Modeling Assignment 4:
Building Linear Regression Models – Diagnostics and Transformations

Assignment Overview

In this assignment we will begin building regression models to predict the response variable home sale price (SALEPRICE) using the remaining variables in the AMES data set as explanatory variables.

Preparatory Work

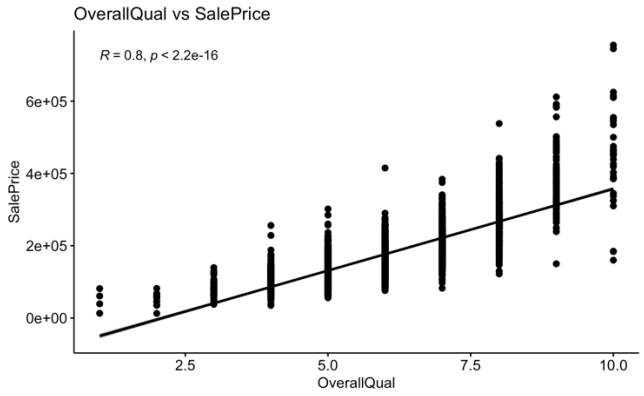
In Modeling Assignment #1, you were exposed to the idea of a Sample Population and the traditions of Exploratory Data Analysis. Every time you start a modeling endeavor, these two tasks need to be completed and formalized.
There is nothing that needs to be written about data preparation for this assignment.

Assignment Tasks

1. Let Y = sale price be the dependent or response variable. Select what you consider to be "the best" continuous explanatory variable from the AMES data set to predict Y. Discuss what criteria you used to select this explanatory variable? Fit a simple linear regression model using your explanatory variable X to predict SALE PRICE(Y). Call this Model 1. To report the results for Model 1, you are to:

OverallQual has the highest correlation with SalePrice: 0.8189. It will be used as "the best" continuous explanatory variable from the AMES data set to predict Y.

   a. Make a scatterplot of Y and X, and overlay the regression line on the cloud of data.



   b. Report the model in equation form and interpret each coefficient of the model in the context of this problem.

Predicted(SalePrice) = -109611.2 + 48480.1 * OverallQual

The constant is -109,611.2 which indicates that a data point with OverallQual = 0 will have a predicted SalePrice of -109,611.2. This value is not plausible, but only provides meaning by setting a baseline to help with the slope of the regression line. The majority of the data has OverallQual values that are relatively higher than 0.

The coefficient for OverallQual is 48,480.1 which means that for every additional point of OverallQual, the predicted SalePrice will increase by 48,480.1. Higher quality properties typically have higher sale prices.

c. Report and interpret the R-squared value in the context of this problem.

R2 = 0.6706. In context, 67.06% of variation in SalePrice can be explained by OverallQual.

d. Report the coefficient and ANOVA Tables.

```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq F value    Pr(>F)
OverallQual    1 9.1626e+12 9.1626e+12  4084.7 < 2.2e-16 ***
Residuals   2006 4.4997e+12 2.2431e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

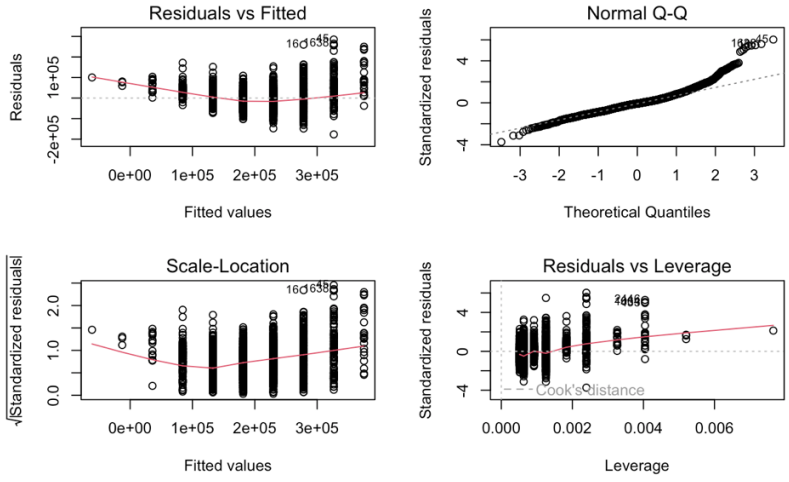
e. Clearly specify the hypotheses associated with each coefficient of the model, as well as the hypothesis for the overall omnibus model. Conduct and interpret these hypothesis tests.

Predicted(SalePrice) = $\beta_0 + \beta_1 * \text{OverallQual}$

Null Hyp	Alt Hyp	Test Stat	P-val
$\beta_0 = 0$	$\beta_0 \neq 0$	T = -22.46	<2e-16
$\beta_1 = 0$	$\beta_1 \neq 0$	T = 63.91	<2e-16
$\beta_1 = 0$	$\beta_1 \neq 0$	F = 4085 on 1 and 2006 DF	<2.2e-16

The T tests for the intercept and coefficient conclude that β_0 and β_1 are not 0. The overall omnibus model also concludes that β_1 is not 0.

f. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met.



The Residual vs Fitted plot has a slight curve up indicating that the relationship between SalePrice and OverallQual may not be linear. The increasing spread of residuals also indicates possible non-independence of errors. The Normal Q-Q plot deviates from the center line on the right end indicating possible non-normality of errors. Scale-Location is also curved up possibly indicating heteroscedasticity. *A transformation for one of the predictors or response should be performed.*

g. Check on leverage, influence and outliers. These points can be identified by several statistics such as DFFITS, Cook's Distance, Leverage, and Influence. Discuss any issues or concerns. Describe what course of action should be taken.

Test	Number of Points	% of population
DFFITS	175	8.72%
Cook's Distance	118	5.88%
Leverage	0	0%
Influence	DFFITS / CooksD	See above

Although many points have been identified here, the next course of action would be to look at plots of leverage, influence, and outliers to if there are any obvious points that should be removed. Removing 100-200 points seems excessive, and comparing how the regression line changes with and without certain influential points can help determine if eliminating them is the right choice. Based on the scatterplot and leverage plot, it doesn't seem like any of the points should be removed, but further investigation can take place if this model is to be tweaked.

2. For Task 2, you will fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). Call this Model 2. The explanatory variables for Model 2 should be the explanatory variable you had in Model 1, plus the OVERALL QUALITY variable. To report the results for Model 2, you are to:

   a. Report the prediction equation and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here relative to the simple linear regression model in Task 1?

Predicted(SalePrice) = -1.257e+05 + 3.377e+04 * OverallQual + 7.060e+01 * GrLivArea

The constant is -1.257e+05 which indicates that a data point with OverallQual = 0 and GrLivArea = 0 will have a predicted SalePrice of -1.257e+05. This value is not plausible, but only provides meaning by setting a baseline to help with the slope of the regression line. The majority of the data has OverallQual values and GrLivArea values that are relatively higher than 0.

The coefficient for OverallQual is 3.377e+04 which means that for every additional point of OverallQual (given all other predictors remain constant), the predicted SalePrice will increase by 3.377e+04. Higher quality properties typically have higher sale prices.

The coefficient for GrLivArea is 7.060e+01 which means that for every additional point of GrLivArea (given all other predictors remain constant), the predicted SalePrice will increase by 37.060e+01. Larger properties typically have higher sale prices.

The interpretations for Model 2 must take into account other predictors. All predictors must be held constant except for the one that is being interpreted (see above).

b. Report and interpret the R-squared value in the context of this problem. Calculate and report the difference in R-squared between Model 2 and Model 1. Interpret this difference.

R2 (Model 1) = 0.6706.
R2 (Model 2) = 0.7758. In context, 77.58% of variation in SalePrice can be explained by OverallQual and GrLivArea.

R2 for Model 2 is 0.1052 higher than R2 for Model 1. Which means that Model 2 is 10.52% better at explaining the variation in SalePrice than Model 1.

c. Report the coefficient and ANOVA Tables.

```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq F value    Pr(>F)
OverallQual    1 9.1626e+12 9.1626e+12 5998.15 < 2.2e-16 ***
GrLivArea      1 1.4369e+12 1.4369e+12  940.67 < 2.2e-16 ***
Residuals   2005 3.0628e+12 1.5276e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

d. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the overall omnibus model. Conduct and interpret these hypothesis tests.

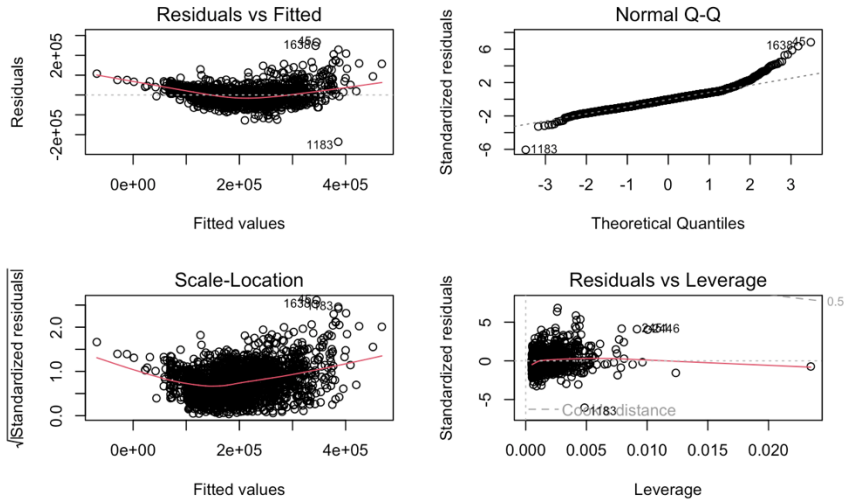
Predicted(SalePrice) = $\beta_0 + \beta_1$ * OverallQual + β_2 * GrLivArea

Null Hyp	Alt Hyp	Test Stat	P-val
$\beta_0 = 0$	$\beta_0 \neq 0$	T = -30.95	<2e-16
$\beta_1 = 0$	$\beta_1 \neq 0$	T = 42.83	<2e-16
$\beta_2 = 0$	$\beta_2 \neq 0$	T = 30.67	<2e-16
$\beta_1 = \beta_2 = 0$	$\beta_1 \neq 0$ or $\beta_2 \neq 0$	F = 3469 on 2 and 2005 DF	<2.2e-16

The T tests for the intercept and coefficient conclude that β_0 and β_1 and β_2 are not 0. The overall omnibus model also concludes that $\beta_1 \neq 0$ or $\beta_2 \neq 0$.

e. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. Check on these underlying assumptions by plotting:
   - Histogram of the standardized residuals
   - Scatterplot of standardized residuals (Y) by predicted values (Y_hat)

   Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.



Same concerns for assumptions as Model 1: *The Residual vs Fitted plot has a slight curve up indicating that the relationship between predictors may not be linear. The increasing spread of residuals also indicates possible non-independence of errors. The Normal Q-Q plot deviates from the center line on the right end indicating possible non-normality of errors. Scale-Location is also curved up possibly indicating heteroscedasticity. A transformation for one of the predictors or response should be performed.*

f. Check on leverage, influence and outliers, and discuss any issues or concerns.

Test	Number of Points	% of population
DFFITS	120	5.98%
Cook's Distance	108	5.38%
Leverage	0	0%
Influence	DFFITS / CooksD	See above

Similar conclusion about influential points as Model 1: *Although many points have been identified here, the next course of action would be to look at plots of leverage, influence, and outliers to if there are any obvious points that should be removed. Removing ~100 points seems excessive, and comparing how the regression line changes with and without certain influential points can help determine if eliminating them is the right choice. Based on the scatterplot and leverage plot, point 1183 seems out of place on most of the assumption plots and should be removed.*

g. Based on the information, should you want to retain both variables as predictor variables of Y? Discuss why or why not.

Both variables should be retained because of the significance of the hypothesis tests pertaining to the coefficient for each of the predictors and higher value of R2.

3. Select any other continuous explanatory variable you wish. Fit a multiple regression model that uses 3 continuous explanatory (X) variables to predict Sale Price (Y). These three variables should be the explanatory variables from Model 2 plus your choice of an additional explanatory variable. Call this Model 3. To report the results for Model 3, you are to:

a. Report Model 3 in equation form and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here to Models 1 and 2?

Predicted(SalePrice) = -9.256e+05 + 2.793e+04 * OverallQual + 7.317e+01 * GrLivArea + 4.210e+02 * YearBuilt

The constant is -9.256e+05 which indicates that a data point with OverallQual = 0 and GrLivArea = 0 and YearBuilt = 0 will have a predicted SalePrice of -9.256e+05. This value is not plausible, but only provides meaning by setting a baseline to help with the slope of the regression line. The majority of the data has OverallQual values and GrLivArea and YearBuilt values that are relatively higher than 0.

The coefficient for OverallQual is 2.793e+04 which means that for every additional point of OverallQual (given all other predictors remain constant), the predicted SalePrice will increase by 2.793e+04. Higher quality properties typically have higher sale prices.

The coefficient for GrLivArea is 7.317e+01 which means that for every additional point of GrLivArea (given all other predictors remain constant), the predicted SalePrice will increase by 7.317e+01. Larger properties typically have higher sale prices.

The coefficient for YearBuilt is 4.210e+02 which means that for every additional point of YearBuilt (given all other predictors remain constant), the predicted SalePrice will increase by 4.210e+02. Newer properties typically have higher sale prices.

The interpretations for Model 3 must take into account other predictors. All predictors must be held constant except for the one that is being interpreted (see above).

b. Report and interpret R-squared value in the context of this problem. Calculate the difference in R-squared between Model 3 and Model 2. How would you interpret this difference? Does your variable of choice help to improve the model's explanatory ability?

R2 (Model 1) = 0.6706.
R2 (Model 2) = 0.7758.
R2 (Model 3) = 0.7855. In context, 78.55% of variation in SalePrice can be explained by OverallQual and GrLivArea and YearBuilt.

R2 for Model 3 is 0.0097 higher than R2 for Model 2. Which means that Model 3 is 0.97% better at explaining the variation in SalePrice than Model 2.

c. Report the coefficient and ANOVA Tables for Model 3.

```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq  F value    Pr(>F)
OverallQual    1 9.1626e+12 9.1626e+12 6266.909 < 2.2e-16 ***
GrLivArea      1 1.4369e+12 1.4369e+12  982.820 < 2.2e-16 ***
YearBuilt      1 1.3281e+11 1.3281e+11   90.837 < 2.2e-16 ***
Residuals   2004 2.9300e+12 1.4621e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

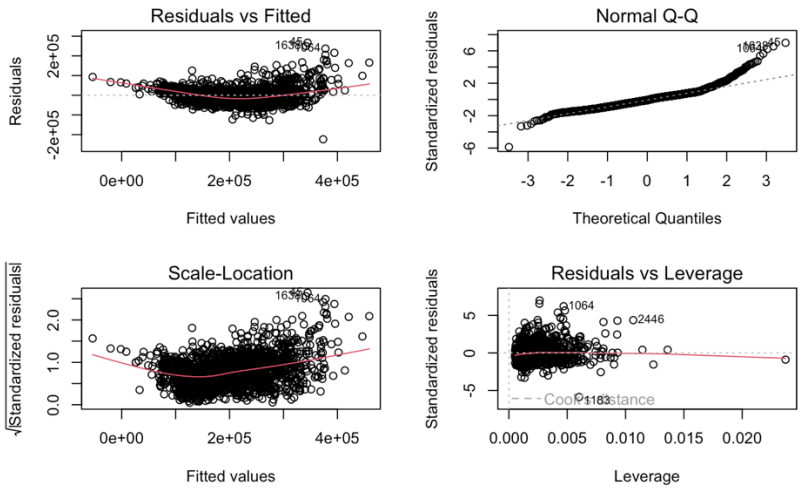
d. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret these hypothesis tests.

Predicted(SalePrice) = $\beta_0 + \beta_1 * \text{OverallQual} + \beta_2 * \text{GrLivArea} + \beta_3 * \text{YearBuilt}$

Null Hyp	Alt Hyp	Test Stat	P-val
$\beta_0 = 0$	$\beta_0 \neq 0$	T = -11.016	<2e-16
$\beta_1 = 0$	$\beta_1 \neq 0$	T = 28.340	<2e-16
$\beta_2 = 0$	$\beta_2 \neq 0$	T = 32.262	<2e-16
$\beta_3 = 0$	$\beta_3 \neq 0$	T = 9.531	<2e-16
$\beta_1 = \beta_2 = \beta_3 = 0$	$\beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$	F = 2447 on 3 and 2004 DF	<2.2e-16

The T tests for the intercept and coefficient conclude that β_0 and β_1 and β_2 and β_3 are not 0. The overall omnibus model also concludes that $\beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$.

e. Check on the underlying assumptions. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.



Same concerns for assumptions as Model 1 and Model 2: *The Residual vs Fitted plot has a slight curve up indicating that the relationship between SalePrice and predictors may not be linear. The increasing spread of residuals also indicates possible non-independence of errors. The Normal Q-Q plot deviates from the center line on the right end indicating possible non-normality of errors. Scale-Location is also curved up possibly indicating heteroscedasticity. A transformation for one of the predictors or response should be performed.*

f. Check on leverage, influence and outliers, and discuss any issues or concerns.

Test	Number of Points	% of population
DFFITS	105	5.22%
Cook's Distance	114	5.68%
Leverage	0	0%
Influence	DFFITS / CooksD	See above

Similar conclusion about influential points as Model 1 and Model 2: *Although many points have been identified here, the next course of action would be to look at plots of leverage, influence, and outliers to if there are any obvious points that should be removed. Removing ~100 points seems excessive, and comparing how the regression line changes with and without certain influential points can help determine if eliminating them is the right choice. Based on the scatterplot and leverage plot, point 1183 seems out of place on most of the assumption plots and should be removed.*

g. Based on this information, should you want to retain all three variables as predictor variables of Y? Discuss why or why not.

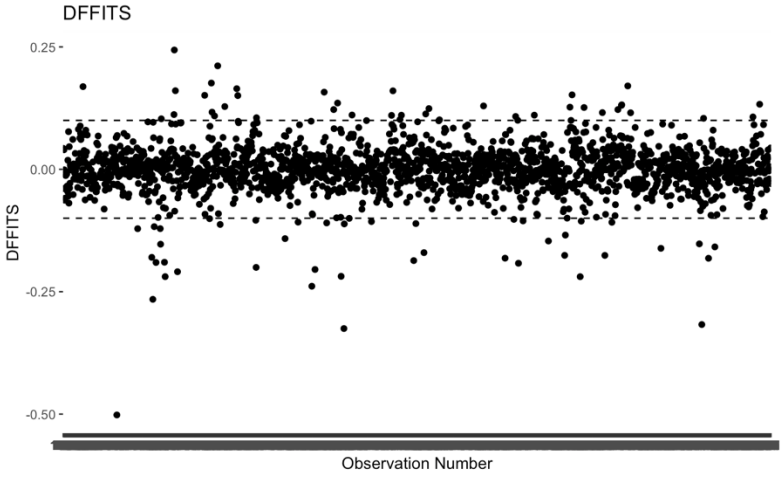
All variables should be retained because of the significance of the hypothesis tests pertaining to the coefficient for each of the predictors and higher value of R2.

4. Refit Model 3 using the Natural Log of SALEPRICE as the response variable. Call this Model 4. This is LOG base e, or LN() on your calculator. You'll have to find the appropriate function using R. Perform an analysis of goodness-of-fit to compare the Natural Log of SALEPRICE model, Model 4, to the original Model 3. Does the transformed model fit better? Provide evidence in your discussion. Discuss if the improvement of model fit justifies the use of the transformed response variable, Log(SALEPRRICE).

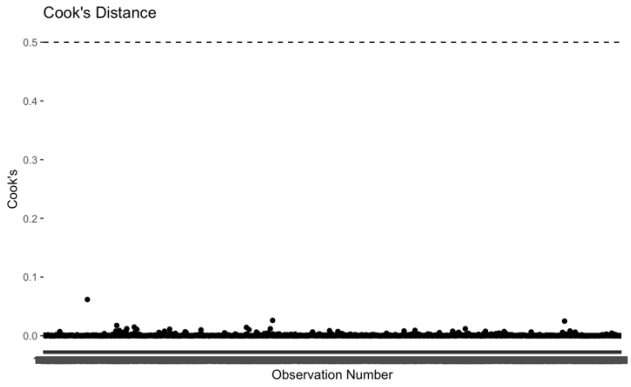
We will use R2 as our goodness of fit measure for linear regression Model 3 and Model 4.
R2 (Model 3) = 0.7855
R2 (Model 4) = 0.8377

The transformed model does fit better because the goodness of fit measure, R2, is higher in Model 4 than Model 3. This difference indicates relatively better performance and justifies the use of log(SalePrice).

5. For either Model 3 or Model 4, your choice, identify the influential, high leverage, or outlier data points. Remove these data points from the dataset, then refit the model after removing the influential points. How many influential points did you find & remove? When you refitted the model, did the model improve? Comment on whether or not you find the improvement of model fit justifies the potential for the modeler biasing the result by removing potentially legitimate data points.



From the graph above, DFFITS indicates that any point outside of the dotted lines is influential. In order to avoid modeler biasing, we want to keep the number of removed data points to a minimum. The one at the bottom (DFFITS = -0.50) is an obvious outlier. I will choose to remove any points outside of the 0.25 DFFITS values which is 4 points total (the 4 at the bottom of the plot: SID = 710, 1183, 1306, 2044)



Cook's Distance however indicates that there are no influential points. For the sake of exploratory analysis, let's test out removing the few points with Cook's Distance that are relatively higher than the rest. There appear to be 3 points in this plot that are slightly higher than the rest: CD > 0.02 (SID = 710, 1183, 2044). All of these points are also listed in the DFFITS analysis.

Here is the summary after removing the outliers:

We will use R2 as our goodness of fit measure for Model 4 with and without influential points.
R2 (Model 4 with influential points) = 0.8377
R2 (Model 4 w/o influential points) = 0.8432

The transformed model does fit better because the goodness of fit measure, R2, is higher in Model 4 without the influential points. This difference indicates relatively better performance and justifies the removal of these influential points. Raising the R2 value by 0.0055 by only removing 4 points improves the model with minimal risk of introducing modeler bias – we are still using 99.8008% of the original population.

6. Use the concept of Change in R-squared, plus anything else you wish, to put together a reasonable approach to find a good, comprehensive multiple regression model to predict SALEPRICE(Y). Any of the continuous variables can be considered fair game as explanatory variables. This can feel like an overwhelming task. You don't need to go overboard, or kill yourself, in doing this. We will learn about automated approaches to do this shortly. But, for now, I'd like you to think about how you would do this by hand.

   Use your approach to identify a good multiple regression model to predict SALEPRICE(Y) from the set of continuous explanatory variables available to you in the AMES dataset. For this task you need to:

   a. Explain your approach

I would first run a correlation plot to find the values that are most highly correlated to SalePrice. I would add them one at a time to a model. I would first test for any multicollinearity using logic (for example OverallQual and OverallCond seem like they would be similar), and try to avoid

including them in the same model, confirming with VIF that there is no multicollinearity. I would check that all predictors have significant coefficients via T-test. I would record R2_adj values because the number of coefficients will var from model to model. The model with the highest R2_adj that does not violate multicollinearity assumptions will be the best model. Models with similar R2_adj values will favor the model with less predictors because that indicates that the model with more predictors may have an insignificant one. Once a good selection of variables is picked, a model may also consider transformations of predictors or response .

Model	All Significant Coefficients (T-test)?	VIF < 5 ?	R2_adj
OverallQual	Yes	n/a	0.6705
+ GrLivArea	Yes	Yes	0.7756
+ GarageCars	Yes	Yes	0.7898
+ GarageArea	No – GarageCars	remove GarageCars	0.7994
+ FirstFlrSF	Yes	Yes	0.8217
+ YearBuilt	Yes	Yes	0.8274
+ FullBath	Yes	Yes	0.8306
+ TotRmsAbvGrd	No – remove TotRmsAbvGrd	Yes	0.8306
+ YearRemodel	Yes	Yes	0.8338
All other predictors have correlations with SalePrice < 0.5.			

Here is the final model according to the methods above:

b. Report the model you determined and interpret the coefficients

```
Predicted(SalePrice) = -1.330e+06
                     + 2.105e+04 * OverallQual
                     + 6.385e+01 * GrLivArea
                     + 5.625e+01 * GarageArea
                     + 4.187e+01 * FirstFlrSF
                     + 2.908e+02 * YearBuilt
                     - 1.469e+04 * FullBath
                     + 3.342e+02 * YearRemodel
```

The coefficient for `OverallQual` is `2.105e+04` which means that for every additional point of `OverallQual` (given all other predictors remain constant), the predicted SalePrice will increase by `2.105e+04`. Higher quality properties typically have higher sale prices.

The coefficient for `GrLivArea` is `6.385e+01` which means that for every additional point of `GrLivArea` (given all other predictors remain constant), the predicted SalePrice will increase by `6.385e+01`. Properties with more living area typically have higher sale prices.

The coefficient for `GarageArea` is `5.625e+01` which means that for every additional point of `GarageArea` (given all other predictors remain constant), the predicted SalePrice will increase by `5.625e+01`. Properties with larger garages typically have higher sale prices.

The coefficient for `FirstFlrSF` is `4.187e+01` which means that for every additional point of `FirstFlrSF` (given all other predictors remain constant), the predicted SalePrice will increase by `4.187e+01`. Properties with more 1st floor living space typically have higher sale prices.

The coefficient for `YearBuilt` is `2.908e+02` which means that for every additional point of `YearBuilt` (given all other predictors remain constant), the predicted SalePrice will increase by `2.908e+02`. Newer properties typically have higher sale prices.

The coefficient for `FullBath` is - `1.469e+04` which means that for every additional point of `FullBath` (given all other predictors remain constant), the predicted SalePrice will increase by - `1.469e+04`. Properties with less baths typically have higher sale prices. This seems counterintuitive and may be due to having too many variables in the model. This variable should be investigated.

The coefficient for `YearRemodel` is `3.342e+02` which means that for every additional point of `YearRemodel` (given all other predictors remain constant), the predicted SalePrice will increase by `3.342e+02`. Recently remodeled properties typically have higher sale prices.

c. Report the coefficient and ANOVA tables.

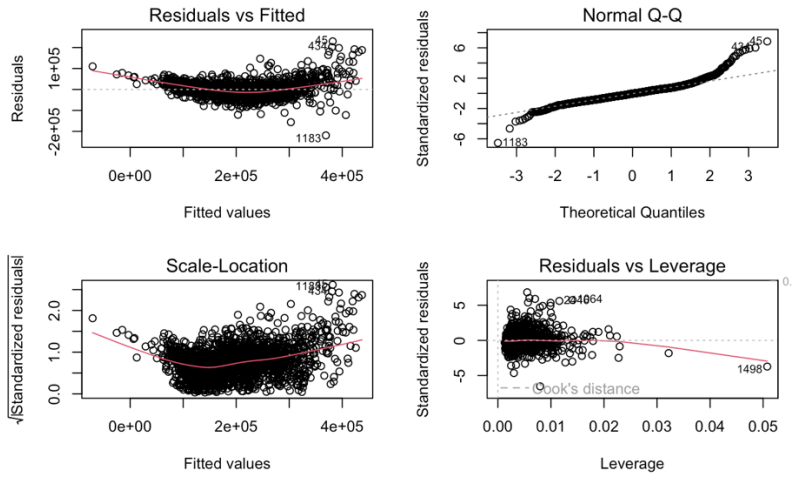
```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq  F value    Pr(>F)
OverallQual     1 9.1626e+12 9.1626e+12 8096.470 < 2.2e-16 ***
GrLivArea       1 1.4369e+12 1.4369e+12 1269.745 < 2.2e-16 ***
GarageArea      1 3.2639e+11 3.2639e+11  288.409 < 2.2e-16 ***
FirstFlrSF      1 3.0533e+11 3.0533e+11  269.800 < 2.2e-16 ***
YearBuilt       1 7.9406e+10 7.9406e+10   70.167 < 2.2e-16 ***
FullBath        1 4.3762e+10 4.3762e+10   38.670 6.093e-10 ***
YearRemodel     1 4.4541e+10 4.4541e+10   39.358 4.312e-10 ***
Residuals    2000 2.2634e+12 1.1317e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

d. Report goodness of fit

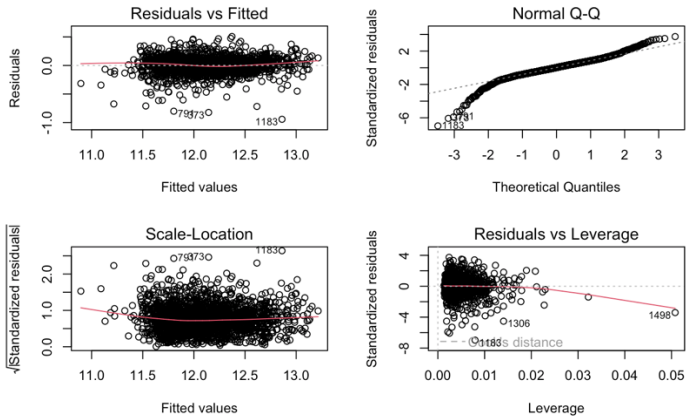
Adjusted R2 = 0.8338. In context, 83.38% of variation in SalePrice can be explained by the 7 predictors in this linear regression model.

e. Check on underlying model assumptions.

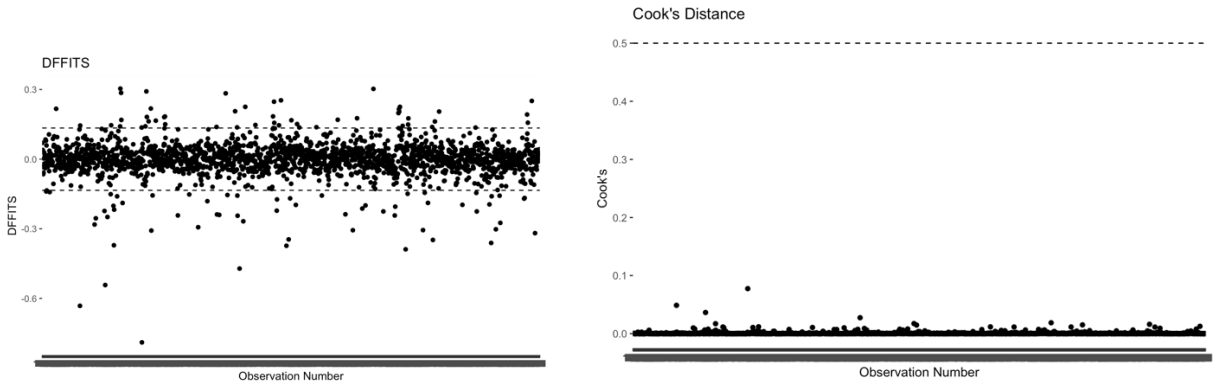


Same concerns for assumptions as Model 1: *The Residual vs Fitted plot has a slight curve up indicating that the relationship between predictors may not be linear. The increasing spread of residuals also indicates possible non-independence of errors. The Normal Q-Q plot deviates from the center line on the right end indicating possible non-normality of errors. Scale-Location is also curved up possibly indicating heteroscedasticity. A transformation for one of the predictors or response should be performed.*

We will use log(SalePrice) and and re-examine the scale plots. From the summary, all p-vals for coefficient significance are < 0.5 and the R2_adj is 0.8744



The Residuals vs Fitted plot has a much more random scatter which indicates independent errors. Let's take a look at outliers:



Although there are no blatant influential points, we will remove the 3 that stand out the most (the bottom 3 from DFFITS and top 3 of Cook's Distance). After removing them, the R2_adj is 0.879 compared to the model with influential points which is 0.8744.

With a slight increase in prediction power after removing the 3 points, it is justified in removing them. And removing so few points minimizes the risk of modeler bias.

7. Please write a conclusion / reflection section that, at minimum, addresses the questions:
    - In what ways do variable transformation and outlier deletion impact the modeling process and the results?
    - Are these analytical activities a benefit or do they create additional difficulties?
    - Can you trust statistical hypothesis test results in regression?
    - What do you consider to be next steps in the modeling process?

Conclusion

The process of manually tweaking linear models for better fit is a long and methodical process. Here are a few best practices that I have learned from this assignment and previous ones.

Prep work: From previous assignments we discovered the importance of identifying the population of interest. In doing so, we are careful not to extrapolate findings to datapoints outside of the population of interest. Defining the population of interest also plays a monumental role in the interpretations of models. The model in this assignment defines the population of interest as single family homes in medium to high residential living zones. The time of data collection was not indicated in the data dictionary which makes it difficult to draw conclusions and is also a point of caution in interpretations. Further waterfall drop statements can be taken, but I have decided that those variables are the only ones important for exclusion criteria in my analyses, and that including more data was more valuable than including more waterfall drop statements. The EDA process is also important in discovering relationships between variables that may uncover the "story" of our population of interest.

Variable Transformation: In linear regression, we assume a linear relationship between predictors and the outcome, independent residuals, normal distribution of errors with a mean of 0, and homoscedasticity. If a model does not satisfy these assumptions, it is our role as data scientists to make modifications to the model so that they are satisfied. One way of doing so is variable transformation. In this assignment, we looked at the log transformation of SalePrice. Log transformations are typically done to outcome variables with a right skew in histogram distribution to bring the tail in and make the distribution more normal. We confirmed in our variable transformation that the Residuals vs Fitted plot became more randomly distributed which indicates that we are satisfying some of those model assumptions. The benefit of variable transformations is not only that the model is useable, but that the model also performs better as indicated by the increase in the goodness of fit statistic, R2, from the models in the assignment. There are many difficulties with variable transformation, like thinking of the many types of transformations that can occur and the many combinations of transformed variables that can be included in a model. Ideally the model assumptions are satisfied, but it may take some time to test optimal transformations that need to be made, and sometimes, there just is no amount of transformations that will guarantee a better model.

Outlier Deletion: From the prep work section, we already know that waterfall drop conditions remove data points. Another way of removing points in order to focus on the population of interest would be outlier deletion. Data points that are considered part of the population of interest may not follow the pattern of prediction that most other data points do, and removing these points allows the model to focus better on those patterns without being distracted by these outliers. In this assignment we used DFFITS and Cook's Distance to determine which points are highly influential and how to remove them from the dataset. With a dataset of 2000+ data points, I opted to remove less than 5 from any models. One difficulty in removing lots of valid data points is that it introduces modeler bias and may push the model to tell a story that is not indicative of the population of interest. Removing fewer points as I did minimized that risk of

introducing bias. From the assignment, we saw that the R2_adj value increased incrementally indicating that the removal of these outliers was indeed justified without going overboard on the number of points deleted. This shows that a benefit of outlier deletion is better performing models.

Statistical Hypotheses: In the intermediary step of checking which variables to include in a model, there are different hypothesis test that we check along the way. From this assignment, as we added variables to the base model, we checked the T-tests of each beta coefficient. This test indicates whether or not the value of the coefficient is 0 or non-0. In other words, does adding this variable to the model improve it more than random chance? Another hypothesis test would be the omnibus model: a hypothesis test that questions whether all of the coefficients are 0 or at least one is non-0. In other words, is random chance better at predicting SalePrice than all of these variables combined? These hypotheses test can be meaningful in whether or not to include these variables, but including them is also dependent on many other factors. For example, a model can have two significant predictors according to p-values, but in reality, the variables may highly overlap and introduce multicollinearity which is a violation of linear regression assumptions. Another case from the assignment is that predictors may be significant according to hypothesis tests, but the coefficient value may be counterintuitive by interpretation. For example, “a house with more baths will sell for less” is a concerning conclusion to see from a model because we would expect the opposite. So yes, by general rule of thumb, hypothesis tests are good indicators of whether or not to include predictors in a model. Because we cannot fully trust them, there are other tests we can use (like VIF for multicollinearity), and logical reasoning that must also be used in interpretations. The hypothesis test is not the end-all-say-all.

Next Steps in Modeling: I believe the next step in modeling would be to incorporate categorical predictors. So far, we have only included continuous predictors which does tell part of the story. But ignoring the categorical predictors is ignoring more parts of the “story.” Other steps in modeling include the different types of models aside from linear modeling. Given the dataset and population of interest in this assignment, linear regression is appropriate. But different types of data, different populations of interest, and different “stories” to tell require different statistical models, and linear regression will not always be appropriate. Aside from categorical predictors and non-linear models, I believe that actual prediction is another next step in modeling which seems fairly simple, but is often used in real-world consultations. Predictions really tie in everything that we have learned from the assignments such as communicating the population of interest and the caution in interpretations of the model. Predictions using modeling can also be coupled with other modeling statistics such as confidence intervals and data visualizations in order to communicate the “story” in multiple ways and indicate with caution what outcomes the model may produce.