Assignment #4

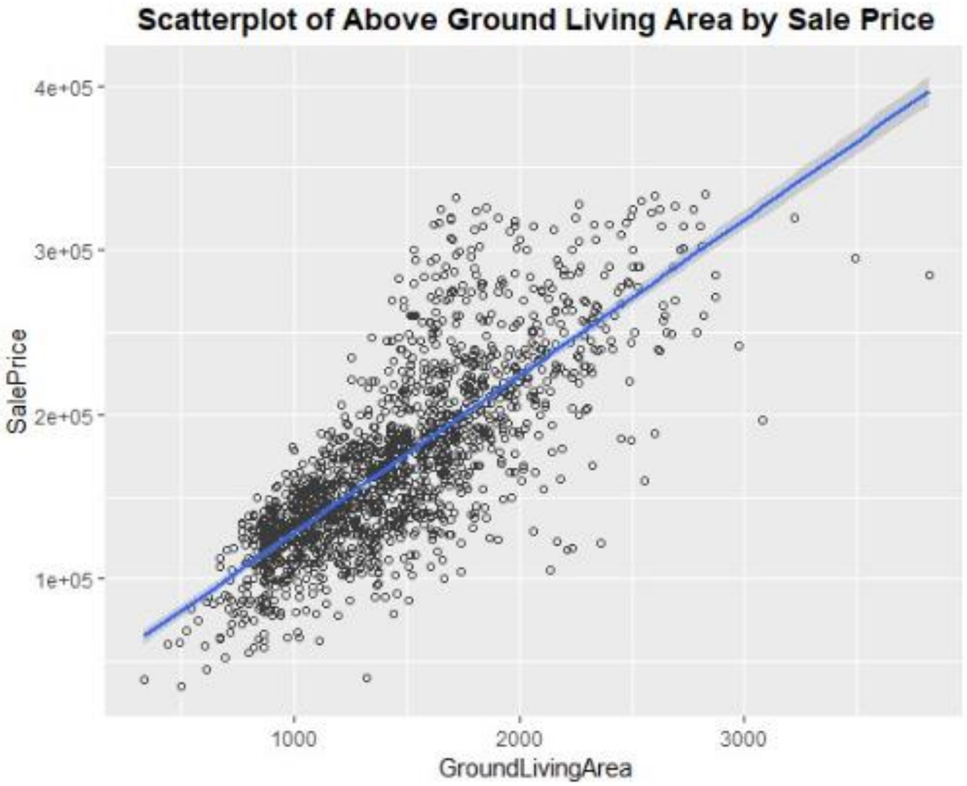
May 2021

Part A)

1)

After conducting analysis on the varying factors, I chose ground above living as the best continuous variable because of its high correlation to sales price.

a)



b)

y_hat = 33418.127 + 95.096X1

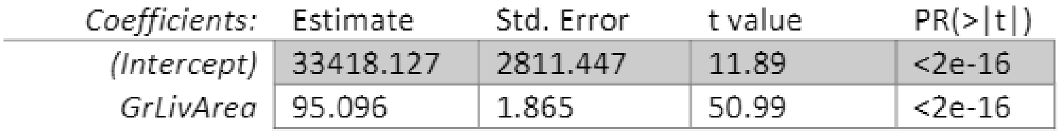
The intercept for X=0 is Y=-33,418.18. The coefficients show that fore every increase in unit for GAR, sales price increases by 95.10

c)

R-Squared value is .601 meaning 60.1% variability for GAR compared to Sales price.

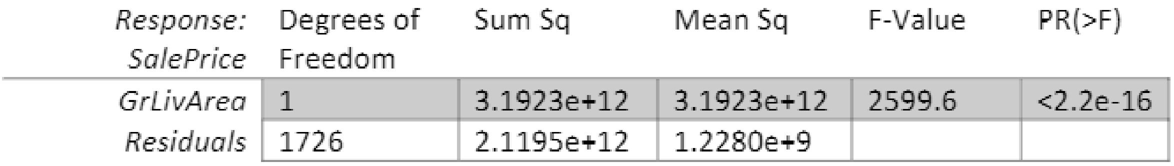
d)

Coefficients Table:



Coefficients:	Estimate	Std. Error	t value	PR(>|t|)
(Intercept)	33418.127	2811.447	11.89	<2e-16
GrLivArea	95.096	1.865	50.99	<2e-16

Anova Table:



Response: SalePrice	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
GrLivArea	1	3.1923e+12	3.1923e+12	2599.6	<2.2e-16
Residuals	1726	2.1195e+12	1.2280e+9		

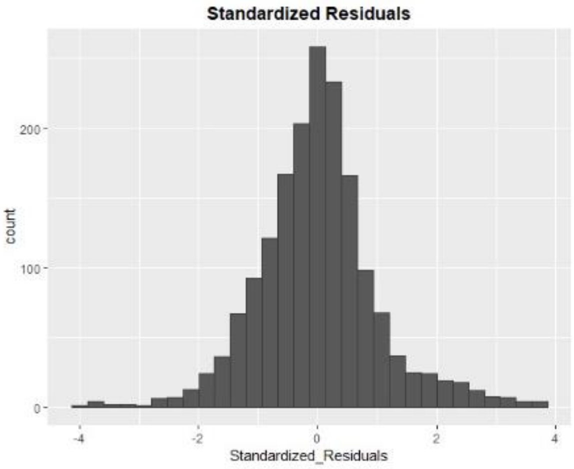
T-Test:

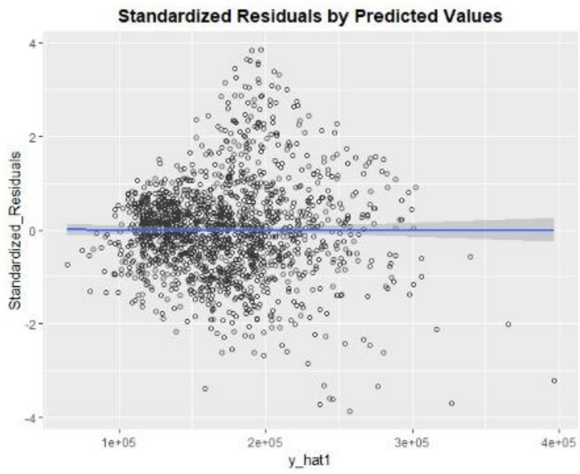
H0B1 = 0, HaB1 = !0; T-Statistic = 11.89, B1 is not equal to zero so we can reject the null hypothesis

F-Test:

H0B1 = 0, HaB1 = !0, F-Statistic = 2599.6, b1 is not equal to zero so we can reject the null hypothesis.

e)

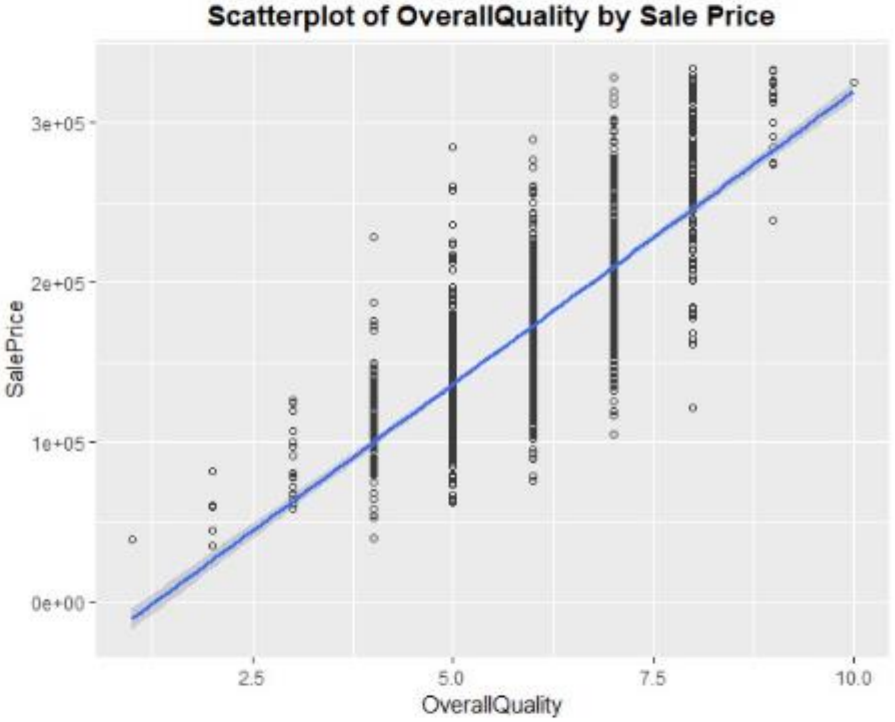




The deviations from normality shows the variance of negative residuals seem to increase with predicted values. There are a lot of outliers especially in the predicted values visualization and these deviations should be further examined. The positive and negative distributions seem to be even and thus it seems our assumptions of the models are true.

2)

a)



b)

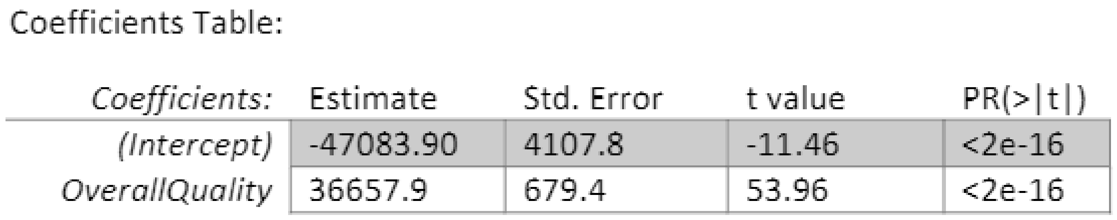
Y_hat = -47083.90 + 36657.9X1. The intercept value here is y=-47,083.90. This is different from model one which had a much smaller slope of 95.10/ This is likely because the units above ground have been scaled up to 4,000 feet.

c)

R-Squared = .6278. this means 62.78 % of variability of response data is explained within the model which is a high amount, and the model is successful.

d)

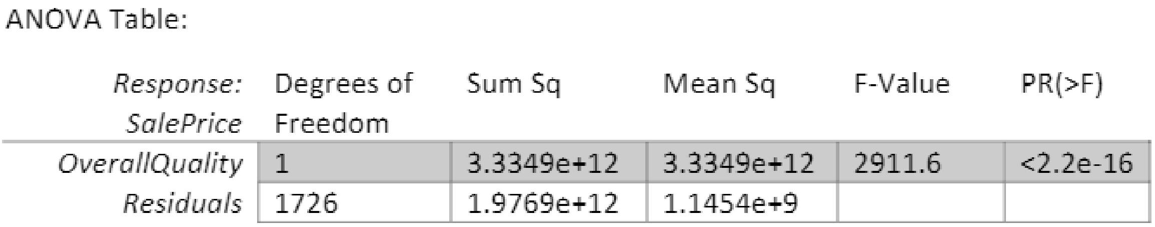
Coefficients Table:



Coefficients Table:

Coefficients:	Estimate	Std. Error	t value	PR(>|t|)
(Intercept)	-47083.90	4107.8	-11.46	<2e-16
OverallQuality	36657.9	679.4	53.96	<2e-16

Anova Table:



ANOVA Table:

Response: SalePrice	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
OverallQuality	1	3.3349e+12	3.3349e+12	2911.6	<2.2e-16
Residuals	1726	1.9769e+12	1.1454e+9		

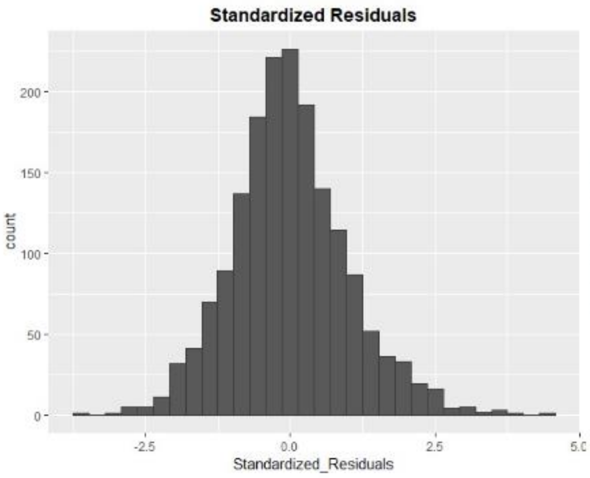
T-test B1 (Overall Quality): H0:B1 vs haB1 !=0

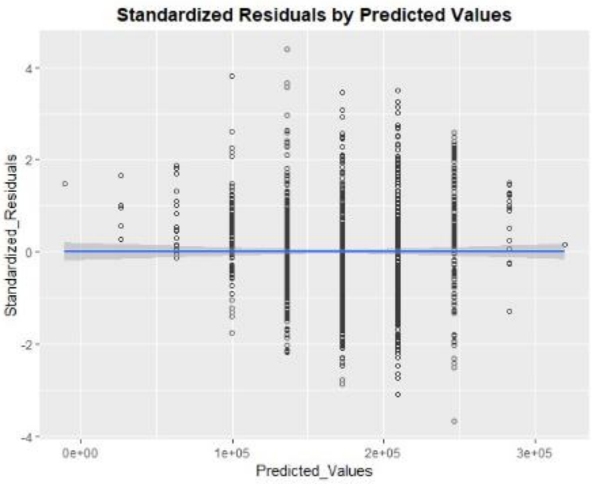
B1 T-test = 53.96 which means we can reject the null hypothesis because it is not equal to zero and we can note that it is significant variable for our model

F-Test (Model 2): H0:B1 vs Ha:b1 1=0

F-test = 2911.6 which means we can reject the null hypothesis and conclude this model provides a better fit than model 1 which only has the intercept.

e)





The results show that both of the negative and positive residuals appear to be normally distributed, with not surprising or differing results. This gives us no indication that the model is false. There are a few outliers but not enough to make a pattern out of them.

3)

Of the two models, Model two seems to have the better fit than model one based on r-squared value and the residual plots. We are assessing model fit on the aforementioned r-squared and residual plots.

Part B)

4)

Y_hat = -5090.159 + 56.875X! + 23499.607X2

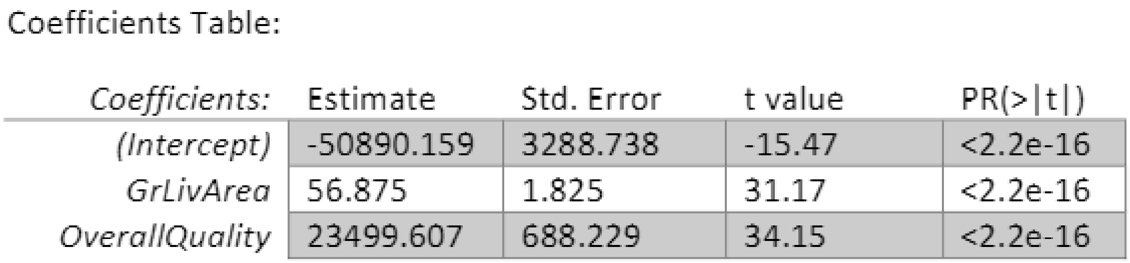
The coefficients are lower in the combined multiple linear regression model than separately in linear regression models because an increase in above ground living area and home quality has a positive result in relationship to sales price. Thus, the intercept starts lower and slopes and beta values in the OLS equation should be smaller with the multiple linear regression model

b)

R-squared: 0.7619 which means that 76.19% of the variability of the response data is explained within the model which is a high amount. This seems to show that the fit for multiple regression is better than simple linear regression because of the higher r-squared score. The difference between model 3 and model 1 is .1608. I interpret the difference as showing that the multiple linear regression model fits better than the simple regression model.

c)

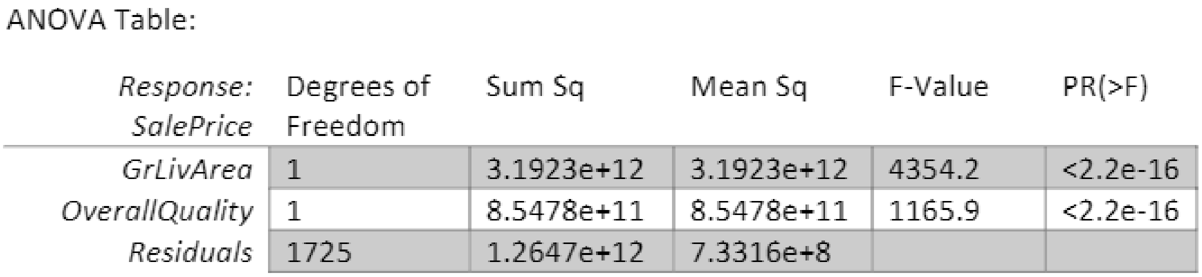
Coefficients table:



Coefficients Table:

Coefficients:	Estimate	Std. Error	t value	PR(>|t|)
(Intercept)	-50890.159	3288.738	-15.47	<2.2e-16
GrLivArea	56.875	1.825	31.17	<2.2e-16
OverallQuality	23499.607	688.229	34.15	<2.2e-16

Anova table:



ANOVA Table:

Response: SalePrice	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
GrLivArea	1	3.1923e+12	3.1923e+12	4354.2	<2.2e-16
OverallQuality	1	8.5478e+11	8.5478e+11	1165.9	<2.2e-16
Residuals	1725	1.2647e+12	7.3316e+8		

T-test for B1 (Above Ground Living Area):

H0:B1 = 0 vs Ha:B1 !=0

T-statistic = 31.17 meaning we can reject the null hypothesis because it does not equal to 0 and the variable is significant to the overall model.

T-test for B2 (Overall Quality):

H0:B2 = 0 vs Ha:B2 !=0

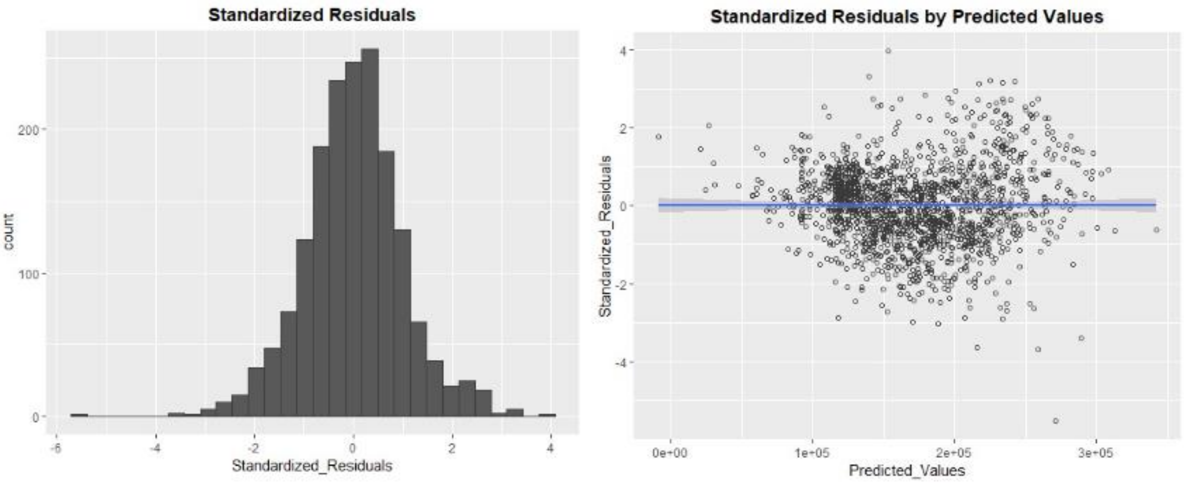
T-statistic = 24.15 meaning we can reject the null hypothesis because it does not equal to 0 and the variable is significant to the overall model

F-test (Model 3):

H0:B1 = B2 = 0 vs Ha:B1 !=0 for E(1,2)

F-statistic = 2760 meaning we can reject the null hypothesis because it does not equal to 0 and the variable is significant to the overall model.

d)



The results show a number of outliers, but not as much in previous models. This suggests that nothing we see in the results disproves any of the assumptions we've made about the model.

e)

The results show based on our various r-squared results, hypothesis tests, and residuals we can keep the predictor variables of model 3.

5)

a)

Y-hat = -62251.445 + 54.697X1 + 18804.602X2 + 42.011X3

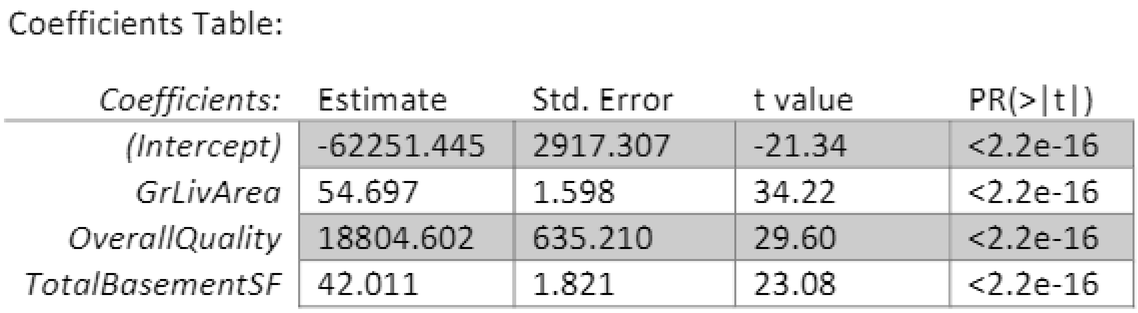
The coefficients in this model are very similar to those in model 3 in that they are lower in the mls model than in the simple linear regression model. The values have been affected by the total basement square footage.

b)

R-squared = .8181. The change is .0562 which shows that the mls is better than the simple linear regression model which we base off of the higher r-squared value.

c)

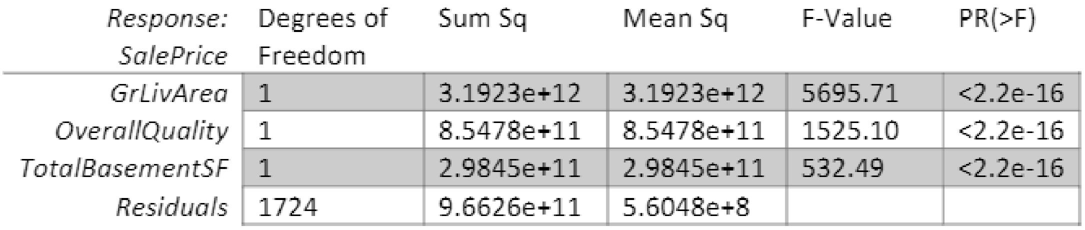
Coefficients table:



Coefficients Table:

Coefficients:	Estimate	Std. Error	t value	PR(>|t|)
(Intercept)	-62251.445	2917.307	-21.34	<2.2e-16
GrLivArea	54.697	1.598	34.22	<2.2e-16
OverallQuality	18804.602	635.210	29.60	<2.2e-16
TotalBasementSF	42.011	1.821	23.08	<2.2e-16

Anova table:



Response: SalePrice	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
GrLivArea	1	3.1923e+12	3.1923e+12	5695.71	<2.2e-16
OverallQuality	1	8.5478e+11	8.5478e+11	1525.10	<2.2e-16
TotalBasementSF	1	2.9845e+11	2.9845e+11	532.49	<2.2e-16
Residuals	1724	9.6626e+11	5.6048e+8		

T-test for B1 (Above Ground Living Area):

H0:B1 = 0 vs Ha:B1 !=0

T-statistic = 34.22 meaning we can reject the null hypothesis because it does not equal to 0 and the variable is significant to the overall model.

T-test for B2 (Overall Quality):

H0:B2 = 0 vs Ha:B2 !=0

T-statistic = 29.60 meaning we can reject the null hypothesis because it does not equal to 0 and the variable is significant to the overall model

T-test for B3 (Total Basement Square Footage):

H0:B3 = 0 vs Ha:B3 !=0

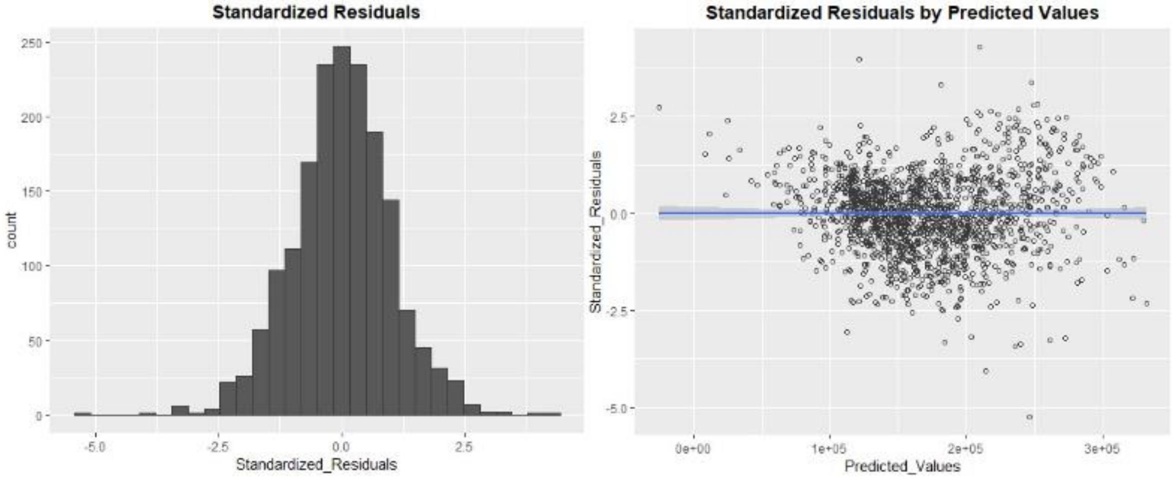
T-statistic = 23.08 meaning we can reject the null hypothesis because it does not equal to 0 and the variable is significant to the overall model

F-test (Model 4):

H0:B1 = B2 = =B3 = 0 vs Ha:B1 !=0 for E(1,2, 3)

F-statistic = 2584 meaning we can reject the null hypothesis because it does not equal to 0 and the variable is significant to the overall model.

d)

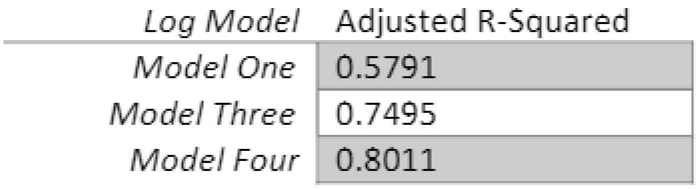


The results show a number of outliers, but not as much in previous models. This suggests that nothing we see in the results disproves any of the assumptions we've made about the model.

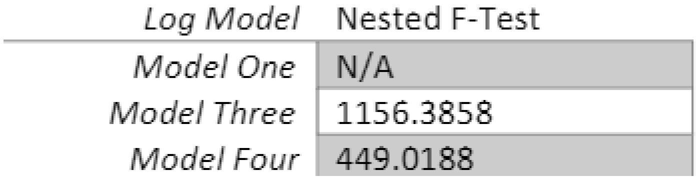
e) The results show based on our various r-squared results, hypothesis tests, and residuals we can keep the predictor variables of model 3.

Part C)

6)

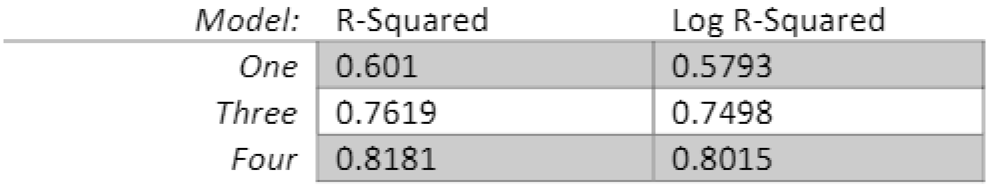


Log Model	Adjusted R-Squared
Model One	0.5791
Model Three	0.7495
Model Four	0.8011



Log Model	Nested F-Test
Model One	N/A
Model Three	1156.3858
Model Four	449.0188

r-squared is still the best indicator for accounting for population size and other variables compared to sales price and is applicable across all the models. The highest r-squared results show that model 4 fits best with a small bump above model 3. We can again reject the null hypothesis for both these models as they don't equal to zero.



Model:	R-Squared	Log R-Squared
One	0.601	0.5793
Three	0.7619	0.7498
Four	0.8181	0.8015

The log models are lower across the board for all models, albeit slightly, thus we should go with the original models as they are still the better fit than the log of sales price.

7)

The interpretation of the log of sales price is different from the sales price models because it is measuring unit changes compared to the percentage change in sale price. I would not keep the model because they score worse than the original models, although its slightly less. Unless they can provide more value, I would keep them out.

Part D)

8)

I found 60 influential points above the .1077 Dfits cutoff value and thus removed the 60 influential values of the refitted model of 1626 total values. The refitting of the model increases the r-squared value to .8469 with an increase of 2.88 variability over model 4. Due to this increase result, the model justifies biasing the result by removing the influential points.

Part E)

9)

a)

For this model I will continue to add more continuous variables and hope the trend of the previous models holds, which is a higher r-squared value. I added the variables of lot frontage, lot area, masonry area and garage are as compared to sales price to see which has the best positive relationship. I chose lot frontage as the variable to display. We can view r-squared and adjusted r-squared to see the results.

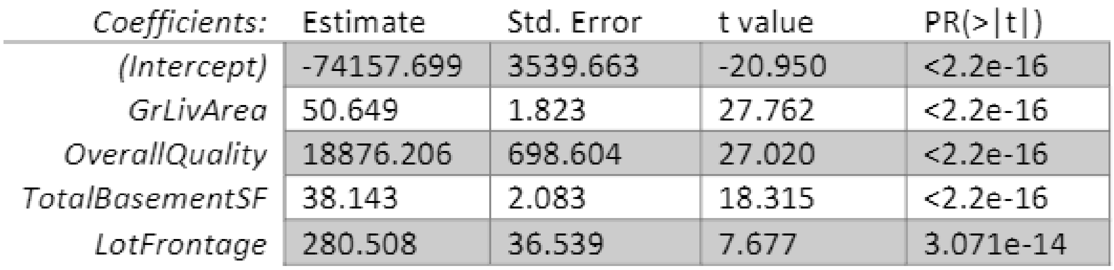
Lot Frontage(Model 5):

Y_hat = -74157.699 + 50.649X1 + 18876.206X2 +38.413X3

b)

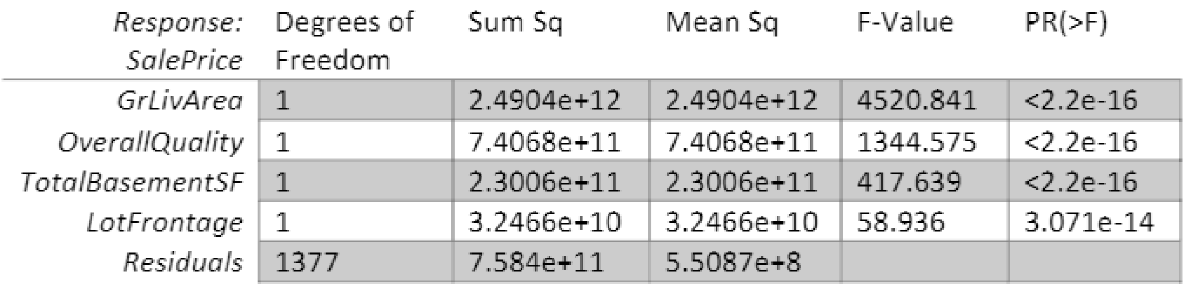
Coefficients table:

c)



Coefficients:	Estimate	Std. Error	t value	PR(>|t|)
(Intercept)	-74157.699	3539.663	-20.950	<2.2e-16
GrLivArea	50.649	1.823	27.762	<2.2e-16
OverallQuality	18876.206	698.604	27.020	<2.2e-16
TotalBasementSF	38.143	2.083	18.315	<2.2e-16
LotFrontage	280.508	36.539	7.677	3.071e-14

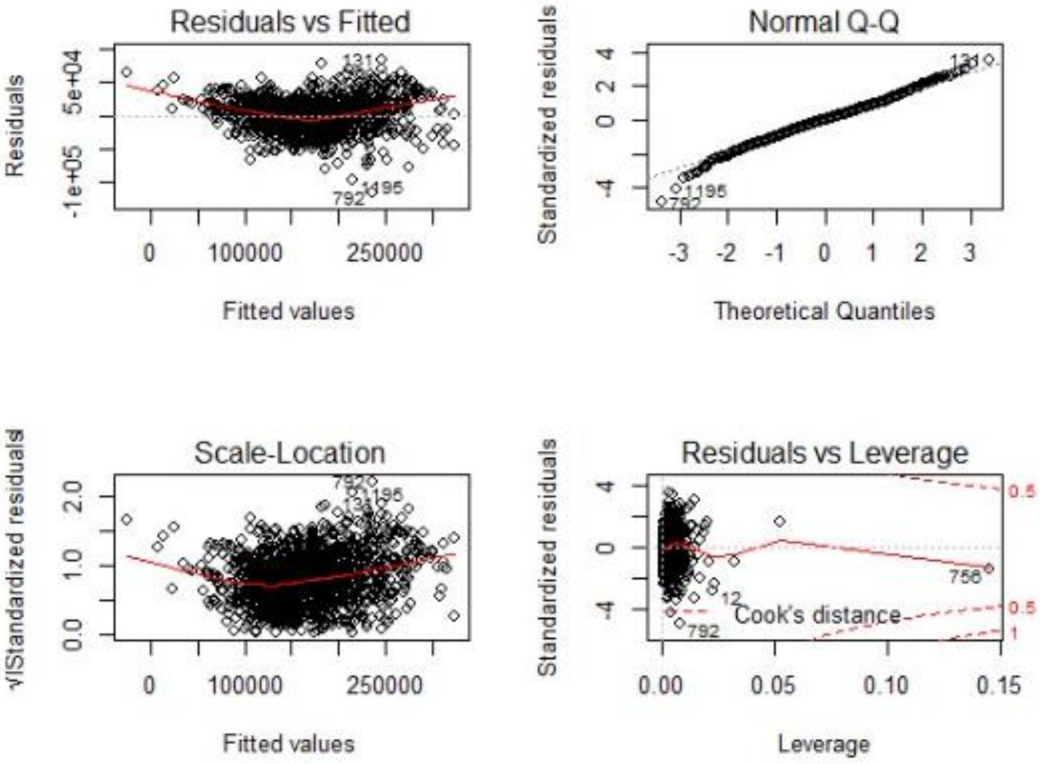
Anova table:



Response: SalePrice	Degrees of Freedom	Sum Sq	Mean Sq	F-Value	PR(>F)
GrLivArea	1	2.4904e+12	2.4904e+12	4520.841	<2.2e-16
OverallQuality	1	7.4068e+11	7.4068e+11	1344.575	<2.2e-16
TotalBasementSF	1	2.3006e+11	2.3006e+11	417.639	<2.2e-16
LotFrontage	1	3.2466e+10	3.2466e+10	58.936	3.071e-14
Residuals	1377	7.584e+11	5.5087e+8		

d)

Goodness of fit & underlying assumptions:



e)

The results show the plot is more randomized and less of a pronounced trend line with the additional variables.

Conclusion/Reflection)

After conducting the last part of the assignment, and adding the continuous variables and deleting the outliers, the model had a higher r-squared value. Doing so, however, is a tough task and can bring up unforeseen or additional challenges to the outcome and results. Outliers should thus only be deleted after a full analysis of improving the value of the results of the model but could be considered as a legitimate choice to improve results. I don't believe one can trust the statistical hypothetical test in regression but you can increase trust by checking the underlying assumptions of the model. This have shown to only be applicable when the residuals mean are zero and the underlying assumptions have not been proven false. Checking underlying assumptions and testing the statistical hypothesis should be done to trust the model more. The next steps for this modeling process should be to include the other categorical explanatory variables that have not been added to the model. There are quite a few other categorical variables that might affect the model further.