Housing Price Prediction Project

Haitie Liu

Anishka Peter

Introduction

For this project we will be looking at the sale prices of houses in Ames, Iowa. First, we will create a predictive model for Century 21 Ames about the sale prices of houses in the North Ames, Edwards and Brookside neighborhoods that will provide insights into how sale prices are related to the square footage of the living area in each neighborhood. We will also include a comparison of potential models and the criteria that we used to pick the best model. Next, we will look at all the houses in Ames, Iowa and create 3 models that predict the sale price of a house based on any of the variables in our data such as square footage, neighborhood, or building type. We will compare these 3 models based on their adjusted R^2 and their CV press score and determine which model is best at predicting future sale prices of homes in Ames.

Data Description:

In this data, we have 81 columns that consist of different types of information to describe each house, such as square footage of the house, different types of building structures, sale price in dollars, Style of dwelling etc and 1460 rows. To making data analysis easier, we have divided the columns into categorical and numeric, for example sales price would be numeric and neighborhood names would be categorical.

Furthermore, to run a model on categorical values, we also dived into the categorical values into ordinal, and non-ordinal. The ordinal data means that the different levels provide an intrinsic meaning of that category. For example, for "Lot Shape": General shape of property, we have Reg meaning Regular, IR1 meaning Slightly irregular, IR2 meaning Moderately Irregular, IR3 meaning, Irregular. Instead of keeping them as-is, we can modify these columns into numeric values, for example, 1 means regular, 2 means slightly irregular, 3 means moderately irregular, 4 means irregulars, that way our regression model can better capture these numeric values. Lastly for these non-ordinal variables, we just transform them into factor levels as they provide a truly intrinsic value that cannot be simply transformed into numeric values.

Analysis Question 1:

In our first analysis, Century 21 Ames wants to understand the relationship between square footage and Sales Price in the neighborhoods that they sell houses in. So, we are taking a deep dive into how the Sale Price of the house is related to the square footage of the living area of the house (GrLIvArea) specifically in North Ames, Edwards and BrookeSide neighborhoods, and if the SalesPrice and its relationship to square footage depends on which neighborhood the house is located in.

Build and Fit Linear Model on Original Data

First, we filtered the houses that are only in neighborhood NAmes, BrkSide, or Edwards. Next, we analyzed the Sales price according to the living area of the house and created linear regression model using only GrLivArea to predict the Sale price of a house. We got the equation SalesPrice = 78205.578 + 45.979*GrLivArea and the details of finding the equation can be found in graphic 1.

Checking Assumptions on Original Data

- ScatterPlot:

- Looking at graphic 2, we see that there is evidence of a linear relationship between the GrLivArea and SalePrice
- Residuals:
  - Looking at graphic 3, we see the residuals are clustered around the smaller fitted values and there are some very large outliers so we will not be able to assume equal standard deviations.
- QQ-plots:
  - Looking at graphic 4, we see that the QQplot shows that the residuals are normal except for a few outliers which should not cause a problem with the size of the sample
- Standardized residuals with Leverage:
  - Looking at graphic 5, we can see that there are a few points, specifically 339 that have a high Cook's Distance. After looking at observation 339, we see that the sales price is low for a huge area house, it may be entered by an error, but we will keep it in the data and proceed with caution.

Since we are unable to achieve the assumption of equal standard deviations of the residuals, we will proceed with a log-log transformation to the data and re-check the assumptions.

Logging Sales Price and GrLivArea and Checking Assumptions

In graphic 6 we have evidence that there is a linear relationship between the log Sales Price and log Living Area Square Footage based on the scatterplot. In graphic 7 we see that the residuals after the log-log transformation are randomly distributed, indicating equal variance of residuals. The qq-plot of the residuals is in graphic 8 and although the residuals are not perfect, it still indicates normal distribution of residuals. Finally, we will assume all (x,y) pairs of observations are independent of one another. Since we have met all the assumptions with the log-log transformed data, we will proceed to build our models.

Building Model on Log data and Comparing Models

Now that our assumptions are met, we will create our linear regression models. We will compare a model without the neighbor variable, with the neighborhood variable as an indicator variable, and the neighborhood as an interaction variable. First, for the simple linear regression model without the neighborhood variable we have:

Log (Sale Price) = beta0 + beta1 * Log(Living Area)

Log (Sale Price) = 7.75338 + 0.56824*Log(GrLivArea)

Further details on building the model and the parameter estimates are in graphic 9.

In the next model we will use neighborhood and square footage to predict sales prices. The change in the sale price is constant for every neighborhood but there may be a difference in the starting sale price of the house. We do this by adding neighborhood as an indicator variable which can be seen in graphic 10. By doing this we get the following equation:

Log (Sale Price) = beta0 + beta1* Edwards + beta2 * NAmes + beta3 * Log(Living Area)

Log (Sale Price) = 7.76936 – 0.02044*Edwards + 0.13279*Names + 0.55579*Log(GrLivArea)

Finally, we also created a model where we can use neighborhoods and square footage to predict sale price but the amount that the sale price can change can be different for each neighborhood. We did this by adding neighborhood as an interaction variable which can be seen in graphic 11. We got the following regression equation:

Log (Sale Price) = beta0 + beta1 * Edwards + beta2 * NAmes + beta3 * Log(Living Area) + beta4 * Log(Living Area) * Edwards + beta5 * Log(Living Area) * NAmes

Log (Sale Price) = 5.91292 +2.09359* Edwards + 2.57981 * NAmes + 0.81965 * Log(Living Area) - 0.29998 * Log(Living Area) * Edwards - 0.34662 * Log(Living Area) * NAmes

We get the various adjusted R square values from the models seen in graphics 9-11

Model	Adjusted R Square
Without Neighborhoods	41.88
Neighborhood as Indicator	48.57
Neighborhood as Interaction	50.56

We see that adding the neighborhoods as interactions in our model increases the Adjusted R squared of the models. Now to test if adding the neighborhood variable as an indicator significantly improved our model, we use the ANOVA tables in graphics 12 and 13 and we built our ANOVA in graphic 15. Based on the ANOVA we find that there is evidence that there is not 0 difference between the Sales Price between the Neighborhoods (p-value <0.0001).

Now to test if it was significant to allow the change in the sales price to be different for each neighborhood, we will compare the ANOVA tables of our models with Neighborhood as an indicator variable versus an interaction variable which we can see in graphics 13 and 14. We built our own ANOVA which is seen in graphic 16 and we found that there is evidence that the slopes of the increase in sales price based on Neighborhoods and Square footage is not 0 (p-value = 0.000214)

After looking at various metrics we conclude that Sales Price and its relationship to square footage depends on the neighborhood and the change of sales prices dependent on the square footage is not equal for each neighborhood

Parameters

Looking at graphic 11 and graphic 17, we can see the parameter estimates and their confidence intervals for our model that has an interaction variable. The following are the interpretations of the estimates and intervals:

Neighborhood BrkSide:

- Log (Sale Price) = 5.91292 + 0.81965 * Log(GrLivArea)
- Every time the square footage of the living area doubles, there is an estimated multiplicative increase of 1.76497775436 (about 76.5% increase) in the median Sale Price.
- Every time the square footage of the living area doubles, the estimated multiplicative increase in median Sales Price for Brookeside is between 1.60081478829 and 1.94597031156.
- When the square footage of the living area is 0, the estimated median SalePrice 369.78435.
- When the square footage of the living area is 0., the estimated median Sale Price is between 137.10639085 and 997.332584908 in Brookside neighborhood.

Neighborhood North Ames:

- Log (Sale Price) = 8.49273 + 0.47303 * GrLivArea
- Every time the square footage of the living area doubles, there is an estimated multiplicative increase of 1.38802158181, (about 38.8% increase) in the median Sale Price.

- Every time the square footage of the living area doubles, the estimated multiplicative increase in median Sales Price for North Ames is between 1.12147868454 and 1.71789873926.
- When the square footage of the living area is 0, the estimated median SalePrice 4879.16803655 .
- When the square footage of the living area is 0, the estimated median Sale Price is between 556.146417272 and 42805.5779923 in North Ames neighborhood.

Neighborhood Edwards:

- Log (Sale Price) = 8.00651 + 0.51965 * GrLivArea
- Every time the square footage of the living area doubles, there is an estimated multiplicative increase of 1.4336074105 (about 43.36% increase) in the median Sale Price.
- When the square footage of the living area is 0, the estimated median Sale Price 3000.42732748.
- When the square footage of the living area is 0, the estimated median Sale Price is between 312.416333335 and 28815.756004 for houses in the Edwards Neighborhood.
- Every time the square footage of the living area doubles, the estimated multiplicative increase in median Sales Price for Edwards Neighborhood is between 1.14827731299 and 1.78988066094.

Conclusion

We preformed a log-log transform on our data and created 3 models to compare. After looking at the adjusted R squared values and ANOVA tables, we found that the best predictor of sales price is based on the neighborhood and the square footage of the living area where the change in sales price can be different for each neighborhood. We found that houses in North Ames tend to have the highest sale price when the square footage is 0 but Brookside has the largest amount of growth as the square footage doubles. The following regression equations for each neighborhood:

Log(Sale Price_BrkSide) = 5.91292 + 0.81965 * Log(GrLivArea)

Log (Sale Price_NAmes) = 8.49273 + 0.47303 * Log(GrLivArea)

Log (Sale Price_Edwards) = 8.00651 + 0.51965 * Log(GrLivArea)

Rshinny App: https://haitiel.shinyapps.io/git2/

Analysis Question 2:

Problem Statement:

We want to build the most predictive model for sales prices of homes in all of Ames Iowa. We will compare 3 automatic selection technique, forward, backward, stepwise selection, then we will build our custom model to try and get the most accurate model.

Data Transformation:

First, to transform the data, we divided our data into categorical and numeric and within the categorical data, we also sub divided them into ordinal and non-ordinal. We also checked columns with rare observations, for example Utilities has 1459 out of 1460 in one category and only one observation has another category. Since this will not provide much information on predicting the sale price, we choose to remove that column and other columns with similar situations that introduce extra noise in the dataset. We deleted the following columns: Street, Utilities,Condition2, MiscFeature, Electrical.

We used graphic 18 to check for collinearity because columns that are highly correlated with one another also introduce redundancy in the data set, therefore, we choose to delete the following columns that have correlation rate above 0.7: 'X1stFlrSF', 'GarageCars', 'GarageYrBlt', 'FireplaceQu', 'GarageCond', 'PoolQC', 'BsmtFinType1', 'BsmtFinType2'.

Finally, in graphic 19 we see that there are two highly influential points in the dataset, observation 1299 and 524, as we look closely, both observations have disproportionally large living area, but low sale price, we think that it is likely these data points could be introduced by error and practically will not be helpful for the model we are providing, we choose to delete both observation 524 and 1299.

Model Selection

After removing noise and redundancies in the data set, we are ready to run our forward selection. Below, we have the columns that each method selected.

Forward Selection (Columns):

- MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea ExterQual ExterCond BsmtQual BsmtCond BsmtExposure BsmtFinSF1 BsmtFinSF2 TotalBsmtSF GrLivArea BedroomAbvGr KitchenAbvGr KitchenQual GarageArea WoodDeckSF OpenPorchSF X3SsnPorch ScreenPorch MoSold LotConfig Neighborhood Condition1 BldgType MasVnrType Foundation Functional GarageType SaleCondition

Backward Selection (Columns):

- MSSubClass LotFrontage LotArea LandSlope OverallQual OverallCond YearBuilt MasVnrArea ExterQual ExterCond BsmtQual BsmtCond BsmtExposure BsmtFinSF1 BsmtFinSF2 BsmtUnfSF HeatingQC X2ndFlrSF LowQualFinSF GrLivArea FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Fireplaces GarageArea GarageQual WoodDeckSF OpenPorchSF X3SsnPorch ScreenPorch PoolArea MoSold MSZoning LandContour LotConfig Neighborhood Condition1 BldgType HouseStyle RoofMatl Exterior1st MasVnrType Foundation Functional GarageType SaleType SaleCondition

Stepwise Selection (Columns):

- MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea ExterQual ExterCond BsmtQual BsmtCond BsmtExposure BsmtFinSF1 BsmtFinSF2 TotalBsmtSF HeatingQC GrLivArea BedroomAbvGr KitchenAbvGr KitchenQual GarageArea WoodDeckSF OpenPorchSF X3SsnPorch ScreenPorch MoSold LotConfig Neighborhood Condition1 BldgType Exterior1st MasVnrType Foundation Functional GarageType SaleCondition

Checking Assumptions

Forward Selection:

- Residuals (Graphic 20): The residuals are in a random cloud formation indicating equal variance.
- Q-Q plots (Graphic 21): Mostly in a straight line, presume normally distributed residuals.
- Histogram of the residuals (Graphic 22): Histogram looks normally distributed.
- Studentized residuals/Cook's D (Graphic 23): No extreme influential points
- We can see that the residuals plots present itself in a random cloud formation, q-q plot is aligned with a straight line, a little deviation on the end, histogram of the residuals are normally distributed, no extremely influential values looking at Cook's D, therefore all assumptions of regressions are met, finally we assume independence from and between each observation.

Backward Selection:

- Residuals (Graphic 24): The residuals are in a random cloud formation indicating equal variance.

- QQ Plot (Graphic 25) : Mostly in a straight line, presume normally distributed residuals
- Histogram of residuals (Graphic 26): Histogram of the residuals looks normal
- Studentized residuals and cook's d (Graphic 27) : No extremely influential points
- We can see that the residuals plots present itself in a random cloud formation, q-q plot is aligned with a straight line, a little deviation on the end, histogram of the residuals are normally distributed, no extremely influential values looking at Cook's D, therefore all assumptions of regressions are met, finally we assume independence from and between each observation.

Stepwise Selection:

- Residuals (Graphic 28): The residuals are in a random cloud indicating equal variance.
- Q-Q plots ((Graphic 29) : Q-Q plots presents a slight deviation, will proceed with caution, assume normally distributed residuals
- Histogram of Residuals (Graphic 30) : Histogram of the residual looks normally distributed
- Studentized residuals and Cook's D (Graphic 31): No extremely influential points
- We can see that the residuals plots present itself in a random cloud formation, q-q plot is aligned with a straight line, a little deviation on the end, histogram of the residuals are normally distributed, no extremely influential values looking at Cook's D, therefore all assumptions of regressions are met, finally we assume independence from and between each observation.

Comparing Competing Models

Forward Selection:
Diagnostics are in Graphic 32

Model Selection	R square	CV Press	Kaggle Score
Forward	0.9158	9.59E11	0.14739

Backward Selection:
Diagnostics are in Graphic 33

Model Selection	R square	CV Press	Kaggle Score
Backward	0.9194	1.005E12	0.14739

Stepwise Selection:
Performing a Stepwise selection based in SAS. Results are as follows:
Diagnostics are in Graphic 34

Model Selection	R square	CV Press	Kaggle Score
Stepwise	0.9120	9.673E11	0.19016

Model Selection (Custom)

For the final custom model, we combined all the models above and optimized using P-values and the interactions between variables. We did this by re-conducting our forward, backward, and stepwise selections on P values with two-term interactions and three-term interactions. Afterwards, we used trial and error on our variable candidate to select the variables in our final model.
Final Feature Selection

- MSZoning:GrLivArea + LotArea: BsmtFinSF1 + KitchenAbvGr: GarageQual + GrLivArea:LotArea + LotArea + OverallQual + OverallCond + YearBuilt + BsmtExposure + BsmtFinSF1 + BsmtFinSF2 + HeatingQC+BsmtUnfSF + GrLivArea + HalfBath +

KitchenAbvGr + KitchenQual + GarageArea + GarageQual + WoodDeckSF + ScreenPorch + MSZoning + Neighborhood + BldgType + MasVnrType + Foundation + Functional + GarageType + SaleCondition + Fireplaces

Assumptions:

- Residuals plot (Graphic 35): The residuals are in a random cloud formation indicating equal variance
- Q-Q plot (Graphic 36): Sightly deviated in the beginning, but mostly in a straight line, indicating normally distributed residuals
- Histogram of the residuals (Graphic 37): Residuals are normally distributed
- Studentized residuals/Cook's D (Graphic 38): No extremely influential points

Our model has met the assumptions of multiple linear regression so can continue. The parameter estimates and model evaluations are in graphic 39. Our final model had a ranking of #1288 at the time of the submission and a score of 0.13734 which can be seen in graphic 40.

Conclusion

The following is statistical measure from the forward, backward, stepwise and custom selections:

Model Selection	R square	CV Press	Kaggle Score
Forward	0.9158	9.59E11	0.14739
Backward	0.9194	1.005E12	0.14739
Stepwise	0.9120	9.673E11	0.19016
Custom	0.9343	1.03E12	0.13734

After running multiple selection methods, we learned that it is crucial to test different selection metrics, for instance, on AIC, BIC or CV Press. Sometimes a slight change of input will result in very different outcomes. For selection method, we have used our best intuition to pick the least P values that was given by the system, two-term, three-term interactions, and even though some of them had small P values but was deemed not helpful on predicting more accurate Sale price.

We found that MSZoning, Girliving and its two-term interactions with other variables had the most significantly increase in our score. However, currently our testing on three-term interactions has not yet been proved to increase our score (Maybe due to overfitting). We have also found that deleting variables, sometimes significant variables, will help us predict more accurately, we think that this is due to overfitting.

GitHubs

Anishka's: anishkapeter.github.io

Haitie Liu's https://haitieliu.github.io/

Appendix

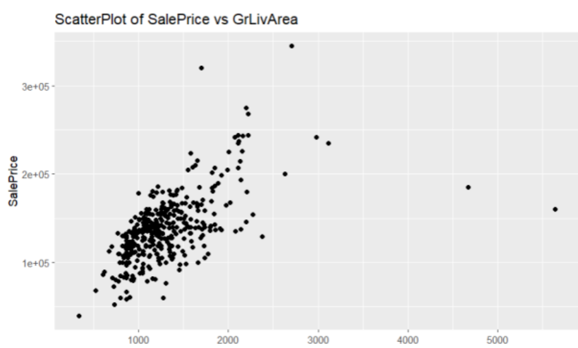
Graphic 1 (Building Model without Interaction or Indicator Var. on Original Data):

```
# Model without Interaction Variable
filtered_neighborhood = data %>% filter (Neighborhood %in% c('NAmes', 'Edwards', 'BrkSide'))

filtered_neighborhood_model= lm(SalePrice ~ GrLivArea, data = filtered_neighborhood)
summary(filtered_neighborhood_model)
```

```
Call:
lm(formula = SalePrice ~ GrLivArea, data = filtered_neighborhood)

Residuals:
    Min      1Q  Median      3Q     Max 
-177619  -17918     919   15227  163722 

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)    
(Intercept) 78205.578   4536.054   17.24 <0.0000000000000002 ***
GrLivArea      45.979      3.265   14.08 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30980 on 381 degrees of freedom
Multiple R-squared:  0.3423,	Adjusted R-squared:  0.3406 
F-statistic: 198.3 on 1 and 381 DF,  p-value: < 0.00000000000000022
```

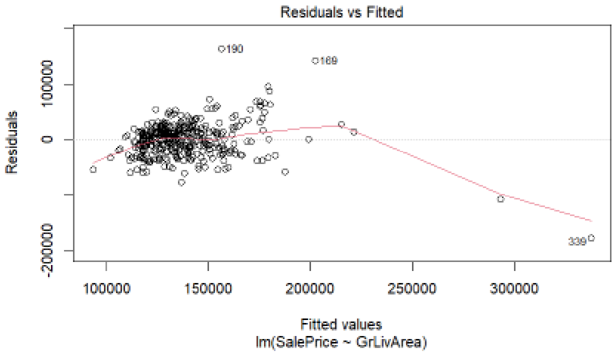
Graphic 2 (Checking Linearity Assumption of Original Data):

```
ggplot(filtered_neighborhood,aes(x =GrLivArea , y =SalePrice )) +
  geom_point() +
  ggtitle("ScatterPlot of SalePrice vs GrLivArea")
```



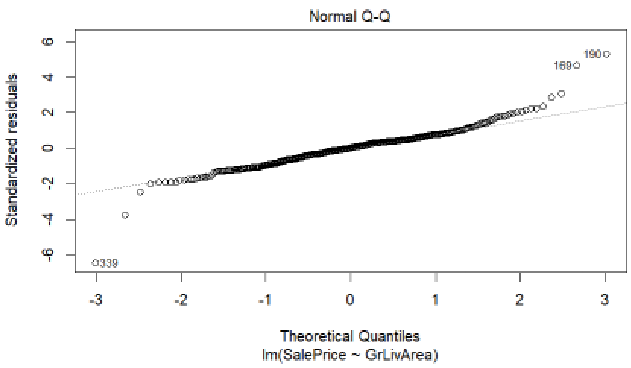
Graphic 3 (Checking residuals with original data):

```
# Check Assumptions Model Without Indicator and Interaction Variable
plot(filtered_neighborhood_model)
```



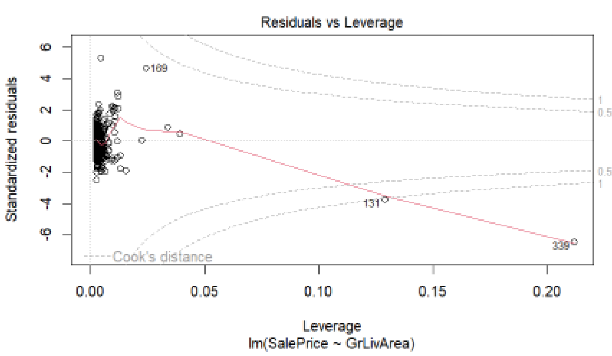
Graphic 4 (Checking Normality with Original Data):

```
# Check Assumptions Model Without Indicator and Interaction Variable
plot(filtered_neighborhood_model)
```



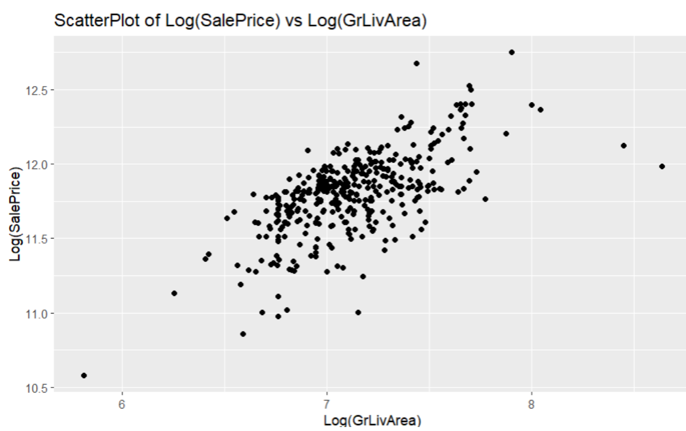
Graphic 5 (Checking Residuals and Cook's Distance of Original Data):

```
# Check Assumptions Model Without Indicator and Interaction Variable
plot(filtered_neighborhood_model)
```



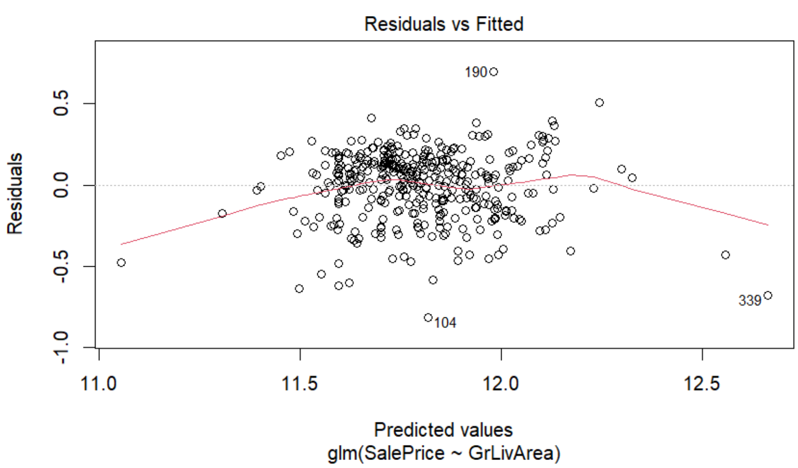
Graphic 6 (Checking Linearity of Log-Log Transformed Data):

```
# Checking Assumptions
ggplot(filtered_data2,aes(x =GrLivArea , y =SalePrice )) +
  geom_point()+ggtitle("ScatterPlot of Log(SalePrice) vs Log(GrLivArea)") +
  xlab("Log(GrLivArea)") +
  ylab("Log(SalePrice)")
```



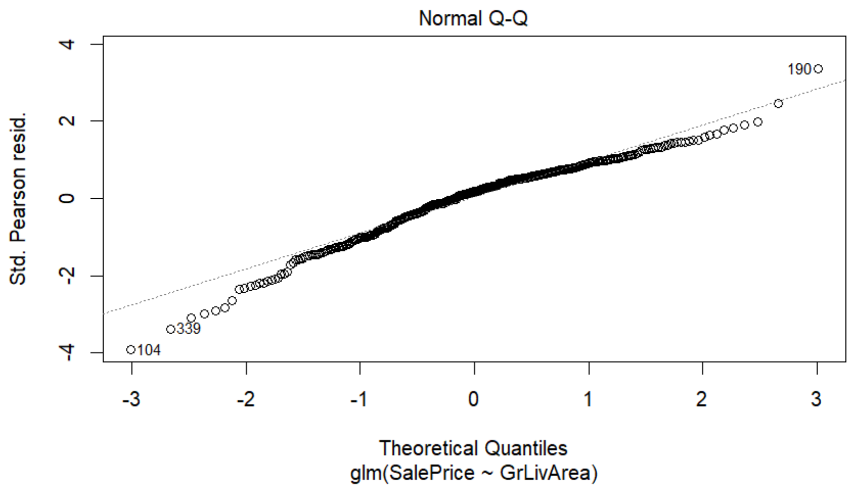
Graphic 7 (Checking Residual Plot of Log-Log Transformed Data):

```
# Checking Assumptions
plot(filtered_data2_model)
```



Graphic 8 (Checking Normality on Log-Log Transformed Data):

```
# Checking Assumptions
plot(filtered_data2_model)
```



Graphic 9 (Log-Log Transform and Model without Neighborhood Variable):

```
# Log log Transform to data
data3=data
data3$SalePrice = log(data$SalePrice)
data3$GrLivArea = log(data$GrLivArea)

# filter the logged data to only include the 3 neighborhoods of interest
filtered_data2 <- data3 %>% filter(Neighborhood %in% c('NAmes', 'Edwards', 'BrkSide'))

# build model on log data without interaction and indicator var
filtered_data2_model = lm(SalePrice ~ GrLivArea, data = filtered_data2)
summary(filtered_data2_model)
```

```
> summary(filtered_data2_model)

Call:
lm(formula = SalePrice ~ GrLivArea, data = filtered_data2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.81505 -0.11973  0.03726  0.14141  0.69657 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept)  7.75338    0.24360   31.83 <0.0000000000000002 ***
GrLivArea    0.56824    0.03418   16.62 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2085 on 381 degrees of freedom
Multiple R-squared:  0.4204,	Adjusted R-squared:  0.4188 
F-statistic: 276.3 on 1 and 381 DF,  p-value: < 0.00000000000000022
```

Graphic 10 (Model on Log-Log Transform With Neighborhood as Indicator) :

```
# Adding Indicator Variable
filtered_data2_model3 <- lm(SalePrice ~  GrLivArea + Neighborhood, data = filtered_data2)
summary(filtered_data2_model3)
```

```
Call:
lm(formula = SalePrice ~ GrLivArea + Neighborhood, data = filtered_data2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72154 -0.10592  0.02469  0.11565  0.79364 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         7.76936    0.22919  33.900  < 2e-16 ***
GrLivArea           0.55579    0.03237  17.171  < 2e-16 ***
NeighborhoodEdwards -0.02044    0.03252  -0.629     0.53    
NeighborhoodNAmes    0.13279    0.02906   4.569 6.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1961 on 379 degrees of freedom
Multiple R-squared:  0.4897,	Adjusted R-squared:  0.4857 
F-statistic: 121.2 on 3 and 379 DF,  p-value: < 2.2e-16
```

Graphic 11 (Model with Neighborhood as Interaction Variable for Log-Log Transformed Data):

```
# Fit the model
filtered_data2_model2 <- lm(SalePrice ~  GrLivArea + GrLivArea * Neighborhood, data =
filtered_data2)

# Summary of the model
summary(filtered_data2_model2)
```

```
Call:
lm(formula = SalePrice ~ GrLivArea + GrLivArea * Neighborhood, 
    data = filtered_data2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.72080 -0.10353  0.02184  0.10586  0.80470 

Coefficients:
                              Estimate Std. Error t value             Pr(>|t|)    
(Intercept)                    5.91292    0.50459  11.718 < 0.0000000000000002 ***
GrLivArea                      0.81965    0.07163  11.443 < 0.0000000000000002 ***
NeighborhoodEdwards            2.09359    0.64589   3.241               0.0013 ** 
NeighborhoodNAmes              2.57981    0.59988   4.301            0.0000217 ***
GrLivArea:NeighborhoodEdwards -0.29998    0.09122  -3.289               0.0011 ** 
GrLivArea:NeighborhoodNAmes   -0.34662    0.08482  -4.087            0.0000535 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1923 on 377 degrees of freedom
Multiple R-squared:  0.5121,	Adjusted R-squared:  0.5056 
F-statistic: 79.14 on 5 and 377 DF,  p-value: < 0.00000000000000022
```

Graphic 12:

```
#ANOVA for model without Neighborhood
anova(filtered_data2_model)
```

```
Analysis of Variance Table

Response: SalePrice
           Df Sum Sq Mean Sq F value               Pr(>F)    
GrLivArea   1 12.008 12.0084  276.32 < 0.00000000000000022 ***
Residuals 381 16.558  0.0435                                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Graphic 13:

```
#ANOVA for model with Neighborhood as Indicator Variable
anova(filtered_data2_model2)
```

```
Analysis of Variance Table

Response: SalePrice
                      Df  Sum Sq Mean Sq  F value    Pr(>F)    
GrLivArea              1 12.0938 12.0938 328.3171 < 2.2e-16 ***
Neighborhood           2  1.8952  0.9476  25.7246 3.381e-11 ***
GrLivArea:Neighborhood 2  0.6192  0.3096   8.4053 0.0002685 ***
Residuals            376 13.8503  0.0368                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Graphic 14:

```
#ANOVA for model with Neighborhood as Interaction Variable
anova(filtered_data3_model3)
```

```
Analysis of Variance Table

Response: SalePrice
              Df  Sum Sq Mean Sq F value    Pr(>F)    
GrLivArea      1 12.0938 12.0938 315.938 < 2.2e-16 ***
Neighborhood   2  1.8952  0.9476  24.755 7.893e-11 ***
Residuals    378 14.4695  0.0383                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Graphic 15:

		DF	Sum of Squares	Mean Square	F-value	P-value
	Model	2	5.0342	2.5171	68.02973	< .00001
Neighborhood as Indicator	Error	379	14.5773	0.037		
Without Neighborhoods	Total	381	19.6115			

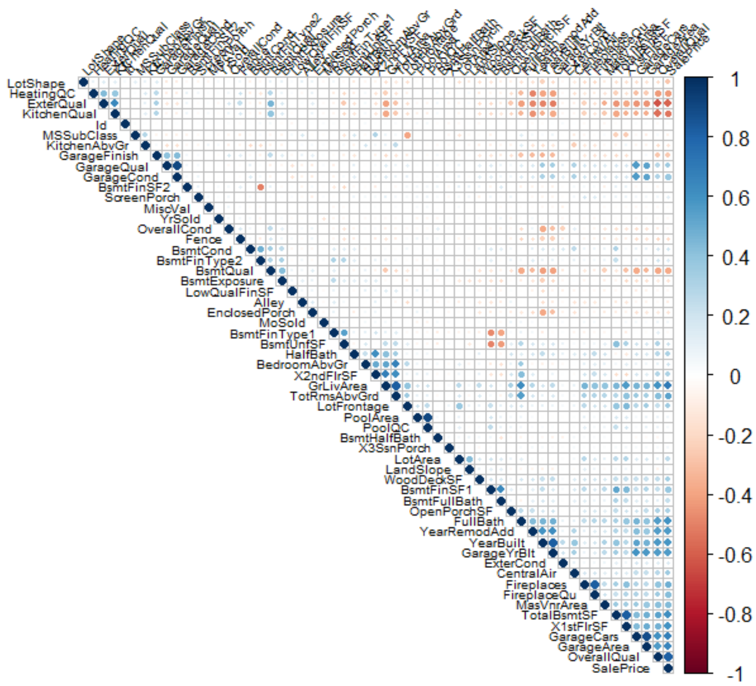
Graphic 16:

		DF	Sum of Squares	Mean Square	F-value	P-value
	Model	2	0.6395	0.31975	8.641892	0.000214
Neighborhood as Interaction	Error	377	13.9378	0.037		
Neighborhood as Indicator	Total	379	14.5773			

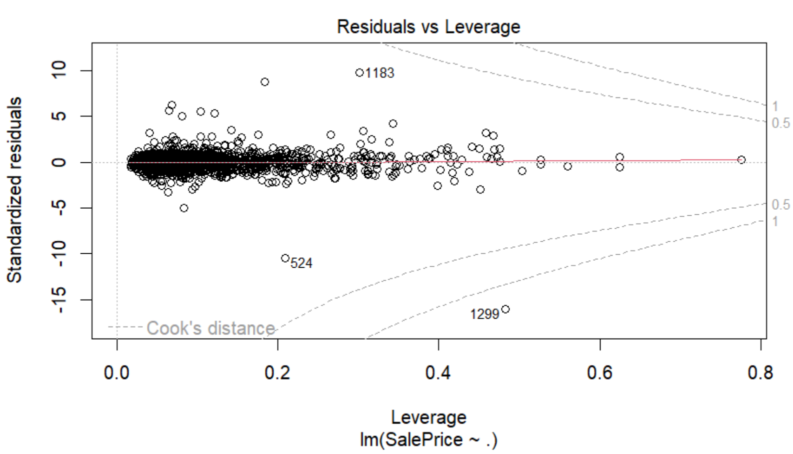
Graphic 17:

```
> confint(filtered_data2_model2)
                                     2.5 %     97.5 %
(Intercept)                      4.9207572  6.9050843
GrLivArea                        0.6788064  0.9604897
NeighborhoodEdwards              0.8235795  3.3635933
NeighborhoodNAmes                1.4002744  3.7593394
GrLivArea:NeighborhoodEdwards   -0.4793353 -0.1206263
GrLivArea:NeighborhoodNAmes     -0.5134042 -0.1798447
```

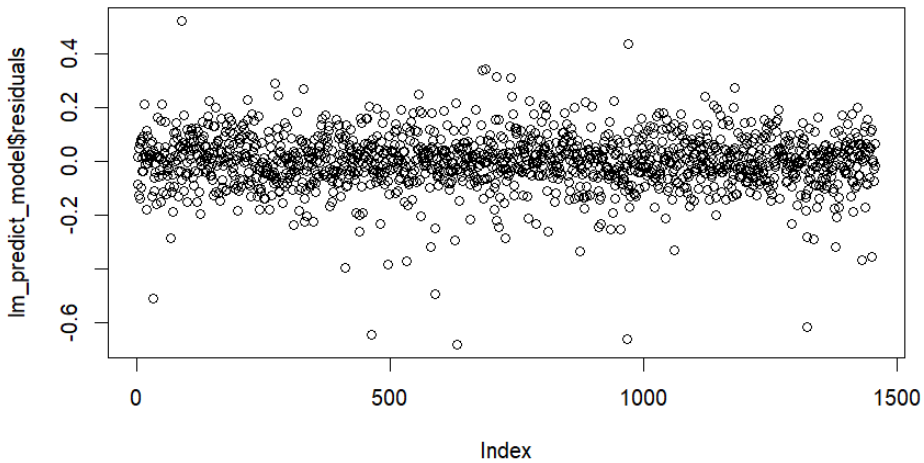
Graphic 18 (Data transformation (Checking for Collinearity)):



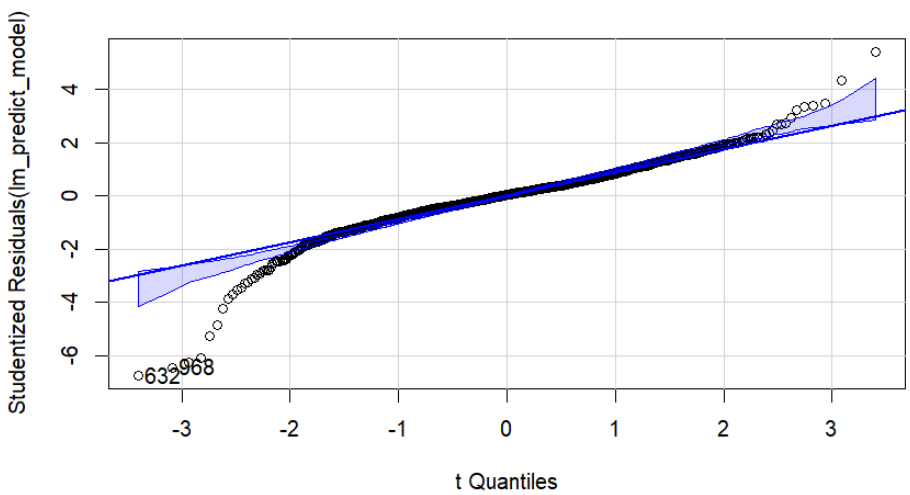
Graphic 19 Data transformation (Checking for influential points):



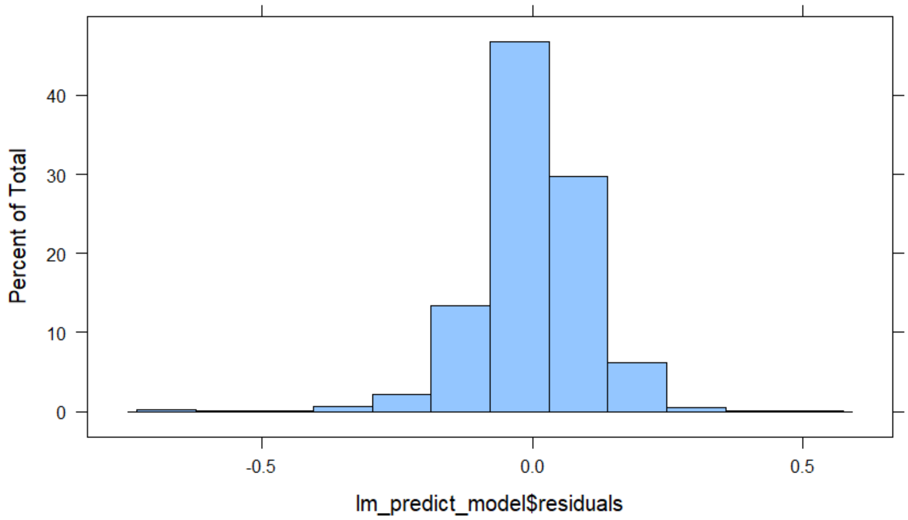
Graphic 20 (Residuals):



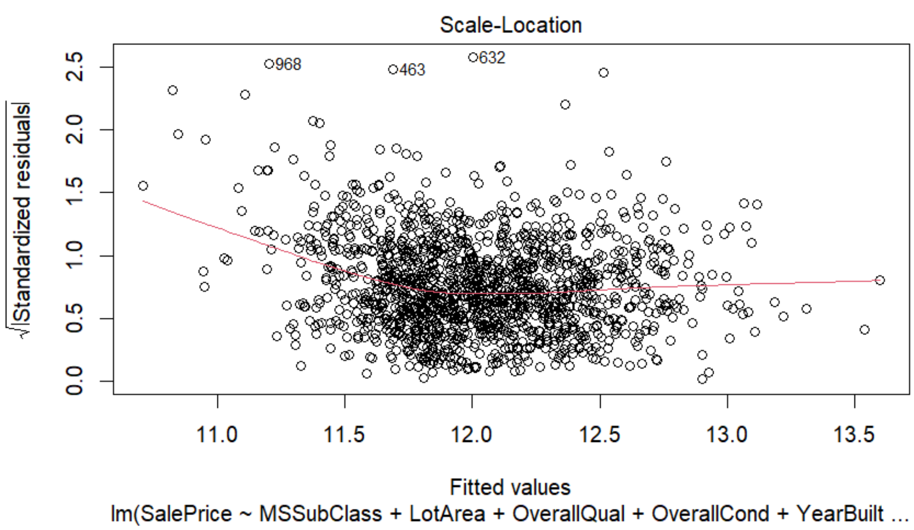
Graphic 21 (QQ-plot):

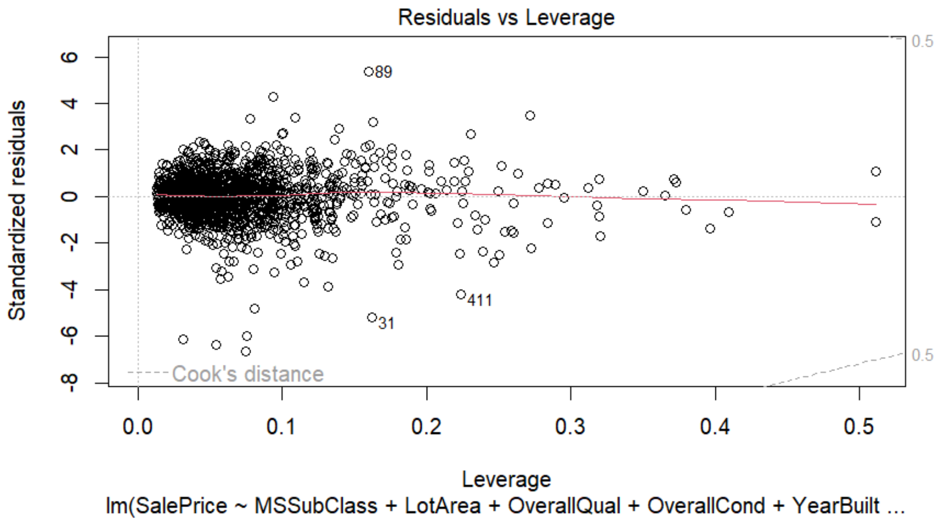


Graphic 22 (Histogram):

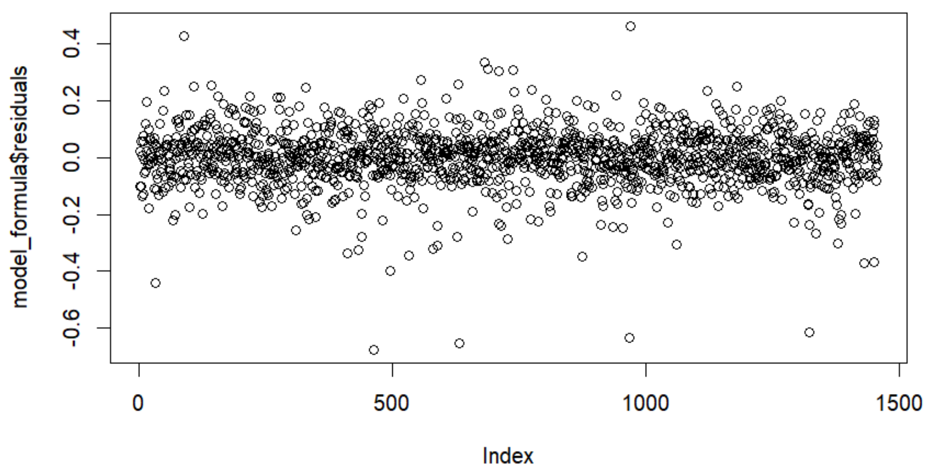


Graphic 23 (Studentized Residual and Cooks D Plot):

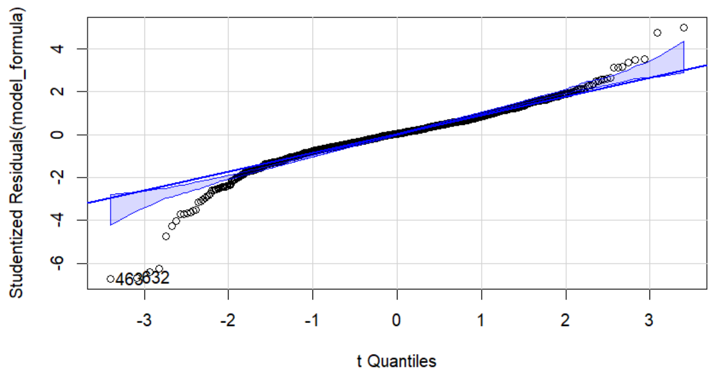




Graphic 24 (Residuals):

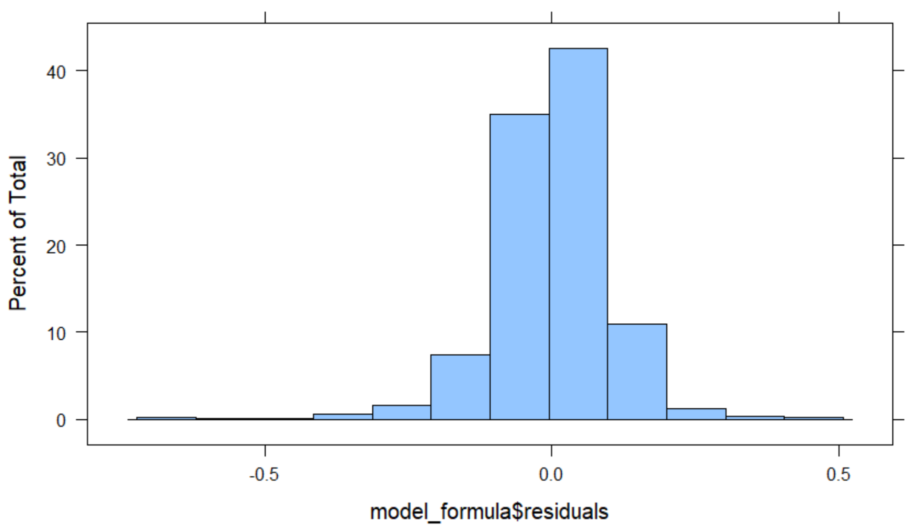


Graphic 25 (QQ-Plot):

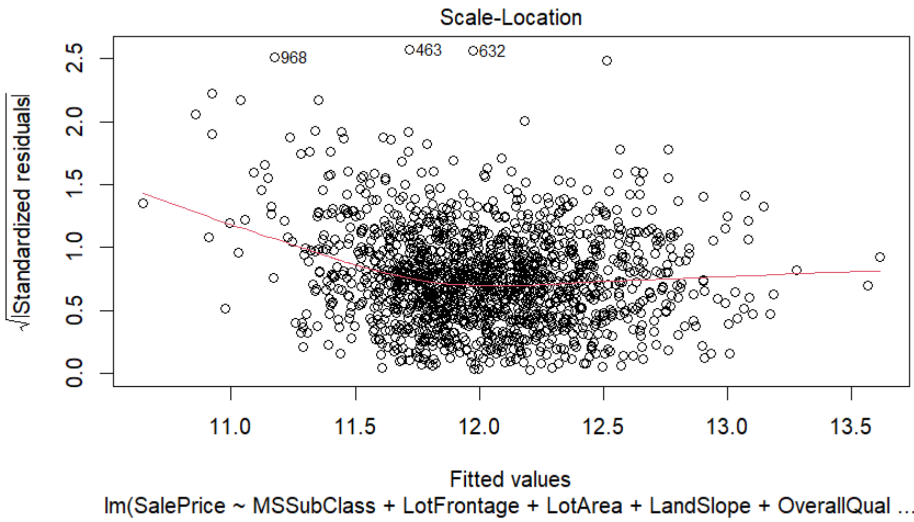


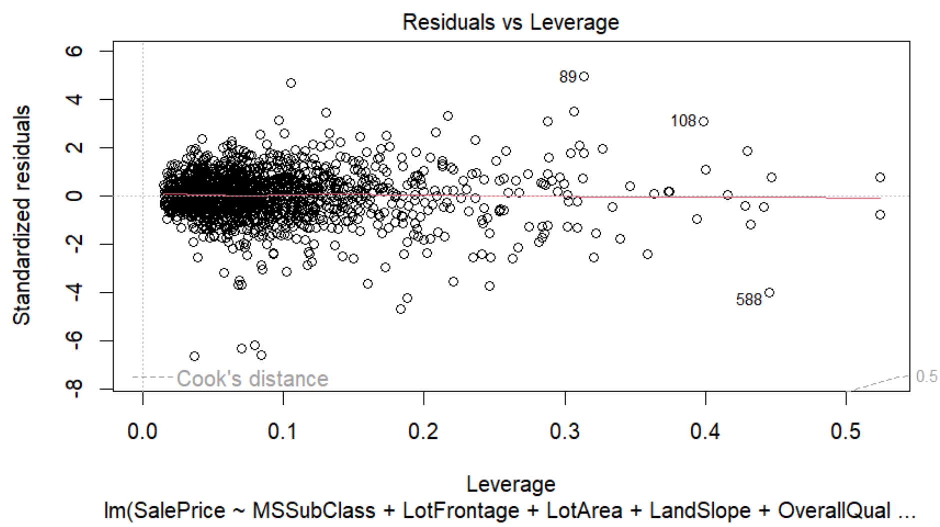
:

Graphic 26 (Histogram):

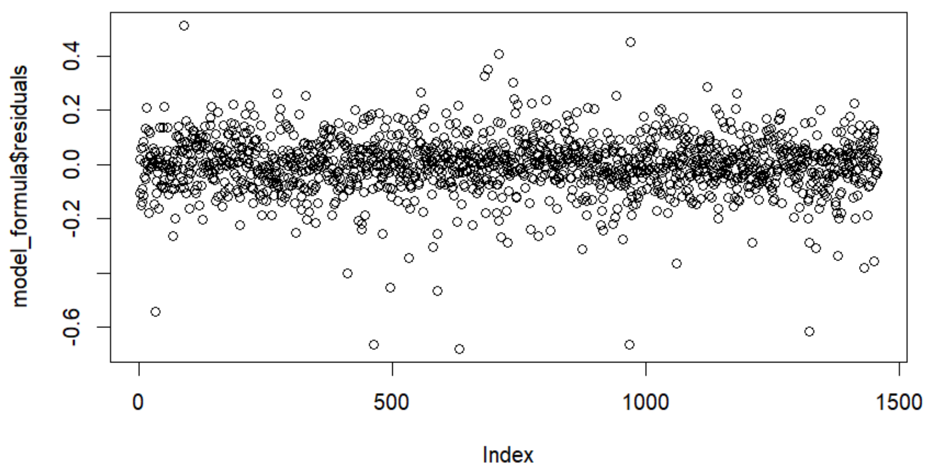


Graphic 27 (Cooks Plot and Studentized Residual):

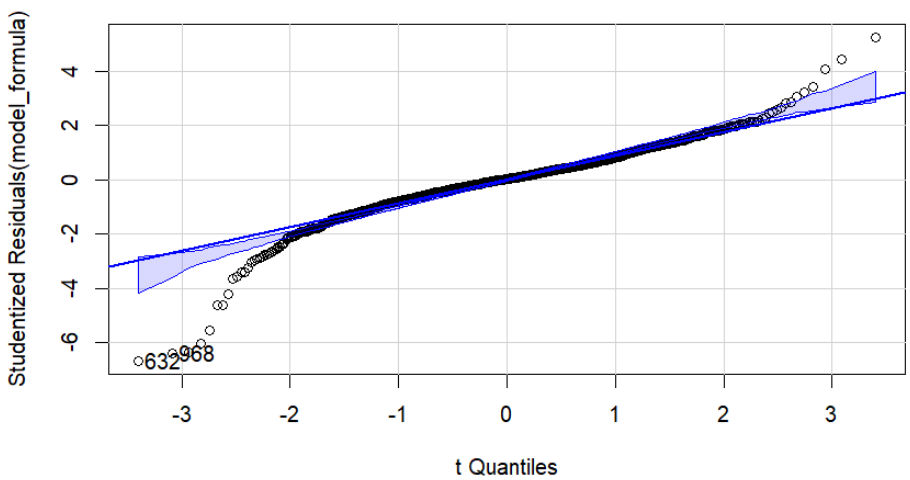




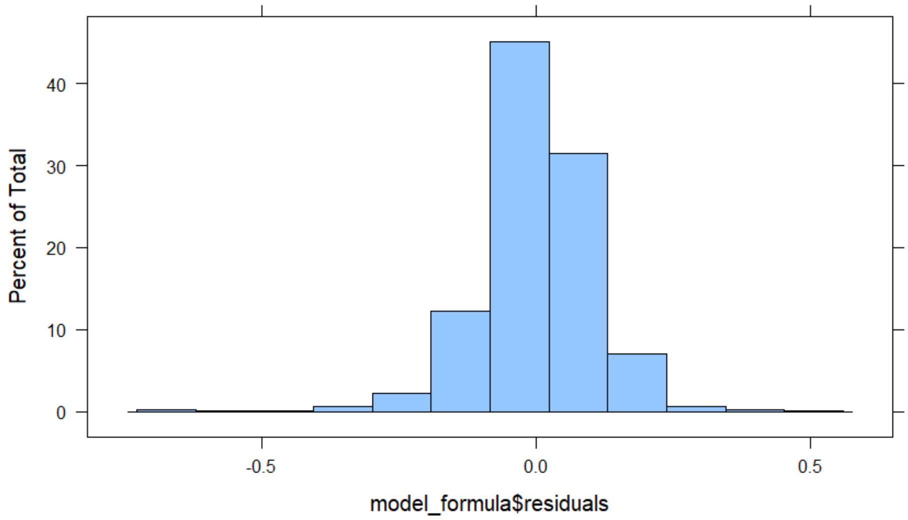
Graphic 28 (Residuals):



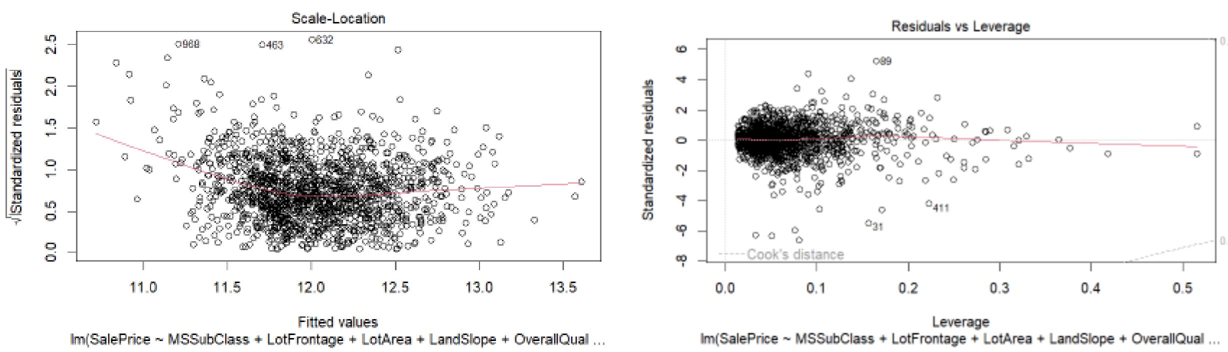
Graphic 29 (QQ-Plot):



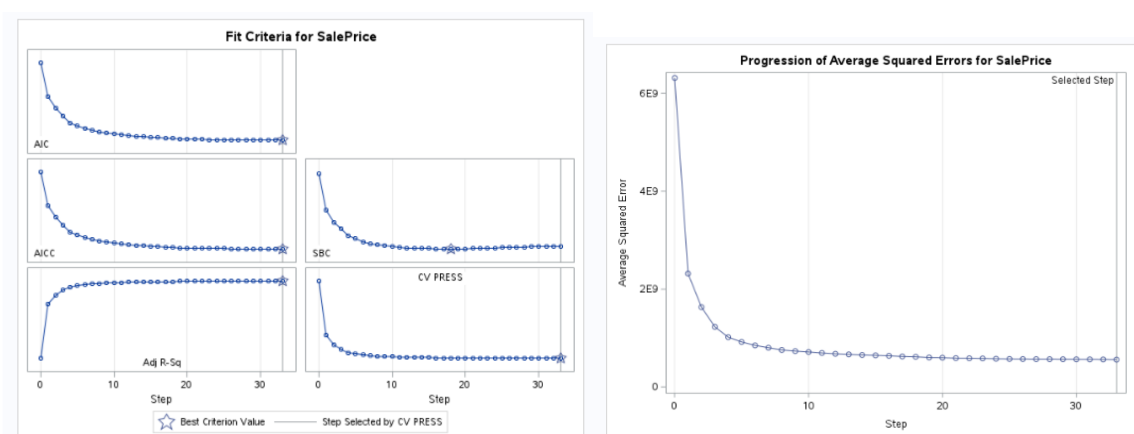
Graphic 30 (Histogram):



Graphic 31 (Studentized Residuals and Cooks D):



Graphic 32 (Forward Selection Diagnostics):



Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	103	8.432577E12	81869675011	143.06
Error	1354	7.748823E11	572291176	
Corrected Total	1457	9.207459E12		

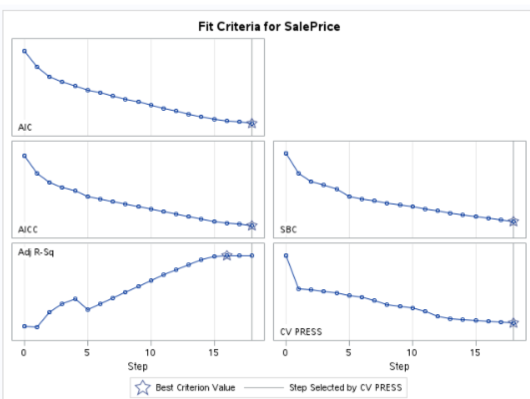
Root MSE	23923
Dependent Mean	180933
R-Square	0.9158
Adj R-Sq	0.9094
AIC	30961
AICC	30977
SBC	30051
CV PRESS	9.590188E11

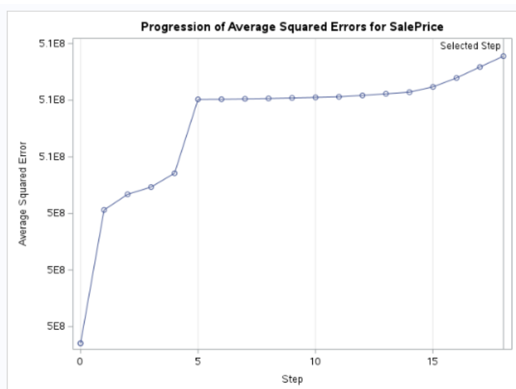
Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1307	151	1.35375E11
2	1302	156	1.34489E11
3	1319	139	7.16042E10
4	1323	135	5.75817E10
5	1292	166	9.39722E10
6	1314	144	6.36725E10
7	1327	131	9.33672E10
8	1325	133	1.02286E11
9	1298	160	1.19108E11
10	1315	143	8.75634E10
Total			9.59019E11

submission_stepwise.csv
Complete · 1h ago

0.19016

Graphic 33 (Backward Selection Diagnostic):





Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	128	8.465081E12	66133448631	118.39
Error	1329	7.423774E11	558598460	
Corrected Total	1457	9.207459E12		

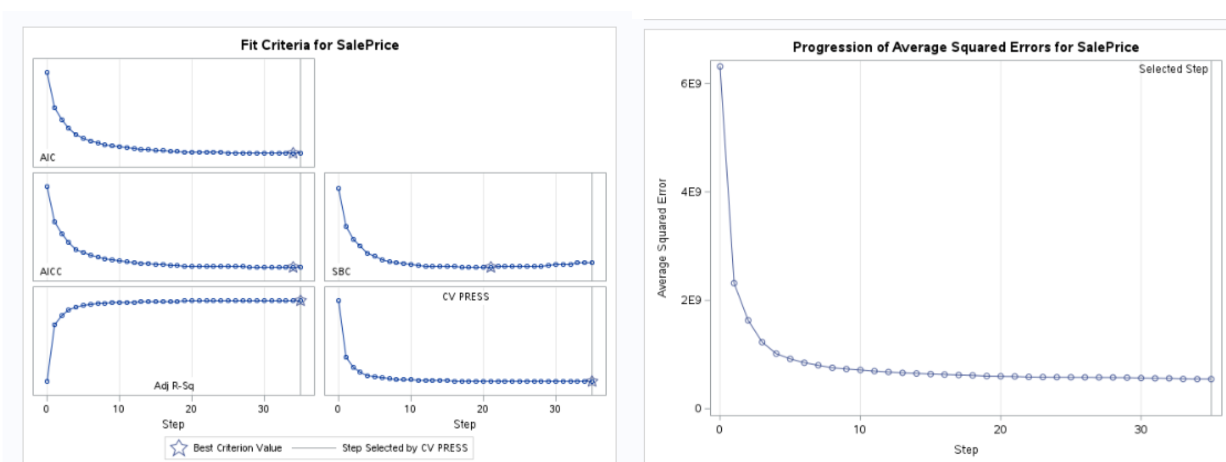
Root MSE	23635
Dependent Mean	180933
R-Square	0.9194
Adj R-Sq	0.9116
AIC	30948
AICC	30974
SBC	30170
CV PRESS	1.005104E12

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1327	131	6.87666E10
2	1303	155	1.07679E11
3	1315	143	8.22222E10
4	1322	136	8.73008E10
5	1313	145	9.11565E10
6	1306	152	1.67624E11
7	1318	140	1.02252E11
8	1313	145	1.20718E11
9	1309	149	1.06042E11
10	1296	162	7.13432E10
Total			1.0051E12

submission_Backward.csv 0.14739

Complete · 1h ago

Graphic 34 (Stepwise Diagnostics):

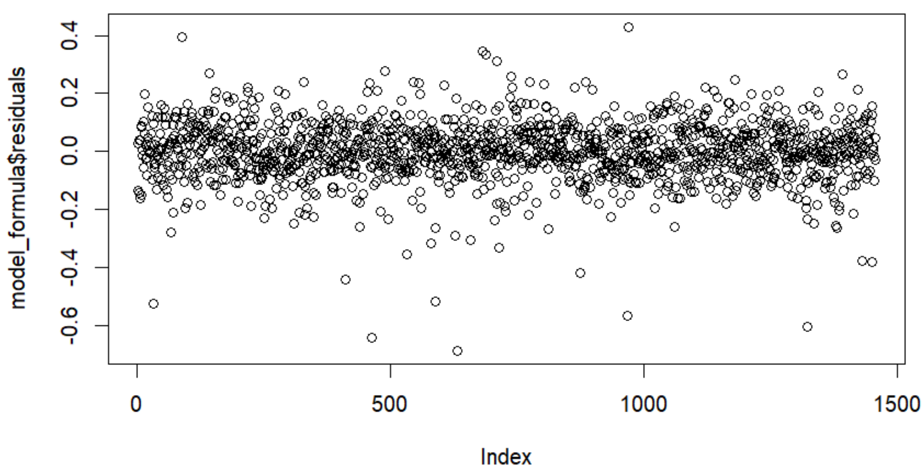


Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	87	8.397425E12	96522128007	163.25
Error	1370	8.100336E11	591265432	
Corrected Total	1457	9.207459E12		

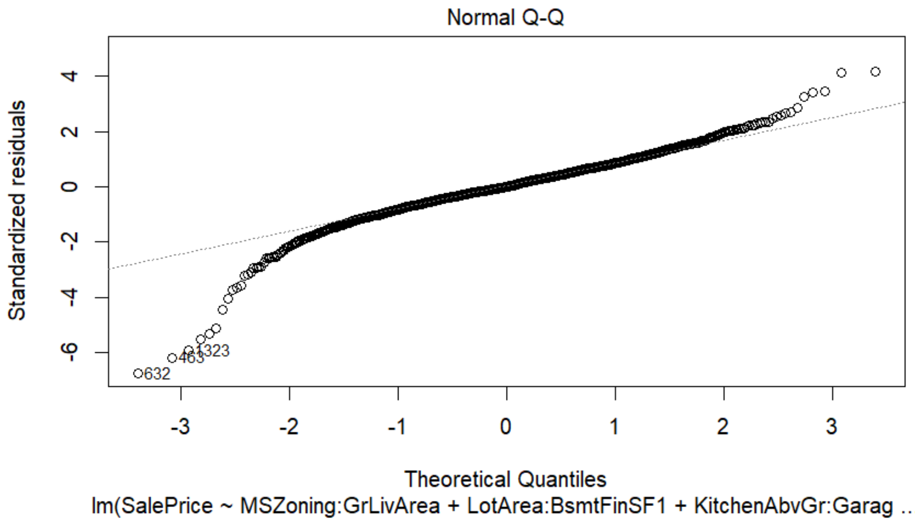
Root MSE	24316
Dependent Mean	180933
R-Square	0.9120
Adj R-Sq	0.9064
AIC	30994
AICC	31005
SBC	29999
CV PRESS	9.673007E11

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1317	141	7.61862E10
2	1326	132	1.09391E11
3	1317	141	8.56306E10
4	1328	130	6.64569E10
5	1303	155	7.15109E10
6	1292	166	8.08206E10
7	1303	155	8.3059E10
8	1315	143	9.25525E10
9	1319	139	8.36981E10
10	1302	156	2.17995E11
Total			9.67301E11

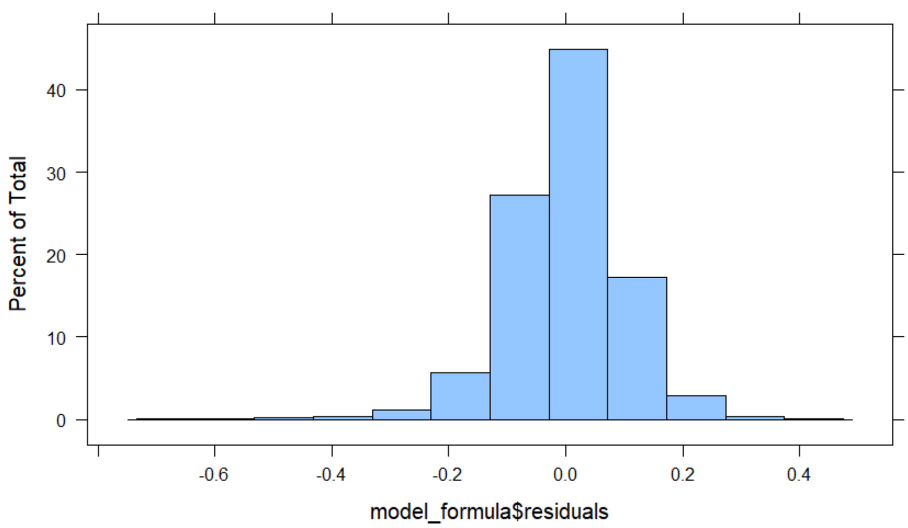
Graphic 35 (Residuals):



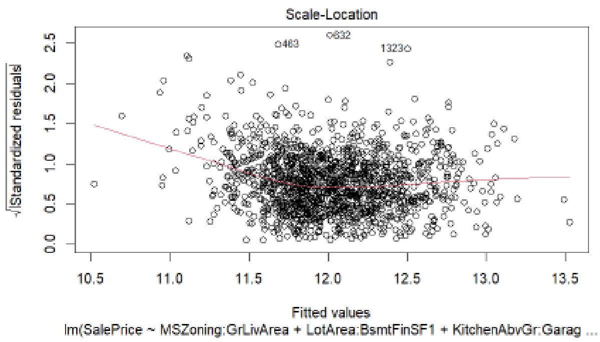
Graphic 36 (QQ-Plot):

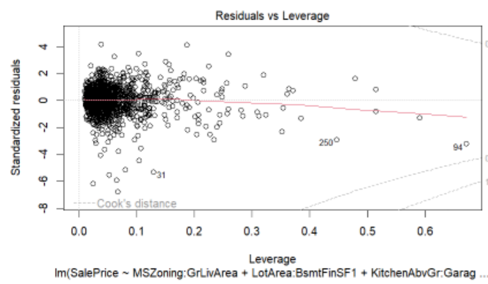


Graphic 37 (Histogram):



Graphic 38 (Studentized Residual and Cook's D) :





Graphic 39 (Custom Model Parameter Estimate):

```
Call:
lm(formula = SalePrice ~ MSZoning:GrLivArea + LotArea:BsmtFinSF1 + 
    KitchenAbvGr:GarageQual + GrLivArea:LotArea + LotArea + OverallQual + 
    OverallCond + YearBuilt + BsmtExposure + BsmtFinSF1 + BsmtFinSF2 + 
    HeatingQC + BsmtUnfSF + GrLivArea + HalfBath + KitchenAbvGr + 
    KitchenQual + GarageArea + GarageQual + WoodDeckSF + ScreenPorch + 
    MSZoning + Neighborhood + BldgType + MasVnrType + Foundation + 
    Functional + GarageType + SaleCondition + Fireplaces, data = data2)
```

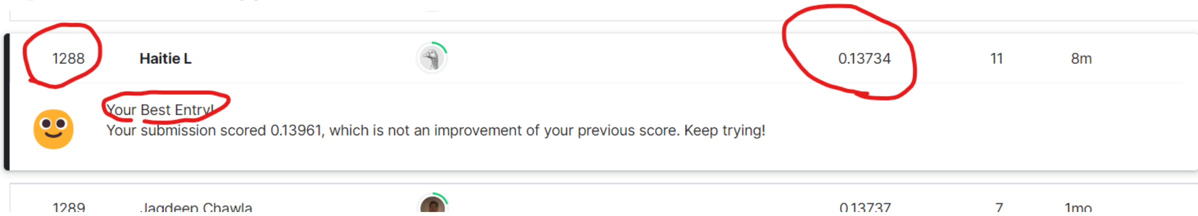
```
Residual standard error: 0.1054 on 1375 degrees of freedom
Multiple R-squared:  0.9343,	Adjusted R-squared:  0.9304 
F-statistic: 238.6 on 82 and 1375 DF,  p-value: < 2.2e-16
```

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	49	8.27213E12	1.68819E11	254.13
Error	1408	9.353288E11	664295993	
Corrected Total	1457	9.207459E12		

Root MSE	25774
Dependent Mean	180933
R-Square	0.8984
Adj R-Sq	0.8949
AIC	31127
AICC	31131
SBC	29932
CV PRESS	1.037438E12

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1307	151	9.4348E10
2	1301	157	8.18579E10
3	1318	140	1.44722E11
4	1324	134	7.48336E10
5	1297	161	8.28589E10
6	1311	147	9.74985E10
7	1303	155	9.46153E10
8	1332	126	7.74615E10
9	1313	145	1.16184E11
10	1316	142	1.73058E11
Total			1.03744E12

Graphic 40 (Final Kaggle Submission):



Appendix (R code)

All R code is included in this document on GitHub

https://github.com/anishkapeter/Stat1Project/blob/main/HouseProjectRmarkdown.Rmd