Assignment #2

Syamala Srinivasan

Introduction:

Exploratory data analysis is utilized to understand the data set to be analyzed as well as ensuring that the data set is clean of any errors and contains the correct desired sample population. The output of the exploratory data analysis includes a clean data set to conduct further trending on, as well as basic and high level trends and hypothesizes. This assignment is the next step after exploratory data analysis. In this assignment the results of the exploratory data analysis will be utilized to determine which variables are the most valuable to conduct simple and multiple regression analysis with. The results of these models will also be utilized to attempt to predict. As with any model, the models created in this assignment will need to be tested and evaluated to ensure that they are sufficient and correct.

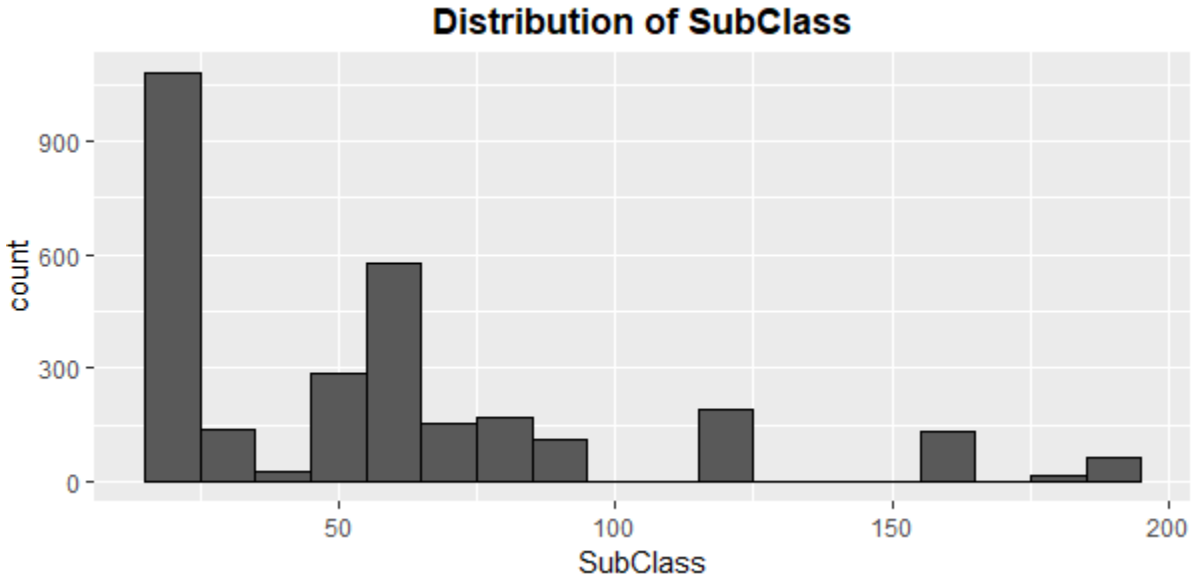
Results:

Section 1: Sample Definition

Task 1:

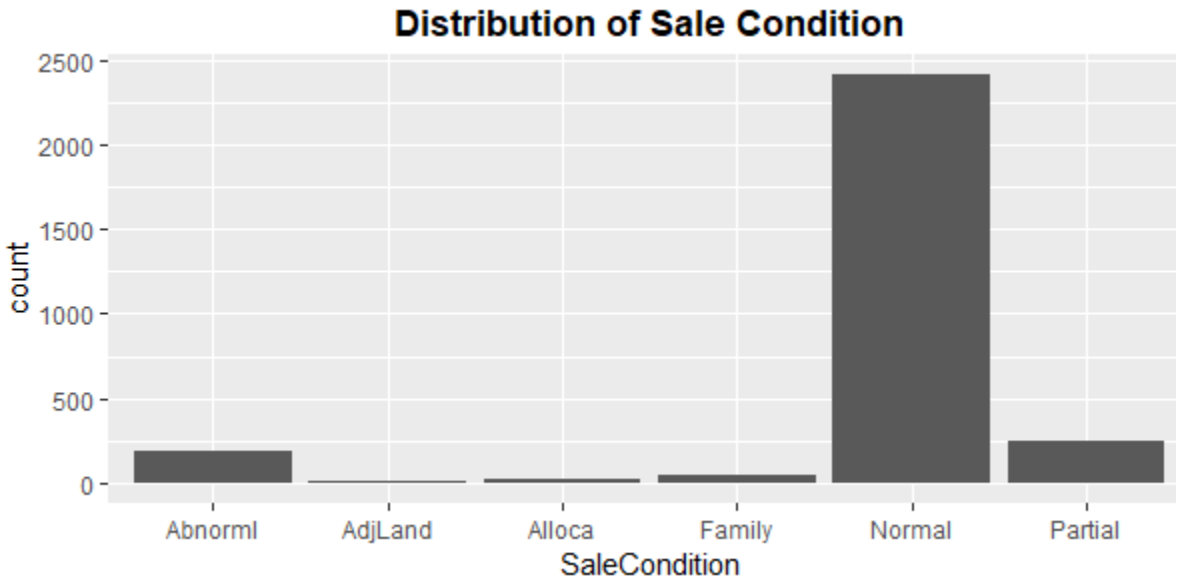
The objective of this task is to provide estimates of home values for typical homes in Ames. A "typical" home in Ames would logically only include homes. Essentially this definition of "home" excludes buildings identified as "PUDs", Duplexes, Agriculture, Commercial lots, etc. The removal of these populations ensures that future trends and analysis are comparing similar sample populations. I have also indicated to remove many of the sale conditions due to the affect that these may have on the price that is not related to the home itself.

The first variable dropped from the analysis will be SubClass.

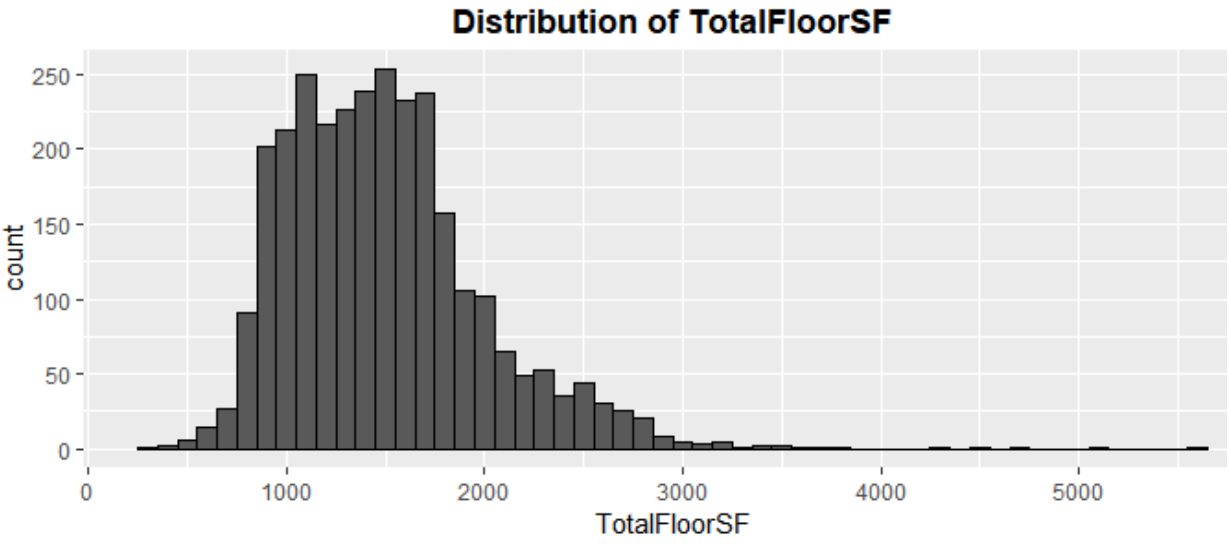


Subclasses between 90 and 180 will be dropped as these sub classes indicate homes that are not "typical" Ames homes. These buildings consist of duplexes and apartments. Since these homes will be significantly different to the typical Ames home and they were sold to buyers with different expectations and intentions, these homes will be excluded from the analysis.

The second variable for dropping consideration will be the condition of the sale taking place.



For the purpose of this analysis, only normal sales will be taken into consideration. The other types of sales that are being excluded include foreclosures, short sales, sales between family members and incomplete homes. Any of these factors will significantly affect the price of the home, typically in a negative way.



Lastly, only five homes have a total square footage of greater than 4,000. Therefore, these outliers will be removed as to reflect more accurately what a "typical" home in Ames is. Since

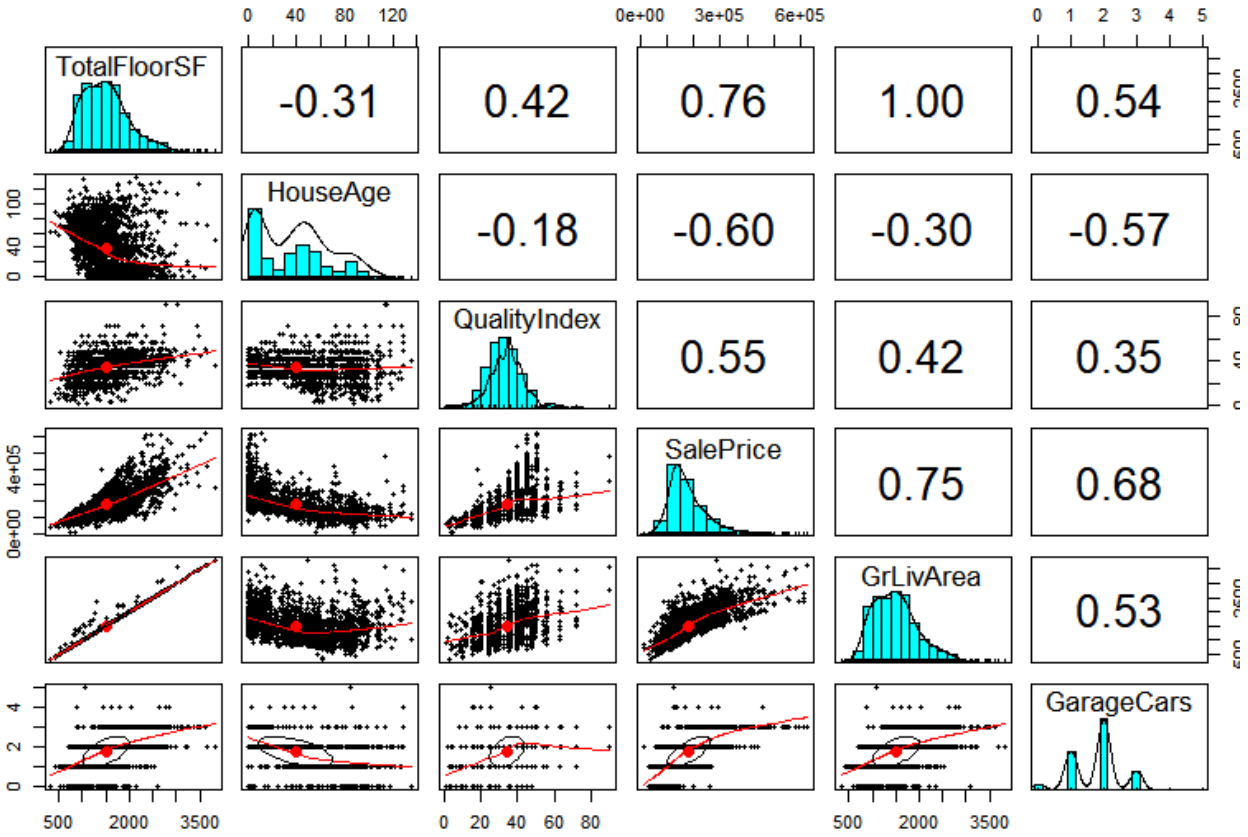
these values are well over 2,500 square feet greater than the mean these are viewed as not typical.

Variable Name	Populations Dropping	Drop Count	Cumulative Drop
SubClass	90,120,150,160,180	448	448
SaleCondition	Abnorma, AdjLand, Alloca, Family, Partial	517	965
Total Square Feet	>4000	5	970

Section 2: Exploratory Data Analysis

Task 2:

The below graphic compares 5 numeric variables to the sales price. These variables include total floor space, age of the home, quality index and the above ground living area. When viewing these variables there are a few things that are quickly evident. Such as, there are a few variables that are skewed left including total square feet, total living square feet and price. Quality Index is almost normally distributed while House Age is relatively flat. This could lead us to an initial assumption that the size of the home may be a good predictor of price. This is confirmed when we look at the correlation coefficients. The largest coefficients for Sale Price are for the total floor space and total living space although neither of the other variables are too far behind.

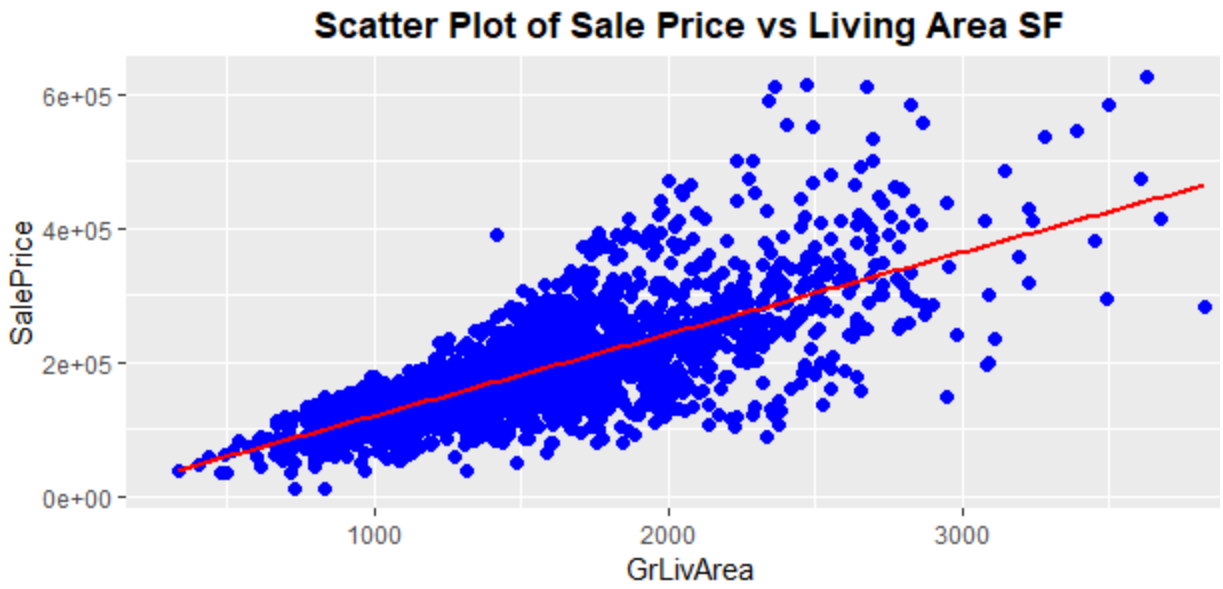


The variable with the next largest correlation coefficient with Sale Price is the capacity of the garage. The scatterplot of garage capacity and sale price (fourth graph on the bottom row) illustrates a steep curve in price with the increase of capacity. Most of the independent variables do not show much of a relationship. The largest exception is the total floor space and living area above ground which has a perfect linear relationship, this however makes sense considering these variables are extremely similar and may be the exact same number for some houses. The only other variables that show a moderate relation to each other is the age of the home and the number of cars. The moderate negative relationship indicates that the younger the home is the more cars it will have. This makes perfect since as some of the homes on this list are over 100 years old.

Of the above five variables, GrLivArea will be selected to continue this analysis as well as Garage Capacity.

Section 3: Simple Linear Regression Models

Section 3.1: Model #1 (Above Ground Living Area)

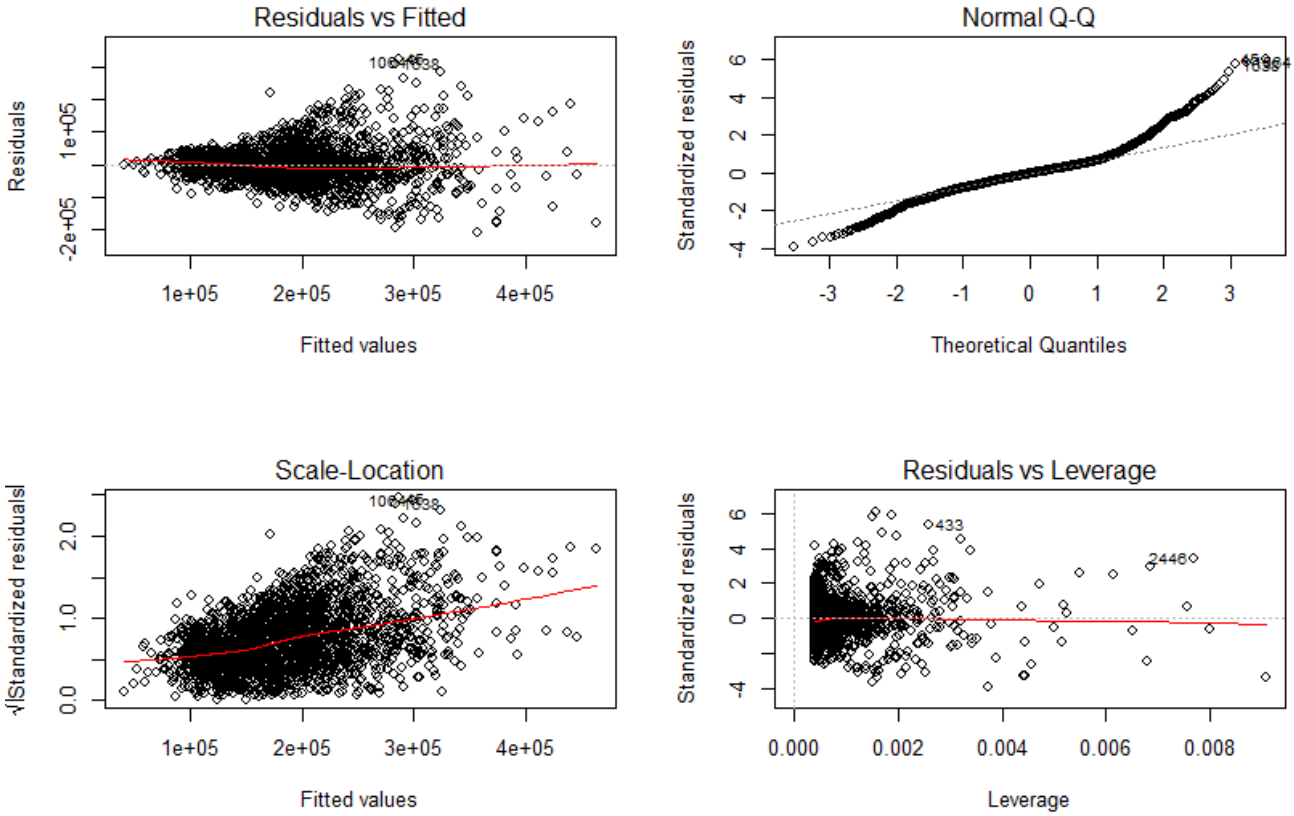


When viewing the scatter plot of the living room area by sale price, the positive linear trend becomes apparent. We can determine due to the small p value that the model is significant and that we should reject the null hypothesis.

```
Residuals:
    Min      1Q  Median      3Q     Max
-207782  -29000   -1235   22044  324534

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -873.537   3432.423  -0.254    0.799
GrLivArea    121.826      2.158  56.453   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53550 on 2464 degrees of freedom
Multiple R-squared:  0.564,     Adjusted R-squared:  0.5638
F-statistic:  3187 on 1 and 2464 DF,  p-value: < 2.2e-16
```

We can also use the information to determine the model equation. This equation ends up being Y=-874 + 121*x1. Therefore, for every additional square foot of living space, the price goes up by $121. Our standard error for this variable is roughly over $53,000 with 56% of variation explained by the model of total living area, this amount represents the average distance from the regression line.

The next step of our regression would be to view the residuals. These four graphs help to identify potential problematic cases in the data set.



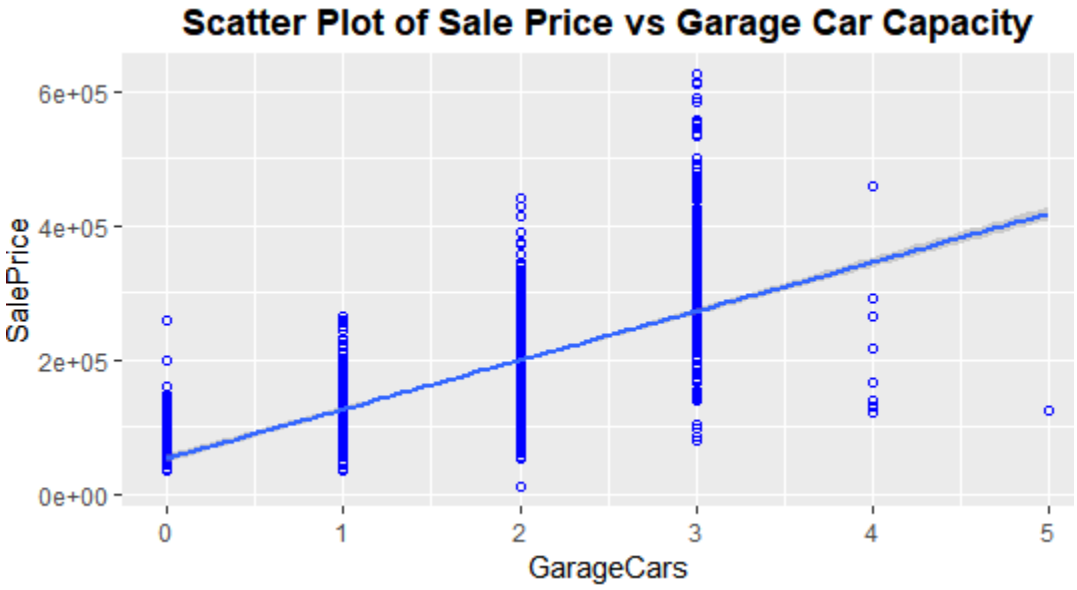
First looking at the residuals versus fits plot to verify the assumption that the residuals are randomly distributed, we can tell that this is not the case due to the increasing trend of variables. There are a few outliers on this graph but not too many, our options would be to transform the response variable. We can also see utilizing the QQ plot that our data is not normally distributed as it strays far from the line. The scale location graph assists us with identifying if the residuals are equally spread along the ranges of predictors. The values appear to be equally and randomly spread as the line is almost horizonal. It is a little concerning that the residuals spread wider towards the right, however the slope of the line is not too steep. The Residuals vs Leverage graph helps to identify if there are any influential cases in our data set. Our graph has a few outliers. SID 45, 1638 and 1064 are appearing on the residual vs fitted graph as outliers because their square feet are 848, 822 and 954 square feet larger than the mean respectively. Removing these outliers adjusts our R squared to .5396, this is actually smaller than the original R squared so removing these outliers is not a method to improve the model. The other outliers I would like to take a look at are SID 433, 1498 and 2446. These are the outliers on the Residuals vs Leverage map which indicates they may be influential to the R squared. These values are significantly larger than the mean at 1,158, 2,304 and 2,111 square feet above the mean. Removing these outliers does not adjust the R squared value at all which indicates that these outliers may not be influential.

Lastly, we can attempt to predict the sale price of the home utilizing the total living space above ground.

	TotalFloorSF	HouseAge	QualityIndex	SalePrice	GrLivArea	GarageCars	fitSLR	fit	lwr	upr
1	1656	50	30	215000	1656	2	200870.1	200870.1	95848.837	305891.5
2	896	49	30	105000	896	1	108282.5	108282.5	3230.834	213334.1
3	1329	52	36	172000	1329	1	161033.1	161033.1	56010.792	266055.4
4	2110	42	35	244000	2110	2	256179.1	256179.1	151128.924	361229.3
5	1629	13	25	189900	1629	2	197580.8	197580.8	92560.148	302601.5
6	1604	12	36	195500	1604	2	194535.2	194535.2	89514.954	299555.4

Utilizing the above predictions we can see that the range between the lower and the upper values is significant. For example, in the first home the range between the lowest estimate and the highest is $210,043. This leads us to assume that the model is not very good.

Section 3.2: Model #2 (Garage Capacity)

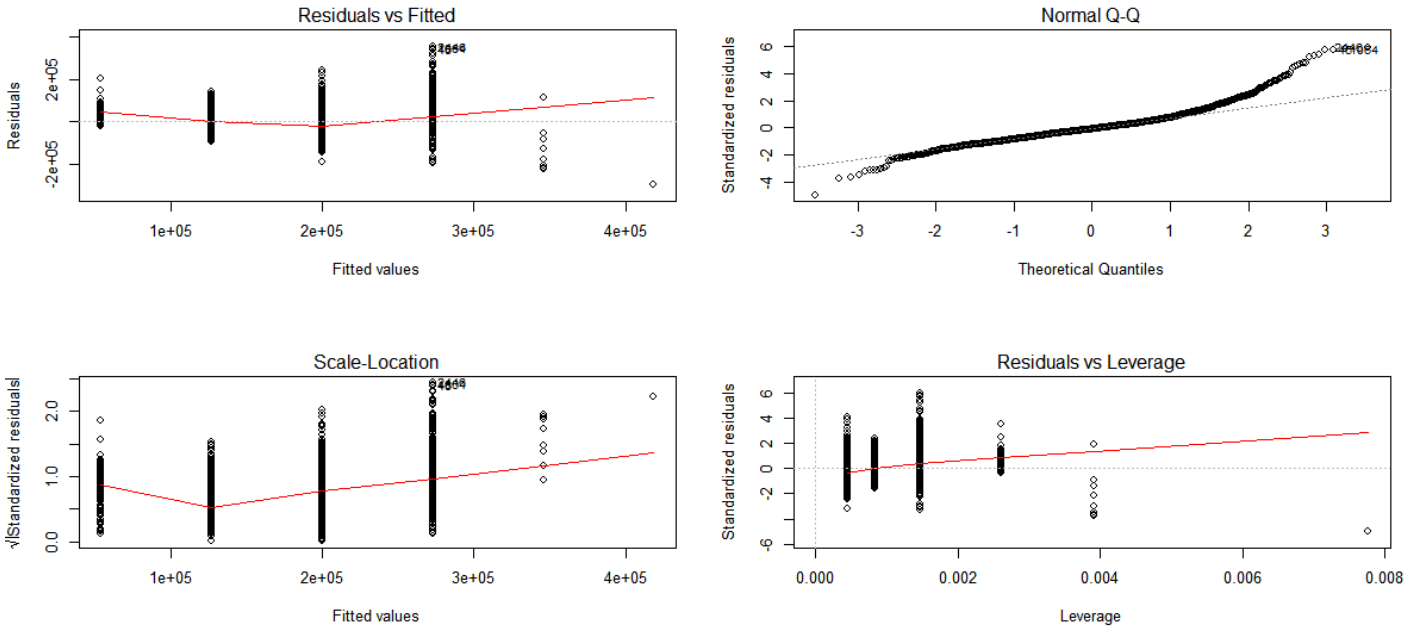


The size and functionality of a garage is a huge convenience and price factor for some home buyers. As seen in the above scatterplot, the larger the garage, the larger the price of the home. However, this does appear to reach a point of diminishing returns after the capacity for 4 cars.

Another aspect to consider regarding garages is the type of garage. Garages that are “built in” are the most valuable. The value of the home appears to decrease the farther the garage is from the home. If the garage is attached but not built on, the price is slightly lower. This decreases drastically when viewing detached garages or carports which are almost as valuable as not having a garage at all.

```
Residuals:
    Min      1Q  Median      3Q     Max
-292265  -34464   -4964   25803  352102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    54097       3030   17.86   <2e-16 ***
GarageCars     72934       1574   46.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59280 on 2463 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4657,    Adjusted R-squared:  0.4654
F-statistic:  2146 on 1 and 2463 DF,  p-value: < 2.2e-16
```

The very small p value allows us to come to the conclusion that the model is significant and the null hypothesis can be rejected. Additionally we can determine the equation of the model to be Y=54097 + 72934x1, thus for each additional car that can fit in the garage, $72,934 is added to the price of the home. The standard error of this model is $59,280 with roughly 47% of the values explained by this model, that isn't a fantastic figure not isn't bad either.



The next step of our regression would be to view the residuals. These four graphs help to identify potential problematic cases in the data set.

First looking at the residuals versus fits plot verify the assumption that the residuals are randomly distributed. This does not appear to be the case as it appears that there is a sort of bell curve with the majority in the middle and less on the sides. This would indicate nonconstant variance. Our options at this point would be to transform the response variable or to identify and remove outliers. When viewing the QQ plot, especially towards the right tail of the graph it

becomes obvious that the data set is not perfectly normalized, there are a few outliers that are preventing this. The residuals in the Scale-Location graph are not spread equally along the ranges of predictors, resulting in a slightly increasing line. Looking at the Residuals vs Leverage plot, the first thing that jumps out at me is the variable in the bottom right. While this is not identified by the graph as an outlier it is certainly far from the other values.

Lastly, we can attempt to predict the sale price of the home utilizing the total living space above ground.

	TotalFloorSF	HouseAge	QualityIndex	SalePrice	GrLivArea	GarageCars	fitSLR	fitSLR	fit	lwr	upr
1	1656	50	30	215000	1656	2	200870.1	200870.1	199964.2	83684.86	316243.5
2	896	49	30	105000	896	1	108282.5	108282.5	127030.6	10729.25	243332.0
3	1329	52	36	172000	1329	1	161033.1	161033.1	127030.6	10729.25	243332.0
4	2110	42	35	244000	2110	2	256179.1	256179.1	199964.2	83684.86	316243.5
5	1629	13	25	189900	1629	2	197580.8	197580.8	199964.2	83684.86	316243.5
6	1604	12	36	195500	1604	2	194535.2	194535.2	199964.2	83684.86	316243.5

Utilizing the above predictions we can see that the range between the lower and the upper values is significant. For example, in the first home the range between the lowest estimate and the highest is $232,559. This leads us to assume that the model is not very good and is slightly worse than the original utilizing the total living area above ground.

Section 4: Multiple Linear Regression Model – Model #3

Task 3:

The next step of our analysis is to conduct a multiple linear regression model on both the living space above ground and the garage capacity.

Analysis of Variance Table

Response: SalePrice

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
GrLivArea	1	9.1411e+12	9.1411e+12	4289.71	< 2.2e-16	***
GarageCars	1	1.8131e+12	1.8131e+12	850.86	< 2.2e-16	***
Residuals	2462	5.2463e+12	2.1309e+09			

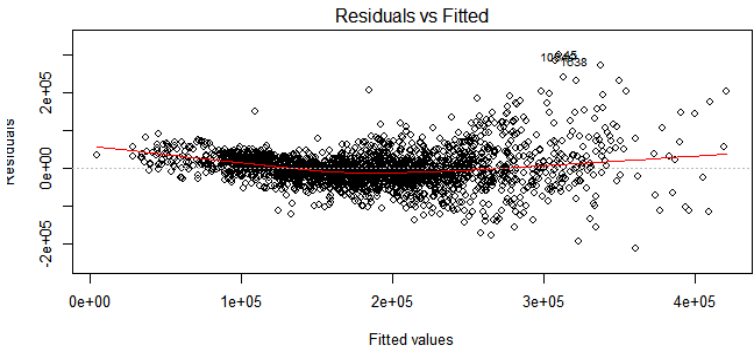
Utilizing anova we can determine that the model is significant from the small p value. We can also reject the null hypothesis that the variables collectively have no effect on the sale price.

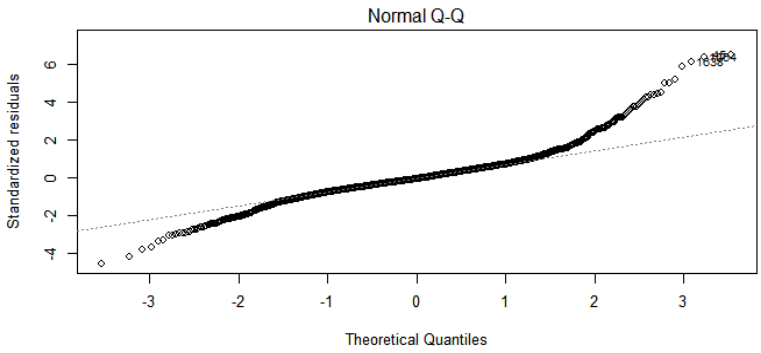
```
Residuals:
    Min      1Q  Median      3Q     Max
-211036  -24133   -2575   21450  301572

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -24177.755   3064.900  -7.889 4.56e-15 ***
GrLivArea       87.846      2.196  40.005  < 2e-16 ***
GarageCars   42198.303   1446.658  29.170  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46160 on 2462 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6762,    Adjusted R-squared:  0.6759
F-statistic:  2570 on 2 and 2462 DF,  p-value: < 2.2e-16
```

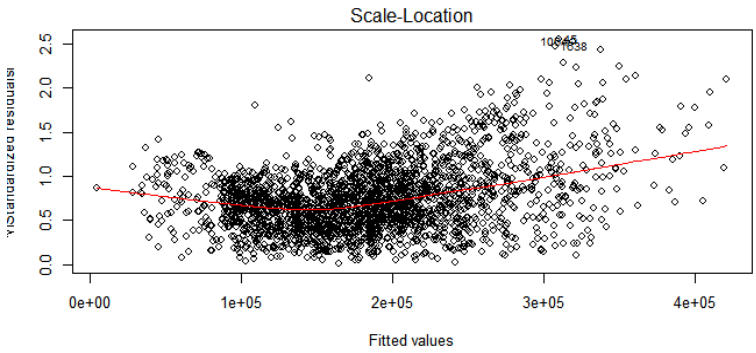
Additionally, from the below we can determine the model equation which is Y= -24177.75 + 87.846*x1 + 4,2198*x2. Utilizing this equation we can determine that if the size of the home is held constant, adding capacity for an additional car adds $42,198 to the price of the home. Additionally, if the capacity of cars is held constant, each additional square foot of living space above ground adds an additional $87 to the sale price.

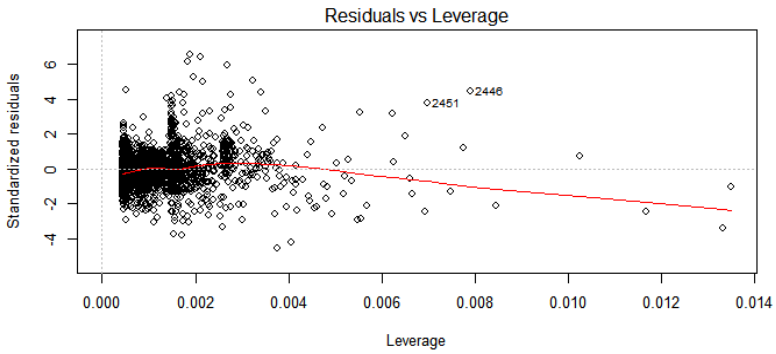
Next we want to evaluate how good the model is. Our predicted error is $46,160 and 67% of our data is explained by the model. This is 21% higher than garage capacity alone and 11% higher than total living space alone.

Next we will review the residuals. There is quite a trend in the fitted graph, very similar to the graphs before. We have outliers in the leverage graph 2451 and 2446. We have already discovered that removing 2446 does not adjust the R squared value. However, this time when removing 2451 and 2446 we see a very small increase in the adjusted R squared from .675 to 0.676. This is not much of a movement at all, so these values will remain in the analysis.









We can again attempt to predict the price of the homes from the multiple linear regression. One example of prediction our model can do is when the total living space above ground is 1,656 square feet and the car capacity is 2 cars. Our model predicts that the price of this home will be $205,692. Since the actual sale price of a home with these parameters is $215,000 our model is not too far off but could be better.

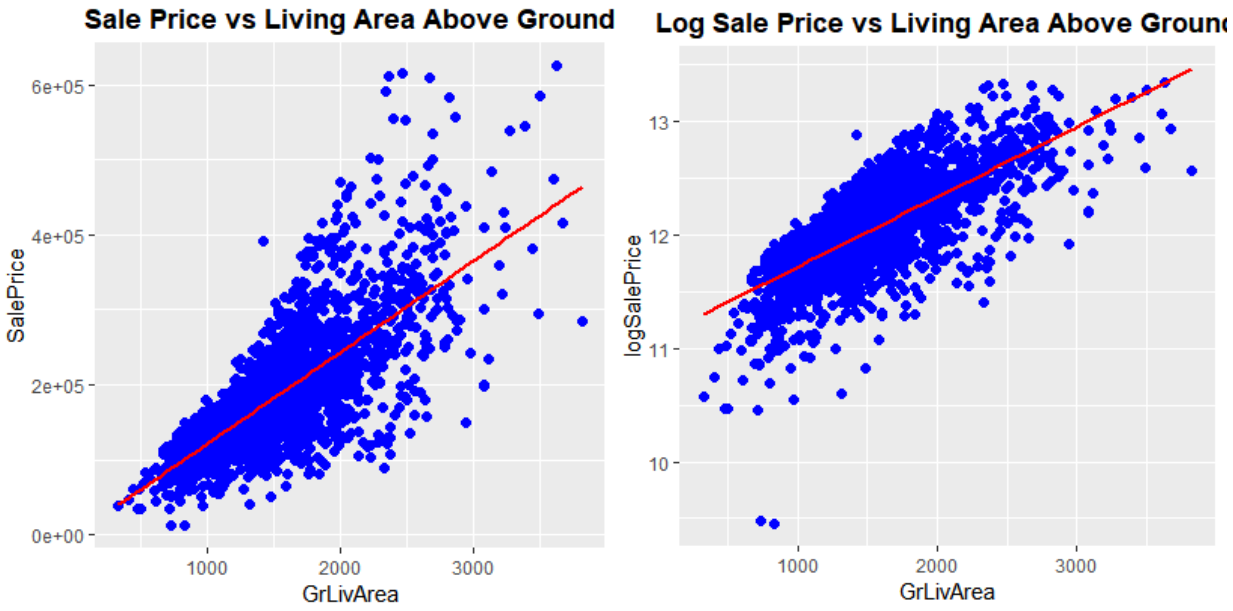
	TotalFloorSF	HouseAge	QualityIndex	SalePrice	GrLivArea	GarageCars	fitSLR	fitSLR.1	fit	fit.1
1	1656	50	30	215000	1656	2	200870.1	200870.1	199964.2	205692.04
2	896	49	30	105000	896	1	108282.5	108282.5	127030.6	96730.68
3	1329	52	36	172000	1329	1	161033.1	161033.1	127030.6	134768.05
4	2110	42	35	244000	2110	2	256179.1	256179.1	199964.2	245574.18
5	1629	13	25	189900	1629	2	197580.8	197580.8	199964.2	203320.19
6	1604	12	36	195500	1604	2	194535.2	194535.2	199964.2	201124.04

Task 4:

Section 5: Log Sale Price Response Models

For this section the sale price will be transformed to log sale price. Log transformation is utilized in this section to assist with confirming to normality and to assist with visualizing the large span of sale prices.

Section 5.1: Model #4 (Total Floor Space Above Ground)

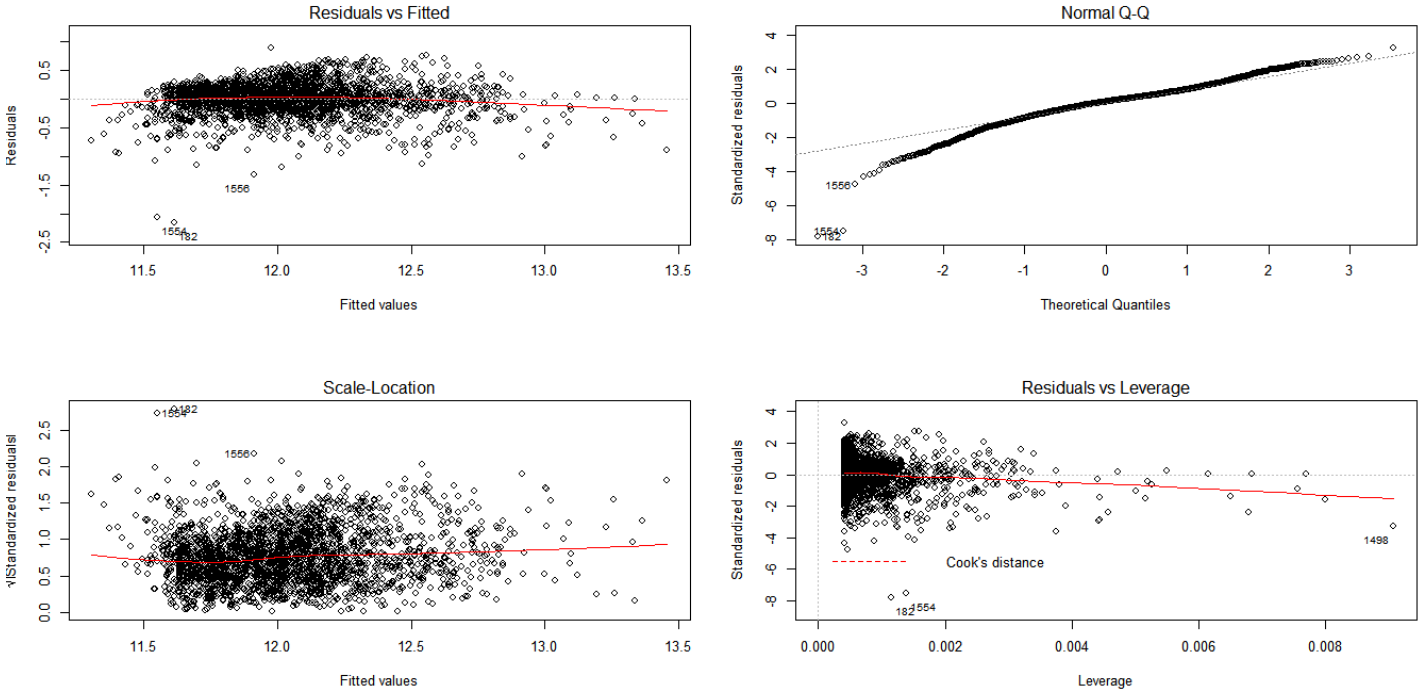


The above graph showcases the distribution of homes by log sale price and the area of the home that is livable above ground. This graph has a few differences from the original graph showing the unadjusted sale price. For example, the two outliers on the bottom left are more noticeable now than they were before. In the original graph, the values were tighter on the left and became more spread out the larger the house became. This graph appears differently. The more expensive values appear to be tighter which the cheaper homes below the line are more spread out. When transforming the data from the original sale price to the log sale price, we also notice that we transition from a cone shape to a more consistent clump of values.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.15650 -0.13695  0.02816  0.15560  0.90357

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.110e+01  1.767e-02   628.1   <2e-16 ***
GrLivArea   6.177e-04  1.111e-05    55.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2757 on 2464 degrees of freedom
Multiple R-squared:  0.5565,    Adjusted R-squared:  0.5563
F-statistic:  3092 on 1 and 2464 DF,  p-value: < 2.2e-16
```

We can also use the information to determine the model equation. This equation ends up being Y=- 1.110e+01*B0+ 6.177e-04*B1. Therefore, for every additional square foot of living space, the price goes up by $ 6.177e-04. Our standard error for this variable is roughly over $0.27 with 55% of variation explained by the model of total living area, this amount represents the average distance from the regression line. This is 1% less than the original sale price, indicating that the not transformed figure may be a better model.

Next we will review the residuals. Starting with the fitted graph, this graph does not appear random at all and appears to be less random than the original sale price graph. Also, there are more outliers on this graph than the original. A positive result from the sale price transformation is that the QQ plot is showing a much more normal distribution. This makes sense as a log transformation is typically conducted with the purpose of making the data more normally distributed. The scale location graph looks better than the original and more random. The leverage diagram indicates three outliers we should consider especially because they fall outside of Cook's distance line. When we remove the outliers (SIDs 1498,182 and 1554), the adjusted R squared value becomes .5561 which is .0002 less than the original making it a very slightly less fitting model so these figures will not be removed from the analysis.

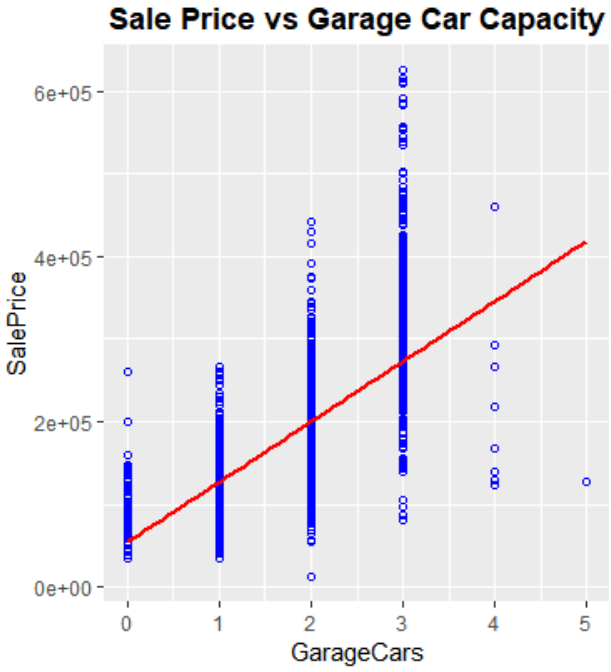


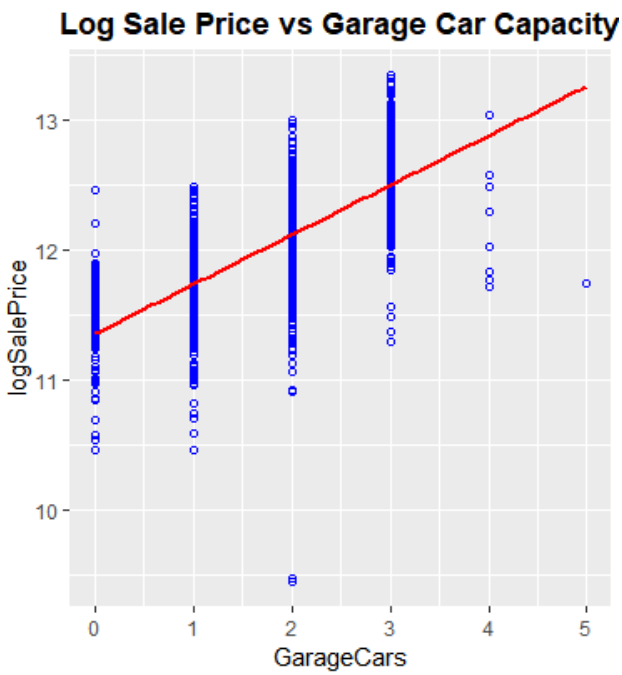
Lastly, we can attempt to predict the sale price of the home utilizing the total living space above ground.

	TotalFloorSF	HouseAge	QualityIndex	SalePrice	GrLivArea	GarageCars	logSalePrice	fit	fit.1	fit	lwr	upr
1	1656	50	30	215000	1656	2	12.27839	12.12185	12.12185	12.12185	11.58118	12.66251
2	896	49	30	105000	896	1	11.56172	11.65237	11.65237	11.65237	11.11155	12.19319
3	1329	52	36	172000	1329	1	12.05525	11.91985	11.91985	11.91985	11.37918	12.46052
4	2110	42	35	244000	2110	2	12.40492	12.40230	12.40230	12.40230	11.86148	12.94311
5	1629	13	25	189900	1629	2	12.15425	12.10517	12.10517	12.10517	11.56451	12.64583
6	1604	12	36	195500	1604	2	12.18332	12.08973	12.08973	12.08973	11.54907	12.63039

This model shows a better prediction as the range between the upper and the lower is only 1.02

Section 5.2 Model #5 (Garage Capacity)





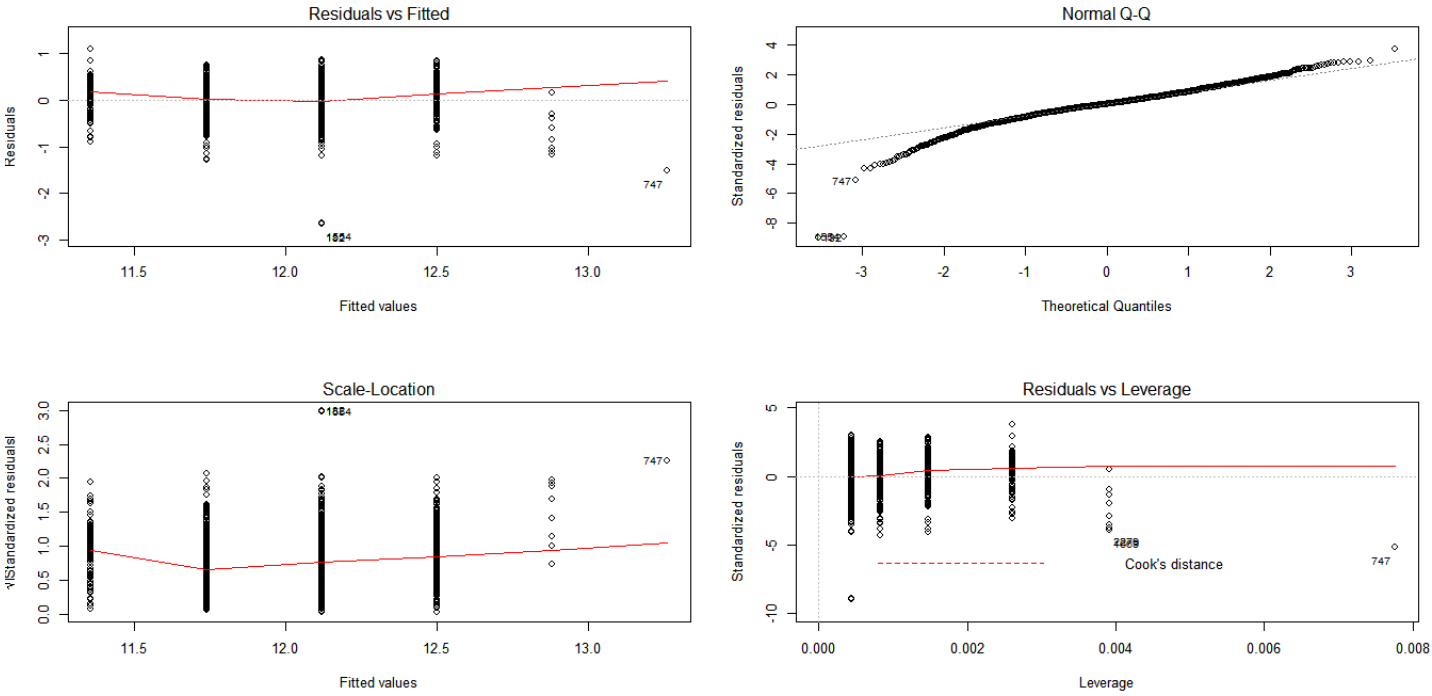
The first differences I notice between the original sale price and the log sale price distributions by car capacity is that there are more cheaper outliers in the log sale price distribution and on the other hand there are more expensive outliers in the original sale price distribution. This could be due to a larger range of values for each car capacity, especially in the cheaper homes coupled with expensive homes that are more similar. There is an improved shape in the graph when adjusting the values to log sale price as the values appear to be more grouped and less spread out.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.66349 -0.14888  0.01098  0.17242  1.11073

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.357707   0.015141  750.11   <2e-16 ***
GarageCars   0.381063   0.007868   48.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2963 on 2463 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4878,    Adjusted R-squared:  0.4876
F-statistic:  2346 on 1 and 2463 DF,  p-value: < 2.2e-16
```

The very small p value allows us to come to the conclusion that the model is significant and the null hypothesis that the car capacity does not have a relationship with price can be rejected. Additionally, we can determine the equation of the model to be Y=11.36*B0 + .381*B1, thus for each additional car that can fit in the garage, $0.38 is added to the price of the home. The standard error of this model is $0.30 with roughly 48% of the values explained by this model.

While 48% is not a good figure for this estimate it is 1% better than the R squared of the original sale price.

The next step in our analysis is to review the residuals.



When reviewing the fitted graph it is initially noticeable that this graph is far more distributed than the original, in fact the line is almost a perfect horizontal line, with only a slight increase on the far right. Again, the QQ plot is showing a far more normalized distribution which is to be expected with a log transformed response variable. The leverage graph shows many outliers but all of the outliers are within Cook's distance line. The three outliers identified by the graph include 747, 2279, and 1669. Removing these three variables results in an adjusted R squared of .4882, this is an adjusted R squared .0006 higher than the original. This leads us to believe that removing the variables makes the model slightly better but not by much at all.

Lastly, we will also attempt to predict the price of the car utilizing this model and adjusted sale price. The difference between the upper and lower predictions is only 1.2.

	TotalFloorSF	HouseAge	QualityIndex	SalePrice	GrLivArea	GarageCars	logSalePrice	fit	fit.1	fit.2	fit	lwr	upr
1	1656	50	30	215000	1656	2	12.27839	12.12185	12.12185	12.12185	12.11983	11.53872	12.70095
2	896	49	30	105000	896	1	11.56172	11.65237	11.65237	11.65237	11.73877	11.15754	12.32000
3	1329	52	36	172000	1329	1	12.05525	11.91985	11.91985	11.91985	11.73877	11.15754	12.32000
4	2110	42	35	244000	2110	2	12.40492	12.40230	12.40230	12.40230	12.11983	11.53872	12.70095
5	1629	13	25	189900	1629	2	12.15425	12.10517	12.10517	12.10517	12.11983	11.53872	12.70095
6	1604	12	36	195500	1604	2	12.18332	12.08973	12.08973	12.08973	12.11983	11.53872	12.70095

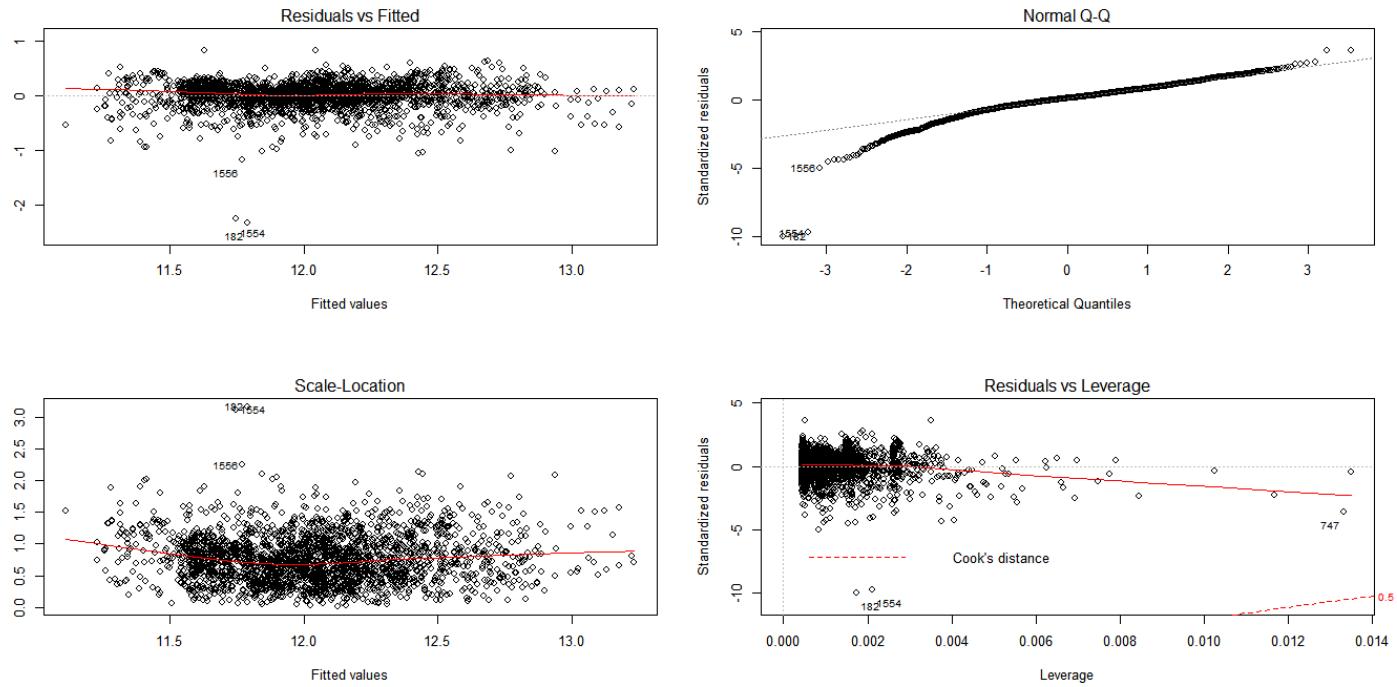
Section 5.2 Model #6 (Multiple Linear Regression Model)

```
Analysis of Variance Table

Response: logSalePrice
             Df  Sum Sq Mean Sq F value    Pr(>F)
GrLivArea     1 235.004 235.004 4335.82 < 2.2e-16 ***
GarageCars    1  53.699  53.699  990.74 < 2.2e-16 ***
Residuals  2462 133.442   0.054
```

When conducting the anova function on the garage capacity and total living space above ground variables, we receive extremely small p values. This indicates that our model is significant and that we can reject the null hypothesis that the values are not related to sale price.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.33511 -0.09936  0.02267  0.14108  0.83941

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.097e+01  1.546e-02  709.83   <2e-16 ***
GrLivArea   4.328e-04  1.107e-05   39.08   <2e-16 ***
GarageCars  2.296e-01  7.296e-03   31.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2328 on 2462 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6839,    Adjusted R-squared:  0.6836
F-statistic:  2663 on 2 and 2462 DF,  p-value: < 2.2e-16
```

Furthermore, the equation of the model can also be determined. The equation of this model is Y=1.097 + 4.32B1 + 2.296B2. Therefore, if the size of the living area above ground is held constant, for each increase in car capacity, the price of a home will increase by $2.296. This is also true for the size of the home, if the car capacity is held constant, for each additional square foot of living space above ground added, the price of the home increases by $4.328. This model accounts for 68% of our data and has a standard error of $.23. This is significantly better than the model with the original sale price which was only at 55%.



The fitted and scale location graph appear much more randomized than the original sale price graphs as the lines in each are almost perfectly vertical. The QQ plot is significantly more normalized than the original sale price QQ plot which again, is to be expected. There are a few outliers on the left tail that may need to be taken into consideration. The leverage graph has many outliers, including a few that are beyond the Cook's distance line. These outliers (182, 1554, and 747) need to be evaluated to see if they hold too much influence over the model. After removing these values, the adjusted R squared becomes 0.6836. This is the exact same as the original R squared so no adjustment is needed.

We next must test our model to see how well it can predict the price of a home. Utilizing our model to again predict the price of a home that has 1,656 square feet of livable space above ground and space for 2 cars, our model predicts a price of $12.15. The log transformation of the actual sale price is 12.28 so our model is fairly accurate at predicting the price.

```
  TotalFloorSF HouseAge QualityIndex SalePrice GrLivArea GarageCars logSalePrice   fitMLR    fit.1
1         1656       50           30    215000      1656          2     12.27839 12.12185 12.12185
2          896       49           30    105000       896          1     11.56172 11.65237 11.65237
3         1329       52           36    172000      1329          1     12.05525 11.91985 11.91985
4         2110       42           35    244000      2110          2     12.40492 12.40230 12.40230
5         1629       13           25    189900      1629          2     12.15425 12.10517 12.10517
6         1604       12           36    195500      1604          2     12.18332 12.08973 12.08973
     fit.2    fit.3    fit.4      res fitMLR.1 fitMLR.2 fitMLR.3 fitMLR.4      fit
1 12.12185 12.11983 12.14805 214987.9 12.14805 12.14805 12.14805 12.14805 12.14805
2 11.65237 11.73877 11.58950 104988.3 11.58950 11.58950 11.58950 11.58950 11.58950
3 11.91985 11.73877 11.77689 171988.1 11.77689 11.77689 11.77689 11.77689 11.77689
4 12.40230 12.11983 12.34453 243987.6 12.34453 12.34453 12.34453 12.34453 12.34453
5 12.10517 12.11983 12.13637 189887.9 12.13637 12.13637 12.13637 12.13637 12.13637
6 12.08973 12.11983 12.12555 195487.9 12.12555 12.12555 12.12555 12.12555 12.12555
```

Of the three models that were transformed, the multiple regression model fits the best with the higher R squared of 68% compared to the original 55%.

Section 6: Summary/Conclusions:

Task 5:

To begin creating regression models a crucial first step was to review the data and ensure that our sample population utilized accurately represented the typical Ames home but did not remove too many values resulting in a skewed data set.

Multiple regression models were conducted to view the influence of total living space above ground and the car capacity on the sale price of the home. These two variables already had a moderate relationship between themselves and overall had a larger relationship with the price.

In the end, both of the variables selected for regression were reliable predictors of price and assisted in creating a model that predicted fairly accurately the sale price of a home. There are certainly additional areas of the model that can be improved such as the removal of outliers and taking into consideration the influence of additional variables.