Assignment #3

Andrew Knight

Introduction:

The sections that follow contain the analysis of categorical variables. These models are based on the Ames Housing Data and build on the analysis performed in Assignment 1 and 2.

I started by using all observations in the mydata set however after running through a few of the tasks I changed direction on this. I decided to use the same drop conditions as in Assignment #2 as basis for my sample data. This was done so that I could compare the resulting models produced below with the fit obtained from the previous assignment.

Data File: ames_housing_data.csv

Calculations File: Knight_Assignment3_code.R

Assignment Tasks

1: Select Categorical Variable

Categorical Variable Chosen: **KitchenQual**

Characteristics of the predictor variable 'KitchenQual': Factor with 5 levels

Find mean of SalePrice for each category:

KitchenQualEx = $337,339

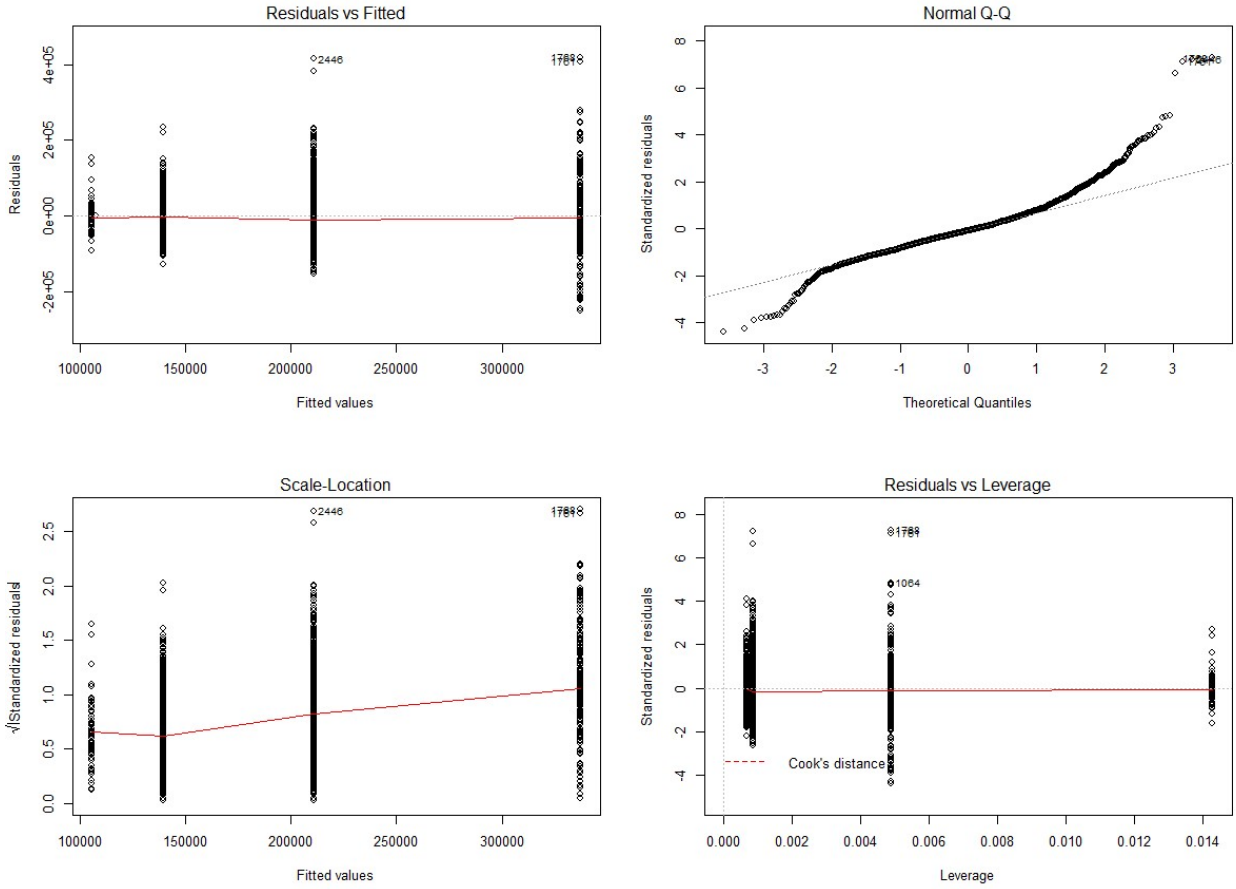
KitchenQualGd = 337,339 – 126,504 = $210,835

KitchenQualTA = 337,339 – 197,789 = $139,550

KitchenQualFa = 337,339 – 231,432 = $105,907

KitchenQualPo = 337,339 – 229,839 = $107,500

The mean values for SalePrice of each category seems to make sense. You can see a clear relationship between the quality of the kitchen and the SalePrice of the house. Note that KitchenQualPo, or Poor quality only lists one single data point so it is difficult to make assumptions about additional listings with category Poor. The mean of the residuals versus fitted values does appear to mostly go through the means of each category, there is slight deviation. This seems acceptable.



2: Create Dummy Variables for Categorical Variable

Basis for interpretation is KitchenQualEx, or Excellent quality.

Table for the five new Indicator Variables:

	kq_ex	**kq_gd**	**kq_ta**	**kq_fa**	**kq_po**
Ex	1	0	0	0	0
Gd	0	1	0	0	0
TA	0	0	1	0	0
Fa	0	0	0	1	0
Po	0	0	0	0	1

kq_ex$_i$ = {1, ith house has Excellent Kitchen Quality; 0 otherwise}

kq_gd$_i$ = {1, ith house has Good Kitchen Quality; 0 otherwise}

kq_ta$_i$ = {1, ith house has Typical/Average Kitchen Quality; 0 otherwise}

kq_fa$_i$ = {1, ith house has Fair Kitchen Quality; 0 otherwise}

kq_po$_i$ = {1, ith house has Poor Kitchen Quality; 0 otherwise}

SP = β0 + β1 * X1 + γ1 * kq_ex + γ2 * kq_gd + γ3 * kq_ta + γ4 * kq_fa + γ4 * kq_po + ε

```
Call:
lm(formula = SalePrice ~ kq_ex + kq_gd + kq_ta + kq_fa + kq_po,
    data = sampledat)

Residuals:
    Min      1Q  Median      3Q     Max
-251339  -32136   -4550   25450  417661

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   107500      57535   1.868   0.0618 .
kq_ex         229839      57675   3.985 6.91e-05 ***
kq_gd         103336      57560   1.795   0.0727 .
kq_ta          32050      57554   0.557   0.5777
kq_fa          -1593      57944  -0.027   0.9781
kq_po             NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57530 on 2925 degrees of freedom
Multiple R-squared:  0.482,     Adjusted R-squared:  0.4813
F-statistic: 680.5 on 4 and 2925 DF,  p-value: < 2.2e-16
```

Looking at the ANOVA table, kq_ex, kq_gd, and kq_ta are all highly significant however kq_fa and kq_po are not significant.

	kq_ex	**kq_gd**	**kq_ta**	**kq_fa**	**kq_po**
RSE	68,380	78,470	69,700	81,130	81,850
F-stat	1017	207	890	43 (not sig)	1 (not sig)
MAE	51,228	53,254	47,999	59,202	59,904

The Fair and Poor categories do not give a statistically significant result and the predicted model does not go through the mean of these categories. These two categories also have much lower number of observations, in fact the poor category only contains one data point.

3: Report on the Hypothesis Tests for each of the Betas

According to the results summary printout above, the following linear equation could be stated.

Let n = number of levels in categorical variable, KitchenQual. n = 5

z = β0 + β1 * X1 + … + βn * Xn then can be written as,

z = β0 + β1 * kq_ex + β2 * kq_gd + β3 * kq_ta + β4 * kq_fa + β5 * kq_po

The coefficient estimates from above give:

z = 107,500 + 229,839 * kq_ex + 103,336 * kq_gd + 32,050 * kq_ta – 1,593 * kq_fa + 0 * kq_po

The RSE gives 58790 which means that for each step up in category of KitchenQual you can expect an increase in the SalePrice of $58,790. The poor quality is considered the control group here.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   107500      58789   1.829   0.0676 .
kq_ex         241307      58962   4.093 4.41e-05 ***
kq_gd         107810      58819   1.833   0.0669 .
kq_ta          36587      58814   0.622   0.5339
kq_fa           6083      59362   0.102   0.9184
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58790 on 2342 degrees of freedom
Multiple R-squared:  0.485,     Adjusted R-squared:  0.4841
F-statistic: 551.4 on 4 and 2342 DF,  p-value: < 2.2e-16
```

The null and alternative hypotheses for each term of the population parameters involved can be stated as H0: the mean of each category is equal to the mean of the whole. The ANOVA summary indicates the typical and fair are not significant and thus we cannot reject the null hypothesis for these.

Because the lowest category, poor is used as the baseline, the estimates provide the amount by which the SalePrice increases for each category. Fair adds an average of $6083 onto Poor category, Typical/Average adds an average of 36,587 to poor category, Good adds an average 107,810 to the SalePrice from poor and Excellent adds 241,307 above poor quality base line.

4: Multiple Linear Regression Model

Model 1: Try MLR model with predictor variables: **GrLivArea** and **TotalBsmtSF**. Dataset used is the same **sampledat** used with drop conditions stated above.

Model 1 is defined by **lm(SalePrice ~ GrLivArea + TotalBsmtSF + KitchenQual)**. MAE is **25,650**.

The output of the MLR Model 1 is:

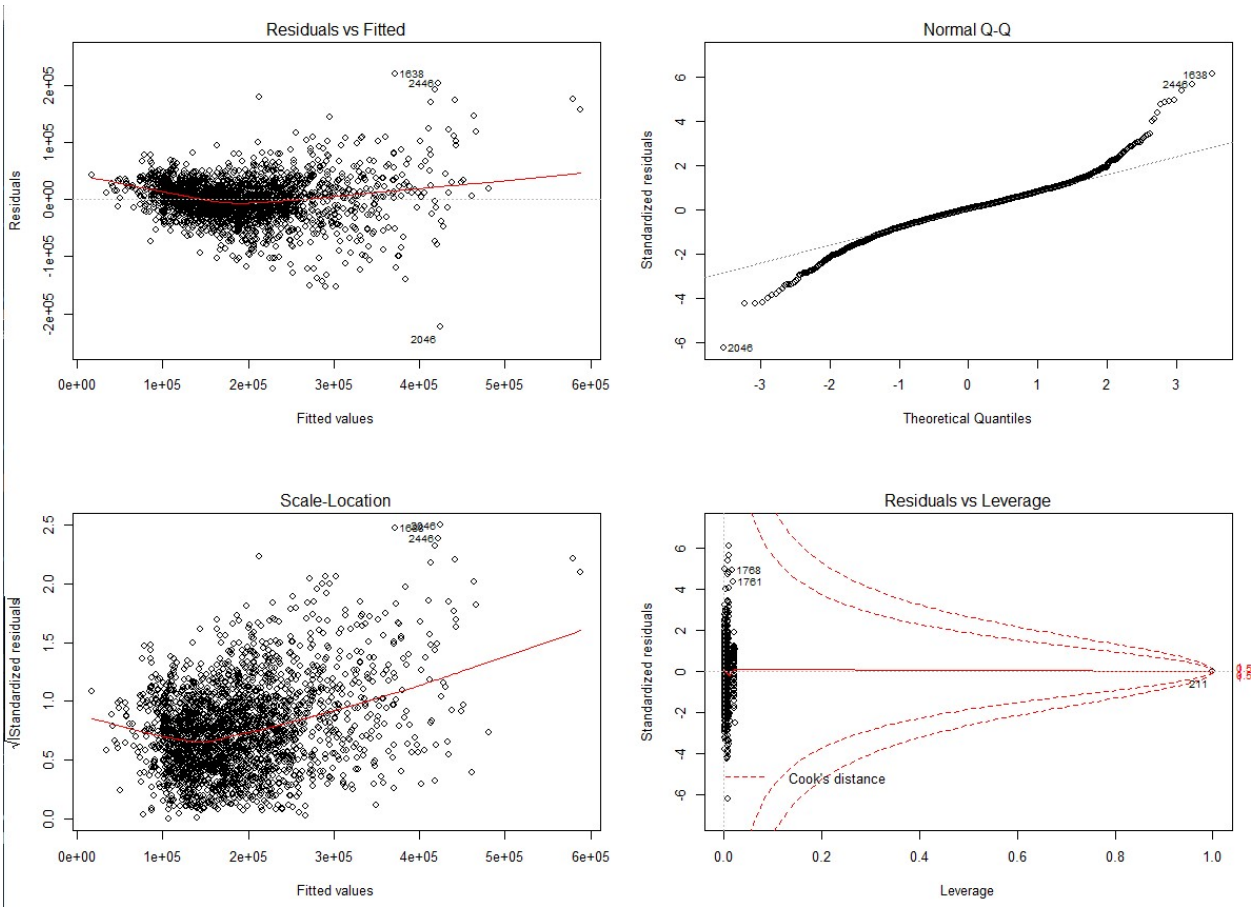
```
Call:
lm(formula = SalePrice ~ GrLivArea + TotalBsmtSF + kq_ex + kq_gd +
    kq_ta + kq_fa + kq_po, data = sampledat)

Residuals:
    Min      1Q  Median      3Q     Max
-223417  -19198    1335   19709  219515

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -36579.413  36146.383  -1.012  0.31165
GrLivArea       78.255      1.763  44.375  < 2e-16 ***
TotalBsmtSF     70.094      2.288  30.639  < 2e-16 ***
kq_ex       106990.232  36250.097   2.951  0.00319 **
kq_gd        39239.935  36109.228   1.087  0.27728
kq_ta        10557.629  36092.354   0.293  0.76992
kq_fa         -213.554  36425.827  -0.006  0.99532
kq_po               NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36070 on 2340 degrees of freedom
Multiple R-squared:  0.8063,    Adjusted R-squared:  0.8058
F-statistic:  1623 on 6 and 2340 DF,  p-value: < 2.2e-16
```

Part of the reason I chose the same two variables as in Assignment #2 is that I wanted to see how the model could be improved using the techniques covered this week. While adding additional predictor variables does not always mean a better fitting model, it can have a positive effect on the fit. If following the assumptions about the predictors, we should be able to show an improved R squared value over the MLR for the two categorical variables selected in Assignment #2.

Comparing the results between Assignment 2 and Assignment 3:

	RSE	**R-squared**	**F-stat**
MLR – Assignment #2	41,980	0.737	3288
MLR – Assignment #3	36,070	0.806	1623

Plot of the Model 1 Residuals:

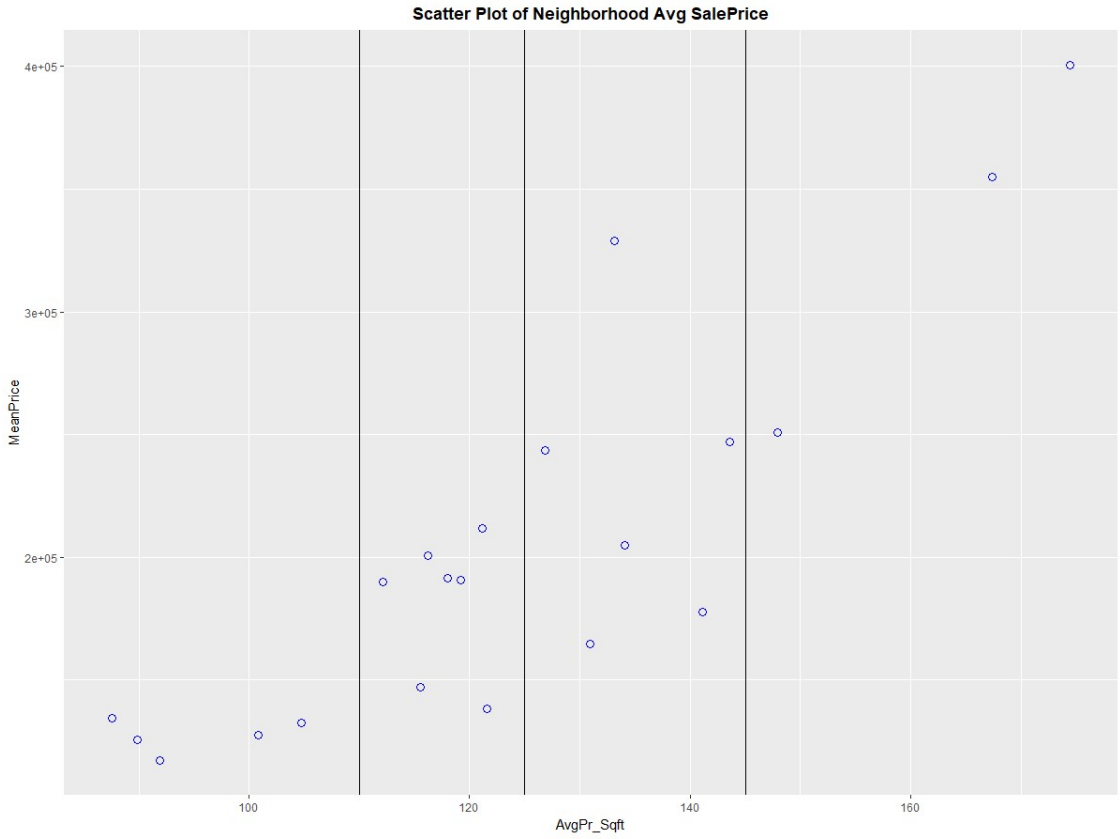


The analysis of Model1, which adds the indicator variables for each category of KitchenQual, gives a better overall R-squared value and lower residual standard error. This points to a better fitting model as a result of adding the indicator variables. A plot of the residuals shows slight improvement over the previous model however some of the same issues persist such as a slightly quadratic residual fit and some problematic outliers that will need to be addressed. Also, the leverage plot shows a data point that is highly influential.

5: Neighborhood Accuracy

The boxplot of the Neighborhoods versus SalePrice shows that neighborhoods seem to have a correlation to SalePrice and furthermore, they appear to fall within predefined ranges. It shows that a few neighborhoods have a much higher average SalePrice, or much lower average SalePrice than the rest but most of the neighborhoods are rather similar. The boxplots indicate quite a bit of movement in terms of average SalePrice by neighborhood which means that this categorical variable has the potential to be a good predictor of price. However, we will first have to clean up categories to be able to use them. We will need to make sure they have sufficient number of data points as well. To do this we need to group them into viable indicator predictors.

Based on my plot of the average SalePrice for each neighborhood, I've decided to split the neighborhoods into four buckets ngrp1, ngrp2, ngrp3, ngrpb4. Average SalePrice range for each neighborhood group is given in dollars below.



ngrp1 = $0-110, ngrp2 = $111 – 125, ngrp3 = $126 – 145, and ngrp4 > $145

Introducing the above groups into my Model1 from above improves my fit further to an R-squared value of 0.888. Based on the boxplots of neighborhoods, there are some neighborhoods that show much more variation than others and neighborhoods with smaller sample sizes may tend to overpredict values. One thing to consider is that the last group still only has three data points however the distribution is more of less even.

I refit the next model, Model 2 as defined below. The KitchenQual Ex is the base category.

Model 2 is **lm(SalePrice ~ GrLivArea + TotalBsmtSF + NbhdGrp, data=sampledat).**

MAE for Model 2 is $18,083 which is about $7500 less than the MAE for Model 1. Based on this value, Model 2 seems to improve on our fit from Model 1. The Model fit comparison stats so far are listed in the following table.

	RSE	**R-squared**	**MAE**
Model #1	36,070	0.806	25,650
Model #2	27,450	0.888	18,083

Next, we will see how we can improve on these values with additional variable and/or transformations.

6: Model Comparison of Y versus log(Y)

In an analysis of two additional models we are focusing on comparison between the standard non-transformed SalePrice model and one in which SalePrice response is log-transformed. The new models 3 and 4 are defined below as follows:

Model 3:

Predictor variables are GrLivArea + TotalBsmtSF + QualityIndex + YearBuilt + NbhdGrp + KitchenQual

These predictor variables consist of four continuous (GrLivArea, TotalBsmtSF, QualityIndex, YearBuilt) and two discrete (NbhdGrp, KitchenQual) variables. The predictors focus on size, location and quality attributes of properties. The response comparison of SalePrice and log(SalePrice) is included.

Model 4:

Predictor variables are the same as model 3.

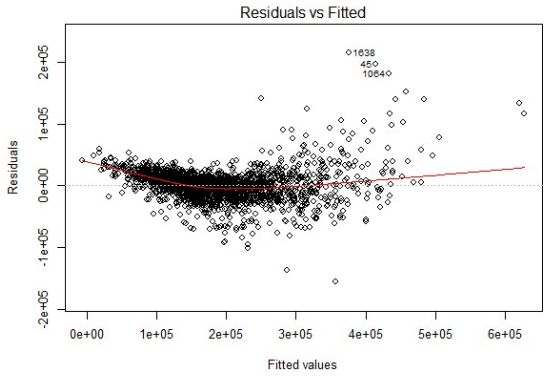
Table giving the model fit comparison between Model 3 and Model 4:

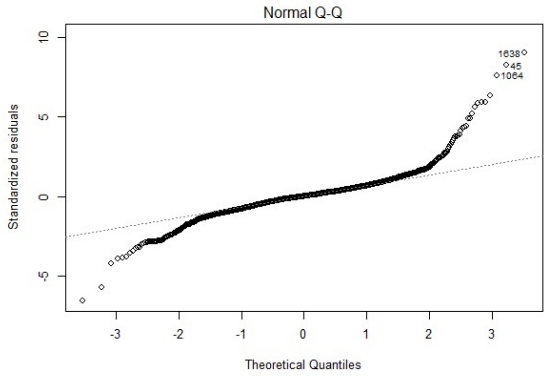
The R-squared metric for Model #3 is 0.916 vs the R-squared metric for Model #4 is 0.920. The RSE and MAE numbers cannot be directly compared due to the log transformation. While the fit score is slightly improved for Model #4, we need to visually compare the residual plots for the full story.

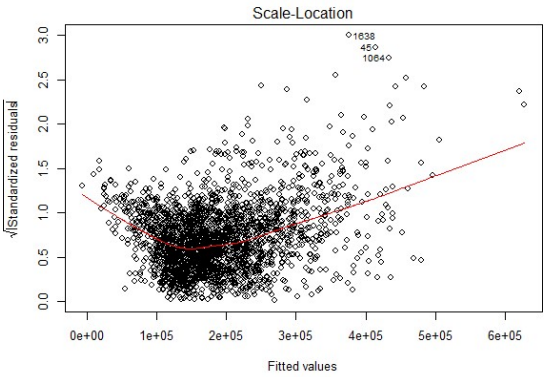
	RSE	**R-squared**	**MAE**
Model #3	24,120	0.916	15,752
Model #4	0.1122	0.920	log(187,245)

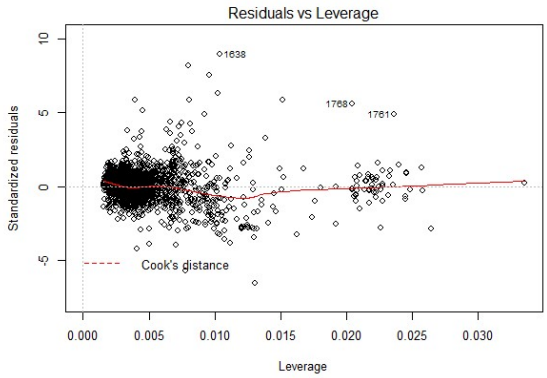
Comparison of the plots (regular results on top vs log transformed results on bottom):

Model #3:

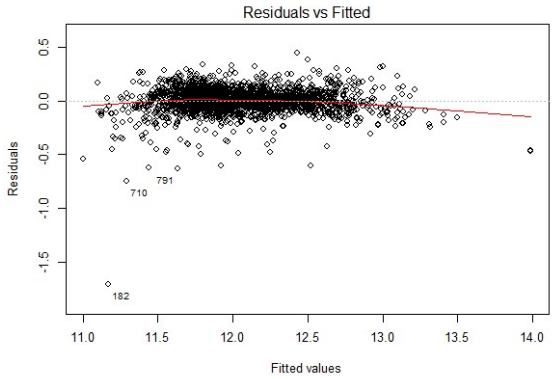

Residuals vs Fitted
Residuals
Fitted values
1638
45
1064

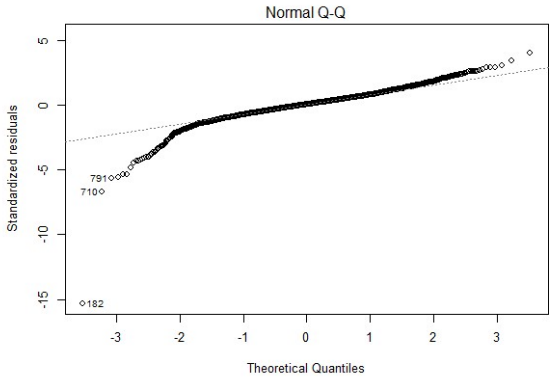

Normal Q-Q
Standardized residuals
Theoretical Quantiles
1638
45
1064

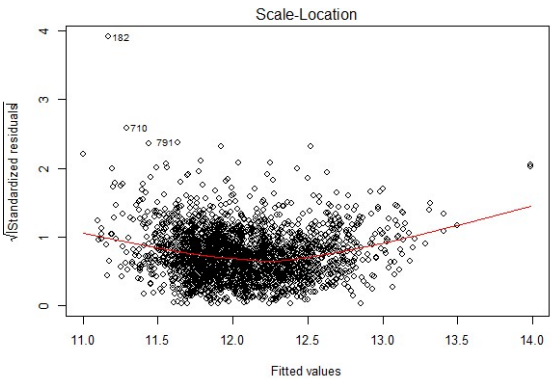

Scale-Location
√|Standardized residuals|
Fitted values
1638
45
1064

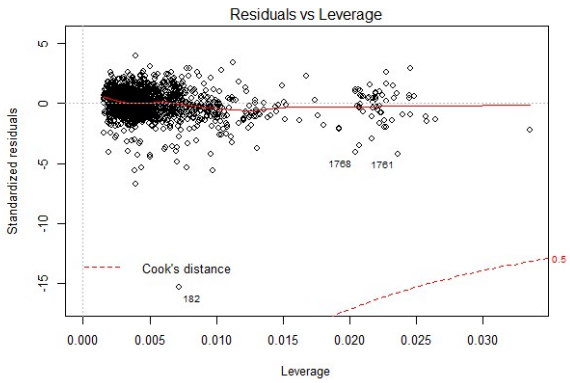

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1638
1768
1761

Model #4:


Residuals vs Fitted
Residuals
Fitted values
791
710
182


Normal Q-Q
Standardized residuals
Theoretical Quantiles
791
710
182

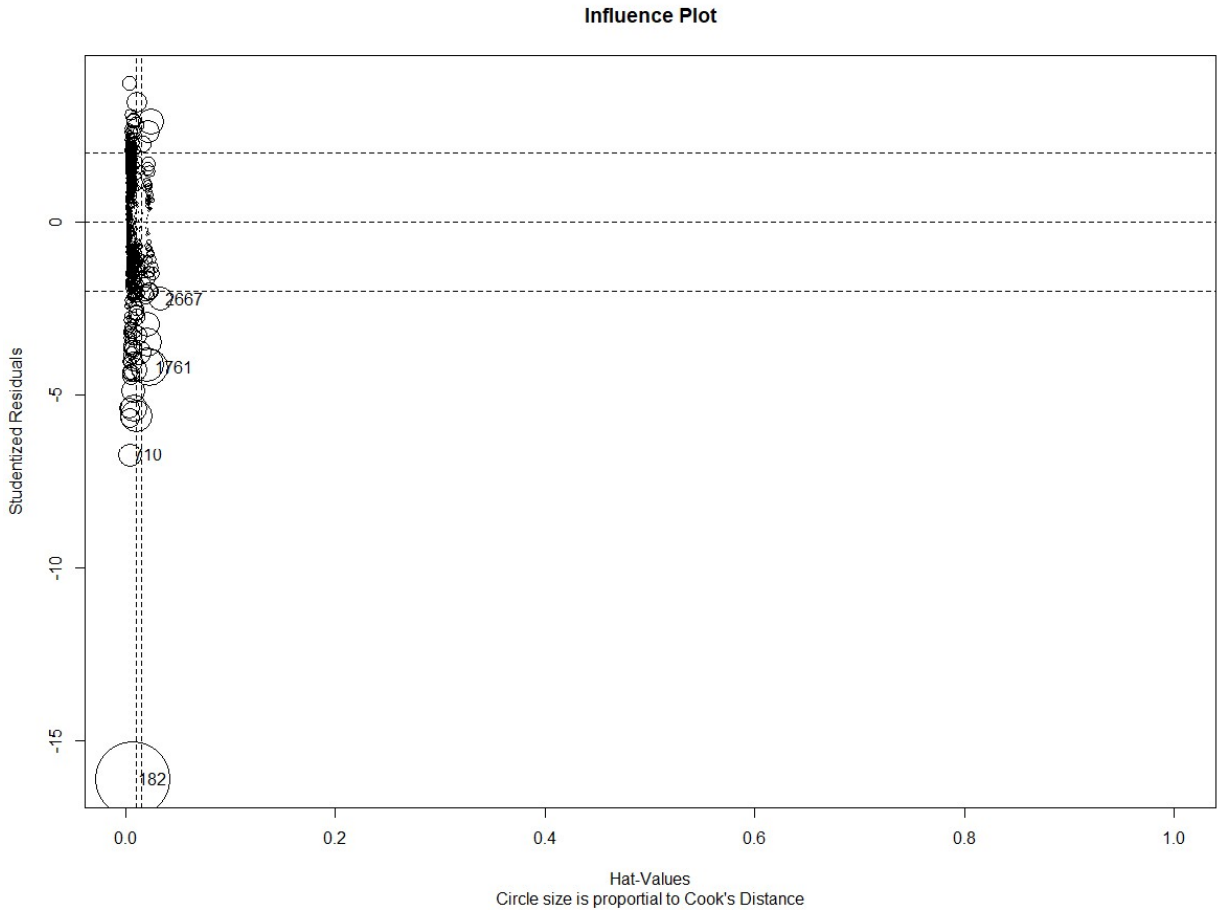

Scale-Location
√|Standardized residuals|
Fitted values
182
710
791


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
182
1768
1761
0.5

The Model #4 residuals versus fitted plot of the logSalePrice on show a more linear fit then the non-transformed model #3 on top left. The Scale-Location plot for Model #4 shows similar corrective behavior over Model #3. I noticed that the Q-Q Plot shows improvement on positive tail in Model #4 but does not appear to have much closer of a normal standardized residual approximation on the lower negative tail between Model 3 and Model 4. The Residuals Leverage plot interestingly shows the main cluster of residual points as expected to the left, but it also shows a small but mostly separate cluster between the .02 and the .025 leverage distance. These will likely need further investigation in the next section regarding influential points. While the logSalePrice residuals are more tightly centered around the mean (a promising improvement), the cluster of points to the right may be causing an undesired change to my model. Even those these are within the 95% confidence interval, we may need to further modify our model below.

7: Influential Points

The initial Influence Plot from Model #4 is shown here.



The Influence plot reveals residuals which have high leverage and/or high influence. When checking the results from my Model #4 above, I see one residual that sticks out with much higher than normal DFITS

value indicating high leverage. The absolute value for most other observations are 0.5 or less while this particular observation is 1.37. This value from observation 182 can be seen here.

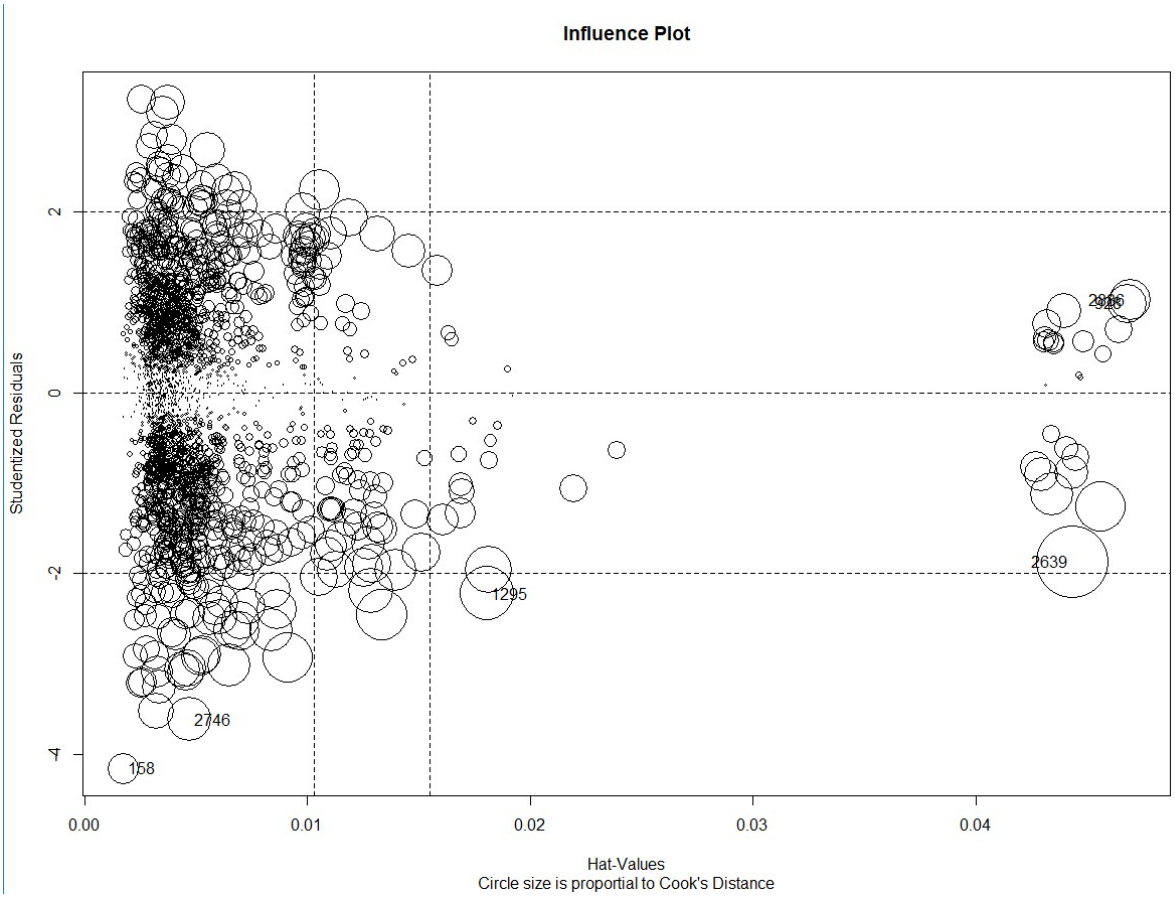
	dfb.KQTA	dffit	cov.r	cook.d	hat
4	0.05	-0.08	1.02_*	0.00	0.01
18	-0.23	0.37_*	0.96_*	0.01	0.01
45	-0.12	0.26_*	0.97_*	0.01	0.01
83	0.02	-0.13	0.97_*	0.00	0.00
131	0.02	-0.22_*	1.02_*	0.00	0.02_*
150	0.00	0.06	1.03_*	0.00	0.02_*
158	-0.01	-0.10	0.98_*	0.00	0.00
170	0.00	-0.30_*	0.95_*	0.01	0.01
179	0.00	0.03	1.03_*	0.00	0.02_*
182	0.38	-1.37_*	0.29_*	0.14	0.01
184	0.00	-0.09	1.02_*	0.00	0.01

It seems this observation should be removed and the model should be retested. Which other points should be closely considered for removal? Based on the DFITS suggested value I obtained from my number of observations (n=2347) and my number of predictors (p=6), I should consider removing values greater than 0.109. Using this to subset my sampledat, I reduce the data down to 2132 observations.

Now if I rerun my model:

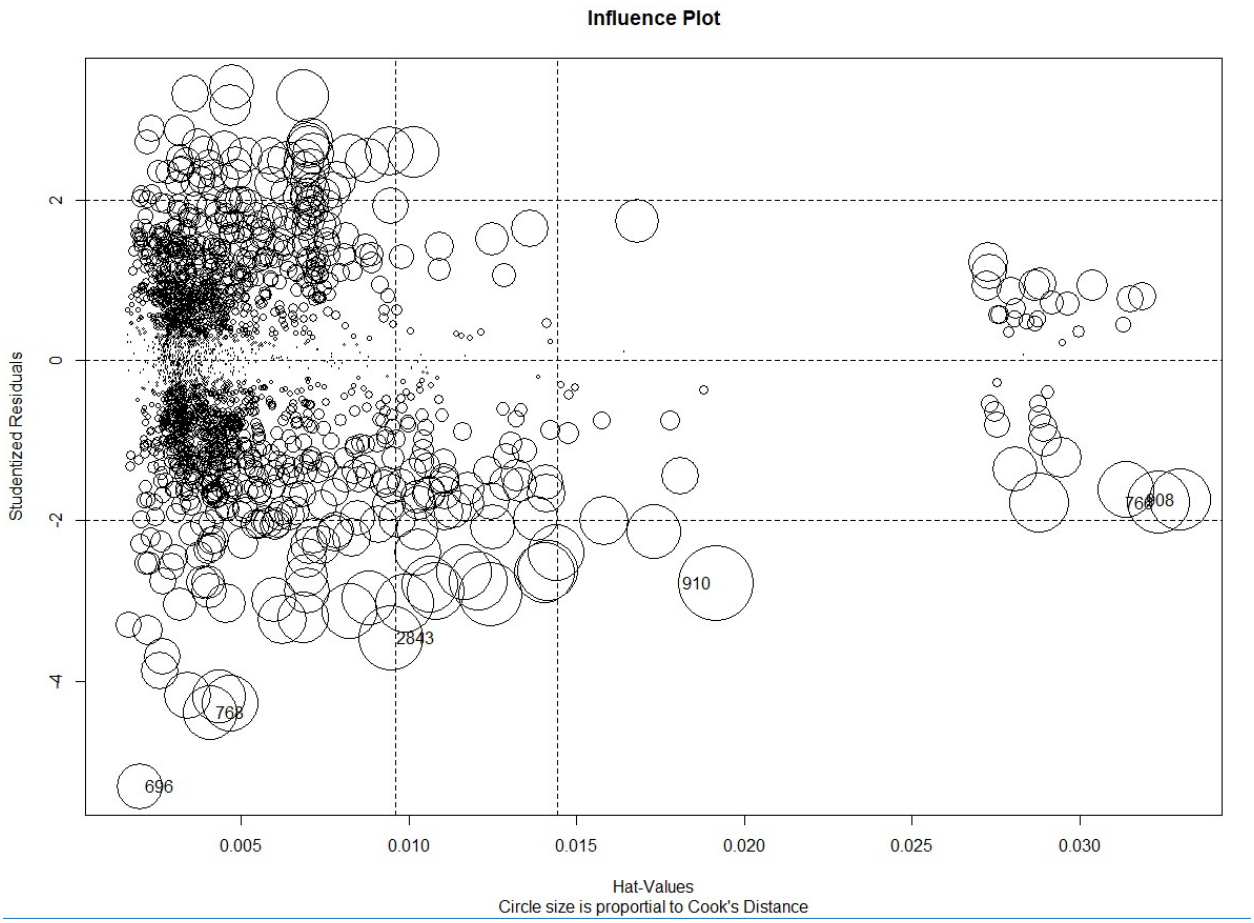
```
MLR7Logresult = lm(logSalePrice ~ GrLivArea + TotalBsmtSF + QualityIndex + YearBuilt +
                   NbhdGrp + KitchenQual, data = subsampledat)
```

I get a new R-squared score of 0.957 and a new influence plot shown here:



Now, from the scale of the graph we can see that the residuals versus the hat values are much more tightly centered in both the x-direction (influence) and the y-direction (leverage) than before the points were removed.

One test that may be worth trying before accepting this improved model is to take the original sampledat set and increase the DFITS threshold (by let's say double), just to see how it affects the overall model performance. In this case we have a resulting cleaned set of 2289 observations after removing influential points greater than calculated DFITS * 2. We once again rerun the MLR Model and find that we achieve a R-squared score of 0.945. As expected, this is higher than Model #4 but slightly lower than the first trial of DFITS residual influential points removed. The Trial #2 Influence Plot is shown below for comparison:



So, which is the right choice? It probably depends on the use case and the objective of the analysis. If we intend to use this for out-of-sample prediction for estimating the SalePrice (especially after a first-run analysis), it might be best to error on the conservative side and use the results from Trial 2, leaving additional influential points in the data for the model. This should provide a little protection against overfitting until we are able to collect more observations and/or more rigorously test the model.

Conclusion & Reflections

In summary our model using the log-transformed SalePrice as the dependent variable with a total of six predictors seems to give the best fit. Removing influential points improves the model furthers but we know that it is not advisable to automatically drop observations without a careful review. As such, the 58 records that were removed should receive adequate checking. Which properties types are they, what were the sale conditions, where are they located?

It was nice to see the model improve as we took steps to increase the number of predictors, test transformation, and remove specific observations as it helps reinforce the concepts that each step provides. Anytime we build a model we should also use an 'intuition test' to help check our results. For example, we expect attributes like size, condition and location of a property to all be positively correlated to estimated price. As we combine several of these factors, we should expect to see the model improve.