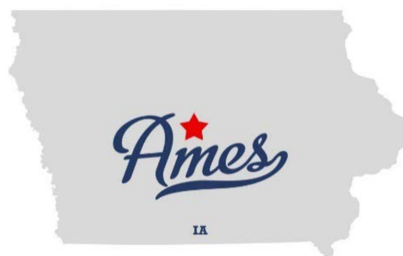
Assignment #5

Automated Variable Selection, Multicollinearity, and Predictive Modeling



Name: Young, Brent

Predict 410 Section #: 57

Quarter: Summer 2017

Introduction

Context

The dataset that we will be working with is called Ames Housing data (includes 2,930 rows) and is observational data collected by Ames Assessor's Office. The data includes houses sold in Ames, Iowa from 2006 to 2010 with SalePrice as the response variable and 81 predictors (includes nominal, ordinal, discrete, and continuous variables). The final goal is to build a Predictive model (e.g., multiple linear regression) to predict SalePrice of a house using other attributes. In order to accomplish this, an iterative regression process focused on statement of the problem, selection of potentially relevant variables, data collection, model specification, parameter estimation, model adequacy checking, model validation and model use will be conducted within the next five weeks.

Objectives/Purpose

The overall purpose/objective of assignment 5 is to begin building regression models for the home sale price by fitting these specific models. We will set up a predictive modeling framework, explore the use of automated variable selection techniques for model identification, assess the predictive accuracy of our model using cross-validation, and compare and contrast the difference between a statistical model validation and an application (or business) model validation. First, a waterfall of my drop conditions with counts will be provided to define the sample data/population of interest that we will want to use for the modeling purpose and ensure that the sample data is representative of the population that we want to model. Second, we will assess model performance by splitting the sample into a 70/30 train/test split, one for in-sample model development and one for out-of-sample model assessment so that we can cross-validate the data (e.g., train each model by estimating the models and test each model by examining predictive accuracy). A table of observation counts for our train/test data partition in our data section will also be shown. Third, we will create a pool of 15-20 candidate predictor variables in combination with using the training data to find the 'best' models using automated variable selection using the techniques: forward, backward, and stepwise variable selection for model identification purposes. We will also make sure that we "like" these variable selection models by using the VIF to assess multicollinearity. We will then determine if the different variable selection procedures selected the same model or different models. Final estimated models and their VIF values for each of these four models will also be displayed. We will then compare the in-sample fit and predictive accuracy of our models and compute adjusted R-Squared, AIC, BIC, mean squared error, and the mean absolute error for each of these models for the training sample and the rank for each model in each metric. Fourth, we will assess how well our model performs (predicts) out-of-sample by computing the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) for the test sample. This will allow us to determine which model fits the best based on these criteria; discussion of these concepts will follow. Fifth, we will validate these models from a business sense using defined cut-off points (e.g., defining PredictionGrades). These prediction grades for the in-sample training data and the out-of-sample test data will be produced so we can determine the accuracy of the models under this definition of predictive accuracy in comparison to our predictive accuracy results (e.g., did model ranking remain the same?). Sixth, after determining the "best" model after all these comparisons, we will then revisit the issues (diagnostics, etc.) and re-fit the "Best" model including all dummy coded variables associated with the categorical variables so that we can report our final model. Lastly, we will reflect upon the challenges presented by the data and the recommendations for improving predictive accuracy.

Section 1: Sample Definition

Figure 1: Boxplot of Sale Price & Building Style

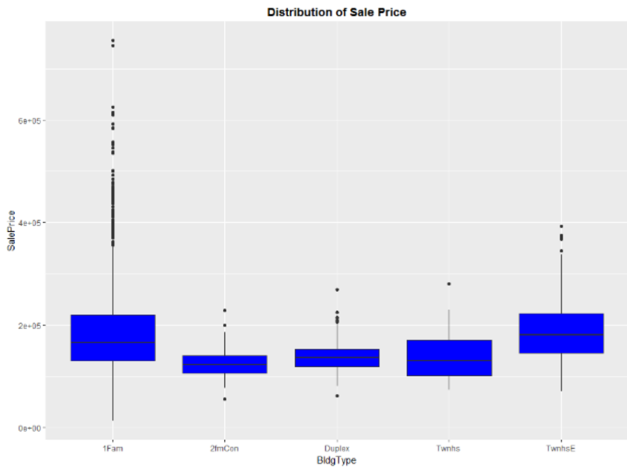


Figure 2: Boxplot of Sale Price & Sale Condition

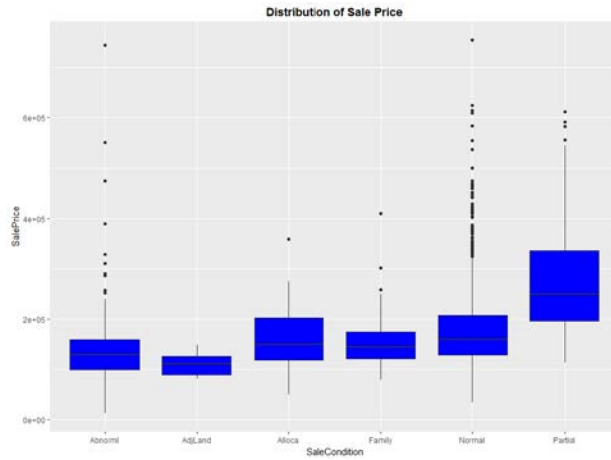


Figure 3: Waterfall of 'Drop Conditions'

```
          1Fam 2fmCon Duplex  Twnhs TwnhsE
 Before   2425     62    109    101    233

          Abnorml AdjLand  Alloca  Family  Normal Partial
Before    190      12      24      46    2413     245

          1Fam 2fmCon Duplex  Twnhs TwnhsE
 After    2002      0      0      0      0

          Abnorml AdjLand  Alloca  Family  Normal Partial
After       0        0       0       0     2002   0
```

Definition of Sample Data & Observations: Figure 1 shows a boxplot of SalePrice & Bldg Type and Figure 2 shows a boxplot of SalePrice & Sale Condition. When comparing figure 1 & 2, 'single-family' homes and 'normal' sale have similar medians as well as the amount and location of the outliers. As a result, based on this, it makes sense for the sample population/data of interest for 'typical' homes in Ames, Iowa to be 'single-family' homes with 'normal' sales in Ames, Iowa. Figure 3 shows the population of interest ('single family' homes and sale condition 'normal' in Ames, Iowa) after the drop conditions were applied, which comes out to 2002 rows and 81 variables.

Section 2: The Predictive Modeling Framework

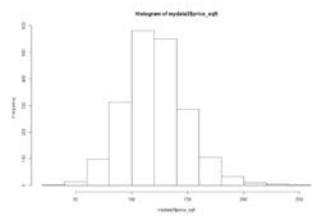
Figure 4: Table of Observation Counts for Train/Test Data

Sample Population	2002
Train	1410
Test	592
Total	2002

Observations: Figure 4 shows a table of observation counts for train/data partition. The training data is comprised of 1410 counts, while the test data is comprised of 592. As a result, the totals add up to 2002, which is my sample population total.

Section 3: Model Identification by Automated Variable Selection

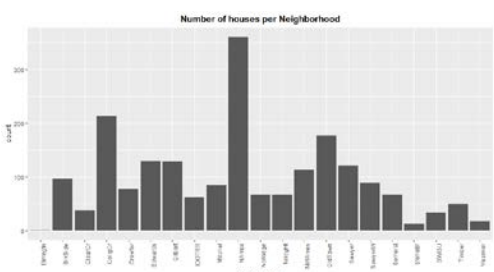
Figure 5: Histogram of Price Per SQFT & Summary Statistics

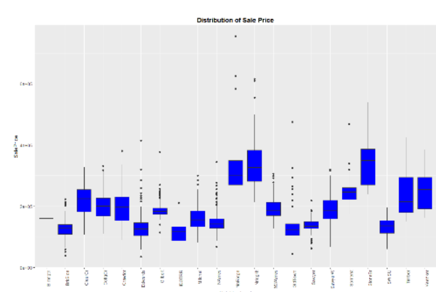


```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
30.37  103.29  119.82  121.13  137.44  248.99
```

Observations: Figure 5 shows a histogram of price per sqft and summary statistics so that we can use it to help define our clusters. The summary statistics indicate that 25% of houses are $103 or less, 25% are between $103 to $120, 25% are between $120 to $137, and another 25% are more than $137.

Figure 6 & 7: Bar plot of Neighborhood and Boxplot of SalePrice & Neighborhood





Observations: Figure 7 shows a boxplot of SalePrice &Neighborhood, which allows us to see if SalePrice is correlated with Neighborhood. The results suggests that there is correlation between SalePrice and Neighborhood because the Average SalePrice is different for these different categories. As a result, the following dummy variables will be created: "NbhdGrp1", "NbhdGrp2", "NbhdGrp3", with a baseline of "NbhdGrp4" (aka: Other).

Figure 10: Pool of Candidate Predictor Variables

```
[1] "OverallQual"   "TotalBsmtSF"   "FirstFlrSF"    "Fireplaces"    "PoolArea"
 [6] "GrLivArea"     "TotRmsAbvGrd"  "Neighborhood"  "GarageCars"    "GarageArea"
[11] "TotalFloorSF"  "HouseAge"      "SalePrice"     "price_sqft"    "QualityIndex
"
[16] "logSalePrice"  "TotalSqftCalc" "NbhdGrp"       "NbhdGrp1"      "NbhdGrp2"
[21] "NbhdGrp3"      "u"
```

Observations: Figure 10 shows a pool of candidate predictor variables. The variables include a mix of discrete and continuous variables. Our next step is to then 'clean' the data so that any missing values are removed.

Figure 11: Specification of Upper Model

```
> summary(upper.lm)

Call:
lm(formula = SalePrice ~ ., data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-81451 -11948  -1273   8986 208045

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -12837.368   5080.273  -2.527   0.0116 *
OverallQual     9992.053   1012.693   9.867  < 2e-16 ***
TotalBsmtSF       13.080      2.733   4.785 1.89e-06 ***
TotRmsAbvGrd    -823.712    745.296  -1.105   0.2693
GarageCars      2603.782   1110.246   2.345   0.0192 *
TotalFloorSF      75.149     11.501   6.534 8.95e-11 ***
HouseAge         -57.783     32.811  -1.761   0.0784 .
QualityIndex      46.434    105.852   0.439   0.6610
TotalSqftCalc     14.690      1.699   8.644  < 2e-16 ***
NbhdGrp1      -68266.573   2604.774 -26.208  < 2e-16 ***
NbhdGrp2      -44453.727   1999.333 -22.234  < 2e-16 ***
NbhdGrp3      -29064.401   1792.357 -16.216  < 2e-16 ***
GrLivArea          6.906     11.468   0.602   0.5471
FirstFlrSF         2.819      2.867   0.983   0.3256
Fireplaces      1986.310   1057.339   1.879   0.0605 .
PoolArea          -2.892     19.158  -0.151   0.8800
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21760 on 1394 degrees of freedom
Multiple R-squared:  0.9163,   Adjusted R-squared:  0.9154
F-statistic:  1018 on 15 and 1394 DF,  p-value: < 2.2e-16
```

Observation: Figure 11 shows a summary of the Multiple Linear Regression Model SalePrice ~ OverallQual+ TotalBsmtSF+ TotRmsAbvGrd+ GarageCars+ TotalFloorSF+ HouseAge+ QualityIndex+ TotalSqftCalc+ NbhdGrp1+ NbhdGrp2+ NbhdGrp3+GrLivArea+ FirstFlrSF+ Fireplaces+PoolArea.

This is considered the full model. Since TotRmsAbvGrd, HouseAge, QualityIndex, GrLivArea, FirstFlrSF, FirePlaces, and PoolArea are insignificant (>0.05) we will delete these variables so that we can have a simpler model. This means that we are making a tradeoff – settling for less accuracy but more precision. The residual standard error of 21760, shows us that when predicting SalePrice, one standard error = $21760. The multiple R-squared value of 0.9163, indicates that 91.6% of the variation in SalePrice is explained by the predictor variables.

Figure 12: Specification of Lower Model

```
> summary(sqft.lm)

Call:
lm(formula = SalePrice ~ TotalSqftCalc, data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-136137  -26531   -4045   23354  208628 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   10998.316   3403.578   3.231  0.00126 ** 
TotalSqftCalc    84.558      1.603  52.747  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43380 on 1408 degrees of freedom
Multiple R-squared:  0.664,	Adjusted R-squared:  0.6637 
F-statistic:  2782 on 1 and 1408 DF,  p-value: < 2.2e-16
```

Observation: Figure 12 shows a summary of the Linear Regression Model SalePrice ~ TotalSqftCalc. The equation of the regression line is: SalePrice = 10998.316 + 84.558 *TotalSqftCalc. Since the t-test of TotalSqftCalc is statistically significant (p<0.001), we can use this equation. This means that for every sqft increase, average TotalSqftCalc increases by $84.56.

Figure 13: Specification of Upper Model *(after insignificant variables are were deleted)*

```
> summary(upper.lm)

Call:
lm(formula = SalePrice ~ ., data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max 
-87568 -11604  -1404   8732 208637 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -19786.299   3577.727  -5.530 3.81e-08 ***
OverallQual    10472.405    703.404  14.888  < 2e-16 ***
TotalBsmtSF       14.795      2.119   6.981 4.50e-12 ***
GarageCars      2932.057   1080.222   2.714  0.00672 ** 
TotalFloorSF      81.225      2.730  29.753  < 2e-16 ***
TotalSqftCalc     15.760      1.657   9.510  < 2e-16 ***
NbhdGrp1      -70162.477   2365.427 -29.662  < 2e-16 ***
NbhdGrp2      -45118.372   1947.390 -23.169  < 2e-16 ***
NbhdGrp3      -29420.654   1771.213 -16.610  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21770 on 1401 degrees of freedom
Multiple R-squared:  0.9158,	Adjusted R-squared:  0.9153 
F-statistic:  1904 on 8 and 1401 DF,  p-value: < 2.2e-16
```

Observation: Figure 13 shows a summary of the Multiple Linear Regression Model SalePrice ~ OverallQual+ TotalBsmtSF+ GarageCars+TotalFloorSF+TotalSqftCalc+ NbhdGrp1+ NbhdGrp2+ NbhdGrp3+ Style1+ Style2. This is the upper model after GrLivArea, TotRmsAbvGrd, HouseAge, and QualityIndex were deleted so that we can have a simpler model. The equation of the regression line is: SalePrice = -19786.299 + 10472.405*OverallQual + 14.795*TotalBsmtSF +2932.057*GarageCars+81.225*TotalFloorSF+15.760*TotalSqftCalc-70162.477* NbhdGrp1-45118.372*NbhdGrp2-29420.654*NbhdGrp3. Since the t-test of all the predictor variables are statistically significant, we can use this equation. The baseline category is NbhdGrp4 (aka: Other houses). The results suggest that when NbhdGrp1 is compared to the Other houses, NbhdGrp1 homes on average, have a SalePrice of 70162.477 less and that it is significant. Furthermore, when NbhdGrp2 is compared to the Other houses, NbhdGrp2 homes on average, have a SalePrice of 45118.372 less and that it is significant. Lastly, when NbhdGrp3 is compared to the Other houses, NbhdGrp3 homes on average, have a SalePrice of $29420.654 less and that it is significant. The residual standard error of 21770, shows us that when predicting SalePrice, one standard error = $21770. The multiple R-squared value of 0.9158, indicates that 91.58% of the variation in SalePrice is explained by the predictor variables.

Figure 14: Forward Selection

```
> summary(forward.lm)

Call:
lm(formula = SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 +
    TotalFloorSF + NbhdGrp2 + NbhdGrp3 + TotalBsmtSF + GarageCars,
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-87568 -11604  -1404   8732 208637

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -19786.299   3577.727  -5.530 3.81e-08 ***
TotalSqftCalc     15.760      1.657   9.510  < 2e-16 ***
OverallQual    10472.405    703.404  14.888  < 2e-16 ***
NbhdGrp1      -70162.477   2365.427 -29.662  < 2e-16 ***
TotalFloorSF      81.225      2.730  29.753  < 2e-16 ***
NbhdGrp2      -45118.372   1947.390 -23.169  < 2e-16 ***
NbhdGrp3      -29420.654   1771.213 -16.610  < 2e-16 ***
TotalBsmtSF       14.795      2.119   6.981 4.50e-12 ***
GarageCars      2932.057   1080.222   2.714  0.00672 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21770 on 1401 degrees of freedom
Multiple R-squared:  0.9158,   Adjusted R-squared:  0.9153
F-statistic:  1904 on 8 and 1401 DF,  p-value: < 2.2e-16
```

Observations: Figure 14 shows a summary of the Multiple Linear Regression Model SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF + NbhdGrp2 + NbhdGrp3 + TotalBsmtSF + GarageCars after forward selection was conducted (note: see appendix for additional details). The AIC's at every step decreased and as a result all the predictors were retained. Additionally, all the predictors in this model are significant. The equation of the regression line is: SalePrice = -19786.299 + 10472.405*OverallQual + 14.795*TotalBsmtSF +2932.057*GarageCars+81.225*TotalFloorSF+15.760*TotalSqftCalc-70162.477* NbhdGrp1-45118.372*NbhdGrp2-29420.654*NbhdGrp3. Since the t-test of all the predictor variables are statistically significant, we can use this equation. The baseline category is NbhdGrp4 (aka: Other houses). The results suggest that when NbhdGrp1 is compared to the Other houses, NbhdGrp1 homes on average, have a SalePrice of 70162.477 less and that it is significant. Furthermore, when NbhdGrp2 is compared to the Other houses, NbhdGrp2 homes on average, have a SalePrice of 45118.372 less and that it is significant. Lastly, when NbhdGrp3 is compared to the Other houses, NbhdGrp3 homes on average, have a SalePrice of $29420.654 less and that it is significant. The residual standard error of 21770, shows us that when predicting SalePrice, one standard error = $21770. The multiple R-squared value of 0.9158, indicates that 91.58% of the variation in SalePrice is explained by the predictor variables. *This is the same as what we saw in figure 13.*

Figure 15: Backward Selection

```
> summary(backward.lm)

Call:
lm(formula = SalePrice ~ OverallQual + TotalBsmtSF + GarageCars +
    TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3,
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-87568 -11604  -1404   8732 208637

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -19786.299   3577.727  -5.530 3.81e-08 ***
OverallQual    10472.405    703.404  14.888  < 2e-16 ***
TotalBsmtSF       14.795      2.119   6.981 4.50e-12 ***
GarageCars      2932.057   1080.222   2.714  0.00672 **
TotalFloorSF      81.225      2.730  29.753  < 2e-16 ***
TotalSqftCalc     15.760      1.657   9.510  < 2e-16 ***
NbhdGrp1      -70162.477   2365.427 -29.662  < 2e-16 ***
NbhdGrp2      -45118.372   1947.390 -23.169  < 2e-16 ***
NbhdGrp3      -29420.654   1771.213 -16.610  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21770 on 1401 degrees of freedom
Multiple R-squared:  0.9158,   Adjusted R-squared:  0.9153
F-statistic:  1904 on 8 and 1401 DF,  p-value: < 2.2e-16
```

Observations: Figure 15 shows a summary of the Multiple Linear Regression Model SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF + NbhdGrp2 + NbhdGrp3 + TotalBsmtSF + GarageCars after backward selection was conducted (note: see appendix for additional details). The AIC showed that if you eliminate none, AIC is going to be equal to 28176. However, if you eliminate GarageCars, AIC will increase, hence we should eliminate no variables. Additionally, all the predictors in this model are significant. The equation of the regression line is: SalePrice = -19786.299 + 10472.405*OverallQual + 14.795*TotalBsmtSF +2932.057*GarageCars+81.225*TotalFloorSF+15.760*TotalSqftCalc-70162.477* NbhdGrp1-45118.372*NbhdGrp2-29420.654*NbhdGrp3. Since the t-test of all the predictor variables are statistically significant, we can use this equation. The baseline category is NbhdGrp4 (aka: Other houses). The results suggest that when NbhdGrp1 is compared to the Other houses, NbhdGrp1 homes on average, have a SalePrice of 70162.477 less and that it is significant. Furthermore, when NbhdGrp2 is compared to the Other houses, NbhdGrp2 homes on average, have a SalePrice of 45118.372 less and that it is significant. Lastly, when NbhdGrp3 is compared to the Other houses, NbhdGrp3 homes on average, have a SalePrice of $29420.654 less and that it is significant. The residual standard error of 21770, shows us that when predicting SalePrice, one standard error = $21770. The multiple R-squared value of 0.9158, indicates that 91.58% of the variation in SalePrice is explained by the predictor variables. *This is the same as what we saw in figure 13 and 14.*

Figure 16: Stepwise Selection

```
> summary(stepwise.lm)

Call:
lm(formula = SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + 
    TotalFloorSF + NbhdGrp2 + NbhdGrp3 + TotalBsmtSF + GarageCars, 
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max 
-87568 -11604  -1404   8732 208637 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -19786.299   3577.727  -5.530 3.81e-08 ***
TotalSqftCalc     15.760      1.657   9.510  < 2e-16 ***
OverallQual    10472.405    703.404  14.888  < 2e-16 ***
NbhdGrp1      -70162.477   2365.427 -29.662  < 2e-16 ***
TotalFloorSF      81.225      2.730  29.753  < 2e-16 ***
NbhdGrp2      -45118.372   1947.390 -23.169  < 2e-16 ***
NbhdGrp3      -29420.654   1771.213 -16.610  < 2e-16 ***
TotalBsmtSF       14.795      2.119   6.981 4.50e-12 ***
GarageCars      2932.057   1080.222   2.714  0.00672 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21770 on 1401 degrees of freedom
Multiple R-squared:  0.9158,	Adjusted R-squared:  0.9153 
F-statistic:  1904 on 8 and 1401 DF,  p-value: < 2.2e-16
```

Observations: Figure 16 shows a summary of the Multiple Linear Regression Model SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF + NbhdGrp2 + NbhdGrp3 + TotalBsmtSF + GarageCars after backward selection was conducted (note: see appendix for additional details). The AIC's at every step decreased and as a result all the predictors were retained. Additionally, all the predictors in this model are significant. The equation of the regression line is: SalePrice = -19786.299 + 10472.405*OverallQual + 14.795*TotalBsmtSF +2932.057*GarageCars+81.225*TotalFloorSF+15.760*TotalSqftCalc-70162.477* NbhdGrp1-45118.372*NbhdGrp2-29420.654*NbhdGrp3. Since the t-test of all the predictor variables are statistically significant, we can use this equation. The baseline category is NbhdGrp4 (aka: Other houses). The results suggest that when NbhdGrp1 is compared to the Other houses, NbhdGrp1 homes on average, have a SalePrice of 70162.477 less and that it is significant. Furthermore, when NbhdGrp2 is compared to the Other houses, NbhdGrp2 homes on average, have a SalePrice of 45118.372 less and that it is significant. Lastly, when NbhdGrp3 is compared to the Other houses, NbhdGrp3 homes on average, have a SalePrice of $29420.654 less and that it is significant. The residual standard error of 21770, shows us that when predicting SalePrice, one standard error = $21770. The multiple R-squared value of 0.9158, indicates that 91.58% of the variation in SalePrice is explained by the predictor variables. *This is the same as what we saw in figure 13, 14, and 15. As a result, the different variable selection procedures selected the same model.*

Figure 17: Junk Model

```
> summary(junk.lm)

Call:
lm(formula = SalePrice ~ GarageCars + TotalBsmtSF, data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-210937  -30810   -3452   24855  309395 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10218.591   3897.746   2.622  0.00884 ** 
GarageCars  47273.453   1897.994  24.907  < 2e-16 ***
TotalBsmtSF    84.749      3.535  23.974  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47200 on 1407 degrees of freedom
Multiple R-squared:  0.6024,	Adjusted R-squared:  0.6018 
F-statistic:  1066 on 2 and 1407 DF,  p-value: < 2.2e-16
```

Observations: Figure 17 shows a summary of the Multiple Linear Regression Model SalePrice ~ GarageCars + TotalBsmtSF for model comparison purposes. The equation of the regression line is: SalePrice = 10218.591 + 47273.453*GarageCars+84.749*TotalBsmtSF. Since the t-test of all the predictor variables are statistically significant, we can use this equation. The residual standard error of 47200 is a lot greater than the residual standard error of 21770 that we saw in the other models. This means than when predicting SalePrice, one standard error = $47200. The multiple R-squared value of 0.6024 is also a lot worse than the R-squared value of 0.9158 that we saw in the other models. This indicates that 60.24% of the variation in SalePrice is explained by the predictor variables.

Figure 18: VIF Values for the Variable Selection Models

```
> sort(vif(forward.lm),decreasing=TRUE)
 TotalFloorSF TotalSqftCalc      NbhdGrp1   OverallQual      NbhdGrp2   TotalBsmtS
F    GarageCars
     5.495046      4.242142      2.996339      2.792065      2.202382      2.09376
6      1.887037
     NbhdGrp3
     1.728122
> sort(vif(backward.lm),decreasing=TRUE)
 TotalFloorSF TotalSqftCalc      NbhdGrp1   OverallQual      NbhdGrp2   TotalBsmtS
F    GarageCars
     5.495046      4.242142      2.996339      2.792065      2.202382      2.09376
6      1.887037
     NbhdGrp3
     1.728122
> sort(vif(stepwise.lm),decreasing=TRUE)
 TotalFloorSF TotalSqftCalc      NbhdGrp1   OverallQual      NbhdGrp2   TotalBsmtS
F    GarageCars
     5.495046      4.242142      2.996339      2.792065      2.202382      2.09376
6      1.887037
     NbhdGrp3
     1.728122
> sort(vif(junk.lm),decreasing=TRUE)
 GarageCars TotalBsmtSF
    1.23933     1.23933
```

Observations: Figure 18 shows us the VIF values for the variable selection models. A VIF of 1 would mean that no multicolinearity exists at all, while a large VIF number (e.g., 10) would indicate serious multicolinearity issues. As a result, since the VIF for all the predictors above are low, this concludes that we don't have serious multicolinearity issues.

Figure 19: Model Comparison for Training Sample

Model Name	**Adj R-Squared**	**Rank**	**AIC**	**Rank**	**BIC**	**Rank**	**MSE** (Residual Standard Error)	**Rank**	**MAE**	**Rank**
forward.lm	0.9153	1	32179.51	1	32232.02	1	21770	1	14321.37	1
backward.lm	0.9153	1	32179.51	1	32232.02	1	21770	1	14321.37	1
stepwise.lm	0.9153	1	32179.51	1	32232.02	1	21770	1	14321.37	1
junk.lm	0.6018	2	34355.79	2	34376.8	2	47200	2	34250.58	2

Observations: Figure 19 shows the model comparisons so that we can compare the in-sample fit and predictive accuracy of our models. The results above show the computations for adjusted R-Squared, AIC, BIC, mean squared error, and the mean absolute error for each of these models for the training sample. Each of these metrics represents some concept of 'fit'. Models: forward.lm, backward.lm, and stepwise.lm ranked #1 in each metric, while junk.lm ranked #2 in each metric. This isn't surprising since the variable selection procedures selected the same model. However, it's important to note that a model that is #1 in one metric, may not be #1 in other metrics. As a result, we shouldn't expected each metric to give us the same ranking of model 'fit'. Evidence of this can be seen in this week's Special Topic Lecture: Likelihood Function, where we saw differences in the metrics for the 5 models.

Section 4: Predictive Accuracy

Figure 20: MSE & MAE for Out-of-Sample

Model Name	MSE (Residual Standard Error)	MAE
forward.lm2	21770	14321.37
backward.lm2	21770	14321.37
stepwise.lm2	21770	14321.37
junk.lm2	47200	33627.06

Observations: Figure 20 shows the MSE and MAE for the out-of-sample test data. Based on the criteria, forward.lm2, backward.lm2, and stepwise.lm2 fit the best based on this criteria because the MSE and MAE are low. We also saw the same conclusion in the in-sample test as well. It's also interesting to point out that the MAE for the junk model in the out-of-sample test data decreased slightly. Both the MAE and MSE are valuable metrics to assess model fit so we do not necessarily have to have a preference, especially since the purpose of using these metrics are for estimation and prediction. If a model has a better predictive accuracy in-sample then it does out-of-sample, it means that our MSE of prediction went down (e.g., reduce variance Y or reduced bias).

Section 5: Operational Validation

Figure 21: Mean Absolute Percent Error for Training Data

```
> MAPE <- mean(forward.pct)
> MAPE
[1] 0.09495122

> MAPE <- mean(backward.pct)
> MAPE
[1] 0.09495122

> MAPE <- mean(stepwise.pct)
> MAPE
[1] 0.09495122

> MAPE <- mean(junk.pct)
> MAPE
[1] 0.2095764
```

Observations: Figure 21 shows Mean Absolute Percent Error for training data for each of the models. The results show that the MAPE using forward selection, backward selection, and stepwise method is 9.5%, while the junk model is 21%. MAPE or PredictionGrade is a metric that translates more easily to the development of a business policy than MSE or MAE.

Figure 22: Mean Absolute Percent Error for Test Data

```
> MAPE <- mean(forward.testPCT)
> MAPE
[1] 0.08257221

> MAPE <- mean(backward.testPCT)
> MAPE
[1] 0.08257221

> MAPE <- mean(stepwise.testPCT)
> MAPE
[1] 0.08257221
> MAPE <- mean(junk.testPCT)
> MAPE
[1] 0.1974362
```

Observations: Figure 22 shows Mean Absolute Percent Error for the test data for each of the models. The results show that the MAPE using forward selection, backward selection, and stepwise method is 8.3%, while the junk model is 19.7%. This is lower than what we saw in the training data. MAPE or PredictionGrade is a metric that translates more easily to the development of a business policy than MSE or MAE.

Figure 23: Prediction Grades for Training Data

```
forward.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]
          0.68226950           0.14326241           0.12056738
    Grade 4: (0.25+]
          0.05390071


backward.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]
          0.68226950           0.14326241           0.12056738
    Grade 4: (0.25+]
          0.05390071


stepwise.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]
          0.68226950           0.14326241           0.12056738
    Grade 4: (0.25+]
          0.05390071


junk.PredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]
           0.3347518            0.1276596            0.2368794
    Grade 4: (0.25+]
           0.3007092
```

Observations: Figure 23 shows the prediction grades using the training data for each of the models. The results show that forward.lm, backward.lm, and stepwise.lm are the most accurate. The results show that on the training data set, 68% of houses we can predict within +/-10% error, 14.3% of houses we can predict within +/-10 to 15% error, 12% of houses we can predict within +/-15 to 25% of error, and about 5% houses we can predict within +/-25% of error. In comparison to the junk model, these prediction grades were a lot better. Overall, the prediction grades of the forward.lm, backward.lm, and stepwise.lm models validate the low MAE, MSE, and MAPE that we saw in our predictive accuracy results. Additionally, our model ranking remained the same since forward.lm, backward.lm, and stepwise.lm all had the same prediction grades.

Figure 24: Prediction Grades for Test Data

```
forward.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]
          0.75168919          0.10304054          0.09290541
    Grade 4: (0.25+]
          0.05236486


backward.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]
          0.75168919          0.10304054          0.09290541
    Grade 4: (0.25+]
          0.05236486


stepwise.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]
          0.75168919          0.10304054          0.09290541
    Grade 4: (0.25+]
          0.05236486


junk.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]
           0.3344595           0.1418919           0.2601351
    Grade 4: (0.25+]
           0.2635135
```

Observations: Figure 24 shows the prediction grades using the test data for each of the models. The results show that forward.lm, backward.lm, and stepwise.lm are the most accurate, which is similar to what we saw in figure 23 as well. The results show that on the test data set, 75% of houses we can predict within +/-10% error, 10.3% of houses we can predict within +/-10 to 15% error, 9.3% of houses we can predict within +/-15 to 25% of error, and 5.2% houses we can predict within +/-25% of error. In comparison to the junk model, these prediction grades were a lot better. Overall, the prediction grades of the forward.lm, backward.lm, and stepwise.lm models validate the low MAE, MSE, and MAPE that we saw in our predictive accuracy results. Additionally, our model ranking remained the same since forward.lm, backward.lm, and stepwise.lm all had the same prediction grades. It's also interesting to note that the prediction grades in the test data set improved compared to the training data set. In conclusion, this shows that forward.lm, backward.lm, and stepwise.lm are all underwriting quality since the model is accurate within 10% more than 50% perfect of the time (Grade 1: 75%).

Section 6: Best Model

Best Model without Transformation

Figure 25: Analysis of Variance for SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

```
Analysis of Variance Table

Response: SalePrice
                Df     Sum Sq    Mean Sq  F value    Pr(>F)    
OverallQual      1 6.9449e+12 6.9449e+12 14673.06 < 2.2e-16 ***
TotalBsmtSF      1 7.7879e+11 7.7879e+11  1645.42 < 2.2e-16 ***
GarageCars       1 3.3492e+11 3.3492e+11   707.61 < 2.2e-16 ***
TotalFloorSF     1 9.4463e+11 9.4463e+11  1995.79 < 2.2e-16 ***
TotalSqftCalc    1 2.2144e+11 2.2144e+11   467.85 < 2.2e-16 ***
NbhdGrp1         1 2.5851e+11 2.5851e+11   546.17 < 2.2e-16 ***
NbhdGrp2         1 1.7370e+11 1.7370e+11   366.99 < 2.2e-16 ***
NbhdGrp3         1 1.6305e+11 1.6305e+11   344.49 < 2.2e-16 ***
Residuals     1993 9.4331e+11 4.7331e+08                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

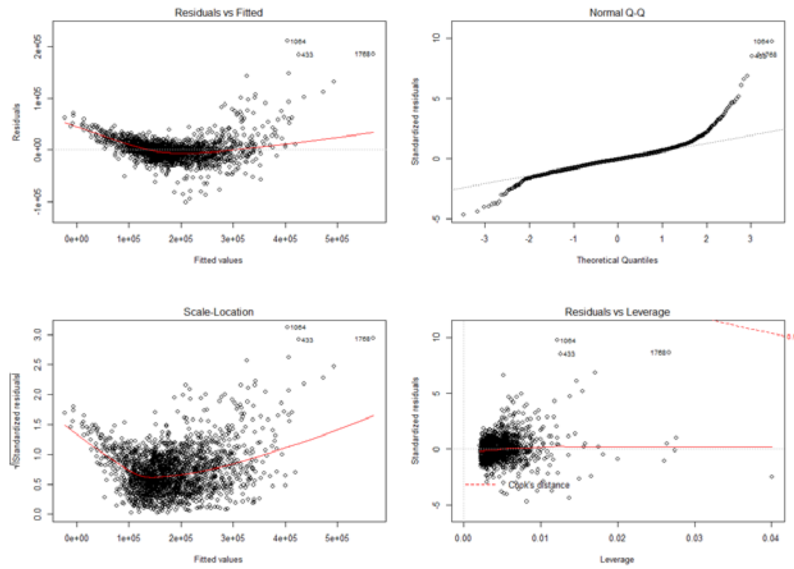
Observations: Figure 25 shows an ANOVA for SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + Total FloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3. A statistically significant result was obtain ed overall as indicated by the F-statistic which is 2593 with a p-value = < 2.2e-16. This indicates the model has produced statistically significant results to be investigated. The AVOVA tables shows that NbhdGrp1, N bhdGrp2 and NbhdGrp3 all have significant difference when compared to the Others group.

Figure 26: Multiple Linear Regression Model SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

```
Call:

lm(formula = SalePrice ~ OverallQual + TotalBsmtSF + GarageCars +
    TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3,
    data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-101052  -11141   -1159    8617  210949

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -21341.800   3041.157  -7.018 3.08e-12 ***
OverallQual    10337.475    599.600  17.241  < 2e-16 ***
TotalBsmtSF       14.995      1.742   8.609  < 2e-16 ***
GarageCars      4002.265    923.353   4.334 1.53e-05 ***
TotalFloorSF      81.783      2.318  35.283  < 2e-16 ***
TotalSqftCalc     14.808      1.376  10.759  < 2e-16 ***
NbhdGrp1      -69813.617   1986.238 -35.149  < 2e-16 ***
NbhdGrp2      -43732.949   1655.507 -26.417  < 2e-16 ***
NbhdGrp3      -27819.069   1498.832 -18.560  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21760 on 1993 degrees of freedom
Multiple R-squared:  0.9124,   Adjusted R-squared:  0.912
F-statistic:  2593 on 8 and 1993 DF,  p-value: < 2.2e-16
```

Observations: Figure 26 shows a summary of the Multiple Linear Regression Model SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3. The equation of the regression line is: SalePrice = -21341.800 + 10337.475*OverallQual + 14.995*TotalBsmtSF +4002.265*GarageCars+81.783*TotalFloorSF+14.808*TotalSqftCalc-69813.617* NbhdGrp1-43732.949*NbhdGrp2-27819.069*NbhdGrp3. Since the t-test of all the predictor variables are statistically significant, we can use this equation. The baseline category is NbhdGrp4 (aka: Other houses). The results suggest that when NbhdGrp1 is compared to the Other houses, NbhdGrp1 homes on average, have a SalePrice of 69813.617 less and that it is significant. Furthermore, when NbhdGrp2 is compared to the Other houses, NbhdGrp2 homes on average, have a SalePrice of 43732.949 less and that it is significant. Lastly, when NbhdGrp3 is compared to the Other houses, NbhdGrp3 homes on average, have a SalePrice of $27819.069 less and that it is significant. The residual standard error of 21760, shows us that when predicting SalePrice, one standard error = $21760. The multiple R-squared value of 0.9124, indicates that 91.24% of the variation in SalePrice is explained by the predictor variables.

Figure 27: Scatterplots with Residuals & QQ-Plot of Residuals



Observations: Figure 27, shows scatterplots with residuals and qq-plots of residuals so that we can check to make sure the model is meeting all the assumptions. The QQ plot reveals that the density distribution is non-normal. This is present in the plot where some of the data points are progressively departing from the line in the upper right hand corner of the plot. This indicates non-normality and shows us that it does not correspond relatively well to a standard normal distribution. The scatterplot of residuals vs. fitted shows us that there is “funnel shaped” pattern with heteroscedasticity and a few outliers. By comparison, a healthy normal probability plot of the residuals would be relatively linear and would have a random scatter of data over the range of values for the independent variable. In addition, the residual vs. leverage plot shows that there are some influential points on the right side of the graph. It’s important to note that it is highly desirable for the residuals to conform to a normal distribution with few to no outliers. As a result, in order to correct the problems of non-linearity, non-constant variance, non-normality, and influential points we are going to transform SalePrice by creating a new variable called logSalePrice and possibly conduct transformation on the predictor variables as well.

Figure 28: Predictions: MLR Model SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

	fit	lwr	upr
1	206478.94	163759.87	249198.0
2	99447.83	56706.89	142188.8
3	178830.81	136088.02	221573.6
4	266507.61	223683.20	309332.0
5	177591.15	134853.26	220329.0
6	198599.11	155877.26	241321.0

Observations: Figure 28 shows us that the predicted value of the first house is $206478.94. Additionally, the lower and upper confidence bands shows $163759.87 and $249198.0, respectively. This means that the 95% confidence band on this predicted value is $163759.87 and $249198.0.

Best Model with Transformation & Comparison

Figure 25: Analysis of Variance for SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

```
Analysis of Variance Table

Response: SalePrice
                Df     Sum Sq    Mean Sq  F value    Pr(>F)    
OverallQual      1 6.9449e+12 6.9449e+12 14673.06 < 2.2e-16 ***
TotalBsmtSF      1 7.7879e+11 7.7879e+11  1645.42 < 2.2e-16 ***
GarageCars       1 3.3492e+11 3.3492e+11   707.61 < 2.2e-16 ***
TotalFloorSF     1 9.4463e+11 9.4463e+11  1995.79 < 2.2e-16 ***
TotalSqftCalc    1 2.2144e+11 2.2144e+11   467.85 < 2.2e-16 ***
NbhdGrp1         1 2.5851e+11 2.5851e+11   546.17 < 2.2e-16 ***
NbhdGrp2         1 1.7370e+11 1.7370e+11   366.99 < 2.2e-16 ***
NbhdGrp3         1 1.6305e+11 1.6305e+11   344.49 < 2.2e-16 ***
Residuals     1993 9.4331e+11 4.7331e+08                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 29: Analysis of Variance for L_SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + L_TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

```
Analysis of Variance Table

Response: SalePrice
                Df     Sum Sq    Mean Sq  F value    Pr(>F)    
OverallQual      1 6.9449e+12 6.9449e+12 14673.06 < 2.2e-16 ***
TotalBsmtSF      1 7.7879e+11 7.7879e+11  1645.42 < 2.2e-16 ***
GarageCars       1 3.3492e+11 3.3492e+11   707.61 < 2.2e-16 ***
TotalFloorSF     1 9.4463e+11 9.4463e+11  1995.79 < 2.2e-16 ***
TotalSqftCalc    1 2.2144e+11 2.2144e+11   467.85 < 2.2e-16 ***
NbhdGrp1         1 2.5851e+11 2.5851e+11   546.17 < 2.2e-16 ***
NbhdGrp2         1 1.7370e+11 1.7370e+11   366.99 < 2.2e-16 ***
NbhdGrp3         1 1.6305e+11 1.6305e+11   344.49 < 2.2e-16 ***
Residuals     1993 9.4331e+11 4.7331e+08                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparison/Discussion: Both models were statistically significant as indicated by the F-statistic with a p-value = < 2.2e-16. This indicates that both models produced statistically significant results to be investigated. Figure 25 & 29, the AVOVA tables shows that NbhdGrp1, NbhdGrp2, and NbhdGrp3 all have significant difference when compared to the Others group.

Figure 26: Multiple Linear Regression Model SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

```
Call:
lm(formula = SalePrice ~ OverallQual + TotalBsmtSF + GarageCars +
    TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3,
    data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-101052  -11141   -1159    8617  210949

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -21341.800   3041.157  -7.018 3.08e-12 ***
OverallQual   10337.475    599.600  17.241  < 2e-16 ***
TotalBsmtSF      14.995      1.742   8.609  < 2e-16 ***
GarageCars     4002.265    923.353   4.334 1.53e-05 ***
TotalFloorSF     81.783      2.318  35.283  < 2e-16 ***
TotalSqftCalc    14.808      1.376  10.759  < 2e-16 ***
NbhdGrp1     -69813.617   1986.238 -35.149  < 2e-16 ***
NbhdGrp2     -43732.949   1655.507 -26.417  < 2e-16 ***
NbhdGrp3     -27819.069   1498.832 -18.560  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21760 on 1993 degrees of freedom
Multiple R-squared:  0.9124,    Adjusted R-squared:  0.912
F-statistic:  2593 on 8 and 1993 DF,  p-value: < 2.2e-16
```

Figure 30: Multiple Linear Regression Model L_SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + L_TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

```
Call:
lm(formula = L_SalePrice ~ OverallQual + TotalBsmtSF + GarageCars +
    L_TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3,
    data = subdat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98078 -0.03964  0.00285  0.04791  0.30365

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.520e+00  8.412e-02  77.504  < 2e-16 ***
OverallQual    4.596e-02  2.463e-03  18.666  < 2e-16 ***
TotalBsmtSF    3.310e-05  6.784e-06   4.878 1.16e-06 ***
GarageCars     2.566e-02  3.659e-03   7.012 3.21e-12 ***
L_TotalFloorSF 7.198e-01  1.422e-02  50.604  < 2e-16 ***
TotalSqftCalc  6.204e-05  5.162e-06  12.020  < 2e-16 ***
NbhdGrp1      -4.256e-01  8.321e-03 -51.146  < 2e-16 ***
NbhdGrp2      -2.274e-01  6.840e-03 -33.242  < 2e-16 ***
NbhdGrp3      -1.290e-01  6.001e-03 -21.491  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08523 on 1993 degrees of freedom
Multiple R-squared:  0.9487,    Adjusted R-squared:  0.9485
F-statistic:  4607 on 8 and 1993 DF,  p-value: < 2.2e-16
```

Comparison/Discussion: Figure 30 shows a summary of the Multiple Linear Regression Model L_SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + L_TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3. The equation of the regression line is: L_SalePrice = 6.520e+00+ 4.596e-02*OverallQual + 3.310e-05*TotalBsmtSF + 2.566e-02*GarageCars+ 7.198e-01*L_TotalFloorSF+ 6.204e-05*TotalSqftCalc- 4.256e-01* NbhdGrp1- 2.274e-01*NbhdGrp2- 1.290e-01*NbhdGrp3. The results suggest that when NbhdGrp1 is compared to the Other houses, NbhdGrp1 homes on average, have a SalePrice of $62 less and that it is significant. Furthermore, when NbhdGrp2 is compared to the Other houses, NbhdGrp2 homes on average, have a SalePrice of $22 less and that it is significant. Lastly, when NbhdGrp3 is compared to the Other houses, NbhdGrp3 homes on average, have a SalePrice of $12 less and that it is significant. After SalePrice and TotalFloorSF were transformed to L_SalePrice and

L_TotalFloorSF, it appears that the multiple R-squared improved from 0.9124 to 0.9487. The multiple R-squared value of 0.9487 indicates that 94.87% of the variation in SalePrice is explained by the predictor variables.

Figure 27: Scatterplots with Residuals & QQ-Plot of Residuals

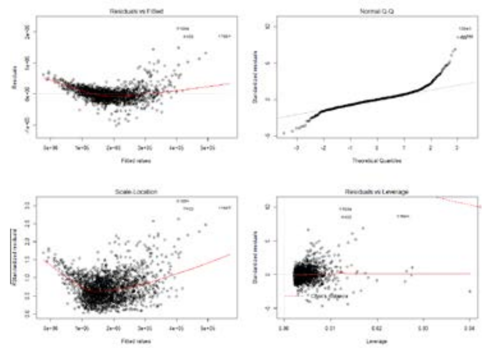
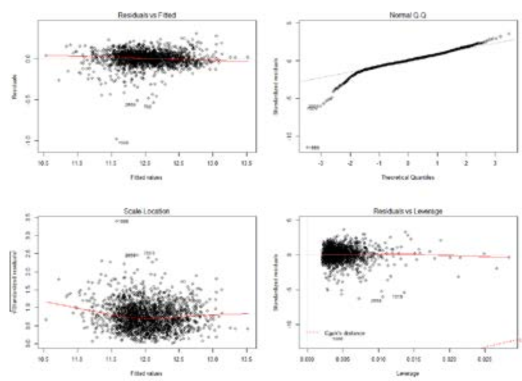


Figure 30: Scatterplots with Residuals & QQ-Plot of Residuals



Observations: After SalePrice and TotalFloorSF were transformed to L_SalePrice and L_TotalFloorSF, it appears, it appears that the QQ plot has improved since the dots have moved closer to the line, indicating that the density distribution has gotten closer to normal compared to Figure 27. Additionally, the scatterplot of residuals vs. fitted shows us that the normal probability plot of the residuals appear to be linear and have a random scatter of data over the range of values for the independent variable compared to Figure 27. In conclusion, given that the assumptions of normality, linearity, and homoscedasticity improved after transformation, our final model is: L_SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + L_TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3. However, it's important to note that there still seems to be outliers/influential points and slight non-normality that needs to be addressed. This can be addressed using outlier deletion/down weighing them in the future.

Reported "Best" Model After Transformation: L_SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + L_TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

Section 7: Reflection/Conclusions

In section 1, we defined the sample population/data of interest for 'typical' homes in Ames, Iowa to be 'single-family' homes with 'normal' sales in Ames, Iowa using drop conditions and boxplots.

In section 2, we assessed model performance by splitting the sample into a 70/30 train/test split, one for in-sample model development and one for out-of-sample model assessment so that we could cross-validate the data. Our train data set had 1410 counts and test data set had 592 counts.

In section 3 and 4, we chose a pool of candidate predictor variables and ran the upper and lower model specifications. This created a full model of Multiple Linear Regression Model SalePrice ~ OverallQual+ TotalBsmtSF+ TotRmsAbvGrd+ GarageCars+ TotalFloorSF+ HouseAge+ QualityIndex+ TotalSqftCalc+ NbhdGrp1+ NbhdGrp2+ NbhdGrp3+GrLivArea+ FirstFlrSF+ Fireplaces+PoolArea. However, we noticed that TotRmsAbvGrd, HouseAge, QualityIndex, GrLivArea, FirstFlrSF, FirePlaces, and PoolArea were insignificant (>0.05) so we deleted these variables so that we can have a simpler model. We then re-ran the code with these variables deleted and created new full model of SalePrice ~ OverallQual+ TotalBsmtSF+ GarageCars+TotalFloorSF+TotalSqftCalc+ NbhdGrp1+ NbhdGrp2+ NbhdGrp3+ Style1+ Style2. All of the variables were now significant. This meant that we made a tradeoff – settling for less accuracy but more precision. We then used the training data to find the 'best' models using automated variable selection using the techniques: forward, backward, and stepwise variable selection for model identification purposes. It was determined that the forward.lm, backward.lm, and stepwise.lm had the best in-sample fit and predictive accuracy. These models had an adjusted R-squared of 0.9153, AIC of 32179.51, BIC of 32232.02, MSE of 21770, and MAE of 14321.37; with no multicollinearity issues. We also saw similar MSE and MAE for these models in the out-of-sample test data as well.

In section 5, we then validated these models from a business sense using defined cut-off points (e.g., defining PredictionGrades). The results show that on the training data set, 68% of houses we can predict within +/-10% error, 14.3% of houses we can predict within +/-10 to 15% error, 12% of houses we can predict within +/-15 to 25% of error, and about 5% houses we can predict within +/-25% of error. In comparison to the junk model, these prediction grades were a lot better. Overall, the prediction grades of the forward.lm, backward.lm, and stepwise.lm models validate the low MAE, MSE, and MAPE that we saw in our predictive accuracy results. Additionally, our model ranking remained the same since forward.lm, backward.lm, and stepwise.lm all had the same prediction grades.

For the test data set, the results show that forward.lm, backward.lm, and stepwise.lm are the most accurate, which is similar to what we saw in the training set. The results show that on the test data set, 75% of houses we can predict within +/-10% error, 10.3% of houses we can predict within +/-10 to 15% error, 9.3% of houses we can predict within +/-15 to 25% of error, and 5.2% houses we can predict within +/-25% of error. In comparison to the junk model, these prediction grades were a lot better. Overall, the prediction grades of the forward.lm, backward.lm, and stepwise.lm models validate the low MAE, MSE, and MAPE that we saw in our predictive accuracy results. Additionally, our model ranking remained the same since forward.lm, backward.lm, and stepwise.lm all had the same prediction grades. It's also interesting to note that the prediction grades in the test data set improved compared to the training data set. This shows that forward.lm, backward.lm, and stepwise.lm are all underwriting quality since the model is accurate within 10% more than 50% perfect of the time (Grade 1: 75%).

In section 6, we determined that Analysis of Variance for SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 was the best model to move forward with. We then conducted diagnostics, etc. The residual standard error of 21760, showed us that when predicting SalePrice, one standard error = $21760. The multiple R-squared value of 0.9124, indicates that 91.24% of the variation in SalePrice is explained by the predictor variables. However, we encountered issues with non-linearity, non-constant variance, non-normality, and influential points. As a result, we decided to transform SalePrice and TotalFloorSF to correct these issues. After transformation was conducted, the multiple R-squared improved from 0.9124 to 0.9487. Additionally, the QQ plot improved since the dots moved closer to the line, indicating that the density distribution has gotten closer to normal compared to Figure 27. Additionally, the scatterplot of residuals vs. fitted shows us that the normal probability plot of the residuals appear to be linear and have a random scatter of data over the range of values for the independent variable compared to Figure 27. In conclusion, given that the assumptions of normality, linearity, and homoscedasticity improved after transformation, our final model is: L_SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + L_TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3. However, it's important to note that there still seems to be outliers/influential points and slight non-normality that needs to be addressed. As a result, our reported "Best" Model After Transformation is: L_SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + L_TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3.

In conclusion, some of the challenges that were presented by the data is that early on we encountered issues with model fit (e.g., low r-squared and large predictor error). We were able to address this by adding more "relevant" predictor variables (both quantitative and categorical). Additionally, we also encountered modeling assumption challenges (e.g., normality, linearity, and homoscedasticity not being met). As a result, we often had to conduct variable transformations (e.g., logSalePrice) to improve model fit and the model itself by making the model assumptions truer than before. For instance, a healthy normal probability plot of the residuals is relatively linear and has a random scatter of data over the range of values for the independent variable. It's also important and highly desirable for the residuals to conform to a normal distribution with few to no outliers. However, after reporting the final model, there still seems to be outliers/influential points and slight non-normality that needs to be addressed. This can be addressed using outlier deletion/down weighing them in the future and possibly. It may also be a good idea to conduct more variable transformations to improve the slight non-normality as well.

In regards to improving predictive accuracy, increasing the sample size of the data and incorporating more relevant predictors can help increase accuracy and precision. This can be addressed by collecting more data within Ames, Iowa or even expanding it to include more cities or possibly states. Additionally, more relevant predictors such as school district would also be an interesting predictor variable to include in the future as well. Lastly, addressing the outliers/influential points can also help improve predictive accuracy. This could be accomplished through robust regression techniques.

Appendix: Forward, Backward, StepWise Selection

```
> forward.lm <- stepAIC(object=lower.lm,scope=list(upper=formula(upper.lm),lower=~
1),
+                        direction=c('forward'));
Start:  AIC=31648.78
SalePrice ~ 1

                Df  Sum of Sq        RSS   AIC
+ TotalSqftCalc  1 5.2349e+12 2.6492e+12 30113
+ OverallQual    1 5.1619e+12 2.7222e+12 30151
+ TotalFloorSF   1 4.9362e+12 2.9479e+12 30264
+ GarageCars     1 3.4686e+12 4.4155e+12 30833
+ TotalBsmtSF    1 3.3671e+12 4.5170e+12 30865
+ NbhdGrp1       1 8.0832e+11 7.0758e+12 31498
+ NbhdGrp2       1 1.6112e+10 7.8680e+12 31648
<none>                        7.8841e+12 31649
+ NbhdGrp3       1 9.4142e+08 7.8832e+12 31651

Step:  AIC=30113.04
SalePrice ~ TotalSqftCalc

               Df  Sum of Sq        RSS   AIC
+ OverallQual   1 1.3309e+12 1.3183e+12 29131
+ GarageCars    1 6.0322e+11 2.0460e+12 29751
+ TotalFloorSF  1 4.4407e+11 2.2051e+12 29856
+ NbhdGrp1      1 3.9726e+11 2.2519e+12 29886
+ TotalBsmtSF   1 2.5533e+11 2.3939e+12 29972
+ NbhdGrp3      1 8.4361e+09 2.6407e+12 30111
+ NbhdGrp2      1 7.9644e+09 2.6412e+12 30111
<none>                       2.6492e+12 30113

Step:  AIC=29130.96
SalePrice ~ TotalSqftCalc + OverallQual

               Df  Sum of Sq        RSS   AIC
+ NbhdGrp1      1 9.1976e+10 1.2263e+12 29031
+ GarageCars    1 9.1198e+10 1.2271e+12 29032
+ TotalFloorSF  1 7.4901e+10 1.2434e+12 29051
+ TotalBsmtSF   1 3.9364e+10 1.2789e+12 29090
+ NbhdGrp2      1 5.1182e+09 1.3132e+12 29128
<none>                       1.3183e+12 29131
+ NbhdGrp3      1 3.0924e+08 1.3180e+12 29133

Step:  AIC=29030.98
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1

               Df  Sum of Sq        RSS   AIC
+ TotalFloorSF  1 2.0714e+11 1.0192e+12 28772
+ GarageCars    1 7.0304e+10 1.1560e+12 28950
+ NbhdGrp2      1 3.7374e+10 1.1889e+12 28989
+ TotalBsmtSF   1 1.8044e+10 1.2083e+12 29012
+ NbhdGrp3      1 6.9098e+09 1.2194e+12 29025
<none>                       1.2263e+12 29031

Step:  AIC=28772.09
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF
```

```
               Df  Sum of Sq        RSS   AIC
+ NbhdGrp2      1 1.7112e+11 8.4804e+11 28515
+ TotalBsmtSF   1 6.6669e+10 9.5248e+11 28679
+ GarageCars    1 3.4401e+10 9.8475e+11 28726
+ NbhdGrp3      1 7.8409e+09 1.0113e+12 28763
<none>                       1.0192e+12 28772

Step:  AIC=28514.93
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF +
    NbhdGrp2

               Df  Sum of Sq        RSS   AIC
+ NbhdGrp3      1 1.5670e+11 6.9133e+11 28229
+ TotalBsmtSF   1 3.8412e+10 8.0962e+11 28452
+ GarageCars    1 1.7433e+10 8.3060e+11 28488
<none>                       8.4804e+11 28515

Step:  AIC=28228.86
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF +
    NbhdGrp2 + NbhdGrp3

               Df  Sum of Sq        RSS   AIC
+ TotalBsmtSF   1 2.3786e+10 6.6755e+11 28182
+ GarageCars    1 4.1768e+09 6.8716e+11 28222
<none>                       6.9133e+11 28229

Step:  AIC=28181.5
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF +
    NbhdGrp2 + NbhdGrp3 + TotalBsmtSF

              Df  Sum of Sq        RSS   AIC
+ GarageCars  1 3492085015 6.6405e+11 28176
<none>                     6.6755e+11 28182

Step:  AIC=28176.1
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF +
    NbhdGrp2 + NbhdGrp3 + TotalBsmtSF + GarageCars

> summary(forward.lm)

Call:
lm(formula = SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 +
    TotalFloorSF + NbhdGrp2 + NbhdGrp3 + TotalBsmtSF + GarageCars,
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-87568 -11604  -1404   8732 208637

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -19786.299   3577.727  -5.530 3.81e-08 ***
TotalSqftCalc     15.760      1.657   9.510  < 2e-16 ***
OverallQual    10472.405    703.404  14.888  < 2e-16 ***
NbhdGrp1      -70162.477   2365.427 -29.662  < 2e-16 ***
TotalFloorSF      81.225      2.730  29.753  < 2e-16 ***
```

```
NbhdGrp2      -45118.372   1947.390 -23.169  < 2e-16 ***
NbhdGrp3      -29420.654   1771.213 -16.610  < 2e-16 ***
TotalBsmtSF       14.795      2.119   6.981 4.50e-12 ***
GarageCars      2932.057   1080.222   2.714  0.00672 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21770 on 1401 degrees of freedom
Multiple R-squared:  0.9158,   Adjusted R-squared:  0.9153
F-statistic:  1904 on 8 and 1401 DF,  p-value: < 2.2e-16

>
> backward.lm <- stepAIC(object=upper.lm,direction=c('backward'));
Start:  AIC=28176.1
SalePrice ~ OverallQual + TotalBsmtSF + GarageCars + TotalFloorSF +
    TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

                Df  Sum of Sq        RSS   AIC
<none>                        6.6405e+11 28176
- GarageCars     1 3.4921e+09 6.6755e+11 28182
- TotalBsmtSF    1 2.3101e+10 6.8716e+11 28222
- TotalSqftCalc  1 4.2869e+10 7.0692e+11 28262
- OverallQual    1 1.0506e+11 7.6912e+11 28381
- NbhdGrp3       1 1.3078e+11 7.9483e+11 28428
- NbhdGrp2       1 2.5443e+11 9.1848e+11 28631
- NbhdGrp1       1 4.1702e+11 1.0811e+12 28861
- TotalFloorSF   1 4.1960e+11 1.0837e+12 28865
> summary(backward.lm)

Call:
lm(formula = SalePrice ~ OverallQual + TotalBsmtSF + GarageCars +
    TotalFloorSF + TotalSqftCalc + NbhdGrp1 + NbhdGrp2 + NbhdGrp3,
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-87568 -11604  -1404   8732 208637

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -19786.299   3577.727  -5.530 3.81e-08 ***
OverallQual    10472.405    703.404  14.888  < 2e-16 ***
TotalBsmtSF       14.795      2.119   6.981 4.50e-12 ***
GarageCars      2932.057   1080.222   2.714  0.00672 **
TotalFloorSF      81.225      2.730  29.753  < 2e-16 ***
TotalSqftCalc     15.760      1.657   9.510  < 2e-16 ***
NbhdGrp1      -70162.477   2365.427 -29.662  < 2e-16 ***
NbhdGrp2      -45118.372   1947.390 -23.169  < 2e-16 ***
NbhdGrp3      -29420.654   1771.213 -16.610  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21770 on 1401 degrees of freedom
Multiple R-squared:  0.9158,   Adjusted R-squared:  0.9153
F-statistic:  1904 on 8 and 1401 DF,  p-value: < 2.2e-16

>
```

```
> stepwise.lm <- stepAIC(object=sqft.lm,scope=list(upper=formula(upper.lm),lower=~
1),
+                        direction=c('both'));
Start:  AIC=30113.04
SalePrice ~ TotalSqftCalc

                Df  Sum of Sq        RSS   AIC
+ OverallQual    1 1.3309e+12 1.3183e+12 29131
+ GarageCars     1 6.0322e+11 2.0460e+12 29751
+ TotalFloorSF   1 4.4407e+11 2.2051e+12 29856
+ NbhdGrp1       1 3.9726e+11 2.2519e+12 29886
+ TotalBsmtSF    1 2.5533e+11 2.3939e+12 29972
+ NbhdGrp3       1 8.4361e+09 2.6407e+12 30111
+ NbhdGrp2       1 7.9644e+09 2.6412e+12 30111
<none>                        2.6492e+12 30113
- TotalSqftCalc  1 5.2349e+12 7.8841e+12 31649

Step:  AIC=29130.96
SalePrice ~ TotalSqftCalc + OverallQual

                Df  Sum of Sq        RSS   AIC
+ NbhdGrp1       1 9.1976e+10 1.2263e+12 29031
+ GarageCars     1 9.1198e+10 1.2271e+12 29032
+ TotalFloorSF   1 7.4901e+10 1.2434e+12 29051
+ TotalBsmtSF    1 3.9364e+10 1.2789e+12 29090
+ NbhdGrp2       1 5.1182e+09 1.3132e+12 29128
<none>                        1.3183e+12 29131
+ NbhdGrp3       1 3.0924e+08 1.3180e+12 29133
- OverallQual    1 1.3309e+12 2.6492e+12 30113
- TotalSqftCalc  1 1.4039e+12 2.7222e+12 30151

Step:  AIC=29030.98
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1

                Df  Sum of Sq        RSS   AIC
+ TotalFloorSF   1 2.0714e+11 1.0192e+12 28772
+ GarageCars     1 7.0304e+10 1.1560e+12 28950
+ NbhdGrp2       1 3.7374e+10 1.1889e+12 28989
+ TotalBsmtSF    1 1.8044e+10 1.2083e+12 29012
+ NbhdGrp3       1 6.9098e+09 1.2194e+12 29025
<none>                        1.2263e+12 29031
- NbhdGrp1       1 9.1976e+10 1.3183e+12 29131
- OverallQual    1 1.0256e+12 2.2519e+12 29886
- TotalSqftCalc  1 1.4517e+12 2.6780e+12 30130

Step:  AIC=28772.09
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF

                Df  Sum of Sq        RSS   AIC
+ NbhdGrp2       1 1.7112e+11 8.4804e+11 28515
+ TotalBsmtSF    1 6.6669e+10 9.5248e+11 28679
+ GarageCars     1 3.4401e+10 9.8475e+11 28726
+ NbhdGrp3       1 7.8409e+09 1.0113e+12 28763
<none>                        1.0192e+12 28772
- TotalFloorSF   1 2.0714e+11 1.2263e+12 29031
- NbhdGrp1       1 2.2422e+11 1.2434e+12 29051
- TotalSqftCalc  1 2.9598e+11 1.3151e+12 29130
```

```
- OverallQual    1 4.4466e+11 1.4638e+12 29281

Step:  AIC=28514.93
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF + 
    NbhdGrp2

                Df  Sum of Sq        RSS   AIC
+ NbhdGrp3       1 1.5670e+11 6.9133e+11 28229
+ TotalBsmtSF    1 3.8412e+10 8.0962e+11 28452
+ GarageCars     1 1.7433e+10 8.3060e+11 28488
<none>                        8.4804e+11 28515
- TotalSqftCalc  1 1.6566e+11 1.0137e+12 28765
- NbhdGrp2       1 1.7112e+11 1.0192e+12 28772
- OverallQual    1 2.4905e+11 1.0971e+12 28876
- TotalFloorSF   1 3.4089e+11 1.1889e+12 28989
- NbhdGrp1       1 3.8021e+11 1.2282e+12 29035

Step:  AIC=28228.86
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF + 
    NbhdGrp2 + NbhdGrp3

                Df  Sum of Sq        RSS   AIC
+ TotalBsmtSF    1 2.3786e+10 6.6755e+11 28182
+ GarageCars     1 4.1768e+09 6.8716e+11 28222
<none>                        6.9133e+11 28229
- TotalSqftCalc  1 9.3584e+10 7.8492e+11 28406
- OverallQual    1 1.4550e+11 8.3683e+11 28496
- NbhdGrp3       1 1.5670e+11 8.4804e+11 28515
- NbhdGrp2       1 3.1998e+11 1.0113e+12 28763
- TotalFloorSF   1 4.5180e+11 1.1431e+12 28936
- NbhdGrp1       1 5.3671e+11 1.2280e+12 29037

Step:  AIC=28181.5
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF + 
    NbhdGrp2 + NbhdGrp3 + TotalBsmtSF

                Df  Sum of Sq        RSS   AIC
+ GarageCars     1 3.4921e+09 6.6405e+11 28176
<none>                        6.6755e+11 28182
- TotalBsmtSF    1 2.3786e+10 6.9133e+11 28229
- TotalSqftCalc  1 4.2445e+10 7.0999e+11 28266
- OverallQual    1 1.1539e+11 7.8294e+11 28404
- NbhdGrp3       1 1.4208e+11 8.0962e+11 28452
- NbhdGrp2       1 2.7701e+11 9.4456e+11 28669
- TotalFloorSF   1 4.7215e+11 1.1397e+12 28934
- NbhdGrp1       1 4.7508e+11 1.1426e+12 28937

Step:  AIC=28176.1
SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + TotalFloorSF + 
    NbhdGrp2 + NbhdGrp3 + TotalBsmtSF + GarageCars

                Df  Sum of Sq        RSS   AIC
<none>                        6.6405e+11 28176
- GarageCars     1 3.4921e+09 6.6755e+11 28182
- TotalBsmtSF    1 2.3101e+10 6.8716e+11 28222
- TotalSqftCalc  1 4.2869e+10 7.0692e+11 28262
- OverallQual    1 1.0506e+11 7.6912e+11 28381
```

```
- NbhdGrp3       1 1.3078e+11 7.9483e+11 28428
- NbhdGrp2       1 2.5443e+11 9.1848e+11 28631
- NbhdGrp1       1 4.1702e+11 1.0811e+12 28861
- TotalFloorSF   1 4.1960e+11 1.0837e+12 28865
> summary(stepwise.lm)

Call:
lm(formula = SalePrice ~ TotalSqftCalc + OverallQual + NbhdGrp1 + 
    TotalFloorSF + NbhdGrp2 + NbhdGrp3 + TotalBsmtSF + GarageCars, 
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max 
-87568 -11604  -1404   8732 208637 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -19786.299   3577.727  -5.530 3.81e-08 ***
TotalSqftCalc     15.760      1.657   9.510  < 2e-16 ***
OverallQual    10472.405    703.404  14.888  < 2e-16 ***
NbhdGrp1      -70162.477   2365.427 -29.662  < 2e-16 ***
TotalFloorSF      81.225      2.730  29.753  < 2e-16 ***
NbhdGrp2      -45118.372   1947.390 -23.169  < 2e-16 ***
NbhdGrp3      -29420.654   1771.213 -16.610  < 2e-16 ***
TotalBsmtSF       14.795      2.119   6.981 4.50e-12 ***
GarageCars      2932.057   1080.222   2.714  0.00672 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21770 on 1401 degrees of freedom
Multiple R-squared:  0.9158,	Adjusted R-squared:  0.9153 
F-statistic:  1904 on 8 and 1401 DF,  p-value: < 2.2e-16

> junk.lm <- lm(SalePrice ~ GarageCars + TotalBsmtSF, data=train.clean)
> summary(junk.lm)

Call:
lm(formula = SalePrice ~ GarageCars + TotalBsmtSF, data = train.clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-210937  -30810   -3452   24855  309395 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10218.591   3897.746   2.622  0.00884 ** 
GarageCars  47273.453   1897.994  24.907  < 2e-16 ***
TotalBsmtSF    84.749      3.535  23.974  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47200 on 1407 degrees of freedom
Multiple R-squared:  0.6024,	Adjusted R-squared:  0.6018 
F-statistic:  1066 on 2 and 1407 DF,  p-value: < 2.2e-16
```