenges to the modeling process, as we do not want to instill bias in our results by removing several observations. Not all variability in the data set is bad, but without having additional context as to why we were seeing these outliers, it makes it difficult to know if we are making the right decision by removing these points. Moreover, after removing our influential points, the residuals demonstrated a clear quadratic pattern, which alludes to adding a quadtric term to our model to see how that improves model fit. Lastly, we chose to evaluate outliers on the basis of DFFITS, but there are other methods as well. Comparison could be done against other methods to determine if the deletion results would be different.

Therefore, transformations and outlier deletion add complexity to the model creation and validation process, but they are both necessary steps in order to form the best model for a given set of data.

While this assignment focused on model building and adequacy-checking, it is clear that more work is needed to find an adequate model. Once this adequate model is found, then hypothesis testing should be performed to determine if all variables in this model are needed (e.g., partial F tests). Once this adequate model is found, then the validation step is entirely dependent on the intended use of the model. For instance, if the purpose of this sale price model is for statistical inference, then maybe two adequate models are selected and fitted to the data and their residuals are compared to one another to determine which one has the 'best' fit to the 2006-2010 data. However, if the purpose of this model is for predictive purposes, then applying this model to a new data set, possibly Ames housing data from 2011+, would be beneficial. This process of using out-of-sample data for validation would help reveal if there is any new behavior in the Ames housing market and would also tell us how well the model fits this new data. This step is vitally important if the model is to be used to predict future sale prices of homes.