**Century 21 Ames Study on Home Sale Prices vs Gross Living Area for BrookSide, Edwards & N Ames Neighborhoods**

**Question 1: Problem Statement**

In the N Ames, Edwards and BrkSide neighborhoods estimate how the SalePrice of the house is related to the square footage of the living area of the house (GrLIvArea). Also, determine if the SalesPrice (and its relationship to SqFt) depends on which neighborhood the house is located in.

H0: = =

Ha: != !=

**Methodology: Build and Fit the Model: Step #1 Plot the Data:**

**a.** Initial scatterplot plot identified need for transformation and suspicious observations (outliers/influential observations) were identified. Initial plots and regression model values are in Appendix A. We then used indicator variables to represent neighborhoods as referenced below:

|  |  |  |
| --- | --- | --- |
| **Neighborhood Name** | **Neighborhood Ind 1** | **Neighborhood Ind 2** |
| BrkSide | 1 | 0 |
| Edwards | 0 | 1 |
| NAmes | 0 | 0 |

**b.** Step-by-step we removed suspicious outlier values that do not fit the data and prevent assumptions being valid. Once outliers removed the data appeared to be equally distributed around linear regression line. No transformation deemed necessary.

The outliers were caused by sale price that were out of proportion to the size of gross living area or the house overall cost and size were too low or too high compared with the study data. Examples causing the outliers were unfinished property and tiny house compared with the study properties.

Initial 383 homes in study neighborhoods. Identified 19 homes that were suspicious as described above and removed from the final study. Ran the regression after each suspicious outlier and determine there were changes to p-values and betas. There were additional values that may be classified suspicious from the charts but excluding them were not influential in the analysis of data. As part of the data review up to 37 identified homes were removed but since were not identified as influential they were added back to the study.

**Step #2 – Construct a Model**

Compare the non-interaction model vs. the interaction model:

**Non-Interaction Model:**μ{Sale Price│〖ft〗^2,Neighborhood}= β\_0+ β\_1 〖ft〗^2 + β\_2 Neighborhood\_ind\_1+ β\_3 Neighborhood\_ind\_2

vs.

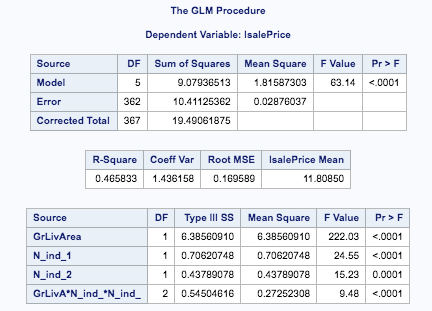
**Interaction Model:**

+ + + +

**Neighborhood Dummy Variables**

|  |  |  |
| --- | --- | --- |
| **Neighborhood Name** | **Neighborhood\_ind\_1** | **Neighborhood\_ind\_2** |
| BrkSide | 1 | 0 |
| Edwards | 0 | 1 |
| NAmes | 0 | 0 |

**Interaction Model --- Type III Sum of Squares test**

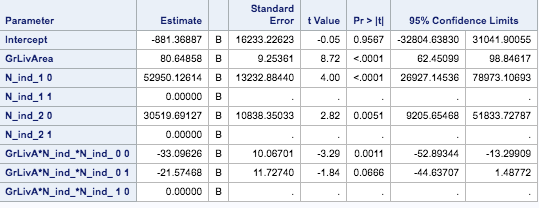


Pvalues are all <.0001. This model is a good fit for explaining the linear relationship Sales Price vs Sq Footage for the three neighborhoods in study.

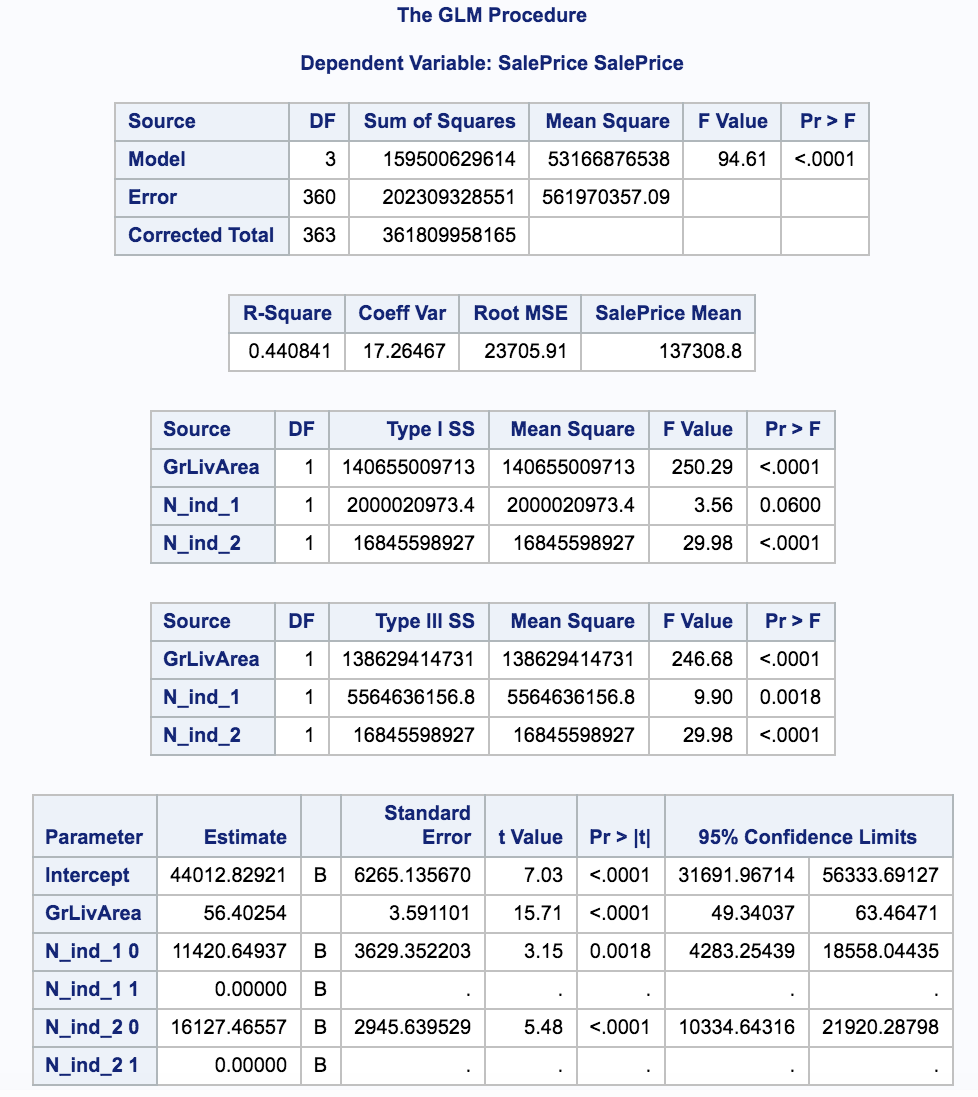
**Step #3 – Fit the Model**

Using SAS output the following model was produced:

**Interaction Model SAS Output:**



In the above results, one of the Interaction Model p-values (0.67) is greater that alpha and rejecting this model. In the analysis below, using the non-Interaction model with the p-values (<0.0001 to 0.0018), Type III SS p-values are (<0.0001 to 0.0018), and scatterplot.



**Using Non-Interaction Model to Fit Model:**

μ{Sale Price│〖ft〗^2,Neighborhood}= 44012.83 + 56.40〖ft〗^2 + 11420.65 \*Neighborhood\_ind\_1+ 16127.47\* Neighborhood\_ind\_2

55,433.48 + 56.40\*sqft

60,140.30 + 56.40\*sqft

44,012.83 + 56.40\*sqft

**2. Checking Assumptions**

**Step #4: Check the Assumptions:**

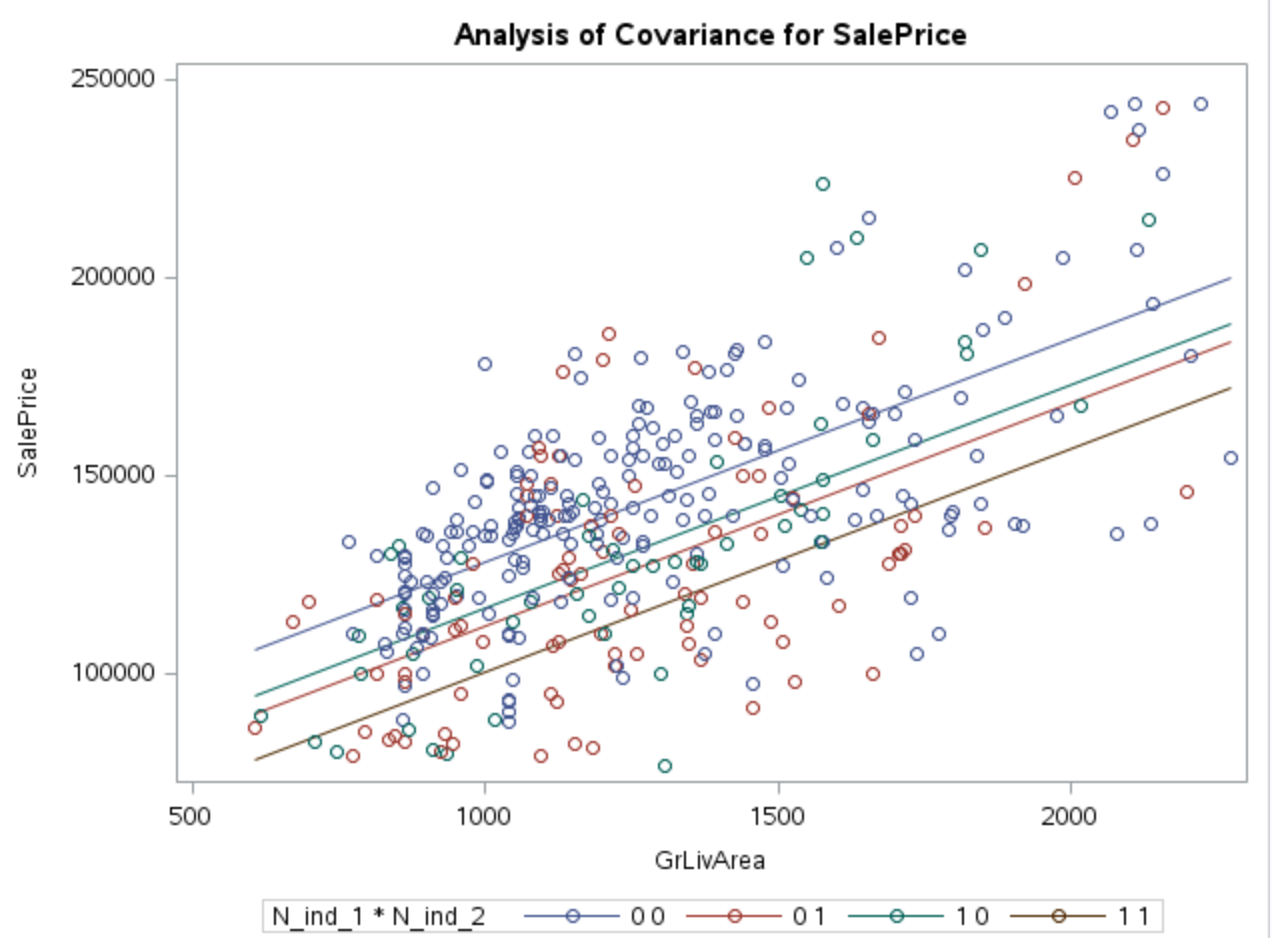
Residuals: Upon removing up to 37 outliers and settling on 15 finally removed there is minimal evidence of changing variance nor model misfit. There are homes with larger sqft but they are not excessive outliers when viewing scatter plot - reference residuals charts below.

Influence: Leverage Cook’s D and Leverage value ranges are low and no evidence of extreme or influential points. Reference Cook D and Leverage charts below.

SD Variance: There is little evidence from the scatter plots of unequal variance (heteroscedasticity)

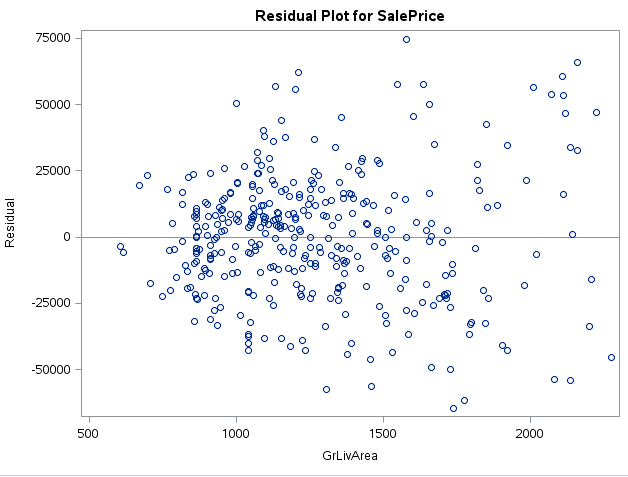
Normality: No evidence of any issue. Reference QQ plot and histogram below.

**Scatter plot (**Ignore 1 1 brown line**):**

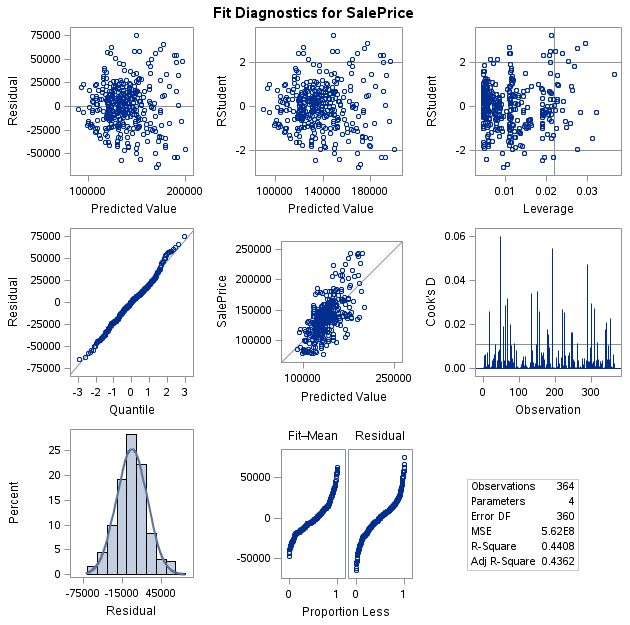


|  |  |
| --- | --- |
| **Neighborhood Name** | **Line Colors** |
| BrkSide | Green |
| Edwards | Blue |
| NAmes | Red |

**Residual Plot:** The residuals are sufficiently random. The residuals reduce as the home sizes get larger but maintain random order and reviewing the plots in next chart all appear to indicate that no extreme outlier is impacting the regression model.

****

**Cook’s D, Histogram, QQ Plot, Leverage :**



3. **Comparing Competing Models**

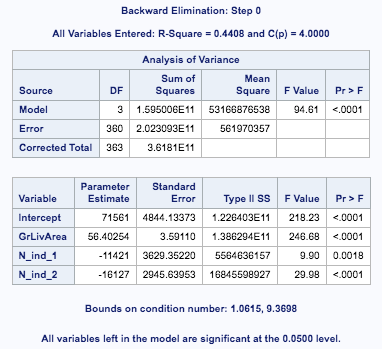
Both Forward Stepwise regression tests produce equivalent models. As shown below the intercepts (71561) and the slope (56.40) are same. Factoring in neighborhoods there is an intercept range of (55434 – 71561) with the same slope. The code is Appendix A.

Comparing Adj R Squared and Internal CV Press models, the AIC for the Adj R Squared is lower that the Internal CV Press. The lower AIC makes Adj R Squared a better model test for this analysis.

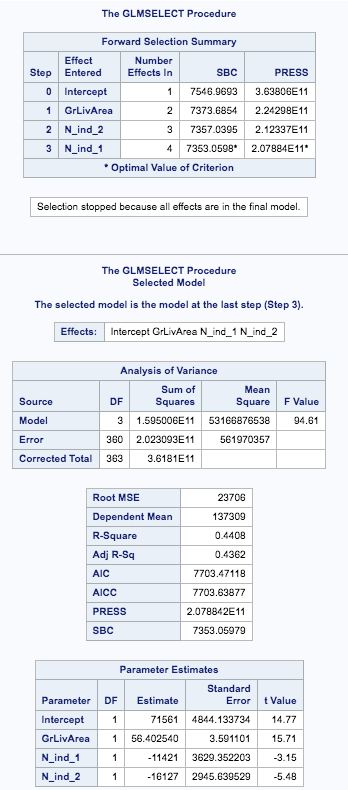
**Adj R Squared:**



**R Squared prediction:**



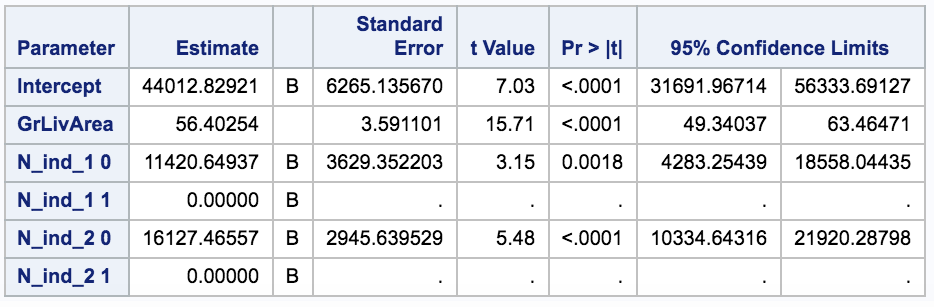
**Internal CV Press**



**4. Parameters Step#5: Interpretation & Estimates with Confidence Intervals:**

Upon analysis we determine that there is significant evidence of an association between sqft and sale price (p-value <.0001) after taking into account neighborhood.  There is insufficient evidence that Sale Price values per 100 sq ft vary between the neighborhoods (BrkSide, Edwards, NAmes).

We estimate after accounting for three neighborhoods a 100 sq ft increase in the size of house is associated with a increase in mean sale price of $5640 (95% CI of [4934 to 6346]).   Confidence Intervals below.



**5. Conclusion:** For the BrkSide and N Ames neighborhoods, there is insufficient statistical evidence that the SalesPrice (and its relationship to square footage) depends on which neighborhood the house is located in. In the Interaction Model the p-value was borderline (0.67) that further study by focusing on higher value properties and all others be completed as separate studies. For BrkSide, Edwards and Names neighborhoods, these homes were not selected at random. This was not a randomized study. Since the home sales data is observational no causal relationship can be determined.

**Question 2: Restatement of Problem**

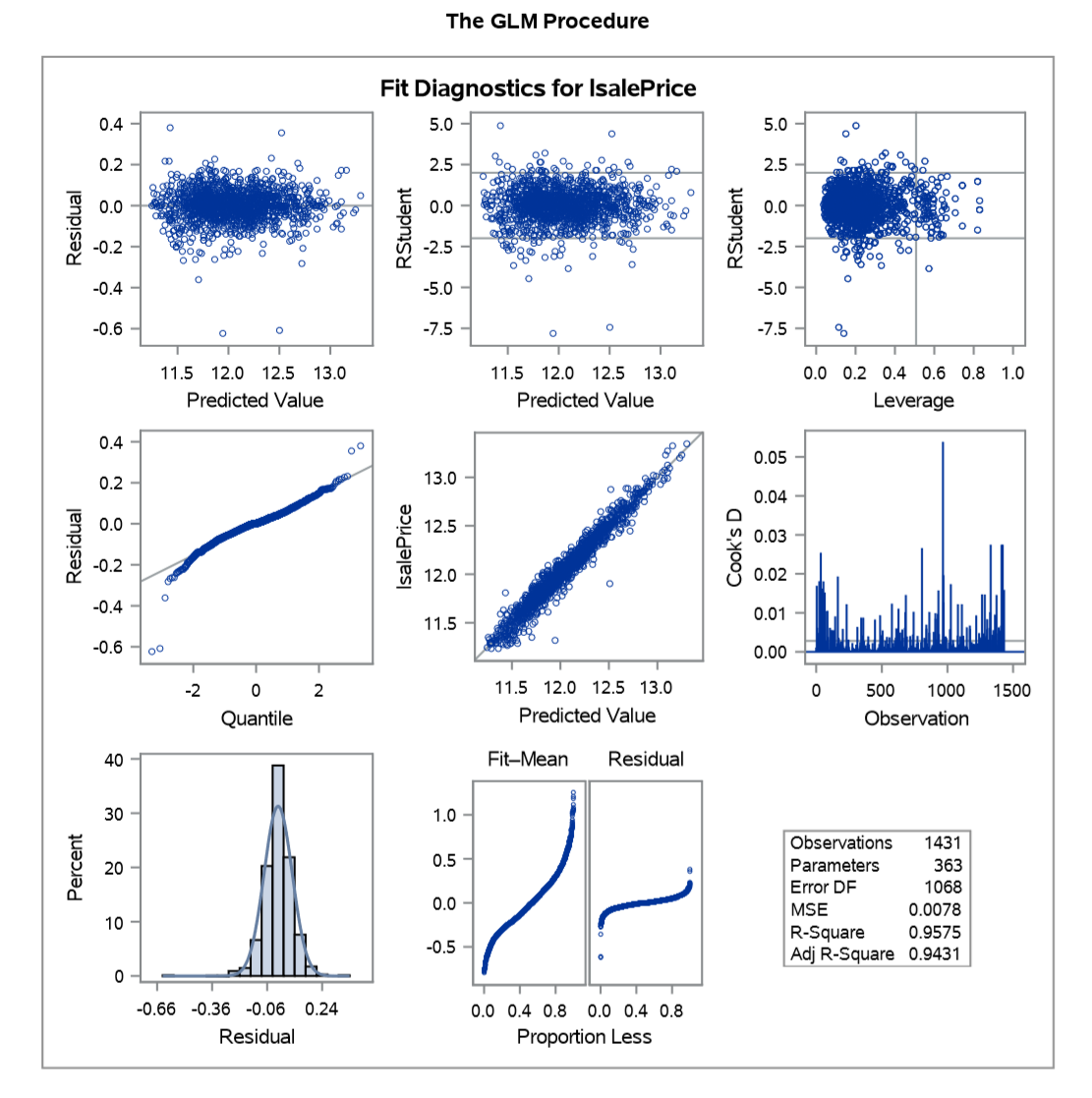
We have been instructed to build a model which can best predict the sales prices of homes in the Ames Iowa area. We were to do this given the information or techniques provided to us in MSDS 6371. In this endeavor, we were instructed to come up with four models, utilizing three different selection models and then a modified selection. The data used for these models and this project was provided to us by [https://www.kaggle.com/](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview).

**Model Selection -** Type of Selection

It is understood that although some types of selection appear to be more sophisticated and we would presume that would provide us the best model, this is not always true. No one model best fits every set of data. Thus, we have proceeded to analyze how each of the following methods would impact the predictions that we are attempting to make given the train and test data sets for this particular housing market.

* Stepwise – This selection method has multiple stages and first adds an explanatory variable and then at each stage (after adding the variable) performs a backwards elimination. These steps would continue for every predictor, until all terms that are statistically significant are found in the model.
* Forward – This selection method initializes with no predictors in the model and proceeds to fit simple linear regression models procedurally (one at a time) for all potential predictors. (This is based on the following equation (). The selection method evaluates if it is statistically significant and then proceeds to leave that predictor in the model. It continues this process until it discovers a predictor that manages to spike the significantly.
  + The method is most likely the easiest to understand and was simple to implement and perform, however it can sometimes give a bad initial fit and doesn’t guarantee the best subset selections especially with collinear variables. It does tend to follow a more cautious model, as it would leave us with a model with less predictors in most cases.
* Backward – This selection method started with all of our predictors in the model and runs an equation () to drop variables that do not have statistical significance or are not causative to a large change in the .
  + The method starts with a good fit especially with larger σ2 , but cannot guarantee the best subset selections as collinear variables could be deleted randomly, making it highly subjective to the seed or process especially when utilized with imputation.
* Custom – This method we chose to select a specific set of predictors after running a scatterplot which consisted of a matrix of variables. We attempted to see if we could select some of the transformed variables we felt would hold good correlations and thus present us with good predictions.

**Checking Assumptions**



* Linearity: The data is spread out along the x axis and is mostly linear, except for a few outliers.
* Normality: The original models looked fine, however there were a decent number fewer outliers after the log model was processed, from there it looks slightly better.
* Equal standard deviations: The standard deviations here look very good from the residual scatter plots.
* Independence: From the data provided we assume that each variable that is contained in this data is independent. We have moved through several selection methods to attempt to ensure this as best we can.
* Outliers: There are a few outliers although with the data given they don’t appear to have much influence. There are two points of concern after the transformation that could be seen as moderately high influence with high leverage, but with the given value there’s no need to attempt to modify them. The other points that exist in our plots are of high leverage but low influence and again the rest of the data allows our model to be very normalized.

**Comparing Competing Models**

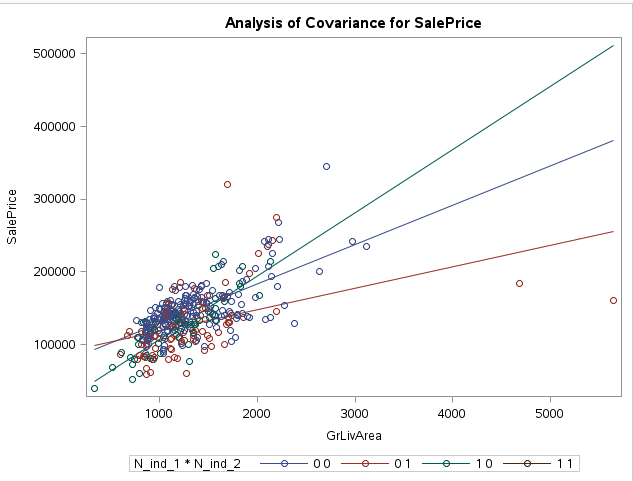
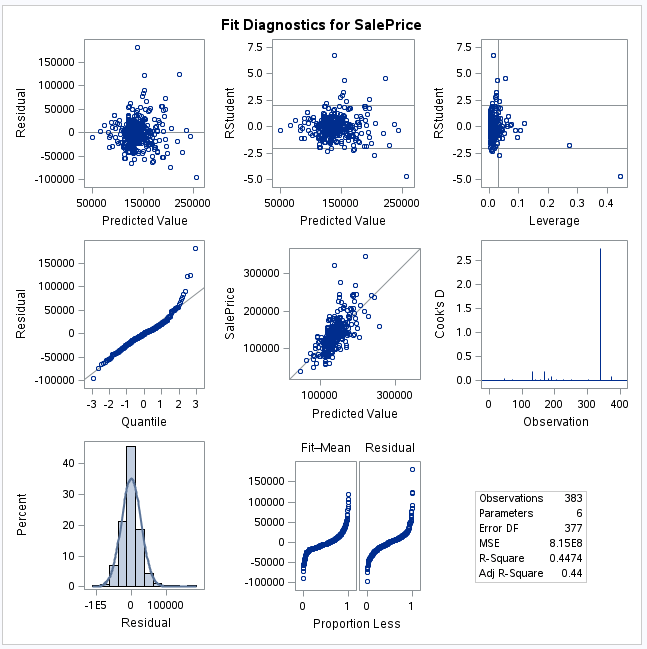
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictive Models | Adjusted R2 | CV Press | Kaggle Score | AIC |
| Forward | 0.8103 | 1.87E+12 | 0.1525 | 32021 |
| Backward | 0.9296 | 2.60E+12 | 0.25738 | 30844 |
| Stepwise | 0.8683 | 1.35E+12 | 0.20705 | 31529 |
| Custom | 0.9053 | 5.41E+11 | 0.48312 | 31148 |

**Conclusion: A short summary of the analysis.**

We began to touch on the conclusion of our model and why we had moved in the direction that we had in the Model Selection section of this analysis. Our best model was found through trial and error, as it seems most statistical processes undergo. Before we could dive into the predictors we would utilize, we needed to first clean up the data provided. Some columns included mixed variables (character/strings and numeric values) and others came with missing values. It was decided that we would use a process to impute the missing values so that we had a complete dataset. After this, we ran several the prescribed methods for variable selection. We were sure to specify a seed for each of these imputations to create a baseline for reproducible results. From there we discovered that forward selection provided us the best predictive model. This was decided after carefully evaluating the AIC, Adjusted and CV Press values. The AIC helped us interpret the maximized likelihood of all data being used versus the number of free parameters, where a higher score means that the model would be more restricted (meaning we would have greater number of free parameters). Thus, our forward model was more restrictive in its variable selection process than the other three methods of model selection. We felt that this model, as it had a close CV Press score but lower Adjusted would be best given the AIC value. From here we compared our work to the Kaggle website and confirmed that our forward model after imputation was our strongest predicting model.

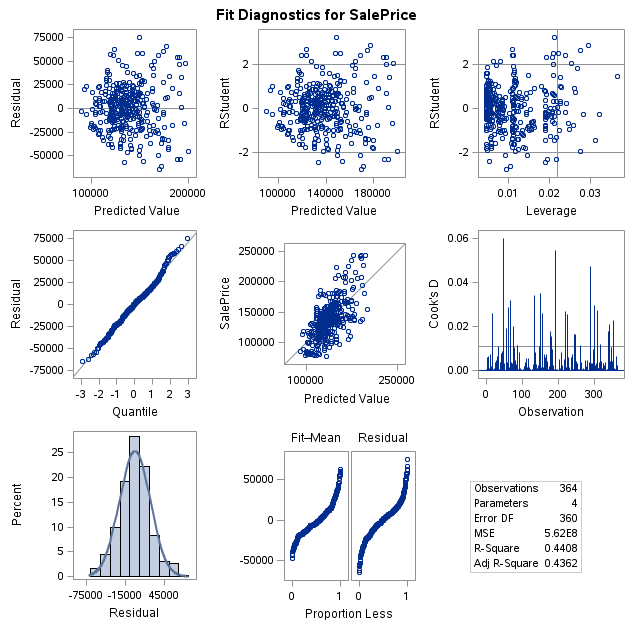
**Appendix A:**

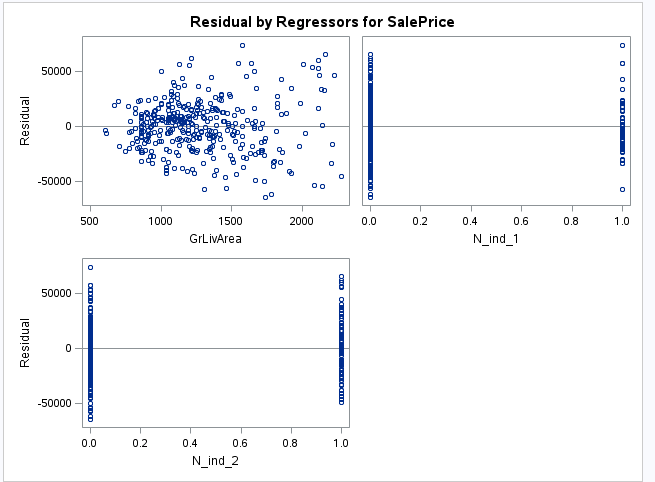
Q1 Initial regression scatterplot and regression model values:



**Q1 R Squared Plots (charts are labeled next to each plot):**





Q1 Code:

FILENAME REFFILE '/home/tomgianelle0/Project/train.xlsx';

PROC IMPORT DATAFILE=REFFILE

REPLACE

DBMS=XLSX

OUT=housing;

GETNAMES=YES;

RUN;

proc print data=housing (obs=10);

run;

DATA t1data;

SET housing;

IF Neighborhood = "BrkSide" | Neighborhood = "Edwards" | Neighborhood = "NAmes" ;

RUN;

data idata;

set t1data;

if Neighborhood = "BrkSide" THEN N\_ind\_1 = 1;

if Neighborhood = "Edwards" THEN N\_ind\_1 = 0;

if Neighborhood = "NAmes" THEN N\_ind\_1 = 0;

if Neighborhood = "BrkSide" THEN N\_ind\_2 = 0;

if Neighborhood = "Edwards" THEN N\_ind\_2 = 1;

if Neighborhood = "NAmes" THEN N\_ind\_2 = 0;

if GrLivArea < 2350;

if SalePrice < 250000;

if SalePrice > 75000;

lsalePrice = log(SalePrice);

run;

/\* Non - Interaction Model \*/

PROC GLM DATA = idata PLOTS=ALL;

CLASS N\_ind\_1 N\_ind\_2;

MODEL SalePrice = GrLivArea N\_ind\_1 N\_ind\_2 / SOLUTION CLparm alpha=.05;

OUTPUT OUT = fitted PREDICTED = muHat;

RUN;

/\* Interaction Model \*/

proc glm data = idata PLOTS=All;

CLASS N\_ind\_1 N\_ind\_2;

model SalePrice=GrLivArea N\_ind\_1 N\_ind\_2 GrLivArea\*N\_ind\_1\*N\_ind\_2 / SOLUTION CLPARM ;

run;

/\* Adj R^2 \*/

proc reg data=idata outest=est1 rsquare;

model SalePrice = GrLivArea N\_ind\_1 N\_ind\_2 / ADJRSQ selection = forward AIC slstay = 0.05;

run;

proc print data=est1;

run;

/\* Internal CV Press \*/

proc glmselect data=idata;

model SalePrice = GrLivArea N\_ind\_1 N\_ind\_2/selection=forward(stop=PRESS);

run;

Q2 Code:

PROC IMPORT DATAFILE='/folders/myfolders/sasuser.v94/train.csv'

DBMS=CSV

OUT=train;

GETNAMES=YES;

RUN;

PROC IMPORT DATAFILE='/folders/myfolders/sasuser.v94/test.csv'

DBMS=CSV

OUT=test;

GETNAMES=YES;

RUN;

data train\_clean;

SET train;

logSalePrice = log(SalePrice);

logGrLIvArea = log(GrLIvArea);

run;

/\* Adding SalePrice Column To Test Set\*/

data test;

set test;

SalePrice = .;

run;

/\* Evaluate missing values (train)\*/

proc means data=train\_clean NMISS N; run;

/\* Clean the data (in train: replace missing values with mean)\*/

proc hpimpute data = train out = train\_clean;

input MasVnrArea GarageYrBlt;

impute MasVnrArea / method=mean;

impute GarageYrBlt / method=mean;

ID MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual smtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition SalePrice LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical

KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

/\* Reasses missing values (train)\*/

proc means data=train\_clean NMISS N; run;

/\* Appending the test and train set\*/

data complete;

set train test;

run;

/\* Cleaning the Appended Data set \*/

proc HPImpute data = complete out = complete\_clean;

input MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF BsmtFullBath

BsmtHalfBath GarageYrBlt GarageCars GarageArea;

impute MasVnrArea / method=mean;

impute BsmtFinSF1 / method=mean;

impute BsmtFinSF2 / method=mean;

impute BsmtUnfSF / method=mean;

impute TotalBsmtSF / method=mean;

impute BsmtFullBath / method=mean;

impute BsmtHalfBath / method=mean;

impute GarageYrBlt / method=mean;

impute GarageCars / method=mean;

impute GarageArea / method=mean;

ID ID SalePrice MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature

SaleType SaleCondition LotArea OverallQual OverallCond YearBuilt YearRemodAdd \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold MsSubClassMSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

proc means data = complete\_clean mean; run;

/\* Forward Selection (All Variables from Training Data Set) \*/

proc glmselect data = train\_clean seed=555;

class MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePrice = LotArea OverallQual OverallCond YearBuilt YearRemodAdd M\_MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBathBsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces M\_GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSoldMsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition / selection = forward (select=cv choose=cv stop=cv) CVDETAILS;

run;

/\* Backward Selection (All Variables from Training Data Set) \*/

proc glmselect data = train\_clean seed=333;

class MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePrice = LotArea OverallQual OverallCond YearBuilt YearRemodAdd M\_MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces M\_GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition / selection = backward (select=cv choose=cv stop=cv) CVDETAILS;

run;

/\* StepWise Selection (All Variables from Training Data Set) \*/

proc glmselect data = train\_clean seed=333;

class MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePrice = LotArea OverallQual OverallCond YearBuilt YearRemodAdd M\_MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces M\_GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition / selection = stepwise (select=cv choose=cv stop=cv) CVDETAILS;

run;

/\* Analysis 2: Submitting Kaggle Scores \*/

/\* Using Forward Selection Model for SalePrice Prediction/Kaggle Submission \*/

proc glm data=complete\_clean plots(unpack)=all;

class LandContour Neighborhood RoofMatl BsmtQual BsmtExposure BsmtFinType1 KitchenQual;

model SalePrice= MsSubClass LandContour Neighborhood RoofMatl BsmtQual BsmtExposure BsmtFinType1 KitchenQual OverallQual GrLivArea IM\_GarageCars OverallCond YearBuilt LotArea ScreenPorch / cli solution;

/\* Output the Results \*/

output out= results\_f p=Predict;

run;

proc print data=results\_f; run;

/\* Creating CSV file with predictions \*/

data results\_f2;

set results\_f;

if Predict > 0 then SalePrice = Predict;

/\* Set any intelligent guess for the SalePrice \*/

if Predict < 0 then SalePrice = 181000;

/\* Retain only the SalePrice \*/

keep ID SalePrice;

/\* Separate only values that appear in the test data \*/

where ID > 1460;

run;

proc print data=results\_f2;run;

/\* Exporting CSV File for Submission \*/

proc export data=results\_f2 DBMS=csv outfile="/folders/myfolders/sasuser.v94/results\_f2.csv"; run;

/\* Using Backward Elimination Model for SalePrice Prediction/Kaggle Submission \*/

proc glm data=complete\_clean;

class MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePrice = LotArea OverallQual OverallCond YearBuilt YearRemodAdd IM\_MasVnrArea IM\_BsmtFinSF1 IM\_BsmtFinSF2 IM\_BsmtUnfSF \_1stFlrSF \_2ndFlrSF LowQualFinSF IM\_BsmtFullBath IM\_BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces IM\_GarageYrBlt IM\_GarageCars IM\_GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold MsSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LotFrontage LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition / cli solution;

/\* Output the Results \*/

output out= results\_b p=Predict;

run;

/\* Create CSV File with Predictions \*/

proc print data=results\_b; run;

data results\_b2;

set results\_b;

if Predict > 0 then SalePrice = Predict;

/\* Set any intelligent guess for the SalePrice \*/

if Predict < 0 then SalePrice = 181000;

/\* Retain only the SalePrice \*/

keep ID SalePrice;

/\* Separate only values that appear in the test data \*/

where ID > 1460;

run;

proc print data=results\_b2; run;

/\* Exporting CSV File for Submission \*/

proc export data=results\_b2 DBMS=csv outfile="/folders/myfolders/sasuser.v94/results\_b2.csv"; run;

/\* Predictions from Stepwise Elimination Model \*/

proc glm data=complete\_clean;

class MsSubClass LandContour LotConfig Neighborhood Condition1 BldgType Foundation BsmtQual BsmtExposure BsmtFinType1 KitchenQual Functional Fireplaces GarageQual GarageCond PavedDrive SaleType GarageFinish;

model SalePrice = OverallQual GrLivArea OverallCond YearBuilt YearRemodAdd IM\_BsmtFullBath FullBath HalfBath KitchenAbvGr Fireplaces IM\_GarageCars WoodDeckSF ScreenPorch MsSubClass LandContour LotConfig Neighborhood Condition1 BldgType BsmtQual BsmtExposure BsmtFinType1 KitchenQual Functional EnclosedPorch GarageFinish GarageQual PavedDrive SaleType / cli solution;

/\* Output the Results \*/

output out= results\_sw p=Predict;

run;

proc print data=results\_sw; run;

/\* creating CSV file with predictions \*/

data results\_sw2;

set results\_sw;

if Predict > 0 then SalePrice = Predict;

/\* Set any intelligent guess for the SalePrice \*/

if Predict < 0 then SalePrice = 181000;

/\* Retain only the SalePrice \*/

keep ID SalePrice;

/\* Separate only values that appear in the test data \*/

where ID > 1460;

run;

/\* Exporting CSV File for Submission \*/

proc print data=results\_sw2; run;

proc export data=results\_sw2 DBMS=csv outfile="/folders/myfolders/sasuser.v94/results\_sw2.csv";run;

/\* Analysis 2: Final Model (Using Forward Pred. Model) \*/

/\* Predictions from Forward Selection Model \*/

/\* before transform (Matrix Scatterplot)\*/

proc sgscatter data=complete\_clean;

title "Scatterplot Matrix for Forward Predictive Model (No Transform)";

matrix SalePrice MsSubClass OverallQual GrLivArea IM\_GarageCars OverallCond YearBuilt LotArea ScreenPorch;

run;

/\*Proc GLM before transform \*/

proc glm data=complete\_clean plots(unpack)=all;

class LandContour Neighborhood RoofMatl BsmtQual BsmtExposure BsmtFinType1 KitchenQual;

model SalePrice = MsSubClass LandContour Neighborhood RoofMatl BsmtQual BsmtExposure BsmtFinType1 KitchenQual OverallQual GrLivArea IM\_GarageCars OverallCond YearBuilt LotArea ScreenPorch / cli solution;

/\* Output the Results After Transformation (Log-Log) \*/

output out= results\_f p=Predict; run;

data complete\_clean;

set complete\_clean;

logSalePrice =log(SalePrice);

logMsSubClass = log(MsSubClass);

logOverallQual = log(OverallQual);

logGrLivArea = log(GrLivArea);

logIM\_GarageCars = log(IM\_GarageCars);

logOverallCond = log(OverallCond);

logYearBuilt = log(YearBuilt);

logLotArea = log(LotArea);

logScreenPorch = log(ScreenPorch);

run;

/\* Scatterplot Matrix After Log-Log Transform\*/

proc sgscatter data=complete\_clean;

title "Scatterplot Matrix for Forward Predictive Model(Log-Log)";

matrix logSalePrice logMsSubClass logOverallQual logGrLivArea logIM\_GarageCars logOverallCond logYearBuilt logLotArea logScreenPorch;

run;

/\* Final model after Log-Log Transform \*/

proc glm data=complete\_clean plots=all;

class LandContour Neighborhood RoofMatl BsmtQual BsmtExposure BsmtFinType1 KitchenQual;

model logSalePrice = logMsSubClass LandContour Neighborhood RoofMatl BsmtQual BsmtExposure BsmtFinType1 KitchenQual logOverallQual logGrLivArea logIM\_GarageCars logOverallCond logYearBuilt logLotArea logScreenPorch / cli solution;

run;

proc glm data=complete\_clean plots=all;

class LandContour Neighborhood RoofMatl BsmtQual BsmtExposure BsmtFinType1 KitchenQual;

model logSalePrice = logMsSubClass LandContour Neighborhood RoofMatl BsmtQual BsmtExposure BsmtFinType1 KitchenQual logOverallQual logGrLivArea logIM\_GarageCars logOverallCond logYearBuilt logLotArea logScreenPorch / cli solution;

run;

/\* Predictions from Forward Selection Model (Our Best Predictive Model) \*/

/\* before transform (Matrix Scatterplot) \*/

proc sgscatter data=complete\_clean;

title "Scatterplot Matrix for Forward Predictive Model (No Transform)";

matrix SalePrice MsSubClass OverallQual GrLivArea IM\_GarageCars OverallCond YearBuilt LotArea ScreenPorch;

run;

proc print data=results\_f; run;

/\* Custom Model \*/

proc glm data = train2\_clean plots=all;

class LotArea GarageType GarageFinish GarageQual;

Model logSalePrice = GrLivArea OverallQual OverallCond LotArea GarageType GarageFinish GarageQual / cli solution;

Output out = results p = predict;

run;

data results2;

set results;

if Predict > 0 then SalePrice = Predict;

if Predict <= 0 then SalePrice = 10000;

keep id SalePrice;

where id > 1460;

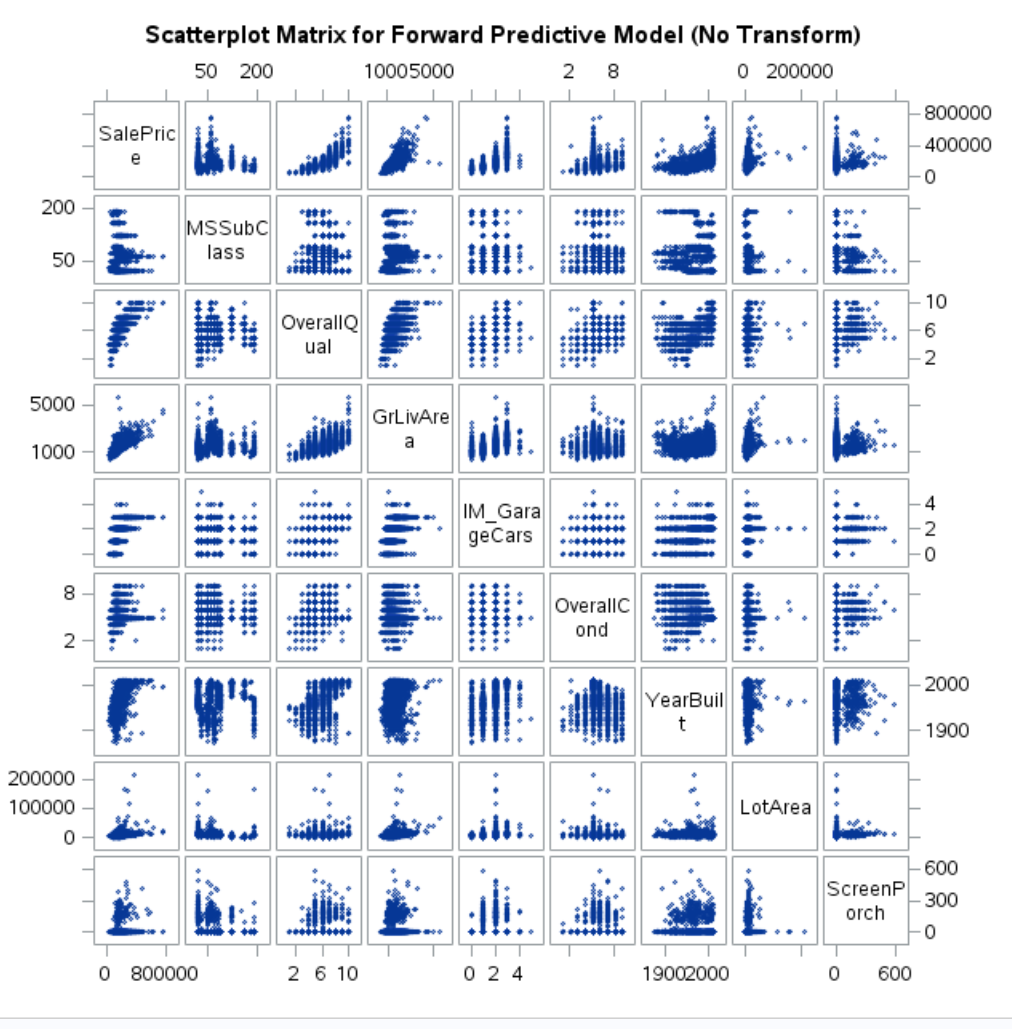
Proc means data = results2;

var SalePrice;

run;

/\* Predictions from Forward Selection Model \*/

/\* before transform (Matrix Scatterplot)\*/



/\* Scatterplot Matrix After Log-Log Transform\*/

