NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Week 4 Assignment - Modeling: Building Linear Regression Models
MSDS 410

In this assignment we will begin building regression models to predict home sale price (SALEPRICE) using the variables in the AMES data. This assignment walks you through a number of modeling experiences, these are delineated below as tasks. Each task should be completed and written about separately. This is not necessarily the sequence of steps of what you should do in a modeling setting, but it is intended to give you perspective on modeling.

Purpose: In this modeling assignment we will begin building linear regression models to predict the home sale price. As such the response variable is: SALEPRICE (Y). We will begin by fitting specific models and looking at diagnostic and model fit information. Models will progressively become more involved and complex over the span of this assignment.

Data: The data for this assignment is the Ames, Iowa housing data set. This data is posted in Canvas.

Explanatory Variables: All continuous variables in the AMES Housing data set

Assignment Tasks:

(0) Define the Sample Population – Exploratory Data Analysis

Please note that since it is part 0 and not graded, I used Dr. Srinivasan's EDA solution guide and my assignment 1 answers.

The Ames Data set is a record of residential home sales in Ames, Iowa from 2006 to 2010. Since we don't have access to data of all homes in Ames, we will use the sample provided by the Assessor's office. The sample data includes 2930 property sales and 80 variables containing property characteristics such as property age, number of rooms, number of additional features, space size, quality, style, and etc. Before we can begin exploring the data, we need to define our problem. We've been tasked with building a model that can provide estimates of home values for "typical" homes in Ames, Iowa. Therefore, our population of interest are all "typical" homes in Ames. The 2930 observations in that data set include values that we would consider typical and atypical; thus, we need to par down the data to only those values that represent typical homes in Ames. We need to explore features of homes and use the relevant ones to predict the future home sale price.

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

As it says above, we are building regression models for the response variable SalePrice(Y). In order to do this, you need to know the Sample Population. Without this, it is not possible to infer results from the sample to the larger population, it makes the notion of hypothesis testing for population parameter values irrelevant, and it makes the process of determining outliers highly problematic. Frankly, it throws your whole purpose for modeling into chaos.

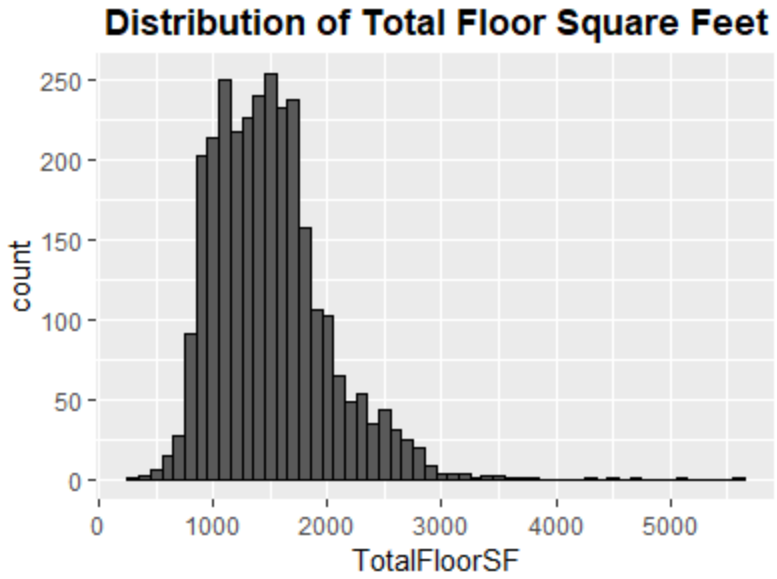
Defining the Sample Population is actually a very powerful tool for you as the modeler. It gives you license to define what aspects of the data are legitimate for you to work with. You don't have to model ALL of the data you are given in one model. You can break the data up into parts and model them separately. Why would you want to do this? Well, are all properties the same? Would we want to include an apartment building in the same sample as a single family residence? Would we want to include a warehouse or a shopping center in the same sample as a single family residence? Would we want to include condominiums in the same sample as a single family residence? Are there certain kinds of properties that are not like the others? Could one be a derelict property such that it is not like the others? Could one be a mansion such that it is not like the other properties in the data set? You get to define this! In doing this, often many records with extreme scores are eliminated from modeling consideration. Define your Target Population and hence the Sample using 'drop conditions'. Create a waterfall for the drop conditions and include it in your report so that it is clear to any reader what you are excluding from the data set when defining your sample population. If you want to use your conditions from Modeling Assignment #1, that is fine. If you feel you need to make changes, now is the time to do so.

Please note that since it is part 0 and not graded, I used Dr. Srinivasan's EDA solution guide and my assignment 1 answers.

Waterfall presentation:

We start with 2930 homes, but data exploration may justify dropping some records from our training data set. After reviewing the data dictionary, there are some areas that warrant further investigation.

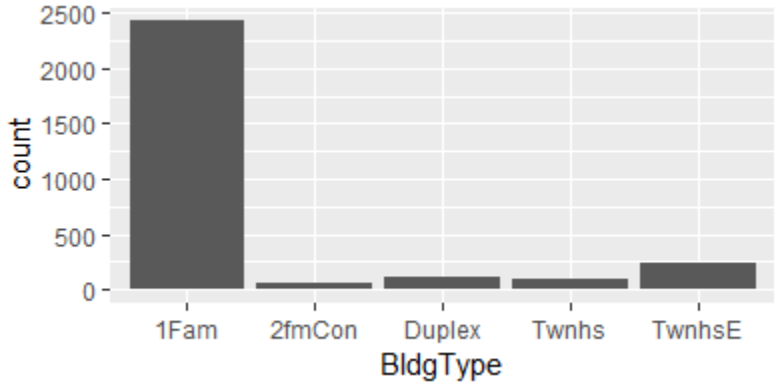
#1: House size
We need to identify the typical house in the market. The histogram below reveals that there are 5 properties that significantly deviate from the center of the distribution of houses. All of these properties are over 4000 square feet and do not well represent the typical size of a house in Ames. As such, our first step will remove those observations with a square footage over 4000 sq. ft.



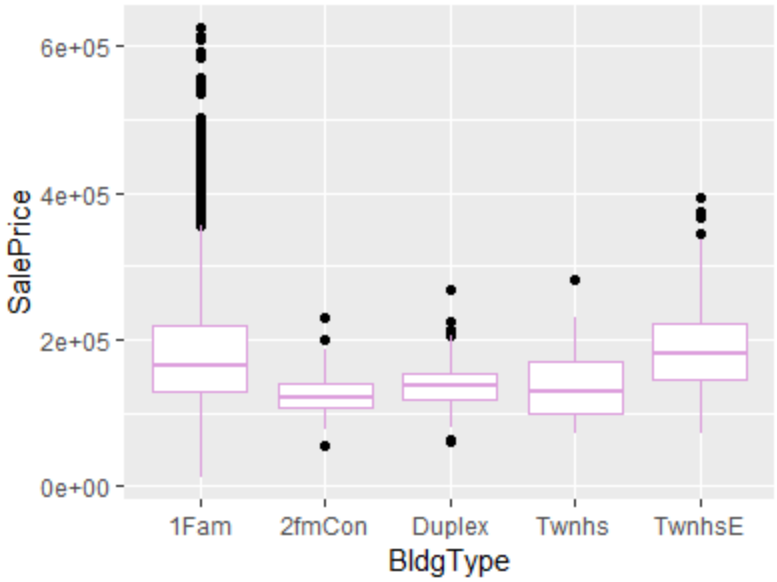
Adjustment: Drop records greater than or equal to 4000 square feet
Old record count: 2930
New record count: 2925

#2: Property type
The data set includes a categorization variable for the type of property that was sold in the transaction. Below is a bar chart of this variable (BldType).

	BldgType	Count	Percentage	AvgSalesPrice
1	1Fam	2420	83	184356
2	2fmCon	62	2	125582
3	Duplex	109	4	139809
4	Twnhs	101	3	135934
5	TwnhsE	233	8	192312



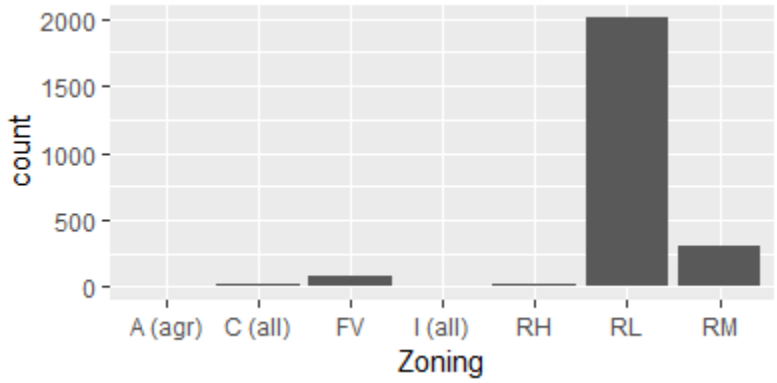
As the chart reveals, most of the home sales are single family (1Fam) homes; however, two-family properties, duplexes, and townhouses are included. By far most sales are single family, which suggests that single-family homes are the “typical” transaction in Ames. It also notable that median price of single-family homes is notably higher than multi-family units, duplex, and townhouses. So those building types are more atypical in frequency and in price. As such, I will exclude them from the training data set.



Adjustment: Reduce sample to only single-family homes
Old record count: 2925
New record count: 2420

#3: Zoning considerations
There are some unexpected and perhaps atypical patterns in the zoning data. In the graph below, we find very few records in Agriculture (A), commercial (C), and Industrial (I). The rest of the zones appear to be different forms of residential zoning. As the typical home is single family, it is advisable to restrict our data to those records with residential zoning.

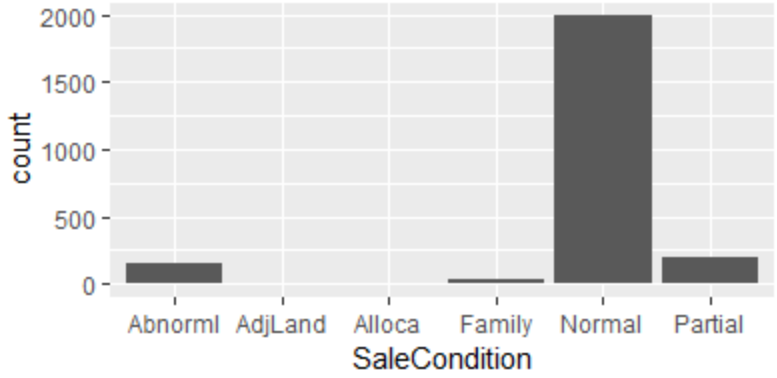
	Zoning	Count	Percentage	AvgSalesPrice
1	A (agr)	2	0	47300
2	C (all)	22	1	73926
3	FV	77	3	243281
4	I (all)	2	0	80312
5	RH	12	0	110786
6	RL	2004	83	193170
7	RM	301	12	123201



Adjustment: Reduce sample to only residential zoning
Old record count: 2420
New record count: 2394

#4: Deal type
In typical real-estate transactions, the buyers are two unrelated parties who negotiate the fair-market value of the home. This is not the only case.

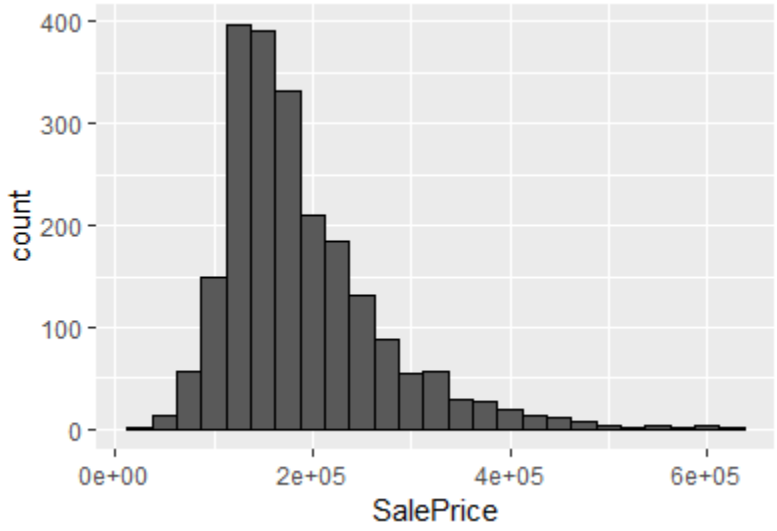
	SaleCondition	Count	Percentage	AvgSalesPrice
1	Abnorml	147	6	143194
2	AdjLand	7	0	105429
3	Alloca	8	0	217140
4	Family	41	2	156524
5	Normal	1987	83	179600
6	Partial	204	9	281623



As we can see, there are several types of unusual sale transactions. Those include sales with adjacent land, multiple allocations (condos), and family transactions. I elected to eliminate these from the training data set as these situations seemed to deviate from typicality. Family transactions are less reliable markers of the real-estate market, and it appears that very few real estate transactions have multiple allocations or adjacent land attached to them. I debated whether I should eliminate abnormal and partial. Both seemed like less than trivial parts of the market, and abnormal transactions (foreclosures) are part of any real-estate market. As foreclosures emerge, it would be helpful if the model were sensitive to the realities of foreclosed home sales. I elected to keep both in the training set.

Adjustment: Restrict sale condition to normal, partial, and abnormal transactions.
Old record count: 2394
New record count: 2191

#5 Sale price
Even after restricting the size of home to below 4000 square feet, there remains a few unusually expensive home sales. It would see that sales about 600,000 dollars are unusual (atypical) in this market.



Adjustment: Drop records greater than or equal to 600,000 sales prices
Old record count: 2191
New record count: 2187

Below is a summary of my waterfall process:

Step	Discussion	Count
0	Original data set	2930
1	Eliminate large house outliers	2925
2	Keep only single-family residents	2420
3	Drop industrial, commercial, agriculture zoning	2394
4	Drop sales between family members, adjoining land deals, condo sales	2191
5	Drop excess sale prices	2187

I slimmed my list down to the following 10 for further analysis based on some of the work of the previous section and information contained in the data displays:

#	Variable	Theory for Inclusion
1	LotArea	Suspected larger lot correlated with house price
2	Neighborhood	Prior research suggests a relationship between local amenities and house price
3	QualityIndex	Higher quality should yield higher price
4	HouseAge	Newer house may be more expensive
5	GarageCars	Larger garages may lead to higher sales price
6	KitchenQual	Kitchens are the focal point of many houses; should correlate positively with price
7	YrSold	Sales may correlate to larger economic trends (Great Recession)
8	SaleCondition	Sales under duress (foreclosures) should lower sales price
9	TotalFloorSF	Total floor sq should correlate with sale price
10	FullBath	Higher number of baths should yield higher price

Once the Sample Population is clearly well defined, and you've selected only those records and fit the Sample Population definition, you can then continue to perform a detailed Exploratory Data Analysis (EDA). Usually, this is broken up into two parts. The first is data preparation (or data cleaning). Here, you concern yourself with any remaining missing values, extreme scores, and outliers.

a. Are there variables with missing values? Should values for these variables be imputed or "fixed"? You can impute values for the missing data points by using a mean or median for the variable. Or, maybe use a decision tree, other contextual information, or models. For variables with large numbers of missing values, you may want to simply eliminate that

variable from the dataset. One option is to not do anything. In R, the default way that missing values are handled is to remove the record with a missing value from the computation, if the variable with the missing value is included in the function. Always keep this fact in mind.

I did not find any missing values in these variables.

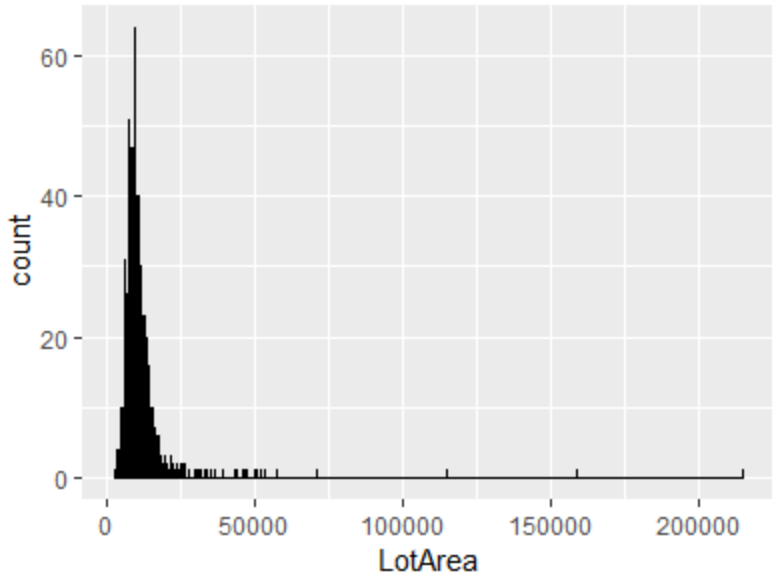
	Null_Value
SalePrice	0
LotArea	0
Neighborhood	0
QualityIndex	0
HouseAge	0
GarageCars	0
KitchenQual	0
YrSold	0
SaleCondition	0
TotalFloorSF	0
FullBath	0

b. Do any of the variables have outliers or extreme values? Should these extreme values be replaced? Fix any extreme values that need fixing. Note: This may be something you do in conjunction with the EDA as you find extreme values.

Quantitative Variables: For quantitative variables (integers, numeric) I constructed a table in R which provides information on measures of center, spread, range, skewness, kurtosis, and correlation 10 varaibles to sales price.

	Min	Q_1st	StDev	Median	Mean	Q_3rd	Max	Range	Skewness	Kurtosis	cv	Corr
SalePrice	35000	134500.0	78895.03	169500	188319.71	224371.5	591587	556587	1.46	2.81	0.42	1.00
LotArea	2500	8241.5	7504.90	9830	10861.91	11968.5	215245	212745	15.07	341.07	0.69	0.29
QualityIndex	1	30.0	8.97	35	34.69	40.0	90	89	0.41	2.76	0.26	0.53
HouseAge	0	7.0	30.38	36	36.52	55.0	136	136	0.55	-0.58	0.83	-0.59
GarageCars	0	1.0	0.74	2	1.81	2.0	5	5	-0.12	0.05	0.41	0.69
YrSold	2006	2007.0	1.30	2008	2007.80	2009.0	2010	4	0.13	-1.14	0.00	-0.03
TotalFloorSF	334	1141.0	492.11	1478	1517.55	1788.0	3820	3486	0.76	0.72	0.32	0.78
FullBath	0	1.0	0.55	2	1.56	2.0	3	3	0.16	-0.98	0.35	0.62

By looking at skewness of summary chart above and the histogram below reveals that there are 21 properties that significantly deviate from the center of the distribution of houses. All of these properties are over 30,000 square feet and do not well represent the typical size of a house in Ames. As such, we will remove those observations with a square footage over 30,000 sq. ft.



Adjustment: Drop records greater than or equal to 30,000 lot area
Old record count: 2187
New record count: 2166

Then, you can turn your attention to understanding the data more deeply. You were exposed to the EDA ideas and traditions in Module 1. For a full blown modeling project, you would want to exam all of the variables in your data set. Some suggestions for things that you could do are:

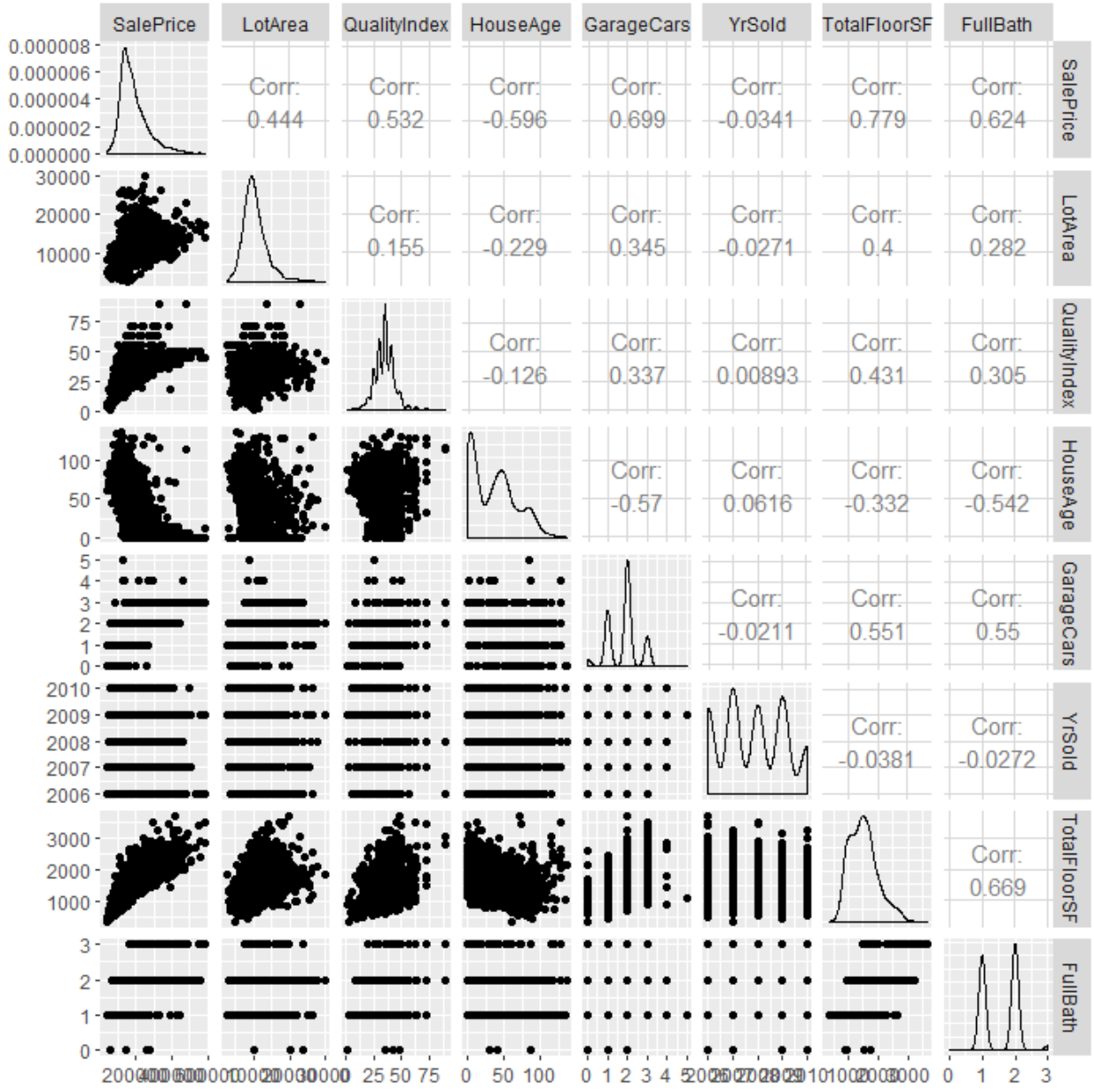
c. Obtain histograms for each continuous variable
d. Obtain summary statistics, such as: Means, standard deviations, minimum, maximum, median for all continuous variables
e. Are the explanatory variables correlated to the response variable?
f. Are the explanatory variables correlated amongst themselves?
g. Obtain scatterplots of explanatory variables with the response variable.
h. Do you want to create new variables to make the analysis more easily interpretable? For example, you might want to create a variable like PRICE per SQR FOOT. This could be a more meaningful response variable than total home sale price. I'm sure with a little bit of google searching you can find other variables that you would want to compute and potentially use. This is totally voluntary on your part. Not at all required. Do this if you have the interest or think such variables might be of value.

Quantitative Variables: For quantitative variables (integers, numeric) I constructed a table in R which provides information on measures of center, spread, range, skewness, kurtosis, and correlation 10 varaibles to sales price.

	Min	Q_1st	StDev	Median	Mean	Q_3rd	Max	Range	Skewness	Kurtosis	cv	Corr
SalePrice	35000	134500.0	78895.03	169500	188319.71	224371.5	591587	556587	1.46	2.81	0.42	1.00
LotArea	2500	8241.5	7504.90	9830	10861.91	11968.5	215245	212745	15.07	341.07	0.69	0.29
QualityIndex	1	30.0	8.97	35	34.69	40.0	90	89	0.41	2.76	0.26	0.53
HouseAge	0	7.0	30.38	36	36.52	55.0	136	136	0.55	-0.58	0.83	-0.59
GarageCars	0	1.0	0.74	2	1.81	2.0	5	5	-0.12	0.05	0.41	0.69
YrSold	2006	2007.0	1.30	2008	2007.80	2009.0	2010	4	0.13	-1.14	0.00	-0.03
TotalFloorSF	334	1141.0	492.11	1478	1517.55	1788.0	3820	3486	0.76	0.72	0.32	0.78
FullBath	0	1.0	0.55	2	1.56	2.0	3	3	0.16	-0.98	0.35	0.62

- Mean: This statistic is helpful in three ways:

    - It helps signal if the data generally conform to what we would expect given the nature of the data set. As this is housing data in Iowa, we have some sense what the average values should be (I would not expect the average home to be 10,000 sq ft).
    - The means may also be used to assess if the data are scaled differently than what we think (data entered as square meters instead of square feet)
    - It also helps familiarize us with the housing market in Ames. For example, on average houses sold in the period of this study were 36.5 years old.
- Standard Deviation: Standard deviation gives us a measure of spread for these variables. One helpful interpretative technique is to use the mean to convert the standard deviation to a coefficient of variation. (See to the right). Having done this, we can see that the variables often post COVs between 0.3-0.4, with some exceptions. Lot area's standard deviation represent is 69% of its mean, and house age's standard deviation is 83% of its mean. This tells us that those variables vary much more than other variables in the data set. Some variation is important in the modeling process, as long as that variation covaries with our response variable.
- Min/Max: Min and max help determine if there are any problems in the range of our data. I would not expect a negative value for any of these variables, and min confirms that the variable values are either zero or positive. A minimum value of 0 for full bath, garage cars, and total basement SF suggests that some properties sold in this period lacked these amenities (basements, garages, full bath). Max helps us see if there were any unusual large data entries.
- Correlation: A column of particular interest is SalePrice, which has reasonably high correlations with a number of the variables in the final 10 factors. Most of these correlations are positive, except for house age, which offers a notable negative correlation with SalePrice. Not surprisingly, TotalFloorSQ, number of cars in the garage (GarageCars), and number of full baths all boast strong positive correlations to SalePrice. The weakest predictor in this set is YrSold, which I thought would be stronger given the time period of the data (spanning some of the great recession).
    - It is also important to look at other relationships in the matrix below. House age is negatively correlated with a number of size dimensions. New homes are larger in overall size (sq ft), number of baths, and size of the garage. This opens up a possible line of inquiry: is it that the homes are old that impacts the selling price negatively, or is it that old homes are small? The scatter plot of size vs. age provides some insight, but perhaps if we combined size, number of baths, and garage into a single variable, the relationship would crystalize.
    - The correlation matrixt also reveals that homes that are larger are more likely to be higher quality, though that correlation (QualityIndex vs. Total SQ Ft) is not as strong.
    - The response variable in this problem is SalePrice as it best represents home value in our data set. When we plot SalePrice against TotalFloorSF (a likely predictor variable), an obvious cone shape emerges to the data cloud.



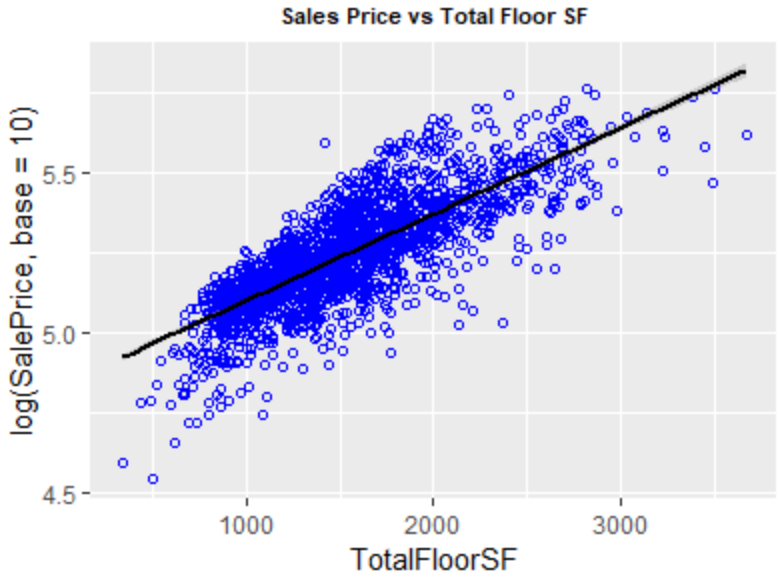
The amount of preparation and EDA is totally up to you. From prior experience, up to 90% of ones time modeling data is spent on data cleaning and preparation issues, depending on the type of data one is working with. Just remember: Garbage in → Garbage out! For this assignment, you want to be sure you have a dataset that you are comfortable working with for the remainder of this assignment. All of the statistics and graphs you produce are for you. They will be helpful as you go through the following steps, but you do not need to report anything here. There is nothing that needs to be written for task (0).

PART A: Simple Linear Regression Models

(1) Let Y = sale price be the dependent or response variable. Select "the best" continuous explanatory variable from the AMES data set to predict Y. What criteria did you use to select this variable? Fit a simple linear regression model using X to predict Y. Call this Model 1. You should:

a. Make a scatterplot of Y and X, and overlay the regression line on the cloud of data.
Before I select "the best" continuous explanatory variable, it is important to solidify my understanding of my response variable. The response variable in this problem is SalePrice as it best represents home value in our data set. When we plot SalePrice against TotalFloorSF (a likely predictor variable), an obvious cone shape emerges to the data cloud.

This cone shape is problematic for regression as it suggests that the assumptions for OLS will not be met. A transformation of the data may be necessary, and plotting the log of sale price vs. total floor square feet reveals an improved shape. Logarithmic scale corresponds to viewing variation through relative or percentage changes, rather than through absolute changes. For example, an increase in house price from, say, $40,000 to $44,000 is an increase of 10%, a substantial increase to the person selling the house, whereas an increase from $300,000 to $304,000, still $4,000 but only a 1.3% increase, is much less consequential.



So, when we build regression models, we assume that the response variable is Normally distributed with equal variances. If these assumptions are not true, then we need to take

corrective actions. Transformation is a technique that is used to transform a skewed variable to a 'symmetric' variable and also take care of the equal variance assumption. Transformation is an important part of the 'feature creation' process.

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

```
model1_lm <- lm(log(SalePrice, base = 10) ~ TotalFloorSF, data
= mydata_7)
Call:
lm(formula = log(SalePrice, base = 10) ~ TotalFloorSF, data = mydata_7)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44053 -0.05772  0.00589  0.06083  0.37863

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)  4.83156726 0.00719907   671.1 <0.0000000000000002
TotalFloorSF 0.00026997 0.00000453    59.6 <0.0000000000000002

Residual standard error: 0.1029 on 2164 degrees of freedom
Multiple R-squared:  0.6214,    Adjusted R-squared:  0.6213
F-statistic:  3552 on 1 and 2164 DF,  p-value: < 0.00000000000000022
```

Y_hat with log base 10 = 4.8316 + 0.00027 * TotalFloorSF + error
The y_intercept (b0) is 4.8316 and the slope (b1) is 0.00027.

The coefficient of total floor square feet has a percentage interpretation when it is multiplied by 100. The sale price increases by 0.027% for every additional square feet of total floor and the p-value verifies statistical significance. The intercept is not very meaningful, because it gives the predicted log10(SalesPrice) when TotalFloorSF = 0.

c. Report and interpret the R-squared value in the context of this problem.
The R-squared shows that TotalFloorSF explains about 62.1% of varaition in log(salesPrice)

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

```
Analysis of Variance Table

Response: log(SalePrice, base = 10)
              Df Sum Sq Mean Sq F value                Pr(>F)
TotalFloorSF    1 37.631  37.631  3552.5 < 0.00000000000000022
Residuals    2164 22.923   0.011
```

We can specify the hypotheses associated with each coefficient of the model by looking at the summary table of 1.b.

t-critical value at alpha 2.5% (two tailed) with more than df = 200 is +/-1.9608
H0: $B_1 = 0$; TotalFloorSF is not a significant predictor
Ha: $B_1 \neq 0$; TotalFloorSF is a significant predictor
t-value = 0.00026997 / 0.00000453 = 59.5960

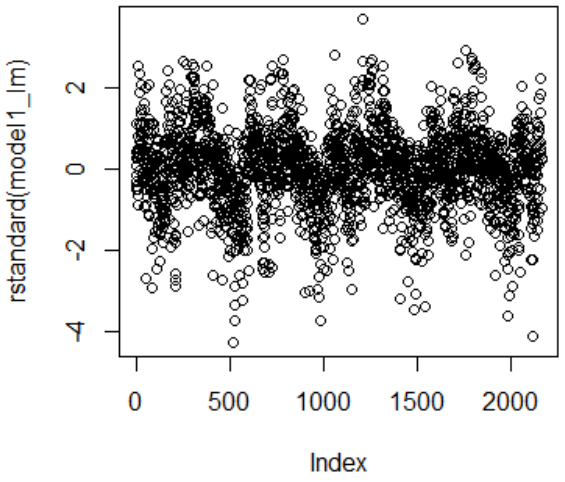
t-value > t-critical and we can also see that R generated p-value is very small.
With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the TotalFloorSF is significantly helping us in predicting the sales price of homes.

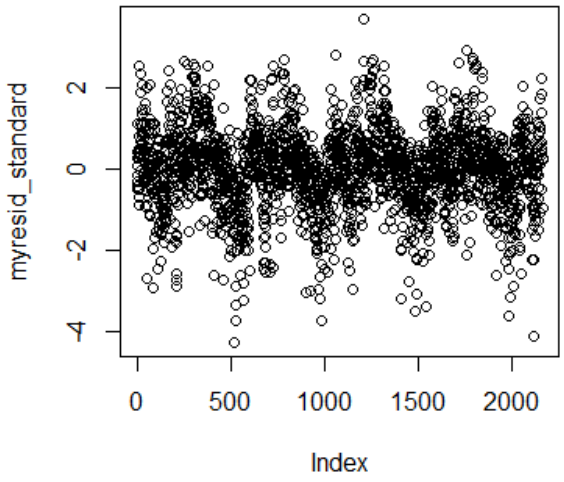
Since we have one predictor, we can skip the overall hypothesis test because we had already specified the hypothesis associated with that coefficient of the model. However, by looking at the ANOVA statistics, the high F value and very low p-value suggests that we reject null hypotheses and conclude that this model is a significant and there is at least one predictor correlates to response variable Y.

e. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. To assess this, use the model from part a) to calculate predicted values for each record. Then use the predicted values to compute residuals. Yes, many of the packages automatically give you the predicted and residuals, but you should know how to code and compute these values. Next standardize the residuals but subtracting off the mean and dividing by the standard deviation for each residual (i.e. you will have to obtain those summary statistics first).

```
myresid <- predict(model1_lm) - log(mydata_7$SalePrice, base = 10
)
myresid_standard <-(mean(myresid)-myresid)/sd(myresid)
```

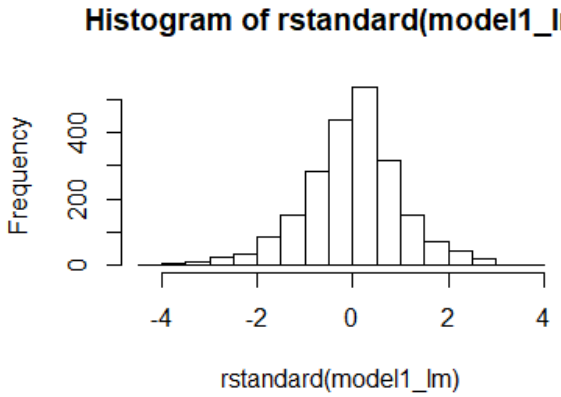
myresid_standard is derived standardized residuals by following the above instructions and rstandard(model1_lm) is computed standardized residuals by R. When I plot them next to each other, they look the same.

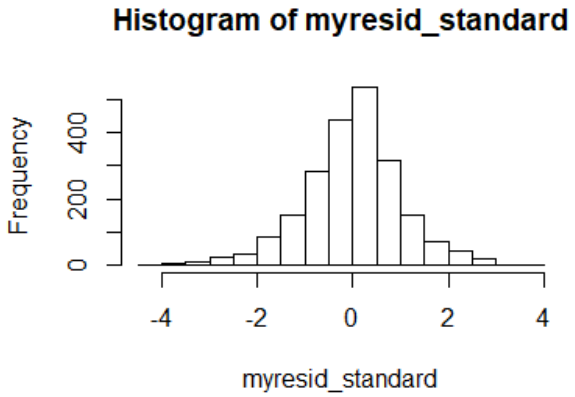




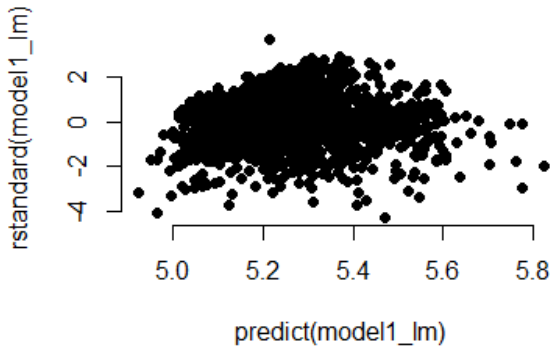
Check on the underlying assumptions by plotting:

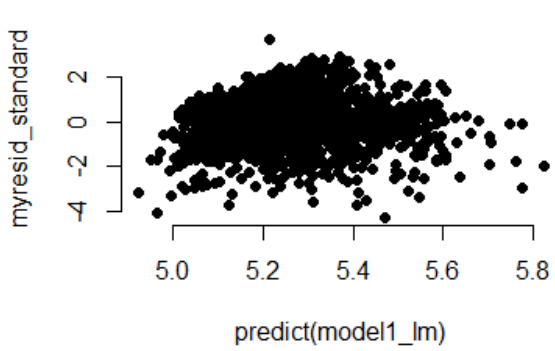
- Histogram of the standardized residuals





- Scatterplot of standardized residuals (Y) by predicted values (X)





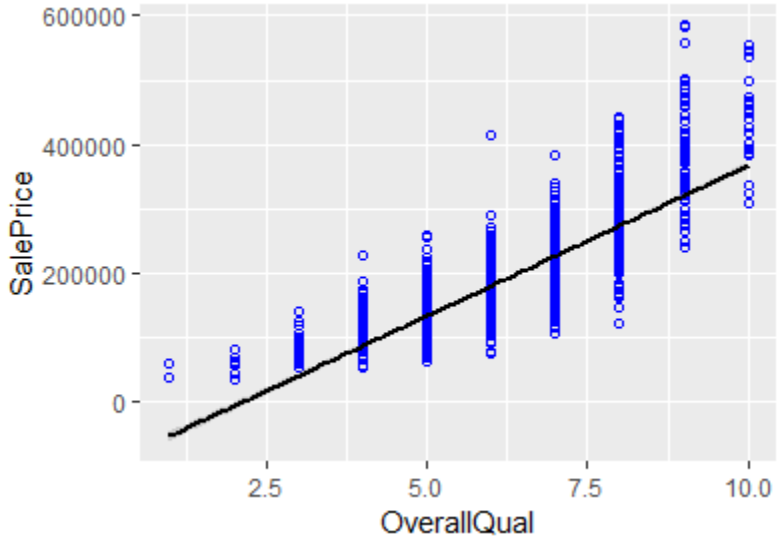
Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity. Do there appear to be outliers or influential points?

After the log transformation of SalesPrice in part 1.a, the residuals are somewhat distributed normally in histogram and scatterplots. If we had not log transform the SalesPrice, we would have seen that the residuals would be distributed as the cone or funnel shapes.

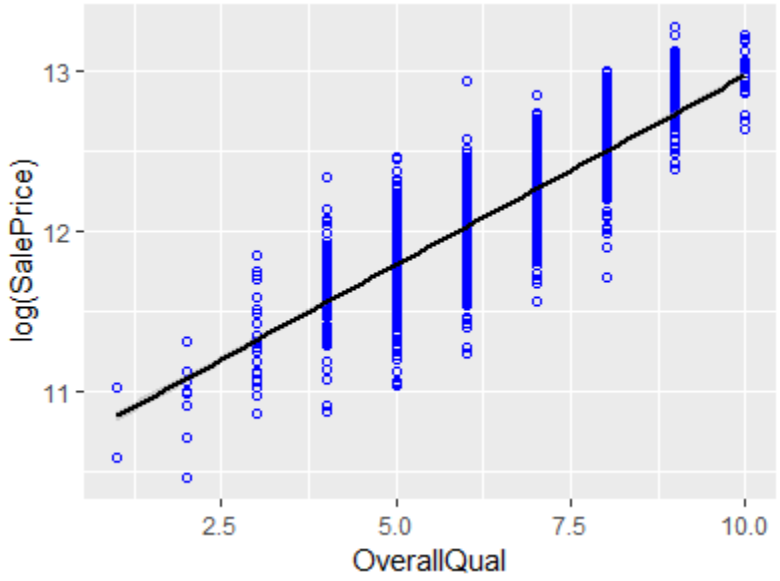
(2) Let Y = sale price be the dependent or response variable. Use the OVERALL QUALITY variable as the explanatory variable (X) to predict Y. Fit a simple linear regression model using X to predict Y. Call this Model 2. You should:

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

When we plot SalePrice against OverallQual, a cone shape emerges to the data cloud.



This cone shape is problematic for regression as it suggests that the assumptions for OLS will not be met. A transformation of the data may be necessary and plotting the log of sale price vs. overall quality reveals an improved shape.



b. Report the model in equation form and interpret each coefficient of the model in the context of this problem. Is there anything different about the interpretation of coefficients here as opposed to those of Model 1? Can you say a 1 unit change in X is the same across all possible values of X?

```
model2_lm <- lm(log(SalePrice, base = 10) ~ OverallQual, data
= mydata_7)
```

```
Call:
lm(formula = log(SalePrice, base = 10) ~ OverallQual, data = mydata_7)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34324 -0.05349  0.00174  0.05354  0.39430

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 4.606202   0.008694  529.80 <0.0000000000000002
OverallQual 0.102924   0.001378   74.69 <0.0000000000000002

Residual standard error: 0.08843 on 2164 degrees of freedom
Multiple R-squared:  0.7205,    Adjusted R-squared:  0.7204
F-statistic:  5579 on 1 and 2164 DF,  p-value: < 0.00000000000000022
```

log10 of Y_hat = 4.6062 + 0.1029 * OverallQual + error
The y_intercept (b0) is 4.6062 and the slope (b1) is 0.1029.

The coefficient of overall quality has a percentage interpretation when it is multiplied by 100. The sale price increases by 10.29% for every additional overall material and finish quality score and the p-value verifies statistical significance. The intercept is not very meaningful in here when the overall quality is absent.

c. Report and interpret the R-squared value in the context of this problem.
The R-squared shows that OverallQual explains about 72.1% of varaition in log(SalesPrice).

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

```
Analysis of Variance Table

Response: log(SalePrice, base = 10)
              Df Sum Sq Mean Sq F value                Pr(>F)
OverallQual    1 43.630  43.630  5579.2 < 0.00000000000000022
Residuals   2164 16.923   0.008
```

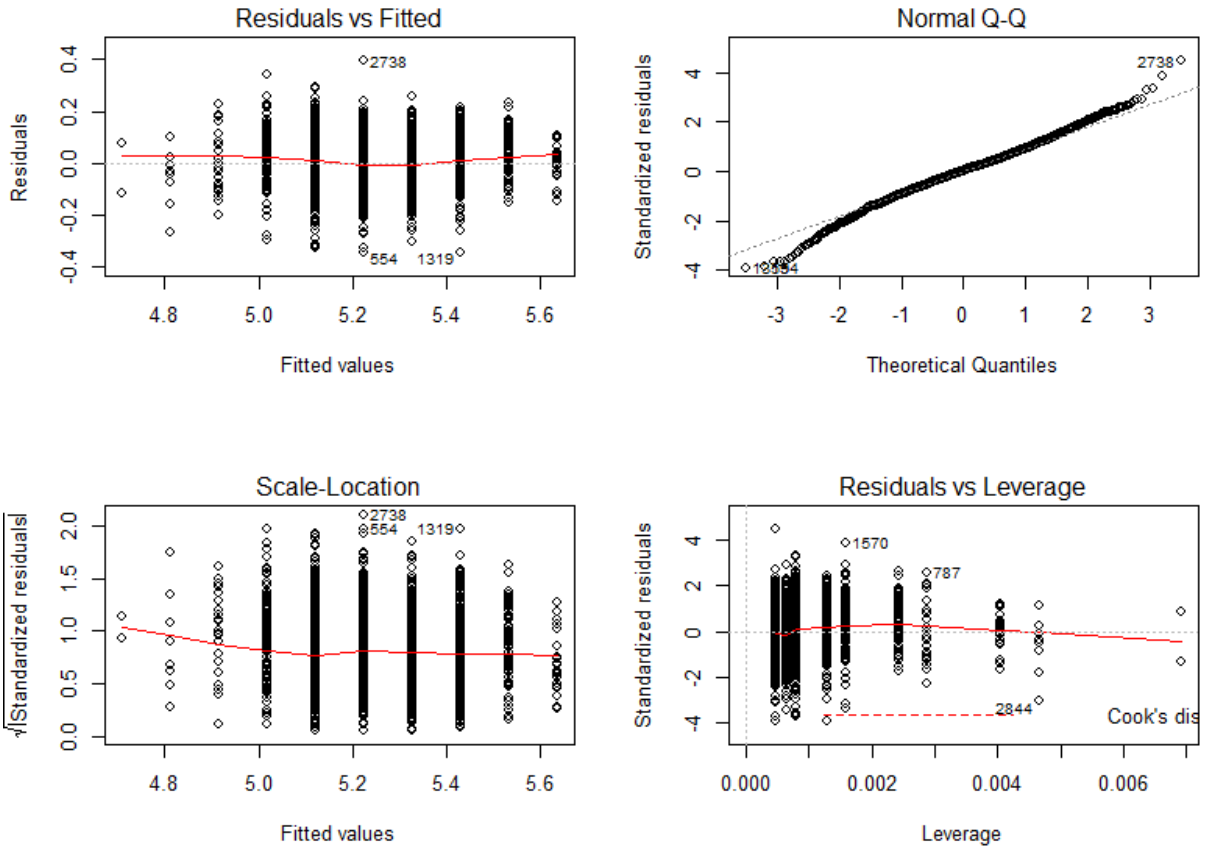
We can specify the hypotheses associated with each coefficient of the model by looking at the summary table of 2.b.

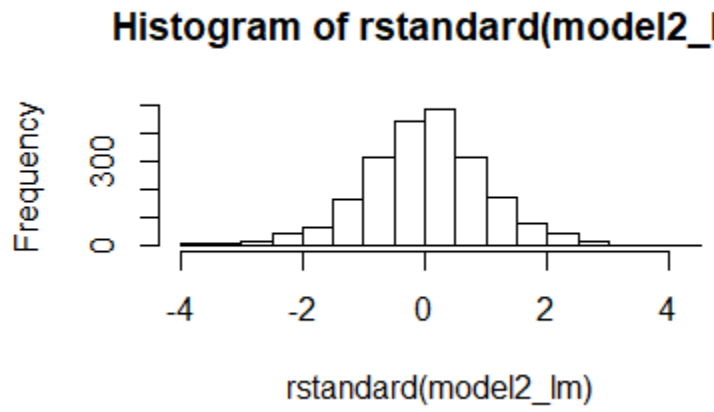
t-critical value at alpha 2.5% (two tailed) with more than df = 200 is +/-1.9608
H0: $B_1 = 0$; OverallQual is not a significant predictor
Ha: $B_1 \neq 0$; OverallQual is a significant predictor
t-value = 0.102924 / 0.001378 = 74.6909
t-value > t-critical and we can also see that R generated p-value is very small.
With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the overall material and finish quality score is significantly helping us in predicting the sales price of homes.

Since we have one predictor, we can skip the overall hypothesis test because we had already specified the hypothesis associated with that coefficient of the model. However, by looking at the ANOVA statistics, the high F value and very low p-value suggests that

we reject null hypotheses and conclude that this model is a significant and there is at least one predictor correlates to response variable Y.

e. Check on the underlying assumptions. You can do this by hand, or use the provided results from one of the regression package functions, like lessR or CAR. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity. Do there appear to be outliers or influential points?





After the log transformation of OverallQual in part 2.a, the residuals are somewhat distributed normally in histogram and scatterplots. If we had not log transform the

OverallQual, we would have seen that the residuals would be distributed as the cone or funnel shapes.

f. Of the above 2 models, which one fits better? On what criteria are you assessing the model fit?

	R-squared	Sum of Square Residuals	Correlation to SalesPrice
Model 1; x = TotalFloorSF	0.6214	22.923	0.779
Model 2; x = OverallQual	0.7205	16.923	0.8274

R-squared value of the model 2 is higher which is better than model 1. In addition, sum of square errors of model 2 is lower than the model 1. Predictor of Model 2 which is OverallQual more correlate than TotalFloorSF. So, I would choose model 2 over Model 1.

PART B: Multiple Linear Regression Models

(3) Fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). These two explanatory(X) variables should be: the explanatory variables from Model 1 and Model 2 above. Call this Model 3. You should:

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 relative to the simple linear regression models above?

log10 of Y_hat = 4.5936 + 0.0706 * OverallQual + 0.00014 * TotalFloorSF + error

```
Call:
lm(formula = log(SalePrice, base = 10) ~ OverallQual + TotalFloorSF,
    data = mydata_7)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37326 -0.04136  0.00436  0.04613  0.25601

Coefficients:
                Estimate  Std. Error t value            Pr(>|t|)
(Intercept)  4.593596665 0.007061210  650.54 <0.0000000000000002
OverallQual  0.070622699 0.001474730   47.89 <0.0000000000000002
TotalFloorSF 0.000139824 0.000004165   33.57 <0.0000000000000002

Residual standard error: 0.07172 on 2163 degrees of freedom
Multiple R-squared:  0.8163,    Adjusted R-squared:  0.8161
F-statistic:  4804 on 2 and 2163 DF,  p-value: < 0.00000000000000022
```

We include OverallQual (overall material and finish quality score) and TotalFloorSF (total floor sqaure footage) in an equation explaining log(SalesPrice).

As in the simple regression case, the coefficients have a percentage interpretation. The only difference in here is that they also have the same interpretation.

The coefficient 0.0706 means that, holding TotalFloorSF variable fixed, the sale price increases by 7.06% for every additional overall material and finish quality score and the p-value verifies statistical significance.

The coefficient 0.00014 means that, holding OverallQual varaible fixed, the sale price increases by 0.01% for every additional square feet of total floor and the p-value verifies statistical significance.

The intercept is not very meaningful in here when the overall quality and total floor square feet are absent.

b. Report and interpret the R-squared value in the context of this problem. Does this multiple linear regression model fit better than the simple linear regression models? How do you know? Calculate the difference between R-squared for Model 3 and R-squared for Model 1. How would you interpret this difference?

The R-squared shows that OverallQual and TotalFloorSF explain about 81.63% of varaition in log(SalesPrice). This multiple linear regression model fit better than the simple linear regression model because of higher R-squared value. Also, the difference between R-squared for Model 3 and R-squared for Model 1 is 0.8163 – 0.6214 = 0.1949. Adding this second variable to the model really improved the explanatory value of the model. This, Model 3 can explain an additional 19.49% of the data.

c. Report the coefficient and ANOVA Tables. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret the hypothesis tests.

```
Analysis of Variance Table

Response: log(SalePrice, base = 10)
              Df Sum Sq Mean Sq F value                 Pr(>F)
OverallQual    1 43.630  43.630  8482.0 < 0.00000000000000022
TotalFloorSF   1  5.797   5.797  1126.9 < 0.00000000000000022
```

We can specify the hypotheses associated with each coefficient of the model by looking at the summary table of 3.b.

t-critical value at alpha 2.5% (two tailed) with more than df = 200 is +/-1.9608

H0: $B_1 = 0$; OverallQual is not a significant predictor

Ha: $B_1 \neq 0$; OverallQual is a significant predictor

t-value = 0.070622699 / 0.001474730 = 47.8886

t-value > t-critical and we can also see that R generated p-value is very small.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the overall material and finish quality score is significantly helping us in predicting the sales price of homes.

H0: $B_2 = 0$; TotalFloorSF is not a significant predictor

Ha: $B_2 \neq 0$; TotalFloorSF is a significant predictor

t-value = 0.000139824 / 0.000004165 = 33.5712

t-value > t-critical and we can also see that R generated p-value is very small.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the total floor square feet is significantly helping us in predicting the sales price of homes.

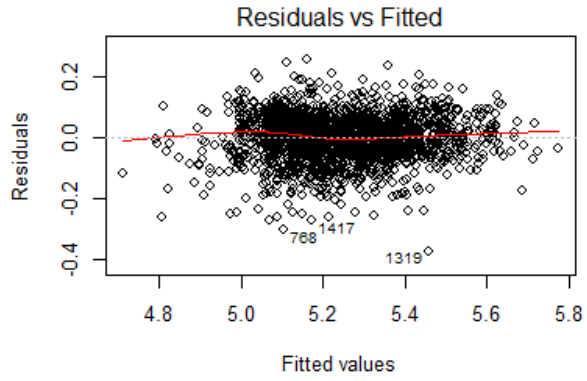
H0: $B_1 = B_2 = 0$; there is no relationship between predictors and a response Y varaible

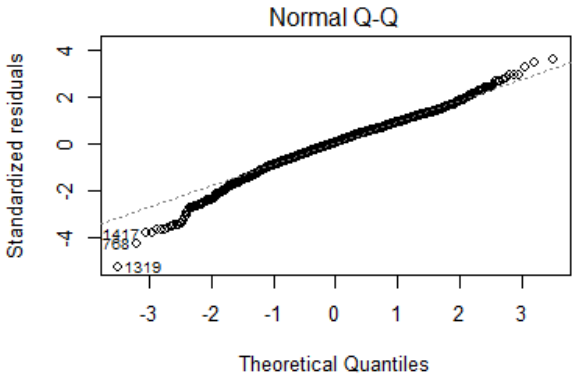
Ha: there is at least 1 inequality

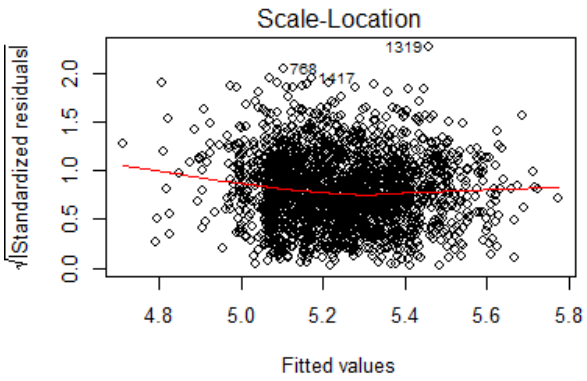
F-statistics = MS regression / MS residual = 4804

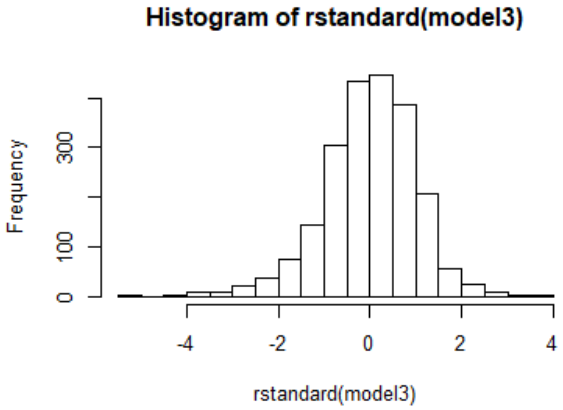
When there is a high F-statitics, there will be very small p-value. We can see that the F-statistics is high which we reject null hypotheses and conclude that this model is a significant and there are at least one predictor correlates to response variable Y.

d. Check on the underlying assumptions. You can do this by hand, or use the provided results from one of the regression package functions, like lessR or CAR. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity. Do there appear to be outliers or points of concern?









If we look at the residuals vs. fitted values, the data satisfy all multiple regression assumptions because we can see that model errors are independent with mean = 0 and it has homogeneous variance, a random scatter around the horizontal line at residual = 0 and there is no any systematic trends. There are very few outliers. Histogram and QQ plot show residuals are somewhat distributed normally but light tailed. We can observe a couple of outliers in the left tail.

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

	Model 1	Model 2
R-Squared of Model 3	0.8163	0.8163
R-Squared of Model 1 and 2	0.6214	0.7205
Change	19%	10%

By looking at the change in R-squared values, both varaibles are good predictors. This multiple linear regression model fit better than the simple linear regression model because of higher R-squared value. Also, the difference between R-squared for Model 3 and R-squared for Model 1 is 0.8163 – 0.6214 = 0.1949. Adding this second variable to the model really improved the explanatory value of the model. This, Model 3 can explain an additional 19.49% of the data. Thus, I can retain both variables as predictor variables of Y.

(4) Select any other continuous 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 your variable of choice plus the explanatory variables from Model 3. Call this Model 4. You should:

a. Report Model 4 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 relative to the simple linear regression models above?

log10 of Y_hat = 4.7641 + 0.0494 * OverallQual + 0.00015 * TotalFloorSF - 0.00146 * HouseAge + error

```
Call:
lm(formula = log(SalePrice, base = 10) ~ OverallQual + TotalFloorSF +
    HouseAge, data = mydata_7)

Residuals:
      Min        1Q    Median        3Q       Max
-0.280266 -0.039624  0.000517  0.039325  0.277387

Coefficients:
                Estimate   Std. Error t value            Pr(>|t|)
(Intercept)   4.764144329  0.008881691  536.40 <0.0000000000000002
OverallQual   0.049376867  0.001510383   32.69 <0.0000000000000002
TotalFloorSF  0.000148744  0.000003634   40.94 <0.0000000000000002
HouseAge     -0.001459355  0.000054968  -26.55 <0.0000000000000002

Residual standard error: 0.0623 on 2162 degrees of freedom
Multiple R-squared:  0.8614,    Adjusted R-squared:  0.8612
F-statistic:  4480 on 3 and 2162 DF,  p-value: < 0.00000000000000022
```

We include OverallQual (overall material and finish quality score), TotalFloorSF (total floor sqaure footage), and HouseAge (age of the house or difference between built and sold) in an equation explaining log(SalesPrice).

The coefficient 0.04938 means that, holding other variables fixed, the sale price increases by 4.938% for every additional overall material and finish quality score and the p-value verifies statistical significance.

The coefficient 0.00015 means that, holding other varaibles fixed, the sale price increases by 0.015% for every additional square feet of total floor and the p-value verifies statistical significance.

The coefficient -0.00146 means that, holding other varaibles fixed, the sale price decreases by 0. 146% for every additional year since it was built (or the house ages) and the p-value verifies statistical significance.

The intercept is not very meaningful in here when the overall quality, house age, and total floor square feet are absent.

b. Report and interpret the R-squared value in the context of this problem. Does this multiple linear regression model fit better than the simple linear regression models? How do you

know? Calculate the difference between R-squared for Model 4 and R-squared for Model 3. How would you interpret this difference? Does your variable of choice help to improve the model's explanatory ability?

The R-squared shows that OverallQual, TotalFloorSF, HouseAge explain about 86.14% of varaition in log(SalesPrice). This multiple linear regression model fit better than the previous models because of higher R-squared value. Also, the difference between R-squared for Model 4 and R-squared for Model 3 is 0.8614 - 0.8163 = 0.0451. Adding the third variable to the model somewhat improved the explanatory value of the model. This, Model 4 can explain an additional 4.51% of the data.

c. Report the coefficient and ANOVA Tables. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret the hypothesis tests.

Analysis of Variance Table

Response: log(SalePrice, base = 10)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
OverallQual	1	43.630	43.630	11242.15	< 0.00000000000000022
TotalFloorSF	1	5.797	5.797	1493.62	< 0.00000000000000022
HouseAge	1	2.736	2.736	704.86	< 0.00000000000000022
Residuals	2162	8.391	0.004		

We can specify the hypotheses associated with each coefficient of the model by looking at the summary table of 4.b.

t-critical value at alpha 2.5% (two tailed) with more than df = 200 is +/-1.9608

H0: $B_1 = 0$; OverallQual is not a significant predictor

Ha: $B_1 \neq 0$; OverallQual is a significant predictor

t-value = 0 0.049376867 / 0.001510383 = 32.69162

t-value > t-critical and we can also see that R generated p-value is very small.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the overall material and finish quality score is significantly helping us in predicting the sales price of homes.

H0: $B_2 = 0$; TotalFloorSF is not a significant predictor

Ha: $B_2 \neq 0$; TotalFloorSF is a significant predictor

t-value = 0.000148744 / 0.000003634 = 40.93121

t-value > t-critical and we can also see that R generated p-value is very small.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the total floor square feet is significantly helping us in predicting the sales price of homes.

H0: $B_3 = 0$; HouseAge is not a significant predictor

Ha: $B_3 \neq 0$; HouseAge is a significant predictor

t-value = -0.001459355 / 0.000054968 = -26.54917

t-value > t-critical and we can also see that R generated p-value is very small.

With 95% confidence (two-tail test), we reject the null hypothesis and conclude that the house age is significantly helping us in predicting the sales price of homes.

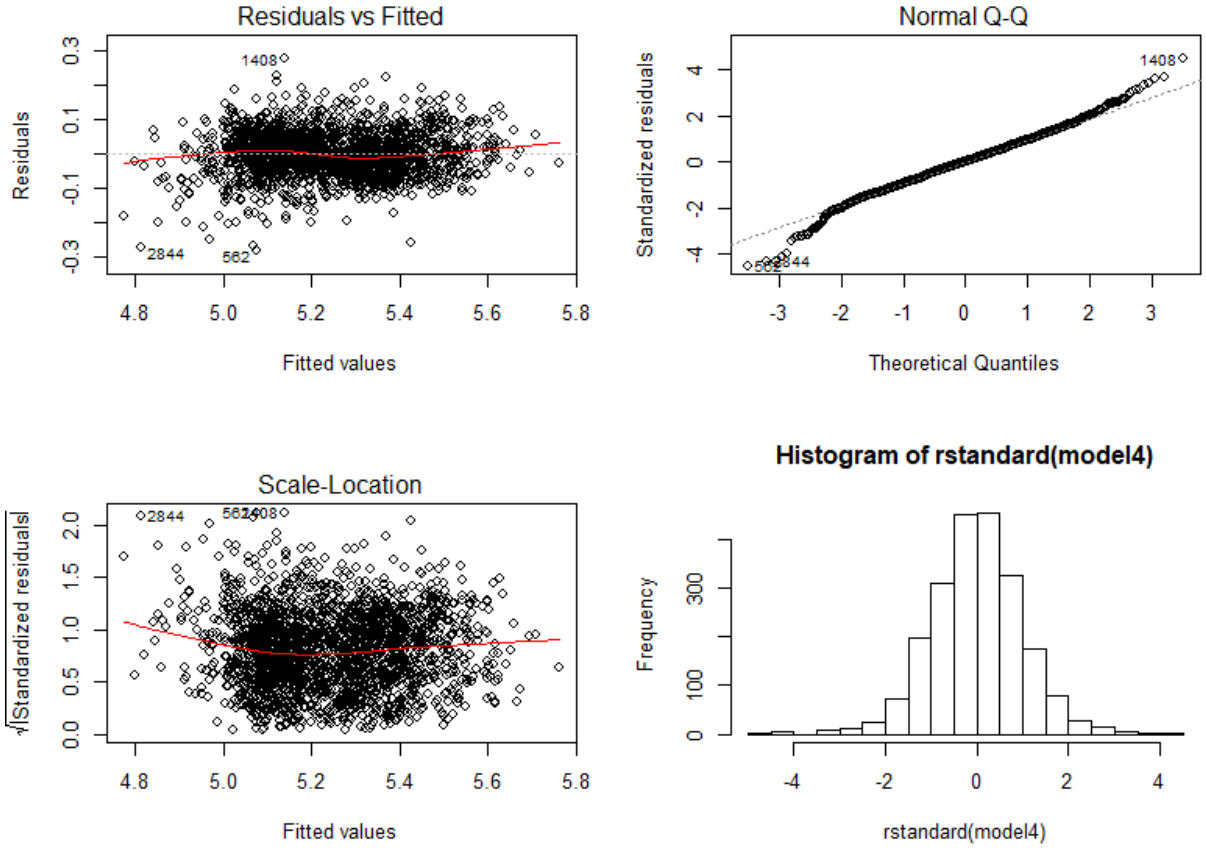
H0: $B_1 = B_2 = B_3 = 0$; there is no relationship between predictors and a response Y varaible

Ha: there is at least 1 inequality

F-statistics = MS regression / MS residual = 4480

When there is a high F-statitics, there will be very small p-value. We can see that the F-statistics is high which we reject null hypotheses and conclude that this model is a significant and there are at least one predictor correlates to response variable Y.

d. Check on the underlying assumptions. You can do this by hand or use the provided results from one of the regression package functions, like lessR or CAR. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity. Do there appear to be outliers or points of concern?



If we look at the residuals vs. fitted values, the data satisfy all multiple regression assumptions because we can see that model errors are independent with mean = 0 and it has homogeneous variance, a random scatter around the horizontal line at residual = 0 and there is no any systematic trends. There are very few outliers. Histogram and QQ plot show residuals are somewhat distributed normally but light tailed. We can observe a couple of outliers in the left tail and an outlier in a right tail.

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

By looking at the change in R-squared values, 3rd varaibles is small at 4.51%, but it is still a positive addition and can be considered a good predictor. Adding this third variable to the model improved the explanatory value of the model by 4.51%. Thus, I can also retain the third variable as a predictor variable of Y.

PART C: Multiple Linear Regression Models on Transformed Response Variable

(5) Refit Model 1, Model 3 and Model 4 using the Natural Log of SALEPRICE as the response variable. 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 models to the original models. Which transformed model fits the best? Do the transformed models fit better than the original models? You do not need to report all of the output like was done in Parts A and B. Rather, you should construct a table to summarize your findings so that the comparisons can be made easily. What is the best way or statistic to use, to make comparisons between models? You may need more than one table to do this adequately, if you have more than 1 criteria.

(6) How is the interpretation of the LN(SalePrice) models different from the SalePrice models? Discuss if the improvement of model fit justifies the use of the Log(SALEPRRICE) response variable, relative to interpretation and explanation to a non-technical audience, like your manager or other executives.

PART D: Multiple Linear Regression and Influential Points

(7) Use Model 4 for this part. Even after you have cleaned your data, still you may have unusually large residuals, which you can see from the residual plots. These are called 'influential' points. Sometimes, we find that a small subset of 'influential' points exerts a disproportionate influence on the model coefficients. These points can be identified by several statistics such as DFFITS, Cook's Distance, Leverage, and Influence. Fit Model 4 using a regression function from one of the comprehensive regression packages (like lessR). Obtain output data with these statistics (DFFITS, etc.) for individual records so that you can identify the influential points. Use the threshold value given in the text book (Like that on Page 112 of Chatterjee and Hadi). 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? The other side of the coin is that if you remove data points due to them being "influential" and not looking like you might want them to look, some would argue that such an action is the modeler biasing the data. 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.

PART E: Beginning to Think About a Final Model

(8) Use Model 4 to start with for this part. So far, we have fit a few models to predict SALEPRICE(Y). But, there are many other continuous variables in the data set. You could use theory, or your background knowledge, to select variables for inclusion in a multiple regression model. Many

modelers do this. It gives a nice place to start the search process. On the technical side, in this assignment, we have been looking at change in R-squared when a new variable has been added to an existing model to isolate the explanatory contribution of that new variable. And, we have been looking at hypothesis tests on the individual coefficients.

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
b. Report the model you determined and interpret the coefficients
c. Report the coefficient and ANOVA tables.
d. Report goodness of fit
e. Check on underlying model assumptions.

CONCLUSION / REFLECTION

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?

Assignment Document:

Results should be presented and discussed in an organized manner, preferably listed by task number and letter. The report should not contain unnecessary results or information. The document should be submitted in pdf format. Name your file Assignment2_LastName.pdf.