Introduction:

In Assignment #2, we will build off of Assignment #1 and investigate the data further by using exploratory data analysis, building multiple regression models and then comparing each with various test statistics including T and F statistics, Goodness of Fit (R-Squared) and analysis/plotting of residuals.

Section 1: Sample Definition

The Ames Housing dataset is comprised of 82 columns including 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. There are also 2 columns (SID and PID) used as observation identifiers. These are not relevant for the purposes of this assignment and were removed. There were also 5 additional variables created. This gives a total of 85 variables used for 2930 residential properties.

It was observed that multiple building types were included in the original data set. It does not make sense to compare one building type to another as each type has a different perception of value in the eyes of a buyer and this is reflected in the different zoning types associated with each building.

The original dataset consists of 5 building types and 8 zoning classifications :

	A (agr)	C (all)	FV	I (all)	RH	RL	RM
1Fam	2	22	77	2	12	2009	301
2fmCon	0	3	0	0	4	30	25
Duplex	0	0	0	0	4	92	13
Twnhs	0	0	19	0	1	28	53
TwnhsE	0	0	43	0	6	114	70

As the purpose of this assignment is to predict the sale price of a “typical” home, it does not make sense to include buildings falling under the zoning classifications industrial (I), agriculture (A) or commercial (C). Residential zoning types including residential high density (RH), residential low density (RL), residential low density park (RP), residential medium density (RM) and floating village residential will be used for further analysis. For building type, we are looking to analyze houses that represent the typical “house” that a family would buy. Duplexes (Duplx) and townhouses (TwnhsE and TwnhsI) are not good representations and are removed leaving single family detached (1Fam) and two family conversion houses (2FmCon) left for analysis.

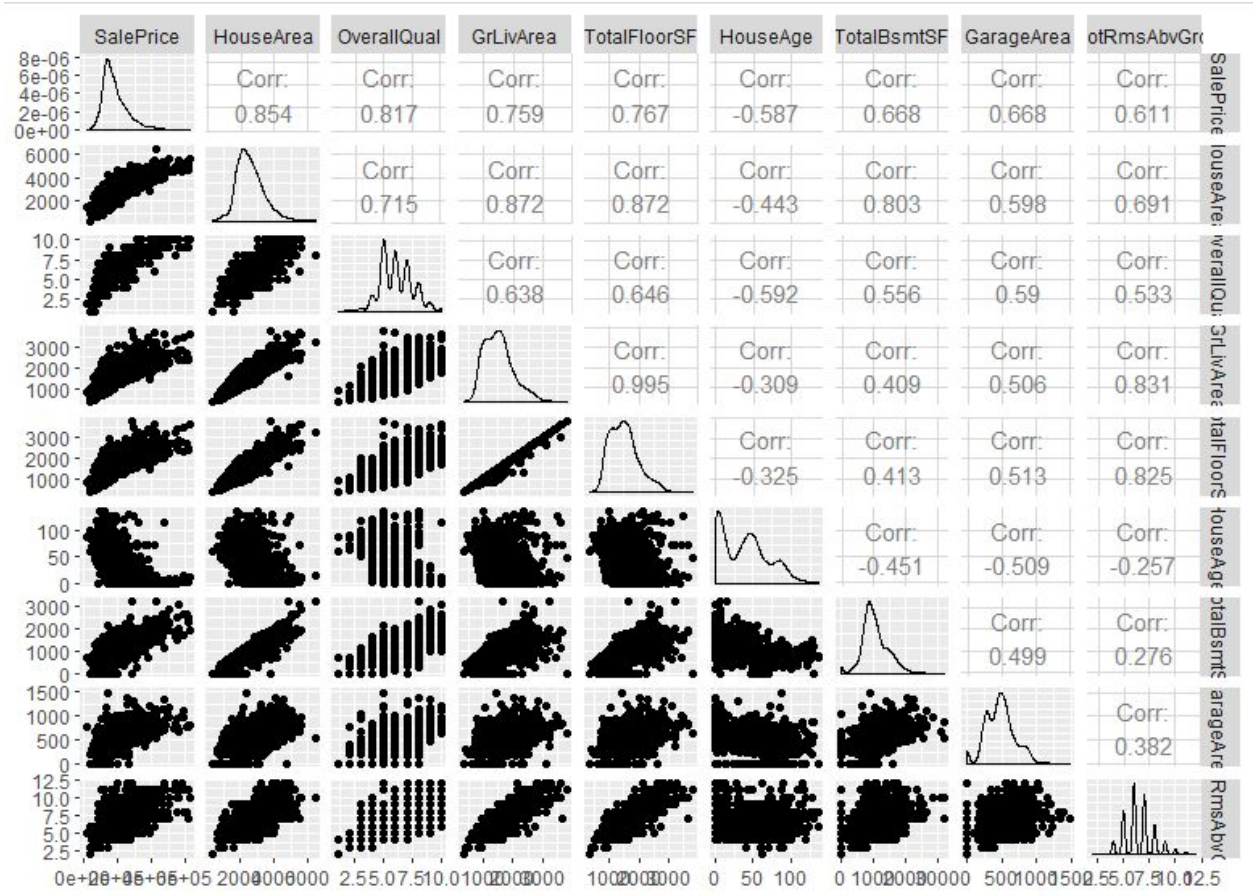
After these conditions are met, there are 2399 observations remaining. Here is a table showing the distribution of homes in the four zoning classifications by building type.

	A (agr)	C (all)	FV	I (all)	RH	RL	RM
1Fam	0	0	77	0	12	2004	301
2fmCon	0	0	0	0	0	0	0
Duplex	0	0	0	0	0	0	0
Twnhs	0	0	0	0	0	0	0
TwnhsE	0	0	0	0	0	0	0

It is evident that none of the properties fall under the two family conversion building type so this can be removed from consideration. While investigating the data further, it was also observed that the majority of properties ranged from 250 total sq ft. to 4000 sq ft. There were 5 properties with over 4000 total sq ft. which were identified as outliers and removed from the sample. This leaves us with a dataset of 85 variables and 2395 properties.

Section 2: Exploratory Data Analysis

In this section, exploratory analysis is performed and statistical graphics are presented to show how the two final predictor variables are chosen. This analysis is performed on the data from assignment #1 with 4 observations being removed due to being an outlier or containing missing values. It should be noted that simple linear regression models require a continuous predictor variable so this will be included as a requirement for choosing variables. Therefore, categorical variables such as bldngtype will not be considered because simple regression models only need 1 predictor variable. Using the EDA from the previous assignment, it was observed that there were 13 numerical variables that showed some relationship with Saleprice. However, since the variables needs to be continuous, some variables needed to be removed. The final 8 variables to be considered are OverallQual, GrLivArea, House Area (GrLivArea + TotalBsmtSF), TotalFloorSF, TotalBsmtSF, GarageArea, TotRmsAbvGrd and HouseAge. GGally is a useful package to visualize the strength of relationship between these variables.



From this graph, it is evident that the two strongest linear relationships with SalePrice are HouseArea and OverallQual. HouseArea has a .854 correlation with SalePrice and OverallQual has a .817 correlation with SalePrice. It is important to check that there is not a big correlation between these two variables because it is not desirable to have collinearity and interaction effects. There appears to be a strong relationship between OverallQual and HouseArea (.715) so OverallQual will not be used. GarageArea has a relatively strong relationship with SalePrice (.668) and less correlation with HouseArea (.598) so it will be used instead.

Variable selection is a very important step in the regression model building process. If the wrong variables are selected, the model will not perform very well. Similarly, it is important to choose variables that are not correlated with each other because this can cause problems for multiple linear regression models.

Section 3: Simple Linear Regression Models

In this section, we will build two separate simple linear regression models for both predictor variables: HouseArea and GarageArea. Statistical software in R is used to calculate the linear equation using the OLS function. Each model will be evaluated using measures like goodness of fit, hypothesis tests (t and F tests) and plotted with statistical graphs to evaluate residuals.

For goodness of fit, the statistic R^2 (also called coefficient of determination) will be used to assess each model. R-squared is a statistical measure of how close the data are to the fitted regression line or in other words, the percentage of the response variable variation that is explained by the model. In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased. Before looking at measures of goodness of fit, it is important to plot the residuals. This can reveal unwanted residual patterns that may indicate biased results.

Section 3.1: Model #1 (HouseArea)

In this model, HouseArea is used as the sole predictor variable. The ANOVA (Analysis of Variance) table shows that the model is statistically significant because it has a very high F-value. This means that we can reject the null hypothesis which is that regression coefficients are equal to zero (meaning the predictor variable is not useful to fit the data). A large F-value also indicates a small P-value. Similar to a a large F- value, a small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis. In this case, the p-value is less than 2.2e^-16.

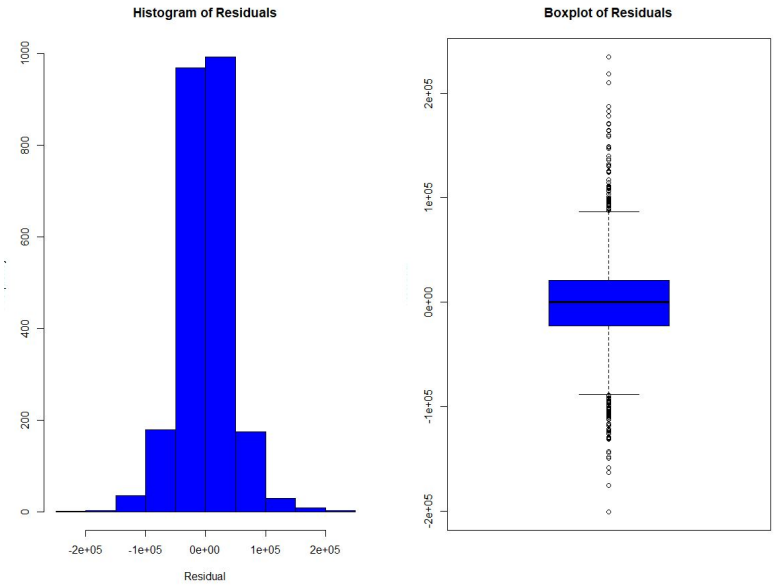
```
Analysis of Variance Table

Response: SalePrice
            Df     Sum Sq    Mean Sq F value    Pr(>F)
HouseArea    1 1.1387e+13 1.1387e+13  6424.9 < 2.2e-16 ***
Residuals 2391 4.2377e+12 1.7723e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

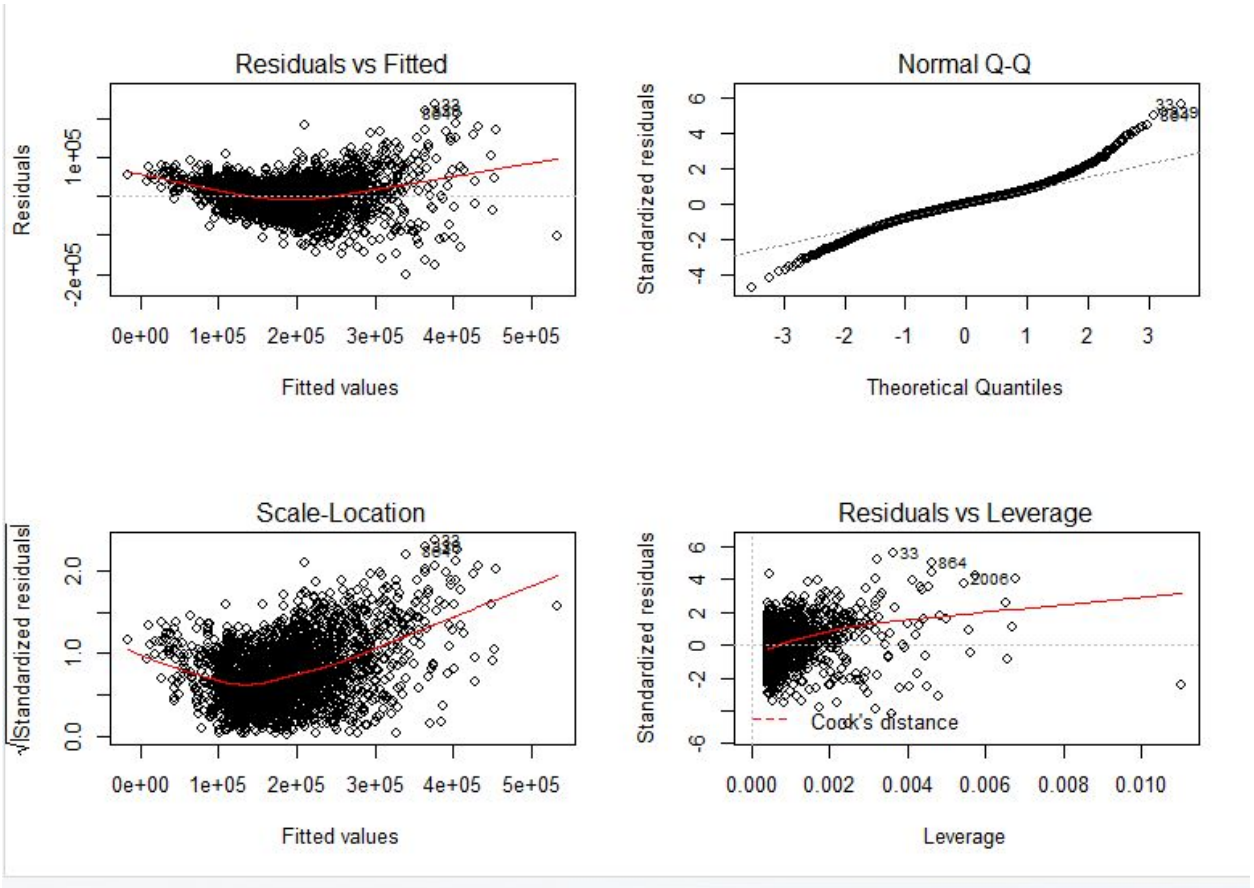
Using the summary() function in R provides a good analysis of the actual equation as well as the coefficients and residuals produced. From the analysis below, the coefficient for HouseArea can be interpreted as “for every 1sqft increase in area there is a 90$

increase in price. The standard error of 1.127 means that if the model was ran multiple times there would be approximately $1.13 worth of variation. A large t-value and subsequent small p-value indicates that the coefficient estimate is statistically significant and provides evidence that there is a strong relationship between HouseArea and SalePrice. The R^2 value of .7288 indicates that approximately 73% of variation in the response variable is accounted for by the predictor variable. The residual standard error of 42100 means that the estimates could be off by $42,100. Given that the mean SalePrice is $185,572, the estimated prices could be off by about 22.6%.

```
Call:
lm(formula = SalePrice ~ HouseArea, data = newdatafinal)

Residuals:
    Min      1Q  Median      3Q     Max
-200779  -22968     342   20964  234868

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -47116.816   3028.370  -15.56   <2e-16 ***
HouseArea       90.308      1.127   80.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42100 on 2391 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7288,    Adjusted R-squared:  0.7287
F-statistic:  6425 on 1 and 2391 DF,  p-value: < 2.2e-16
```

The residuals will now be plotted and visualized below. Based on these plots, there does not seem to be much randomness and residuals seem to be roughly equally distributed above and below the median value . However, the steepness of the histogram may indicate that HouseArea is only a good predictor of house prices that are at these middle prices and may struggle with houses that are extremely high or low in price. This idea is also reflected by the QQ plot which shows variation at both extremes.





Section 3.2: Model #2 (GarageArea)

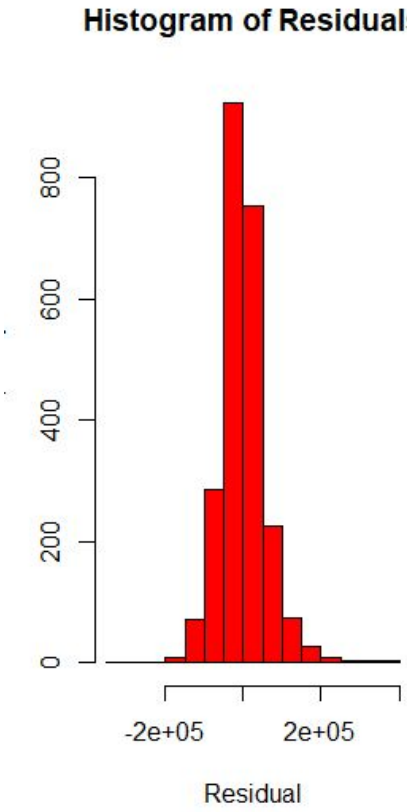
This model will use GarageArea as the sole predictor variable. Based on this, ANOVA table and F-test, the model appears to be statistically significant so the null hypothesis can be rejected.

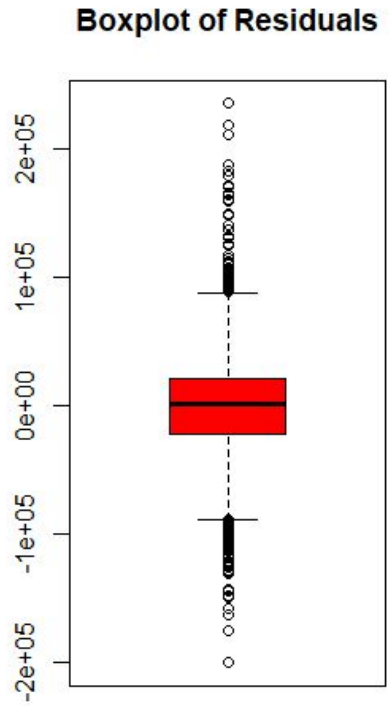
```
Analysis of Variance Table

Response: SalePrice
             Df     Sum Sq    Mean Sq F value    Pr(>F)
GarageArea    1 6.9735e+12 6.9735e+12    1925 < 2.2e-16 ***
Residuals  2391 8.6615e+12 3.6225e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

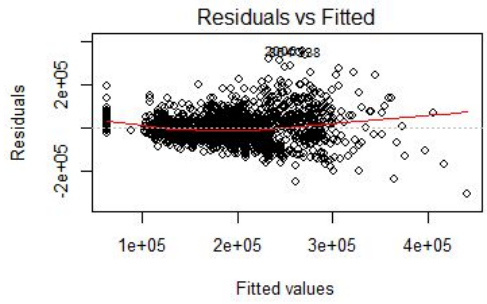
The coefficient estimate of 254.547 indicates that for every sq ft increase in garage area, there is a $254.55 increase in price. The standard error of 5.8 means that if the model is run multiple times there can be $5.80 in variation. One red flag for this model is that the R^2 value is only .446. This means that only 44.6% of variation in sale price is accounted for by garage area. This number is much smaller than the one produced by Model #1. Similarly, the t-value is much smaller than the t-value in Model #1, however, the coefficient estimate is still significant.

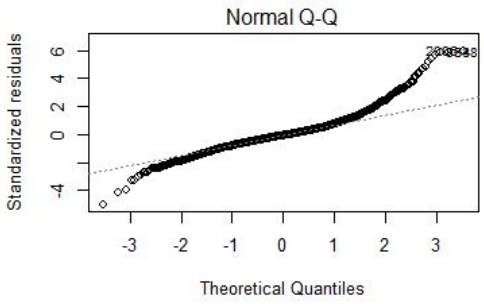
```
Call:
lm(formula = SalePrice ~ GarageArea, data = newdatafinal)

Residuals:
    Min      1Q  Median      3Q     Max
-302381  -32862   -3853   25038  356965

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 62615.132   3060.913   20.46   <2e-16 ***
GarageArea    254.547      5.802   43.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60190 on 2391 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.446,     Adjusted R-squared:  0.4458
F-statistic:  1925 on 1 and 2391 DF,  p-value: < 2.2e-16
```

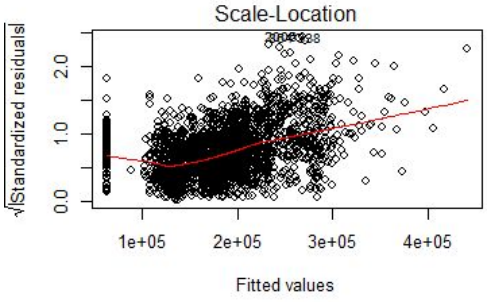
Similar to Model #1, the steepness of the histogram and variation showed in the QQ plot indicates that this model may struggle to predict prices on both extremes.

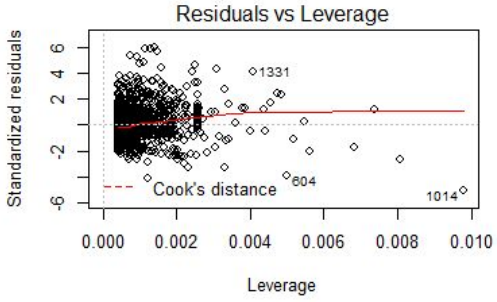

Histogram of Residuals
0
200
400
600
800
-2e+05
2e+05
Residual


Boxplot of Residuals
-2e+05
-1e+05
0e+00
1e+05
2e+05


Residuals vs Fitted
Residuals
-2e+05
2e+05
1e+05
2e+05
3e+05
4e+05
Fitted values


Normal Q-Q
Standardized residuals
-4
0
2
4
6
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
0.0
1.0
2.0
1e+05
2e+05
3e+05
4e+05
Fitted values


Residuals vs Leverage
Standardized residuals
-6
-2
2
4
6
1331
604
1014
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
Leverage

Section 4: Multiple Linear Regression Model – Model #3

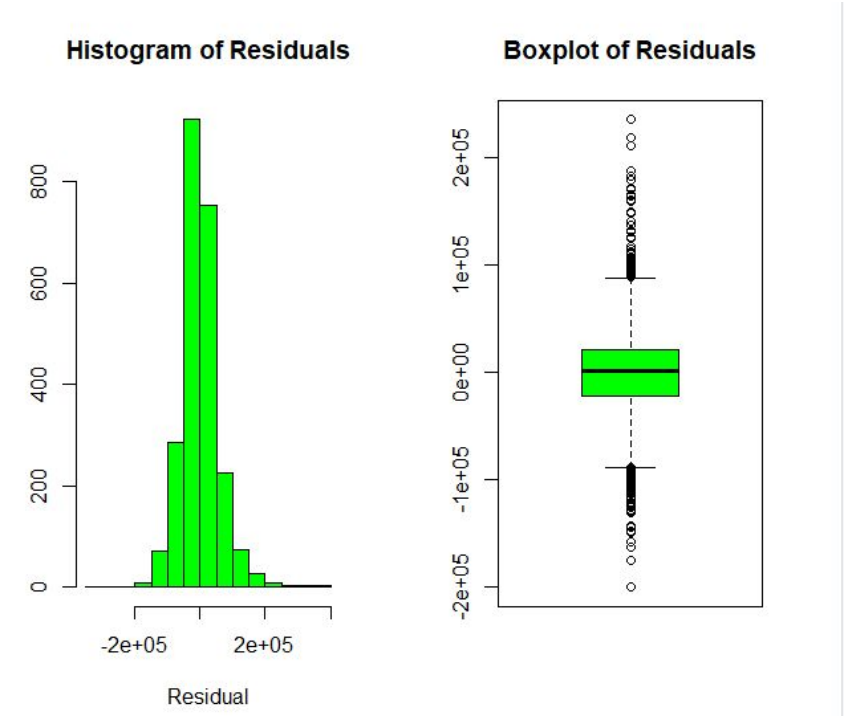
Model #3 is built by combining HouseArea and GarageArea as the predictor variables. Before choosing these variables, the amount of correlation between the predictor variables was minimized as well as taking into account the strength of correlation with SalePrice. Based on the ANOVA table and F-Test, it is evident that the model is statistically significant.

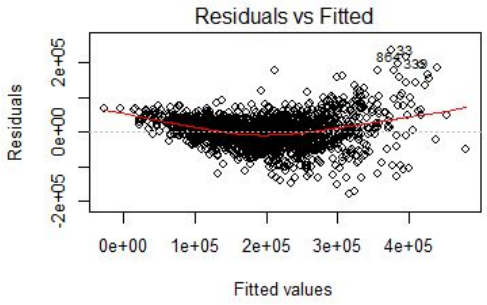
```
Analysis of Variance Table

Response: SalePrice
             Df     Sum Sq    Mean Sq F value    Pr(>F)
HouseArea     1 1.1388e+13 1.1388e+13 7486.35 < 2.2e-16 ***
GarageArea    1 6.0161e+11 6.0161e+11  395.49 < 2.2e-16 ***
Residuals  2389 3.6341e+12 1.5212e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

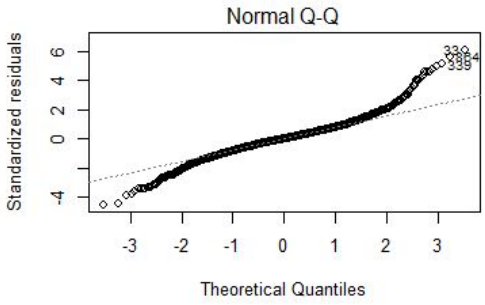
From the analysis below, the t-values indicate that the predictors are significant in predicting the response variable. The estimated coefficient of HouseArea 74.832 means that for every sq ft increase in house area, there is a $74.83 increase in price. Similarly, the estimated coefficient of GarageArea 93.284 means that for every sq ft increase in garage area, there is a $93.28 increase in price. The R^2 value indicates that approximately 77% of variation is accounted for by these variables. The adjusted R^2 is about the same. The residual standard error of 39000 means that on average the predictions will be off by $39,000. This indicates an improvement from both Model #1 and Model #2.

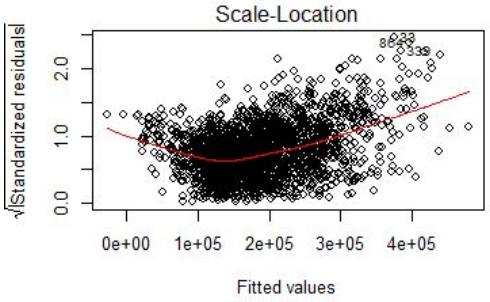
```
Call:
lm(formula = SalePrice ~ HouseArea + GarageArea, data = newdatafinal)

Residuals:
    Min      1Q  Median      3Q     Max
-178747  -20889    -359   20591  236192

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -52287.912   2817.666  -18.56   <2e-16 ***
HouseArea       74.832      1.302   57.47   <2e-16 ***
GarageArea      93.284      4.691   19.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39000 on 2389 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7674,    Adjusted R-squared:  0.7672
F-statistic:  3941 on 2 and 2389 DF,  p-value: < 2.2e-16
```

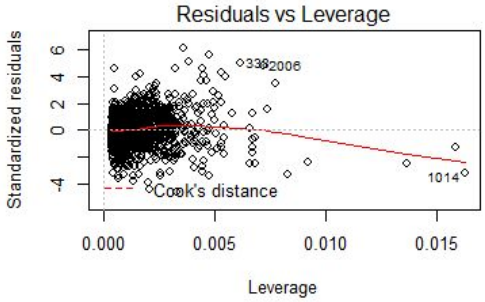
Based on these plots, there is still evidence that the model may struggle to predict sale price values that are on the extreme sides of the price spectrum. There seems to be lots of variation on the higher side of prices as shown in the Residuals vs Fitted values plot below.











Section 5: Log SalePrice Response Models

In these models, the response variable SalePrice will be transformed using log transformation. From Assignment #1, it was evident that a log transformation could help eliminate some of the variation found in each model.

Section 5.1: Model #4 (House Area)

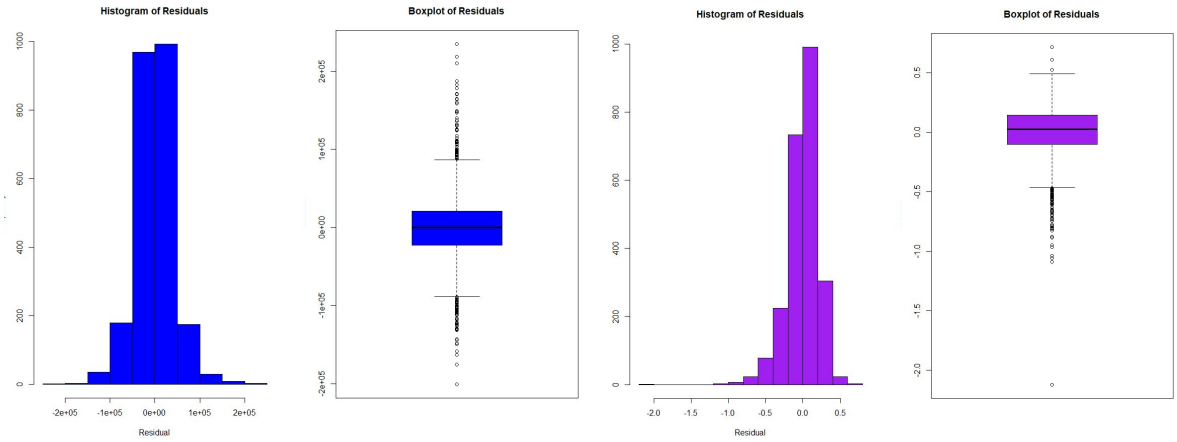
From the ANOVA table below, it is evident that the model is statistically significant and the null hypothesis can be rejected.

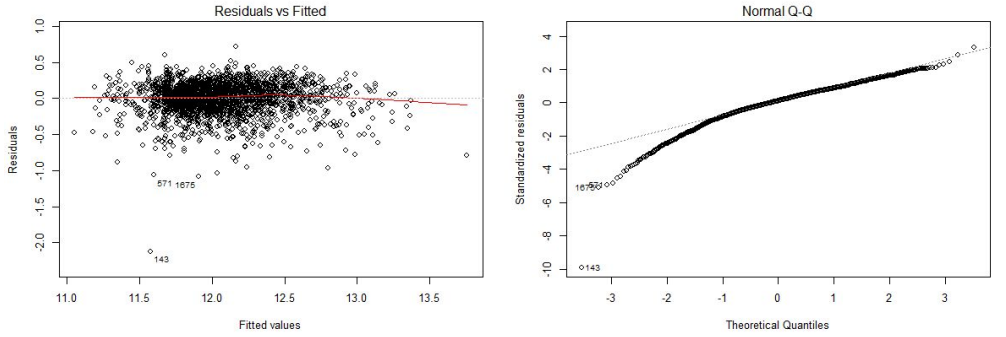
```
Analysis of Variance Table

Response: logSalePrice
            Df Sum Sq Mean Sq F value    Pr(>F)
HouseArea    1 274.46 274.462  5991.6 < 2.2e-16 ***
Residuals 2391 109.53   0.046
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

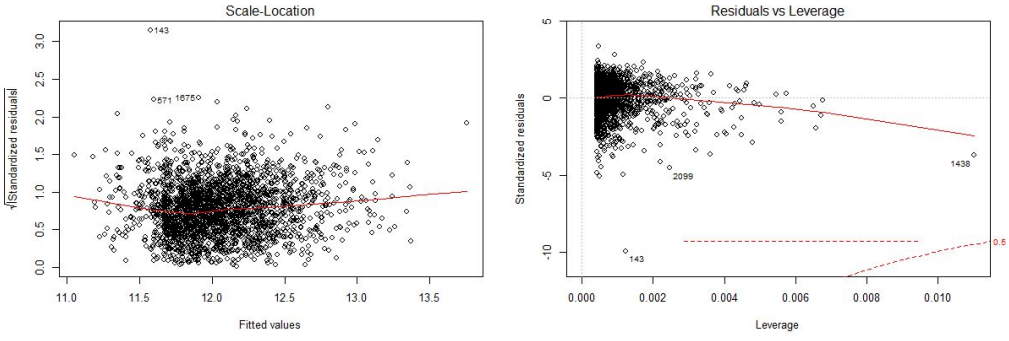
From the analysis below, large t-values indicate estimated coefficients that are statistically significant. Also, an R^2 value of .7148 means that about 71.5% of the variation in sale price is accounted for by this model.

```
Call:
lm(formula = logSalePrice ~ HouseArea, data = newdatafinal)

Residuals:
     Min       1Q   Median       3Q      Max
-2.12001 -0.10250  0.02369  0.14063  0.71388

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.091e+01  1.540e-02  708.42   <2e-16 ***
HouseArea   4.434e-04  5.728e-06   77.41   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.214 on 2391 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7148,    Adjusted R-squared:  0.7146
F-statistic:  5992 on 1 and 2391 DF,  p-value: < 2.2e-16
```

Based on these plots, there appears to be much more variation at the lower end of the price spectrum. However, the model seems to fit the higher price values much better. This may indicate that there are outliers that the model does not seem to be able to deal with. The blue plot on the left is Model #1 with not log transformation and the purple plot is the log transformed model. There appears to be a bigger skew to the left in the log transformed model.







Section 5.2: Model #5 (Garage Area)

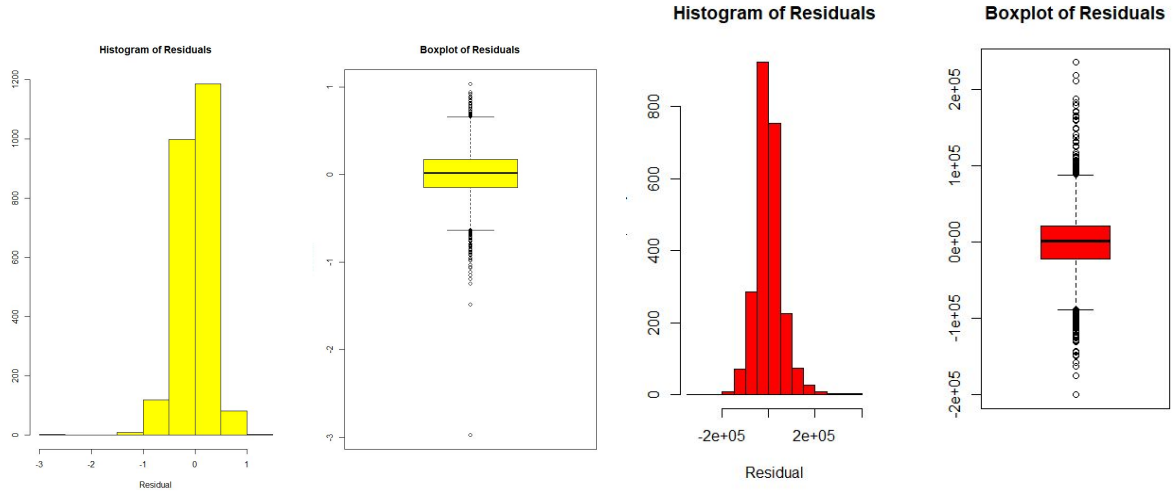
From the ANOVA table below, it is evident that the model is statistically significant and the null hypothesis can be rejected.

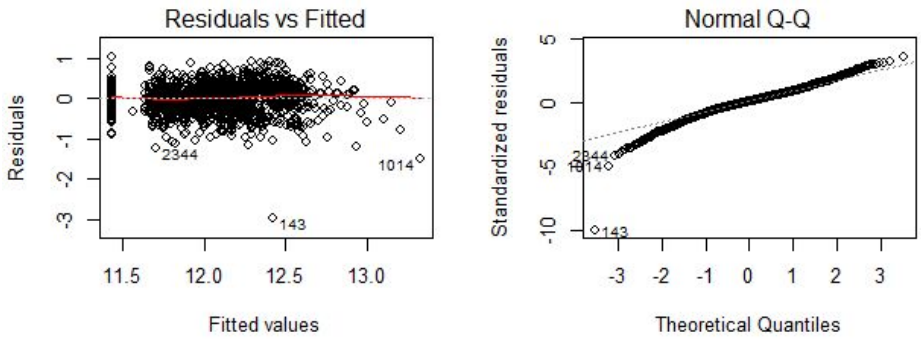
```
Analysis of Variance Table

Response: logSalePrice
             Df Sum Sq Mean Sq F value    Pr(>F)
GarageArea    1 174.01 174.013    1976 < 2.2e-16 ***
Residuals  2391 210.56   0.088
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

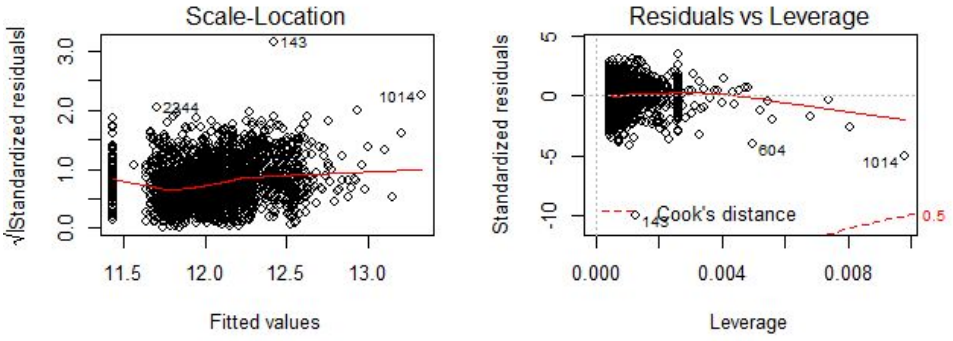
From the analysis below, large t-values indicate estimated coefficients that are statistically significant. However, an R^2 value of .4525 means that about 45.3% of the variation in sale price is accounted for by this model.

```
Call:
lm(formula = logSalePrice ~ GarageArea, data = newdatafinal)

Residuals:
     Min       1Q   Median       3Q      Max
-2.97038 -0.15019  0.01825  0.17390  1.03353

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.143e+01  1.509e-02  757.69   <2e-16 ***
GarageArea  1.272e-03  2.860e-05   44.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2968 on 2391 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4525,    Adjusted R-squared:  0.4523
F-statistic:  1976 on 1 and 2391 DF,  p-value: < 2.2e-16
```

Based on these plots, a log transformation does not seem to have much of an effect on the normality of the model with only a slight improvement in the amount of variation accounted for. The steepness of the histogram and outliers visualized in the boxplot and QQ plot indicates that the model struggles with extreme values. The yellow plot represents Model #5 and the red plot represents the normal plot from Model #2.







Section 5.3: Model #6 (House Area + Garage Area)

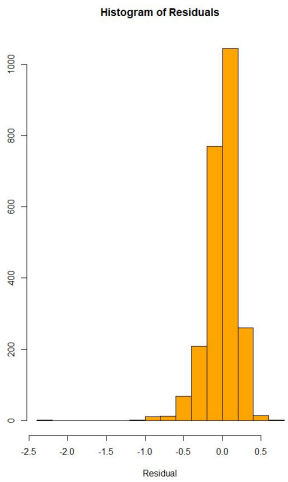
From the ANOVA table below, it is evident that the model is statistically significant and the null hypothesis can be rejected.

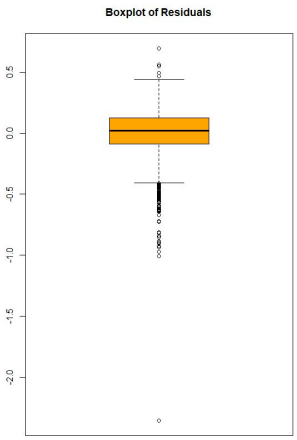
```
Analysis of Variance Table

Response: logSalePrice
             Df  Sum Sq Mean Sq F value    Pr(>F)
HouseArea     1 274.477 274.477 7065.26 < 2.2e-16 ***
GarageArea    1  16.687  16.687  429.53 < 2.2e-16 ***
Residuals  2389  92.810   0.039
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

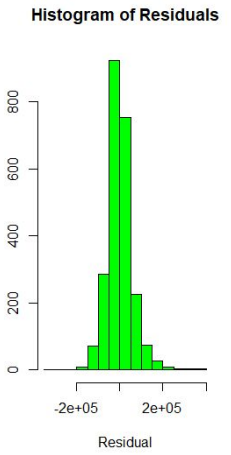
From the analysis below, large t-values indicate estimated coefficients that are statistically significant. However, an R^2 value of .7583 means that about 75.83% of the variation in sale price is accounted for by this model.

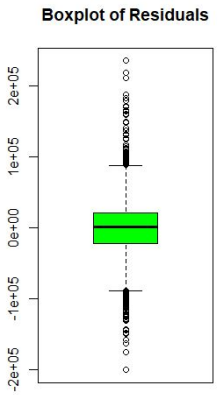
```
Call:
lm(formula = logSalePrice ~ HouseArea + GarageArea, data = newdatafinal)

Residuals:
     Min       1Q   Median       3Q      Max
-2.35288 -0.08791  0.02127  0.12603  0.69390

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.088e+01  1.424e-02  764.05   <2e-16 ***
HouseArea   3.619e-04  6.580e-06   54.99   <2e-16 ***
GarageArea  4.913e-04  2.371e-05   20.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1971 on 2389 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7583,    Adjusted R-squared:  0.7581
F-statistic:  3747 on 2 and 2389 DF,  p-value: < 2.2e-16
```

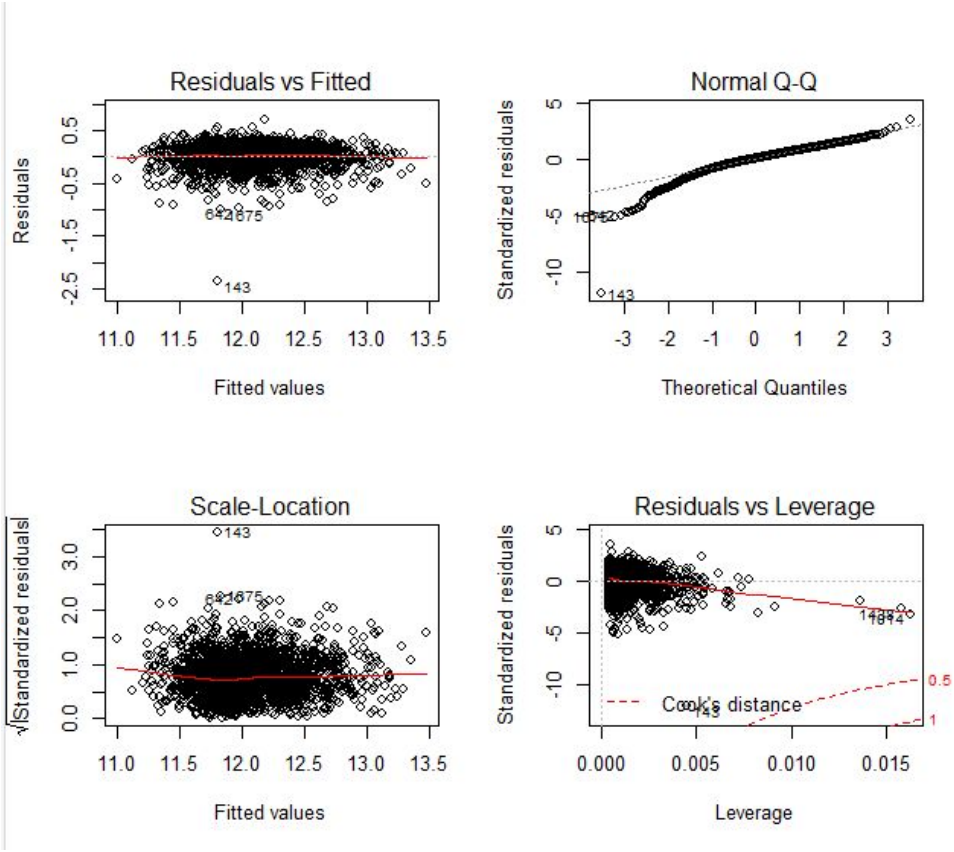
Based on these plots, it is evident that a log transformation may have had an effect on outliers on the higher end on the price spectrum as the model seems to account for much more of this variation. However, it is evident that the model still struggle to account for values that are below the interquartile range. The orange plot represents the log transformed Model # 6 and the green plot represents the normal Model #3.


Histogram of Residuals
Residual


Boxplot of Residuals


Histogram of Residuals
Residual


Boxplot of Residuals


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Section 6: Summary/Conclusions

From Assignment #2, the following conclusions can be made:

- EDA , variable selection, and data cleaning are probably the most important steps in the model building process and can have drastic effects on model performance
- It is also very important to understand the purpose of the assignment which is to build a model for the “typical house”. This relates to variable selection and can also affect the output of a model.
- Regular models struggled with outliers at both ends of the price spectrum and effected the variation accounted for by the predictor variables.
- It appears that log transformations helped account for some of the variation found at higher sale prices, however, log transformed models still struggled with lower price outliers.
- Visualization of residuals can show patterns in variation that are not found in simple descriptive statistics and this is important to improving a model
- The next steps in the model building process might be to build a model with 3 or more variables or to test the accuracy of each model by dividing the data into training and test sets and see how each model performs in prediction.