**Project #1 for MSDS 6372-4023**

**Best Multiple Linear Regression Model**

**for Ames, Iowa Housing Prices**

**Date Due June 8, 2017**

**Submitted by:**

James Hosker

Timothy McWilliams

Leticia Valadez

**Course: MSDS 6372-4023 (Mon, 8:30 to 10:00 PM)**

**Contact Phone Numbers**

James Hosker: (917) 689-7615

Timothy McWilliams: (609) 634-3002

Leticia Valadez: (434) 409-9431

**Best Multiple Linear Regression Model for Ames, Iowa Housing Prices**

This report provides insight into the home sales in Ames, Iowa from 2006-2010. We explore several factors that affect the sale prices of homes such as the usual suspects of neighborhood, square footage, year built and not so typical thought of factors such as a full bathroom in basement. Housing sales data for the Ames, Iowa housing market is publicly available at the following URL [Housing Data for Ames Iowa](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data). The Ames, Iowa Housing dataset was compiled by Dean De Cock of Truman State University[1]. A full description of each data field is given in the “data\_description,txt” text file at the referenced URL. Our team develops a concise model using multiple linear regression techniques for Ames’s housing sale price. The model will help real estate agents, contractors, and prospective buyers gain important insight on the factors that are currently influencing housing sale prices in Ames, Iowa. Our model can be complex due to the substantial number of explanatory variables. However, our model is narrowed down to the most significant factors to keep the model as simple as possible. This is primarily achieved by using the LASSO variable-selection method.

**Problem Statement**

We develop a multiple linear regression model based on an observed training data set of 79 explanatory variables and one sales price response variable for the Ames, Iowa housing market. We then use this model to predict housing sales prices in Ames, Iowa using a different test data set of only explanatory variables.

**Constraints and Limitations**

This is an observational study and not a control randomized study, since the data was collected on housing sales that occurred in Ames, Iowa from 2006 to 2010. Only correlation between explanatory variables and response variable (SalePrice) can be made and no causal relationship can be made between these variables. First, it is possible these data sets do not include some explanatory variables that more appropriately describe the variation in housing sales prices. Second, we do not know if the same or different brokers that may have different biases filled out the explanatory information. The data includes all residential sales in Ames, Iowa; however, approximately 100 homes changed ownership multiple times during the 4-year period. To avoid giving more weight to these 100 homes, only the most recent sales data on these homes were included in the original data sets reduced by Dean De Cock of Truman State University. Any inferences that are made can only apply to the city of Ames and not the whole state of Iowa or any other population.

**Data Set Description and Cleaning Summary**

The data sets train.csv and test.csv were downloaded from the following URL [Housing Data for Ames Iowa](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data). Both data sets each have 79 explanatory variables (36 quantitative/numerical and 43 categorical variables), describing most aspects of residential home sales in Ames, Iowa. The training data set (trian.csv) contains 1460 observations and is used to develop the multiple linear regression model while the test data set (test.csv) contained 1459 observations and is used to predict/forecast housing prices. For further analysis using this data, see the following URL: [Ames, Iowa: Alternative to Boston Housing Data Regression Project.](https://d.docs.live.net/ce543ea648312eb4/JimHDocuments/SMU/MSDS6371ExpmtStatsI/Unit%5eN13/Homework%5eN13/JHosker_GGarza_HW_13_14.docx)

[Appendix I](#Appendix_I)shows all 79 explanatory variables. The following variables are used in the final prediction model: above grade (ground) living area square feet (GrLivArea); first floor square feet (\_1stFlrSF); physical locations within Ames city limits (Neighborhood); rating from 1-worst to 10-best for the overall material and finish of the house (OverallQual); original construction date (YearBuilt); size of garage in car capacity (GarageCars); and a few others.

Described in more detail in [Appendix II](#Appendix_II), the date sets are cleaned by first eliminating the variables Alley, Utilities, PoolQC, Fence, and MiscFeatures due to possessing high frequency of NA entries. Next, some quantitative variables had a few NAs that are replaced with the mean of the variable series (e.g. LotFrontage). A few categorical variables contain NAs, which if numerical are set to 0 (e.g. BsmtHalfBath) and if text are set to category None (e.g. MasVrnType). In a few cases, variables with only one NA are replace the one the most likely value based on the Neighborhood and OverallQual classifications. Finally, two data point are removed that are outliers and highly leveraged points by the method described in [Appendix III](#Appendix_III).

**Exploratory Data Analysis and Variable Screening**

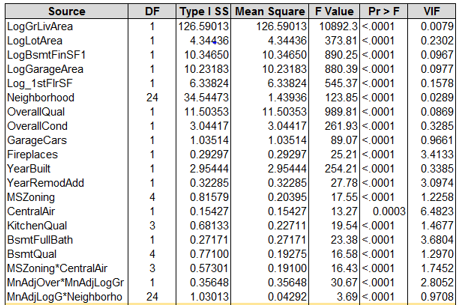
We focus on a subset of the data to get an exposure to the explanatory variables. **Table 1** shows the summary statistics of our quantitative variable after we take their natural log. Highlighted in the red box are the quantitative variables that our model utilizes.

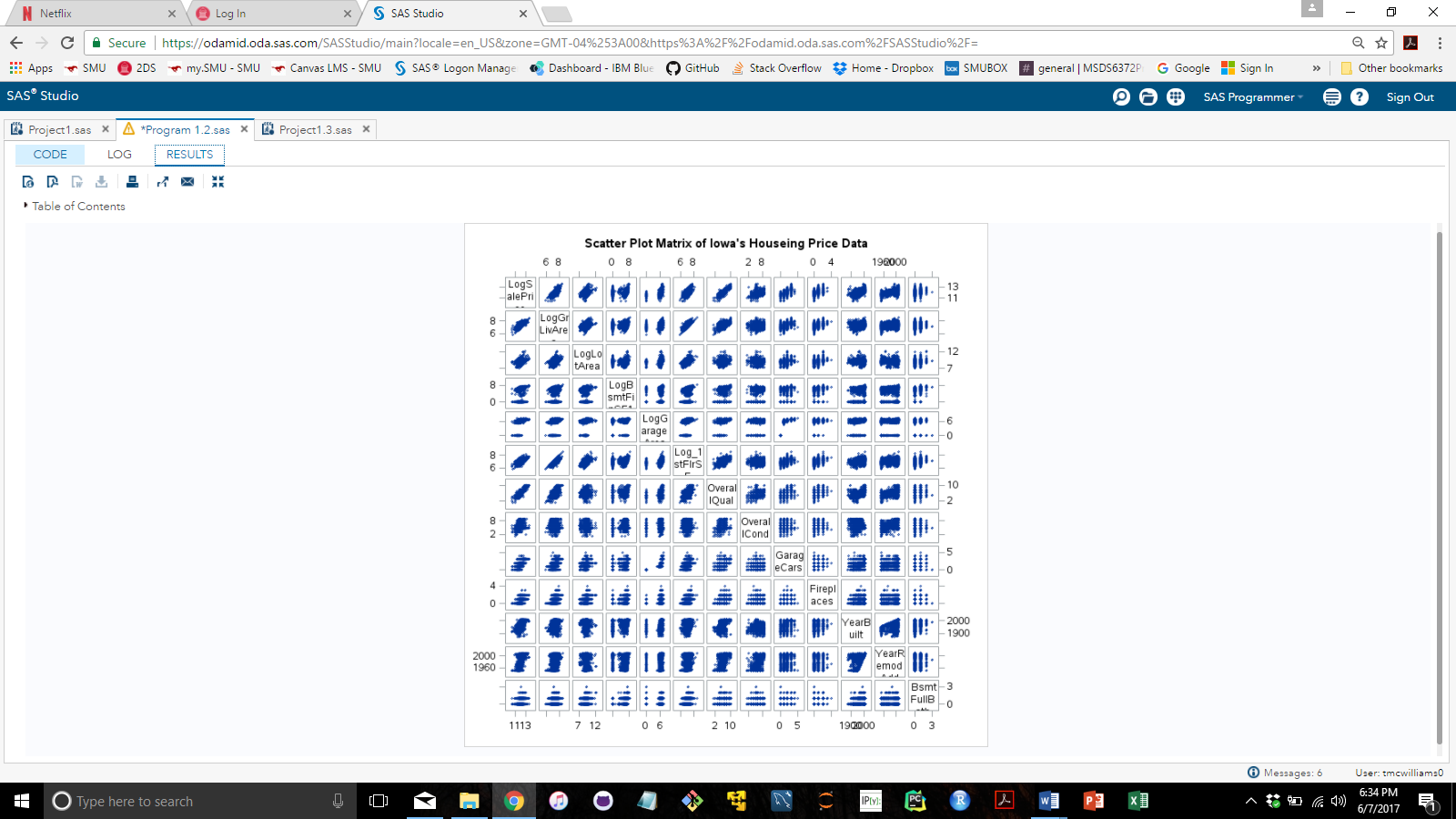
| **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Lower Quartile** | **Median** | **Upper Quartile** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| LogSalePrice  LogGrLivArea  LogGarageArea  LogBsmtFinSF1  LogLotArea  Log\_1stFlrSF  LogLotFrontage  LogMasVnrArea  LogBsmtFinSF2  LogBsmtUnfSF  LogTotalBsmtSF  Log\_2ndFlrSF  LowQualFinSF  LogWoodDeckSF  LogOpenPorchSF  LogEnclosedPorch  Log\_3SsnPorch  LogScreenPorch  LogPoolArea | 1458  1458  1458  1458  1458  1458  1458  1458  1458  1458  1458  1458  1458  1458  1458  1458  1458  1458  1458 | 12.0240  7.2660  5.8065  4.2243  9.1086  7.0067  4.0640  2.1140  0.6563  5.6473  6.7483  2.8588  5.8525  2.4532  2.3037  0.6990  0.0858  0.4112  0.0262 | 0.3997  0.3304  1.4554  2.9905  0.5137  0.3142  0.4985  2.6252  1.8462  1.8550  1.1448  3.2918  48.6560  2.5960  2.1499  1.7283  0.6673  1.4041  0.4084 | 10.4602  5.8111  0.0000  0.0000  7.1709  5.8141  2.3979  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000 | 11.7745  7.0282  5.8021  0.0000  8.9281  6.7833  4.0254  0.0000  0.0000  5.4116  6.6796  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000 | 12.0015  7.2872  6.1748  5.9480  9.1565  6.9912  4.2627  0.0000  0.0000  6.1707  6.8997  0.0000  0.0000  0.0000  3.2385  0.0000  0.0000  0.0000  0.0000 | 12.2737  7.4821  6.3578  6.5695  9.3588  7.2385  4.3307  5.1059  0.0000  6.6958  7.1686  6.5917  0.0000  5.1299  4.2341  0.0000  0.0000  0.0000  0.0000 | 13.5345  8.4065  7.2378  7.6912  12.2795  8.0799  4.6052  7.3784  7.2964  7.7566  8.0731  7.6334  572.0000  6.7546  6.3063  6.3154  6.2324  6.1759  6.6053 |

**Table 1: Summary Statistics of All Quantitative Response and Explanatory Variables**

*Scatterplots and Correlations*

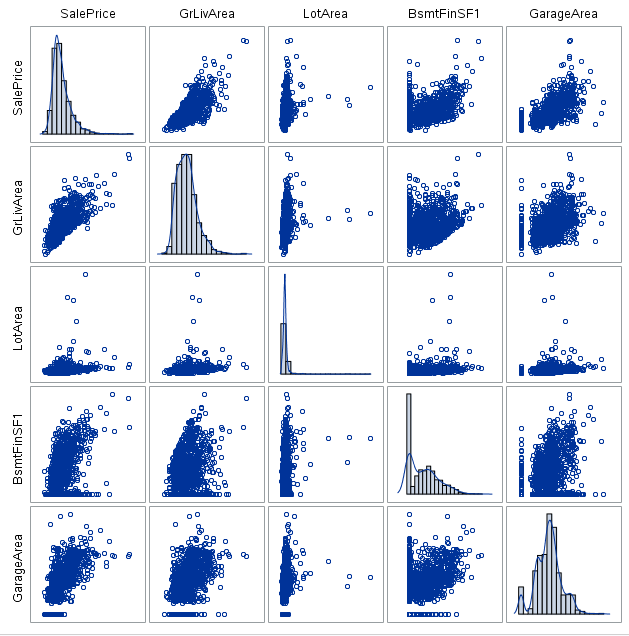
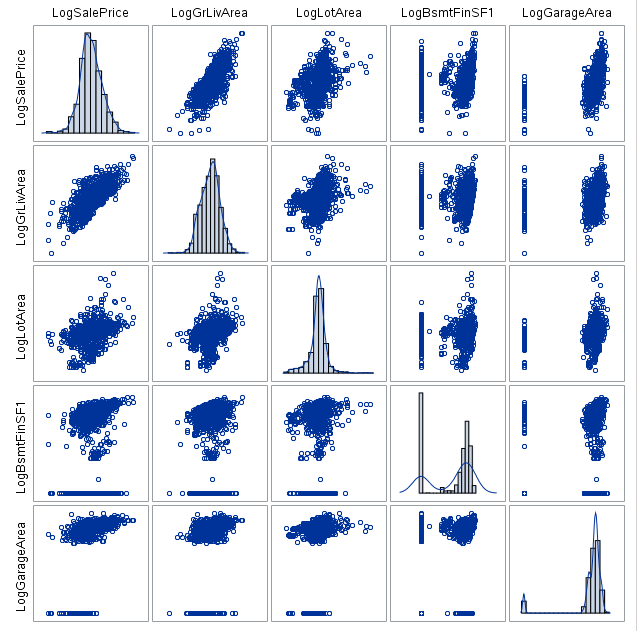
Upon examination of a matrix of scatterplots and Pearson’s coefficients, there are some weak positive correlations (Pearson’s coefficient greater than 0.5 but less than 0.8). **Figure 1** displays the scatter plot matrix of all the quantitative explanatory variables and those highlighted in red have weak correlation. The scatter plots show that there is correlation between the chosen explanatory variables and the SalePrice. We found little to no correlation between the categorical variables and quantitative explanatory variables in our model. In addition, we want to determine if variables are redundant and found that there are some correlations between quantitative explanatory variables. The objective of this analysis is to predict future sale prices rather than the importance of specific variables. Therefore, we keep all variables examined in our model since they exhibit generally weak or no correlations among each other (generally independent). In addition, due to little to no correlation among most of the explanatory variables in our model, the variance inflation factors (VIFs) should be lower than 5 as shown highlighted in red in the table in **Figure 1**. For a more detailed explanation of the correlation, see [Apendix VIII](#Appendix_VIII)**.**





**Figure 1: Scatter Plot Matrix of Numeric Explanatory Variables and Table of Variance Inflation Factors (VIFs)**

**Figure 2** shows that taking the natural log transform of the quantitative response and some of the explanatory variables used in our model provides a more normal distribution of the variables with less skewness. The scatterplot of histograms of the data show the improved normal distribution. In addition, there seems to be a linear relationship between SalesPrice and our selected quantitative variables highlighted in red.

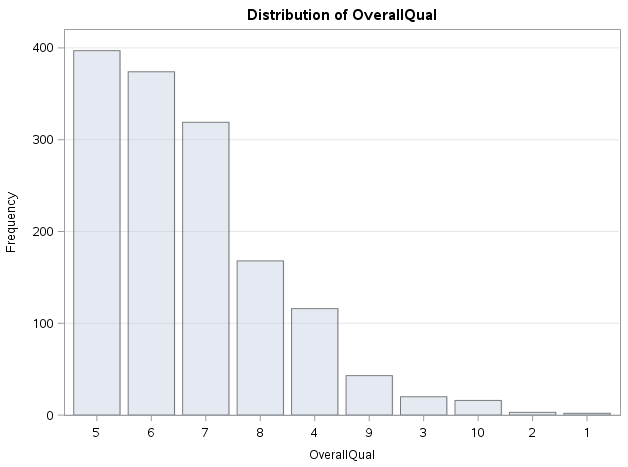
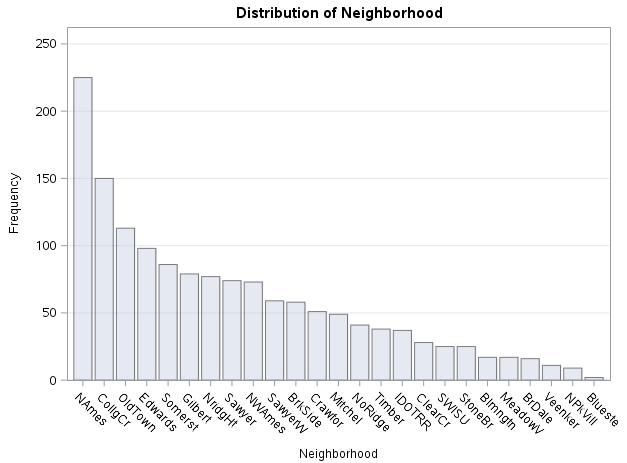
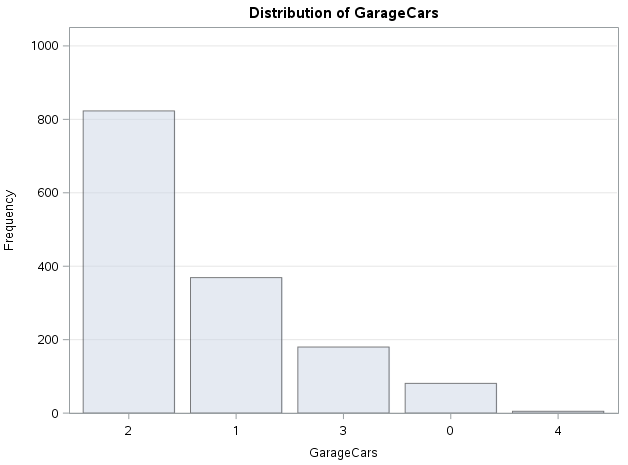
 

Natural Log

Transform

**Figure 2: Scatter Plot Relationship of All Quantitative Variables Pre and Post Log Transformation**

In **Figure 3**, a subset of categorical variables is displayed and provided a distribution of the number of houses sold for each Neighborhood, where Names has many houses sold and Blueste, the least. Also, shown are the distributions of the overall quality (OverallQual) of the houses (most are ranked 5); and the number of cars that can fit in a garage (2 cars seems to be most frequent). See [Appendix IX](#Appendix_IX) for more details.

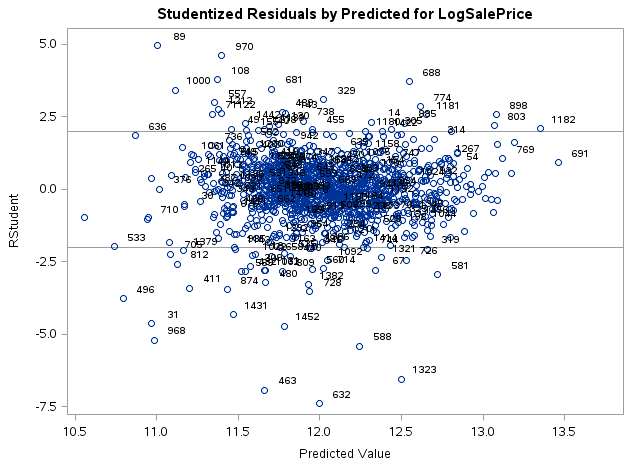
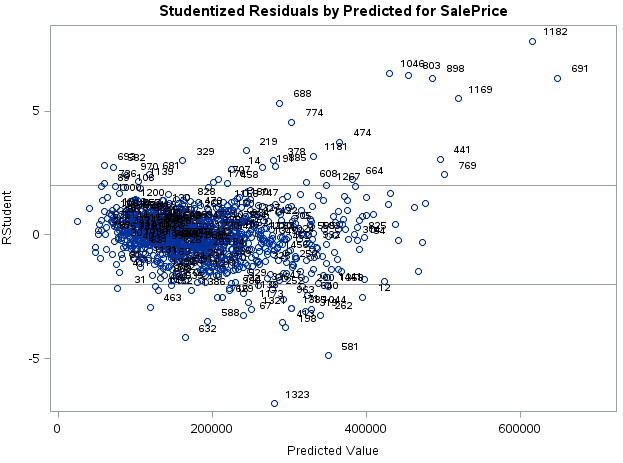
 

**Figure 3: Distribution or Frequency of Sub-categories for Different Categorical Variables.**

*Natural Log Transformation of Variables*

We took the natural log transformation of the response variable (Sales Price) and of several explanatory variables that were in square feet to improve our linear regression, reducing the right (positive) skewness in the data, generating a more normal distributed output for our model and improving constant variance of the residuals and predicted values. Moreover, in our EDA section, we saw that taking the log transformation of the explanatory variables made them more normally distributed. Later in this report, we include interaction terms and by taking the natural log transformation of the explanatory variable, we can reduce the variance inflation factor (VIF). **Figure 4** below shows that our residuals are more constant post the natural log transformation and the model is more normally distributed. For a more detailed analysis of our natural log transformation of the data, see [Appendix IV](#Appendix_IV).

**Studentized Residuals without Ln Transformation Studentized Residuals with Ln Transformation**



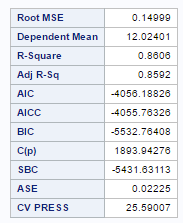
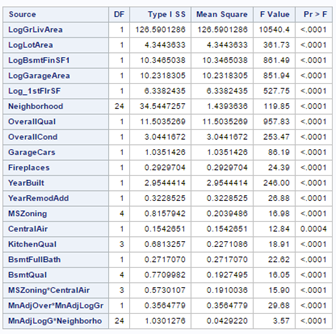
**Figure 4: Plot of Studentized Residuals vs. Predicted Values Pre and Post Natural Log Transformation**

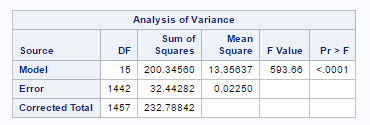
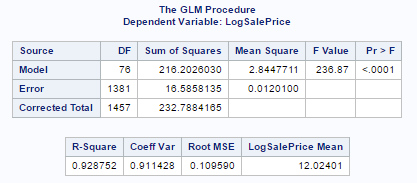
*Multilinear Sequential Variable-Selection Regression Techniques*

We utilize four sequential variable-selection techniques: forward, backward and stepwise selection as well as the Least Absolute Shrinkage and Selection Operator (LASSO) selection process. First, we use the forward, backward and stepwise selection process using the Akaike information criterion (AIC) with a stop of additions or removal based on cross-validation (CV). We want a lower AIC while including penalties for adding more regression coefficients. Second, we run the LASSO model and analyze explanatory interaction terms among the variables in our model. For this test, we increase the penalty for adding more regression coefficients using the Swartz's Bayesian Information Criterion (SBC) because we are now analyzing additional interaction terms that could be correlated with existing terms in the model. For more on the definition and formulation of models used, see [Appendix V](#Appendix_V). After running these variable selection techniques, we then select or remove a final set of explanatory variables base on these model outputs using R2, p-values, residuals and VIF (where VIF = ). For complete details on our model selection process, see [Appendix VI](#Appendix_VI).

**Model Selection**

We provide the original output of the LASSO model in **Figure 5** and our final model regression in **Figure 6.**

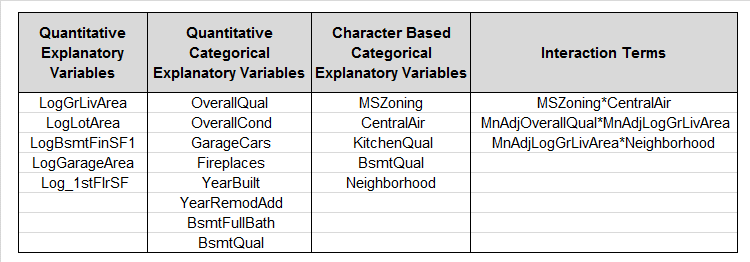
 

**Figure 5. Linear Model Selected Using LASSO Figure 6. Linear Model Selected Including LASSO Output,**

**Adding Neighborhood and Three Interaction Terms.**

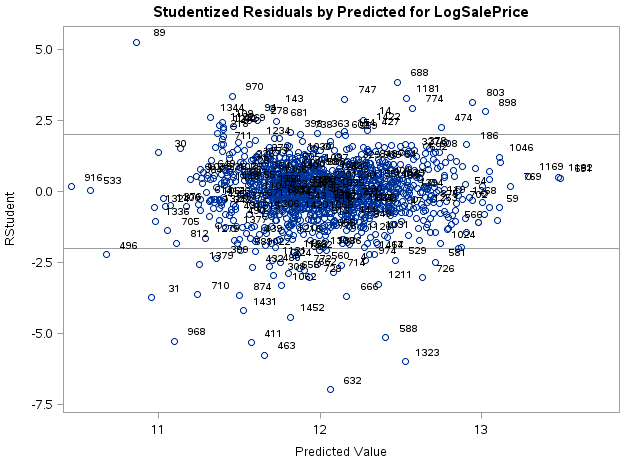
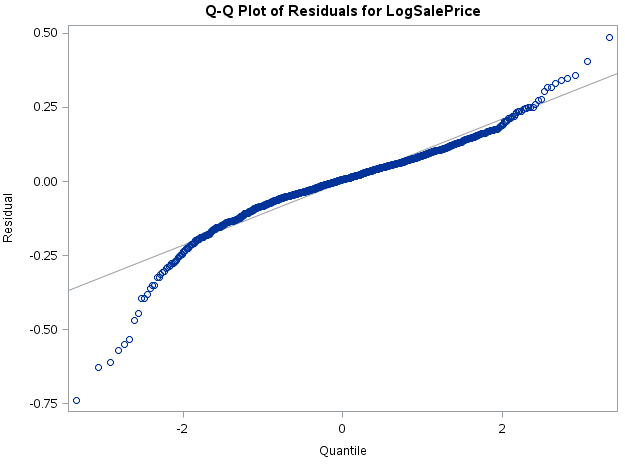
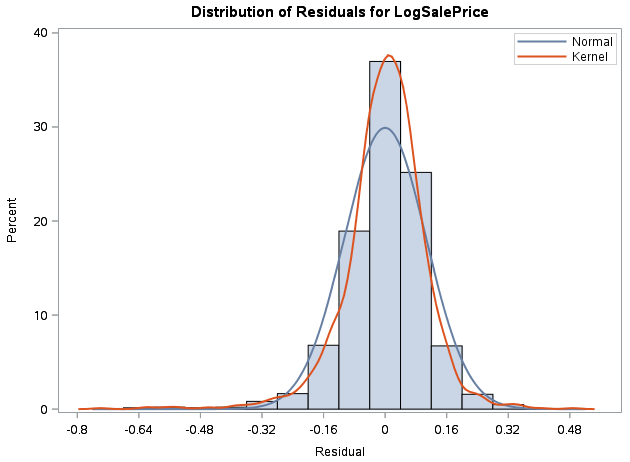
We use LASSO to determine the variables that can be eliminated. The final model has 18 explanatory variables and uses 3 interaction terms to enhance the R2 and fit of the model. All our explanatory variables are chosen by LASSO except one, which appears in two of the alternative variable selection techniques (forward and stepwise). By adding the Neighborhood as the 18th explanatory variable left out by the LASSO criteria, we can explain another 1.4% of variation (R2 = 0.9153 vs R2 = 0.9288). The original linear regression model using LASSO had an R2 of 0.8606 (Figure 5) and then by adding back Neighborhood; by including three interaction terms using LASSO; and by taking out two terms that become insignificant, we increase the final R2 to 0.9288 (Figure 6).

In **Table 2** are explanatory variables that are broken down into four categories after the natural log transformation. *For the equation/model, regression output, and a detailed list of all the regression coefficients under the Estimate output table, see* [Table 1 and Figure 1 of Appendix VII](#Appendix_VII). Note for the explanatory variables name below, a “Log “in a variable name refers to the natural log for our purposes. For interpretation of the coefficients/variables in our regression model, see [Appendix X](#Appendix_X).

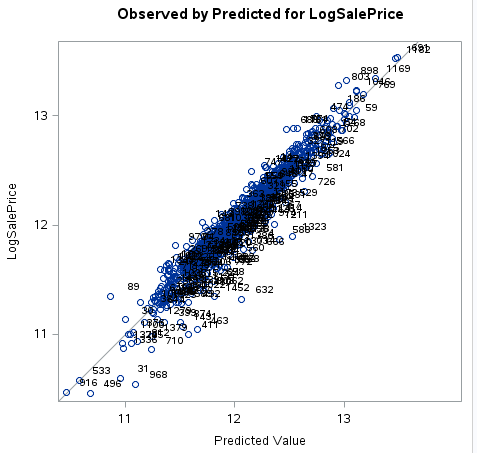
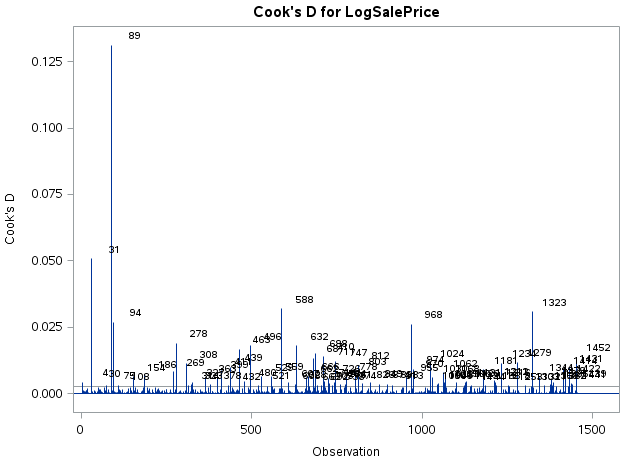
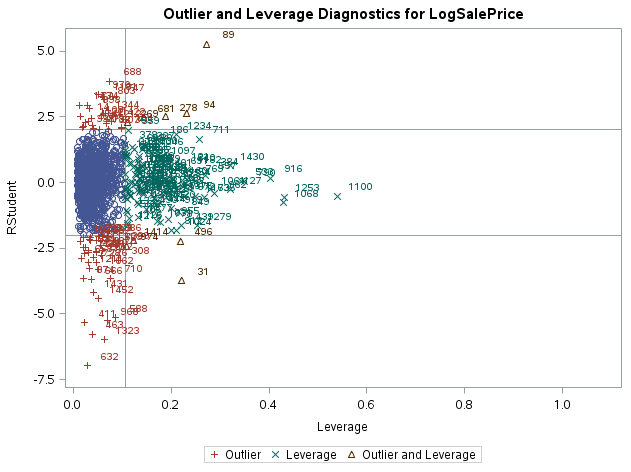


**Table 2. Final Variable Selection Model Using LASSO adding Neighborhood and Interaction Terms**

We check that the assumptions of the regression model are satisfactorily achieved. **Figure 7** shows normality in the studentized residual histogram plot and a mostly normal distribution in Q-Q plot. For some of the lower quantiles in the Q-Q plot, we still see some non-normal behavior but our model has corrected for the right skewness of the upper quantiles in the Q-Q plot. The studentized residual vs. predicted values goes from -5.0 to -5.0 (with some additional points between -5.0 to -7.5) after the transformation of the data, which is much better than the original data set.



**Figure 7. Histogram, Q-Q Plot and Studentized Residual vs. Predicted Values.**



**Figure 8. Leverage vs. Studentized Residuals, Cook’s D Plot and Predicted Values vs. Ln(SalePrice)**

Once again, we investigated if there are any influential observations in the data and remove Id 1299 and 524, see [Appendix III](#Appendix_III) for justification. We have cleaned up the data points nicely but there are still some highly leveraged data points (id 89) but most are between 0 to 0.25 in the Cook’s D plot in **Figure 8**. There are still a few points that are outliers and highly levered but these appeared after we already removed the two data points. Finally, the predicted values vs. the LogSalePrice is linear with lower and a more constant variance.

With the underlying linearity, normality, and constant variance assumptions essentially achieved and with general independence of explanatory variables, this model may be used to predict future sales prices of houses in Ames, Iowa. The final prediction model is statistically significant at the α = 0.05 level with an R2 of 0.9288 (n = 1458 after removing two data points, F = 236.87, p-value < 0.0001). In addition, our Kaggle score is 13.602%, which is the top 13.602%-tile of all submissions (see [Kaggle Website under Submission jhos25](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/submissions?sortBy=date&group=all&page=1)).

**Conclusions**

With extensive analysis and diagnostic testing, a multiple linear regression model has been developed that can be used to predict a sales price of a home in Ames, Iowa that is statistically significant at the α = 0.05 level. The model explains around ~93% (R2 = 0.9288) of the variation in home sales. This is an observational study; therefore, the results have no casual inference. Any correlations made can only be applied to the Ames, Iowa housing market.

References

[[1] Dean De Cock, “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project”, Journal of Statistics Education, Volume 19, Number 3(2011).](http://www.amstat.org/publications/jse/v19n3/decock.pdf)

**APPENDIX I: Explanatory Variable Descriptions**

1. **MSSubClass: Identifies the type of dwelling involved in the sale.**

20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

1. **MSZoning: Identifies the general zoning classification of the sale.**

A Agriculture

C Commercial

FV Floating Village Residential

I Industrial

RH Residential High Density

RL Residential Low Density

RP Residential Low Density Park

RM Residential Medium Density

1. **LotFrontage: Linear feet of street connected to property**
2. **LotArea: Lot size in square feet**
3. **Street: Type of road access to property**

Grvl Gravel

Pave Paved

1. **LotShape: General shape of property**

Reg Regular

IR1 Slightly irregular

IR2 Moderately Irregular

IR3 Irregular

1. **LandContour: Flatness of the property**

Lvl Near Flat/Level

Bnk Banked - Quick and significant rise from street grade to building

HLS Hillside - Significant slope from side to side

Low Depression

1. **Utilities: Type of utilities available**

AllPub All public Utilities (E,G,W,& S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

1. **LotConfig: Lot configuration**

Inside Inside lot

Corner Corner lot

CulDSac Cul-de-sac

FR2 Frontage on 2 sides of property

FR3 Frontage on 3 sides of property

1. **LandSlope: Slope of property**

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

1. **Neighborhood: Physical locations within Ames city limits**

Blmngtn Bloomington Heights

Blueste Bluestem

BrDale Briardale

BrkSide Brookside

ClearCr Clear Creek

CollgCr College Creek

Crawfor Crawford

Edwards Edwards

Gilbert Gilbert

IDOTRR Iowa DOT and Rail Road

MeadowV Meadow Village

Mitchel Mitchell

Names North Ames

NoRidge Northridge

NPkVill Northpark Villa

NridgHt Northridge Heights

NWAmes Northwest Ames

OldTown Old Town

SWISU South & West of Iowa State University

Sawyer Sawyer

SawyerW Sawyer West

Somerst Somerset

StoneBr Stone Brook

Timber Timberland

Veenker Veenker

1. **Condition1: Proximity to various conditions**

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

1. **Condition2: Proximity to various conditions (if more than one is present)**

Artery Adjacent to arterial street

Feedr Adjacent to feeder street

Norm Normal

RRNn Within 200' of North-South Railroad

RRAn Adjacent to North-South Railroad

PosN Near positive off-site feature--park, greenbelt, etc.

PosA Adjacent to postive off-site feature

RRNe Within 200' of East-West Railroad

RRAe Adjacent to East-West Railroad

1. **BldgType: Type of dwelling**

1Fam Single-family Detached

2FmCon Two-family Conversion; originally built as one-family dwelling

Duplx Duplex

TwnhsE Townhouse End Unit

TwnhsI Townhouse Inside Unit

1. **HouseStyle: Style of dwelling**

1Story One story

1.5Fin One and one-half story: 2nd level finished

1.5Unf One and one-half story: 2nd level unfinished

2Story Two story

2.5Fin Two and one-half story: 2nd level finished

2.5Unf Two and one-half story: 2nd level unfinished

SFoyer Split Foyer

SLvl Split Level

1. **OverallQual: Rates the overall material and finish of the house**

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

1. **OverallCond: Rates the overall condition of the house**

10 Very Excellent

9 Excellent

8 Very Good

7 Good

6 Above Average

5 Average

4 Below Average

3 Fair

2 Poor

1 Very Poor

1. **YearBuilt: Original construction date**
2. **YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)**
3. **RoofStyle: Type of roof**

Flat Flat

Gable Gable

Gambrel Gabrel (Barn)

Hip Hip

Mansard Mansard

Shed Shed

1. **RoofMatl: Roof material**

ClyTile Clay or Tile

CompShg Standard (Composite) Shingle

Membran Membrane

Metal Metal

Roll Roll

Tar&Grv Gravel & Tar

WdShake Wood Shakes

WdShngl Wood Shingles

1. **Exterior1st: Exterior covering on house**

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

1. **Exterior2nd: Exterior covering on house (if more than one material)**

AsbShng Asbestos Shingles

AsphShn Asphalt Shingles

BrkComm Brick Common

BrkFace Brick Face

CBlock Cinder Block

CemntBd Cement Board

HdBoard Hard Board

ImStucc Imitation Stucco

MetalSd Metal Siding

Other Other

Plywood Plywood

PreCast PreCast

Stone Stone

Stucco Stucco

VinylSd Vinyl Siding

Wd Sdng Wood Siding

WdShing Wood Shingles

1. **MasVnrType: Masonry veneer type**

BrkCmn Brick Common

BrkFace Brick Face

CBlock Cinder Block

None None

Stone Stone

1. **MasVnrArea: Masonry veneer area in square feet**
2. **ExterQual: Evaluates the quality of the material on the exterior**

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

1. **ExterCond: Evaluates the present condition of the material on the exterior**

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

1. **Foundation: Type of foundation**

BrkTil Brick & Tile

CBlock Cinder Block

PConc Poured Contrete

Slab Slab

Stone Stone

Wood Wood

1. **BsmtQual: Evaluates the height of the basement**

Ex Excellent (100+ inches)

Gd Good (90-99 inches)

TA Typical (80-89 inches)

Fa Fair (70-79 inches)

Po Poor (<70 inches

NA No Basement

1. **BsmtCond: Evaluates the general condition of the basement**

Ex Excellent

Gd Good

TA Typical - slight dampness allowed

Fa Fair - dampness or some cracking or settling

Po Poor - Severe cracking, settling, or wetness

NA No Basement

1. **BsmtExposure: Refers to walkout or garden level walls**

Gd Good Exposure

Av Average Exposure (split levels or foyers typically score average or above)

Mn Mimimum Exposure

No No Exposure

NA No Basement

1. **BsmtFinType1: Rating of basement finished area**

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

1. **BsmtFinSF1: Type 1 finished square feet**
2. **BsmtFinType2: Rating of basement finished area (if multiple types)**

GLQ Good Living Quarters

ALQ Average Living Quarters

BLQ Below Average Living Quarters

Rec Average Rec Room

LwQ Low Quality

Unf Unfinshed

NA No Basement

1. **BsmtFinSF2: Type 2 finished square feet**
2. **BsmtUnfSF: Unfinished square feet of basement area**
3. **TotalBsmtSF: Total square feet of basement area**
4. **Heating: Type of heating**

Floor Floor Furnace

GasA Gas forced warm air furnace

GasW Gas hot water or steam heat

Grav Gravity furnace

OthW Hot water or steam heat other than gas

Wall Wall furnace

1. **HeatingQC: Heating quality and condition**

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

Po Poor

1. **CentralAir: Central air conditioning**

N No

Y Yes

1. **Electrical: Electrical system**

SBrkr Standard Circuit Breakers & Romex

FuseA Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix Mixed

1. **1stFlrSF: First Floor square feet**

1. **2ndFlrSF: Second floor square feet**
2. **LowQualFinSF: Low quality finished square feet (all floors)**
3. **GrLivArea: Above grade (ground) living area square feet**
4. **BsmtFullBath: Basement full bathrooms**
5. **BsmtHalfBath: Basement half bathrooms**
6. **FullBath: Full bathrooms above grade**
7. **HalfBath: Half baths above grade**
8. **Bedroom: Bedrooms above grade (does NOT include basement bedrooms)**
9. **Kitchen: Kitchens above grade**
10. **KitchenQual: Kitchen quality**

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

1. **TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)**
2. **Functional: Home functionality (Assume typical unless deductions are warranted)**

Typ Typical Functionality

Min1 Minor Deductions 1

Min2 Minor Deductions 2

Mod Moderate Deductions

Maj1 Major Deductions 1

Maj2 Major Deductions 2

Sev Severely Damaged

Sal Salvage only

1. **Fireplaces: Number of fireplaces**
2. **FireplaceQu: Fireplace quality**

Ex Excellent - Exceptional Masonry Fireplace

Gd Good - Masonry Fireplace in main level

TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in

1. **basement**

Fa Fair - Prefabricated Fireplace in basement

Po Poor - Ben Franklin Stove

NA No Fireplace

1. **GarageType: Garage location**

2Types More than one type of garage

Attchd Attached to home

Basment Basement Garage

BuiltIn Built-In (Garage part of house - typically has room above garage)

CarPort Car Port

Detchd Detached from home

NA No Garage

1. **GarageYrBlt: Year garage was built**

1. **GarageFinish: Interior finish of the garage**

Fin Finished

RFn Rough Finished

Unf Unfinished

NA No Garage

1. **GarageCars: Size of garage in car capacity**
2. **GarageArea: Size of garage in square feet**
3. **GarageQual: Garage quality**

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

1. **GarageCond: Garage condition**

Ex Excellent

Gd Good

TA Typical/Average

Fa Fair

Po Poor

NA No Garage

1. **PavedDrive: Paved driveway**

Y Paved

P Partial Pavement

N Dirt/Gravel

1. **WoodDeckSF: Wood deck area in square feet**
2. **OpenPorchSF: Open porch area in square feet**
3. **EnclosedPorch: Enclosed porch area in square feet**
4. **3SsnPorch: Three season porch area in square feet**
5. **ScreenPorch: Screen porch area in square feet**
6. **PoolArea: Pool area in square feet**
7. **PoolQC: Pool quality**

Ex Excellent

Gd Good

TA Average/Typical

Fa Fair

NA No Pool

1. **Fence: Fence quality**

GdPrv Good Privacy

MnPrv Minimum Privacy

GdWo Good Wood

MnWw Minimum Wood/Wire

NA No Fence

1. **MiscFeature: Miscellaneous feature not covered in other categories**

Elev Elevator

Gar2 2nd Garage (if not described in garage section)

Othr Other

Shed Shed (over 100 SF)

TenC Tennis Court

NA None

1. **MiscVal: $Value of miscellaneous feature**
2. **MoSold: Month Sold (MM)**
3. **YrSold: Year Sold (YYYY)**
4. **SaleType: Type of sale**

WD Warranty Deed - Conventional

CWD Warranty Deed - Cash

VWD Warranty Deed - VA Loan

New Home just constructed and sold

COD Court Officer Deed/Estate

Con Contract 15% Down payment regular terms

ConLw Contract Low Down payment and low interest

ConLI Contract Low Interest

ConLD Contract Low Down

Oth Other

1. **SaleCondition: Condition of sale**

Normal Normal Sale

Abnorml Abnormal Sale - trade, foreclosure, short sale

AdjLand Adjoining Land Purchase

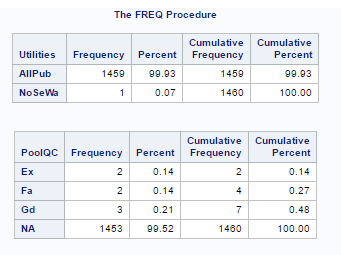
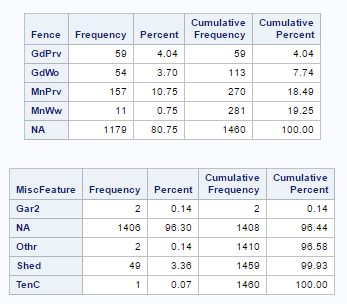
Alloca Allocation - two linked properties with separate deeds, typically condo with a

**APPENDIX II: Data Cleaning**

For both the training and test data set, we performed the following data cleaning.

* First, we eliminated the following variables due to many entries of NA and/or the fact that their p-values in a regression test were to high: Alley, Utilities, PoolQC, Fence, and MiscFeature. Se Figure 1 shows that these variables have mostly NAs or only one different subcategory (Utilities) using the SAS PROC FREQ command:

**Figure 1: High Frequency of NA Category for PoolQC, Fence, and MiscFeature; Only one Difference for Utilities**

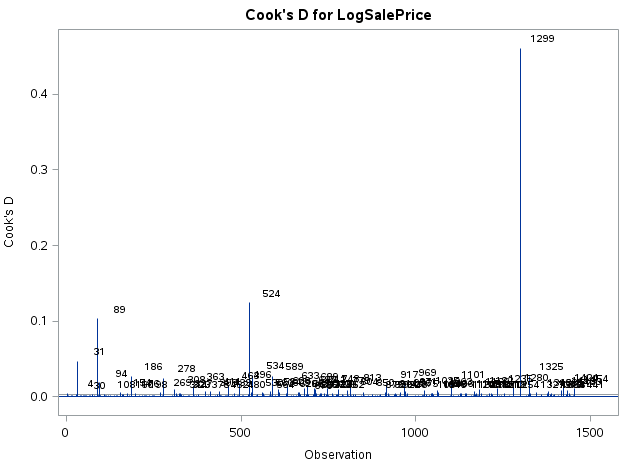
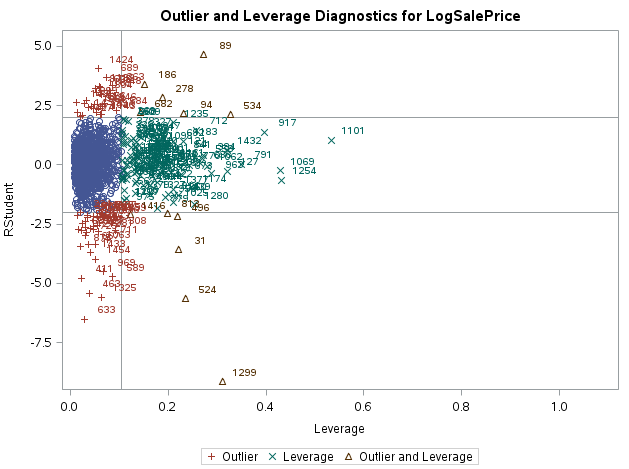
* Second, some quantitative variables had a few NAs, which we replace with the mean of the variable series using the training data series to minimize the effect on the regression and predicted outcomes (e.g. LotFrontage).
  + LotFrontage is read in as characters. Therfore, we rename it to LotFrontageChar and we re-input LotFrontage as input(LogFrontageChar,best.) to convert it to a numeric. Then we dropped LotFrontageChar.
* Third, a few categorical variables had NAs. We set some NAs to zeros (e.g. BsmtHalfBath) if a numerical category and for those with test categories, we replaced NA with the category None (e.g. MasVrnType).
* Finally, in a few one-off cases, we replaced the one NA with what was the most likely category would be based on the neighborhood and/or overall quality classification. We did not use ***imputation*** here because we believe we can come to relevant conclusions by looking at the data broken into subcategories. This includes a very small number of NAs in Exterior1st, Exterior2nd, Electrical, Functional and MSZoning. Rather than creating a separate category, we use inference to fill them in base on their categorization. Below is an example:
  + For example, Exterior1st has a couple of NAs where the houses are all located in the Edwards Neighborhood, which exclusively has almost all wooden shingles so we selected “Wd Shng” to replace the NA. This is also true for Exterior2nd for the same houses located in the Edwards neighborhood.
  + A couple of Electrical NAs are set to “SBkr”, of MSZoning NAs are set to “RL” and of Functional NAs are set to “Typ”.
  + Details of the adjustment can be found in our SAS code.

***For more specific details for each change, see the commented SAS code in*** [*Appendix XI*](#Appendix_XISAS) ***under “Create Filtered Regression Set: TrainReg”; “Create Filtered Regression/Prediction Train2 Set”; and also search for “APPENDIX on Data Removal”.***

**APPENDIX III: Data Removal**

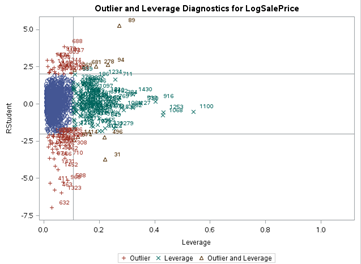
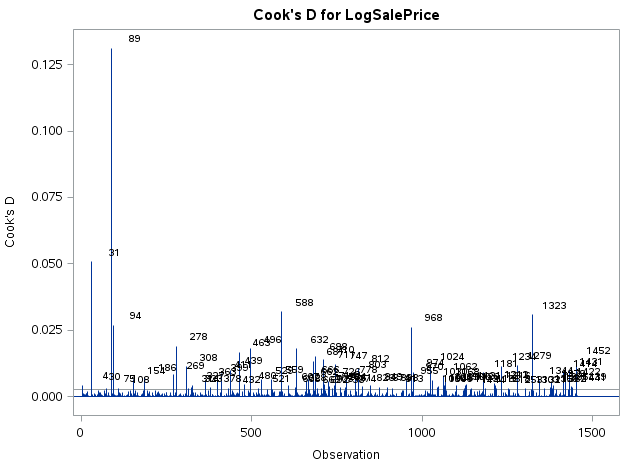
In addition, we removed two data points (Id 1299 and 524) from the training data set due to being highly leverage outliers, shown in **Figure 1** including the two data points versus **Figure 2** removing the two data points. In addition, we help residuals to have more constant variance in **Figure 3**.

**Figure1. With All Data Point**



Remove data points that are outliers and highly levered.

**Figure 2: Without Two Data Points (Ids 1299 and 524)**

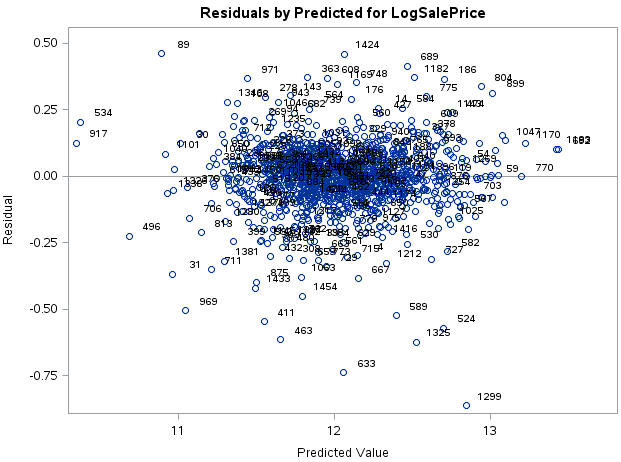
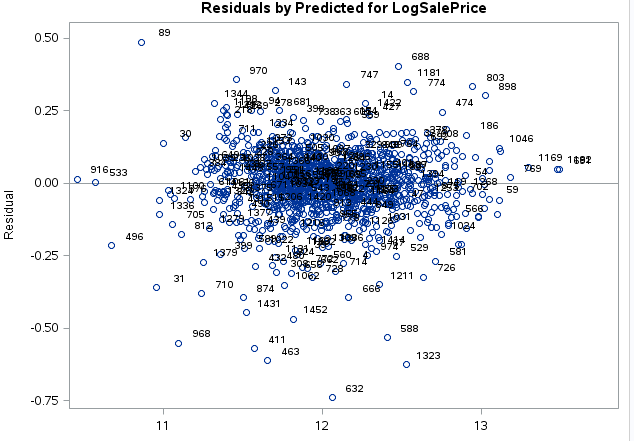
 

**Figure 3: The residuals are in a more constant range by removing these two points.**

**Including both points results in Excluding both points results in**

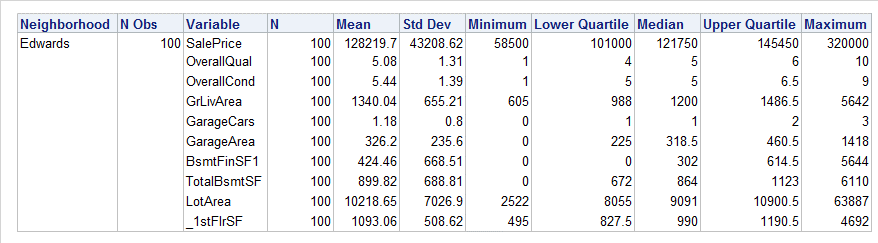
**more variance of residuals more constant variance**

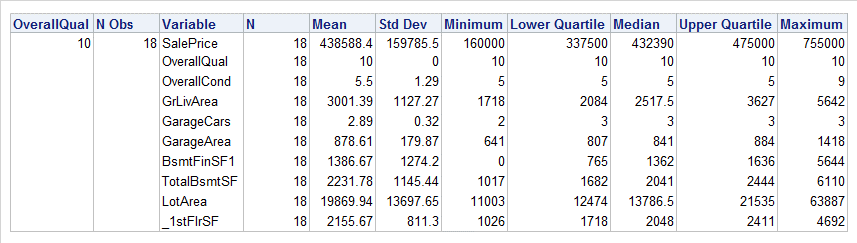
**(below -0.75)**

* Below in **Table 1** are summary statistics for the Edwards Neighborhood for which we will use to remove points, Ids 1299 and 524.

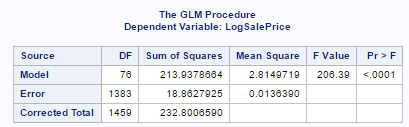
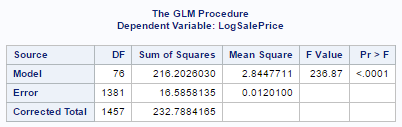
**Table 1. Neighborhood statistics on quantitative variables (100 data points) and Neighborhood statistics with overall house quality of 10 (18 data points).**





* We note, before we begin, that OverallQuality is significant in our model of the SalePrice of a home in Ames, Iowa (0.055859197 \* OverallQual). An Overall Quality of 10 can multiply the SalePrice estimate by or increase by 74.82%. The average is 32.8% with an average quality of 5.08 for the Edwards Neighborhood which we will address later. Therefore, much of the high leverage and outlier status of the residual variance could be explained by this variable in combinations with the Edwards Neighborhood for Ids 1299 and 524.
* **Id 1299** is a 1 family house with 2 stories that sold in the Edwards neighborhood of Ames, Iowa at an above average value of $160K while the mean for Edwards is $128K. It was built and sold in 2008 during the Great Recession as a new house with an overall quality of 10 (extremely high since the average OverallQual is 5.08 for the Edwards neighborhood). It does have an above average amount of square footage for this neighborhood: the living area was 5,642 sqft. We would expect a high price for this house but it is not high enough since overall quality is a significant factor in our model. The overall quality is a rating of the overall material and finish of the house and the house was remodeled externally in 2006. The average price of houses with an overall quality of 10 is $438.49k in our training data set and houses sold with a quality of 10 in 2008 and 2009, sold at an average price of $386.3k even during the middle of the Great Recession. Hence, this data point becomes a residual with a high variance and potentially was mislabeled in the overall quality category.
* **Id 524** is a 1 family house with 2 stories that also sold in the Edwards neighborhood of Ames, Iowa at a very high value of $184.75k as a new home (which was the maximum for the Edwards neighborhood with an overall high quality of 10 since the average OverallQual is 5.08). In addition, the house had a living area of 4,676 sqft. below the mean of 5,159 sqft. and overall quality of 10. It was sold in 2007 at a high price of $184.75k for this neighborhood at the peak of the housing market prior to the Great Recession. However again, that is below the average price for an overall quality of 10 in 2006 and 2007 of $411.38k.
* *By removing these two data points and using PROC GLM, the R2 goes from 0.915 to 0.93 and the mean square error drops from 0.01364 to 0.01201.*

With All Data Point Without Two Data Points

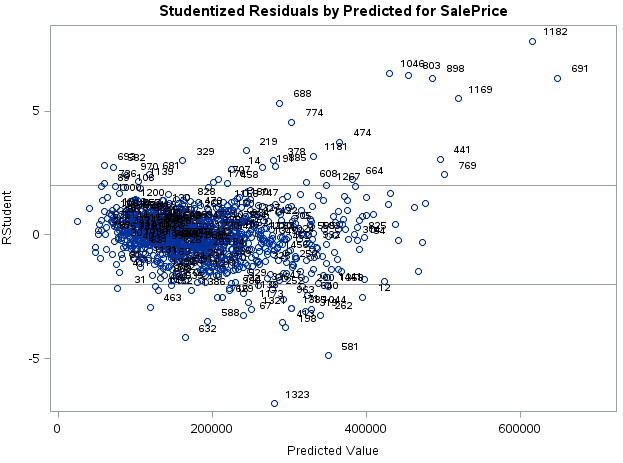
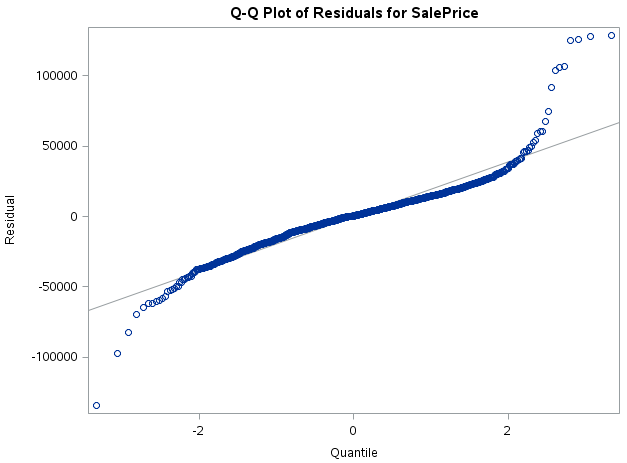
 

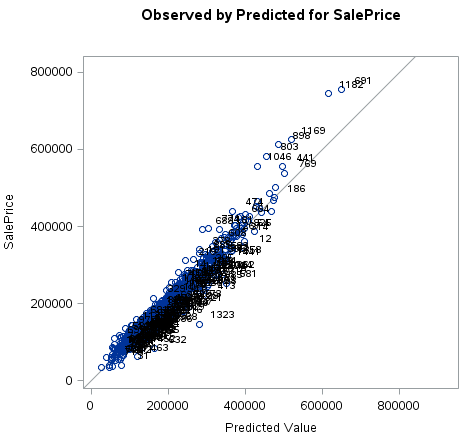
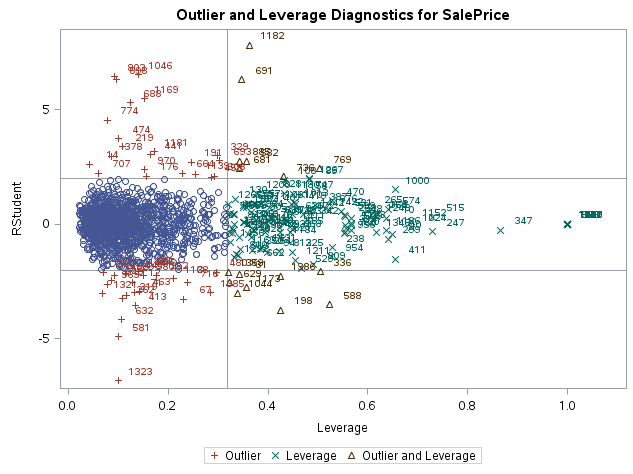
***For SAS code in this appendix, see*** [*Appendix XI*](#Appendix_XISAS) ***in this report and search for “APPENDIX on Data Removal”.***

**[APPENDIX IV: Transformation to Natural Log of Response and Quantitative Explanatory Variables](#_top)**

In **Figure 1**, we used the original loaded training data with sales price as the response variable versus all the original explanatory variables. **Figure 1** shows that the studentized residuals vs. predicted value has a funnel type shaped pattern that then just increases as predicted values increase, which is non-normal. In addition, **Figure 1** shows a non-normal Q-Q plot. We also see large variability in the predicted value versus the house sales price. Finally, even after removing two large outlier and highly leveraged data points, we still have more data points that are highly leverage and/or are outliers.

**Figure 1: Residual, Q-Q, predicted values and Outlier and Leverage plots using original non-transformed data.**

In **Table 1** below and after we natural log transform the response variable of SalePrice and most the quantitative variables that are in square footage, we can get closer to a normally distributed data set for our model. ***In addition, we can see that since some of the explanatory variables may have a value of zero we just add 1 to them to avoid an error when we take the natural log.***

**Table 1: Except of SAS code converting several explanatory and the response variable (SalePrice) to natural log values.**

LotFrontage = input(LotFrontageChar,best.);

LogSalePrice = log(SalePrice);

LogGrLivArea = log(GrLivArea);

LogTotalBsmtSF = log(TotalBsmtSF+1);

LogLotArea = log(LotArea+1);

LogBsmtFinSF1 = log(BsmtFinSF1+1);

Log\_1stFlrSF = log(\_1stFlrSF+1);

Log\_2ndFlrSF = log(\_2ndFlrSF+1);

LogGarageArea = log(GarageArea + 1);

LogBsmtFinSF2 = log(BsmtFinSF2 + 1);

LogBsmtUnfSF = log(BsmtUnfSF + 1);

LogLotFrontage = log(LotFrontage+1);

LogWoodDeckSF = log(WoodDeckSF+1);

LogOpenPorchSF = log(OpenPorchSF+1);

LogMasVnrArea = log(MasVnrArea+1);

LogEnclosedPorch = log(EnclosedPorch+1);

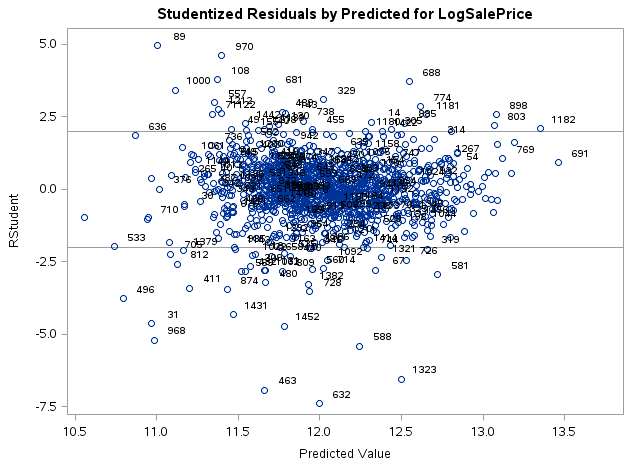
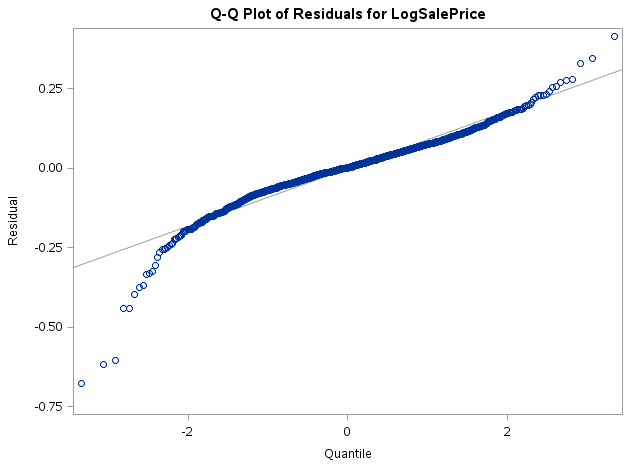
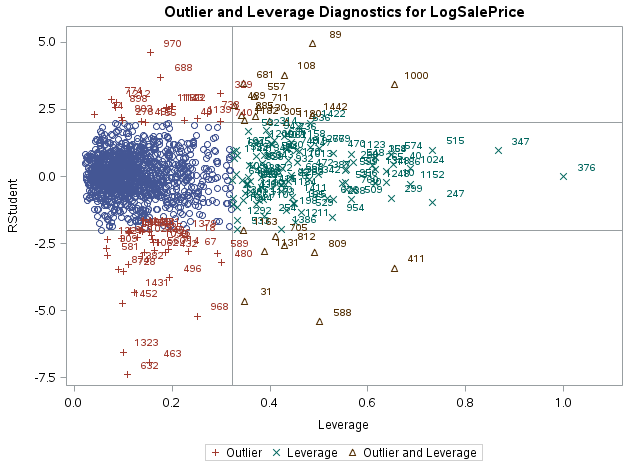
Log\_3SsnPorch = log(\_3SsnPorch+1);

LogScreenPorch = log(ScreenPorch+1);

LogPoolArea = log(PoolArea+1);

After taking the natural log transform of all the variables listed in **Table 1, Figure 2** shows the linear regression output of the residuals seems to have more of a constant variance from 5.0 to -7.5. **Figure 2** shows a more normal or better Q-Q plot for the upper quantiles; and more linear and less dispersed plot of predicted values versus natural log of SalePrice.

**Figure 2: Residual, Q-Q, predicted values and Outlier and Leverage plots using natural log transformed data.**

***For SAS code in this appendix, see*** [*Appendix XI*](#Appendix_XISAS) ***in this report and search for “APPENDIX on Data Transformation to Natural Log”.***

**[APPENDIX V: Definition and Formulation of Models Used](#_top)**

1. **Forward Selection:**

We start with no variables selected and add in the “best” variable at each step.

1. **Stepwise Selection:**

We start with no variables and at each step and then put in the “best” variable. However, we then re-analyze all the variable again with this new variable and take out any previously selected variable if it is no longer helpful.

1. **Backward Selection:**

We start with all the variables and take out the “worst” variable at each step.

1. **Cross-Validation (CV)**

The calculation of cross-validation uses the prediction errors (differences between the actual response values and the predictions) and summarize the predictive ability of the model by the mean squared prediction error (MSPE). The MSPE for the validation assess the predictive ability of the model. Since there are 76 explanatory variables we do not perform the regression using interaction terms. CV calculation is shown below:

= \*

= )2 *where k=n or leave-one out..*

* = where
* *is the leverage*

When there are many predictors, OLS can overfit observed data, leading to large prediction errors. Large prediction errors mean that a model that works for one data set may not work for a second data set, even with same population and predictors. Regression methods with fewer extracted factors can provide better predictability of future observations. Where factors in model more bias in model

1. **Akaike Information Criterion (AIC)**

The Cp statistic takes a measure of the lack of fit of a model and adds a penalty for the number of terms in the model. As more terms are included in the model, lack of fit decreases, but the penalties increase. In general, AIC seems more appropriate if there are not too many redundant and unnecessary X’s in the starting set. We look for small values of AIC to improve the model. You can have many regression coefficients (p) with AIC. AIC has a penalty term for including more parameters in model by adding 2p. The criteria to make AIC smaller is based on the sum of squared residuals (SSRes or also called sum of the squared errors SSE) plus a penalty term for number of terms regression coefficients (p). The formula is given below:

= ln() + 2 \* (p+1)

1. **Swartz's Bayesian Information Criterion (SBC)**

SBC more heavily penalizes models with large p and operates better when predictors are highly correlated or we are running selection on interaction terms.

= + p \* ln(n)

1. **Least Absolute Shrinkage and Selection Operator (LASSO)**

Least absolute shrinkage and selection operator (LASSO) adds and deletes parameters based on a version of ordinary least squares where the sum of absolute regression coefficients is constrained. It minimizes sum of squared residuals according to a constraint which is usually the sum of absolute values of parameter estimates < given constant. The variables added according to correlation with residuals. LASSO is a variable selection technique. LASSO is a shrinkage estimate of β, where the OLS regression coefficients are shrunk toward the origin. The value of c controls the amount of shrinkage. For a given value of *c*, only a subset of the coefficient estimates will have nonzero values. The smaller the value of *c*, the smaller the subset of estimates with nonzero values.

LASSO Performs well when a small-to-medium number of moderate-sized coefficients. However, estimates tend to have large biases. Subset selection performs well only when there are a few nonzero coefficients.

The formulation is below:

* Constrained OLS minimization problem in which:



* is minimized for β subject to:



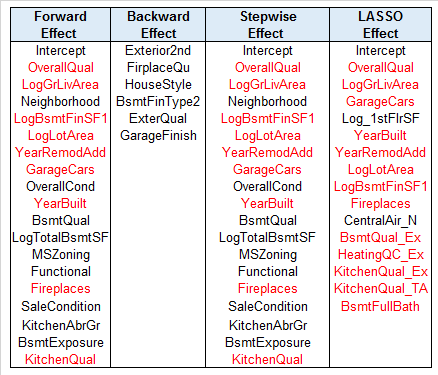
* The problem is solved using complicated quadratic programming methods subject to linear inequality constraints.

**[APPENDIX VI: Model Variable Selection Process](#_top)**

**Below is a summary of our four modeling steps with example outputs A to E in the pages to follow**. We perform the following steps to get to our final model.

1. In **Outputs A to D in the following pages,** we run and provide output for a forward, backward, stepwise selection model with AIC and stop of cross validation. In Item 4, we use a LASSO selection model with SBC and a stop of cross validation.
2. Next, we select a common set of explanatory variables by looking at all the outputs that have significant p-values and overall variable selection results. In comparing all the variable selectin processes, ***we found LASSO to potentially be the best model because we are going to look at interaction terms. Using Swartz's Bayesian Information Criterion (SBC) has a higher penalty for adding a term or regression coefficient.***  This is important because interaction terms can be correlated with existing terms in a regression inflating variance. Below in **Table 1** are all the significant factors of all four models and those highlighted in red show factors that overlap from the different models. The only adjustment made to the output of LASSO model is that due to the high penalty bar it maybe removing the explanatory categorical variable Neighborhood, which is very significant in the Forward and Stepwise process. While many Neighborhood may not be statistical significant in-terms of p-values, there are several Neighborhoods that due add significant value to the regression. **Furthermore, by adding the Neighborhood variable left out by the LASSO criteria, we are only able to explain another 1.4% of variation (R2 = 0.9153 vs R2 = 0.9288).**

**Table 1. Output of Four Variable Selection Models**



1. We now use LASSO to look at all the cross terms. We remove cross-terms if the p-values are not significant. We then remove terms if the variance inflation factors are high for the interaction terms after adjustment. VIFs for PROC GLM and PROC GLMSELECT can be calculated by using the Type I error: VIF = . However, if the VIF is high and one or both terms are quantitative, we adjust the quantitative portion of the interaction term by subtraction off the mean of the series (in this case the natural log of the series, see **Table 2**. **Output E in the following pages** **shows an example of our regressions looking for interaction terms selected by LASSO and checking p-values and the overall regression**. ***Since we run out of memory if we try to check all interaction terms with all variables at the same time,*** ***we execute LASSO looking at a few variables at a time versus all others and then re-run with new variables***. Below is the final set of terms selected by LASSO with interaction terms that are significant:

LogSalePrice = LogGrLivArea LogLotArea LogBsmtFinSF1 LogGarageArea   
Log\_1stFlrSF Neighborhood OverallQual OverallCond GarageCars Fireplaces YearBuilt YearRemodAdd MSZoning CentralAir KitchenQual BsmtFullBath BsmtQual   
MSZoning\*CentralAir MnAdjOverallQual\*MnAdjLogGrLivArea MnAdjLogGrLivArea\*Neighborhood LogTotalBsmtSF MnAdjOverallQual\*Foundation Fireplaces\*CentralAir

**Table 2. Mean Adjustment of Quantitative Variables Used in Regression**

MnAdjOverallQual = OverallQual-6.094;

MnAdjOverallCond = OverallCond - 5.5761;

MnAdjLogGrLivArea = LogGrLivArea - 7.266;

MnAdjGarageCars = GarageCars - 1.7661;

MnAdjLogTotalBsmtSF = LogTotalBsmtSF - 6.7483;

MnAdjLogGarageArea = LogGarageArea - 5.8065;

MnAdjLogBsmtFinSF1 = LogBsmtFinSF1 - 4.2243;

MnAdjLogLotArea = LogLotArea - 9.1086;

MnAdjLogBsmtFinSF1 = LogBsmtFinSF1 - 4.2243;

MnAdjLog\_1stFlrSF = Log\_1stFlrSF - 7.0067;

1. Finally, with the above factors selected and since we had to choose the interaction terms with several partial LASSO interaction regressions, we run a **final** **PROC GLM** to check to make sure that the regression is significant; all factors are significant; and that the VIFs again are not above 10 see **Table 3**. *We likely want to remove the following terms highlighted in yellow (LogTotalBsmtSF, MnAdjOver\*Foundation, Fireplace\*CentralAir) from the model due to high p-values > 0.05 and/or high VIFs after the mean adjustment.* Below is the final model that meets all the criteria after removing the rows highlighted in **Table 3**.

LogSalePrice = LogGrLivArea LogLotArea LogBsmtFinSF1 LogGarageArea   
Log\_1stFlrSF Neighborhood OverallQual OverallCond GarageCars Fireplaces YearBuilt YearRemodAdd MSZoning CentralAir KitchenQual BsmtFullBath BsmtQual   
MSZoning\*CentralAir MnAdjOverallQual\*MnAdjLogGrLivArea MnAdjLogGrLivArea\*Neighborhood

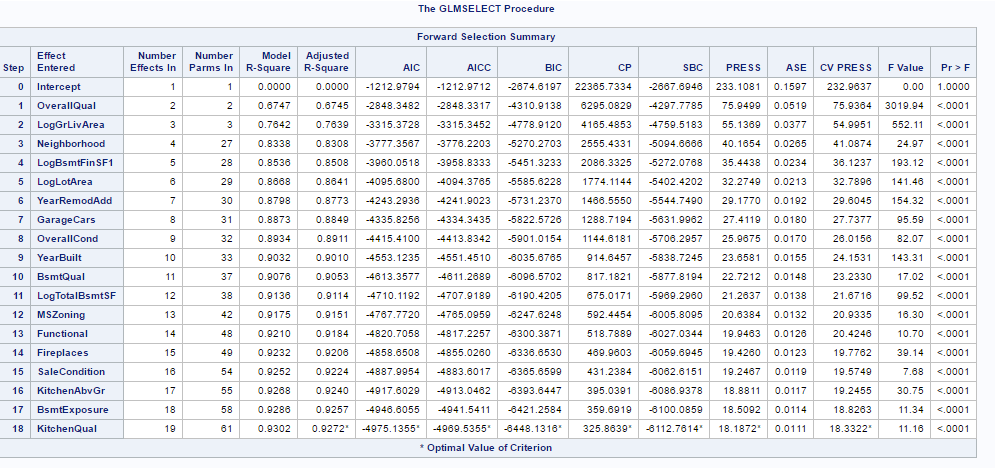
**Table 3. Output of PROC GLM of Final Selected Factors**

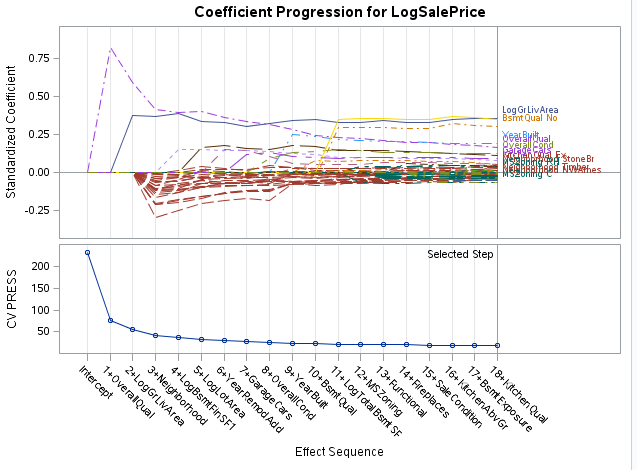
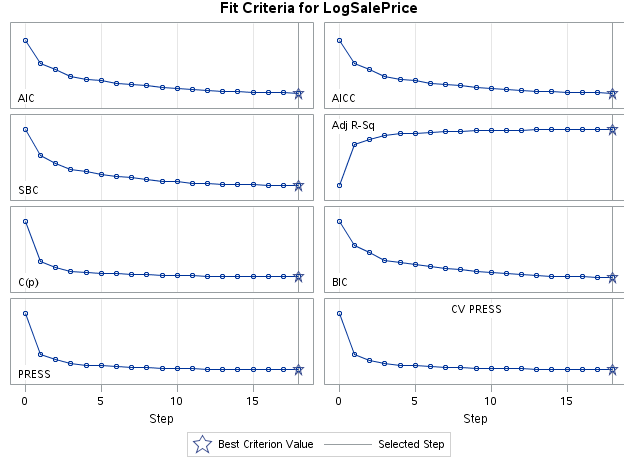


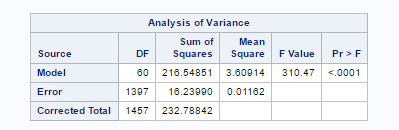
**OUTPUTS A to E**

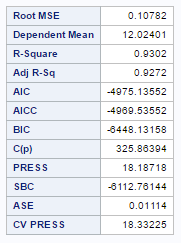
1. **Output A of Forward Selection Process in Figure 1 below:** We perform a forward selection using AIC with a cross-validation stop.

**Figure 1. Output of Forward Selection Process**



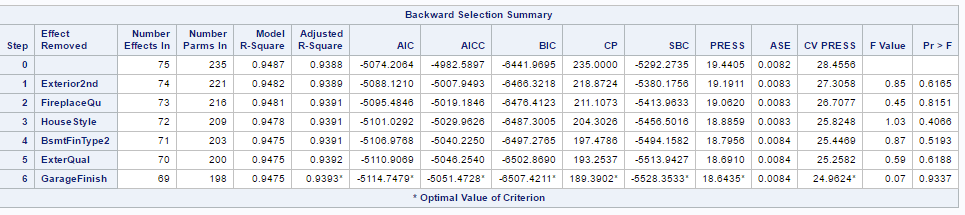


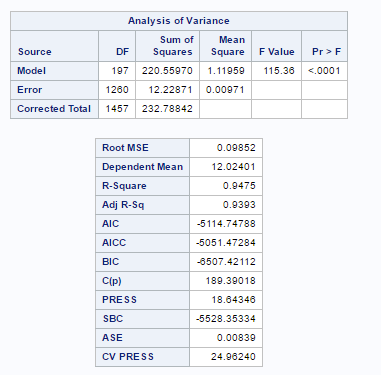


| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | -0.313522 | 0.621006 | -0.50 | 0.6137 |
| **LogGrLivArea** | 1 | 0.429445 | 0.014339 | 29.95 | <.0001 |
| **Neighborhood Blmngtn** | 1 | -0.023435 | 0.044116 | -0.53 | 0.5954 |
| **Neighborhood Blueste** | 1 | -0.014130 | 0.085893 | -0.16 | 0.8694 |
| **Neighborhood BrDale** | 1 | -0.021797 | 0.048141 | -0.45 | 0.6508 |
| **Neighborhood BrkSide** | 1 | 0.034340 | 0.039115 | 0.88 | 0.3801 |
| **Neighborhood ClearCr** | 1 | -0.016642 | 0.039023 | -0.43 | 0.6698 |
| **Neighborhood CollgCr** | 1 | -0.023492 | 0.034733 | -0.68 | 0.4989 |
| **Neighborhood Crawfor** | 1 | 0.095649 | 0.037478 | 2.55 | 0.0108 |
| **Neighborhood Edwards** | 1 | -0.072923 | 0.035940 | -2.03 | 0.0426 |
| **Neighborhood Gilbert** | 1 | -0.040870 | 0.035960 | -1.14 | 0.2559 |
| **Neighborhood IDOTRR** | 1 | -0.021575 | 0.044068 | -0.49 | 0.6245 |
| **Neighborhood MeadowV** | 1 | -0.057335 | 0.046702 | -1.23 | 0.2198 |
| **Neighborhood Mitchel** | 1 | -0.067385 | 0.036924 | -1.82 | 0.0682 |
| **Neighborhood NAmes** | 1 | -0.034215 | 0.034564 | -0.99 | 0.3224 |
| **Neighborhood NPkVill** | 1 | -0.033398 | 0.050463 | -0.66 | 0.5082 |
| **Neighborhood NWAmes** | 1 | -0.072289 | 0.035593 | -2.03 | 0.0424 |
| **Neighborhood NoRidge** | 1 | 0.066272 | 0.037698 | 1.76 | 0.0790 |
| **Neighborhood NridgHt** | 1 | 0.023534 | 0.036595 | 0.64 | 0.5203 |
| **Neighborhood OldTown** | 1 | -0.032066 | 0.039844 | -0.80 | 0.4211 |
| **Neighborhood SWISU** | 1 | -0.005428 | 0.042261 | -0.13 | 0.8978 |
| **Neighborhood Sawyer** | 1 | -0.047474 | 0.036150 | -1.31 | 0.1893 |
| **Neighborhood SawyerW** | 1 | -0.045520 | 0.036314 | -1.25 | 0.2102 |
| **Neighborhood Somerst** | 1 | -0.013161 | 0.041773 | -0.32 | 0.7528 |
| **Neighborhood StoneBr** | 1 | 0.100867 | 0.040344 | 2.50 | 0.0125 |
| **Neighborhood Timber** | 1 | -0.045838 | 0.037433 | -1.22 | 0.2210 |
| **Neighborhood Veenker** | 0 | 0 | . | . | . |
| **MSZoning C** | 1 | -0.313630 | 0.039663 | -7.91 | <.0001 |
| **MSZoning FV** | 1 | 0.094758 | 0.031588 | 3.00 | 0.0027 |
| **MSZoning RH** | 1 | 0.044960 | 0.032785 | 1.37 | 0.1705 |
| **MSZoning RL** | 1 | 0.039234 | 0.015330 | 2.56 | 0.0106 |
| **MSZoning RM** | 0 | 0 | . | . | . |
| **LogLotArea** | 1 | 0.089663 | 0.009083 | 9.87 | <.0001 |
| **OverallQual** | 1 | 0.047468 | 0.004167 | 11.39 | <.0001 |
| **OverallCond** | 1 | 0.044289 | 0.003415 | 12.97 | <.0001 |
| **YearBuilt** | 1 | 0.002502 | 0.000254 | 9.85 | <.0001 |
| **YearRemodAdd** | 1 | 0.001024 | 0.000225 | 4.55 | <.0001 |
| **BsmtQual Ex** | 1 | 0.080827 | 0.017184 | 4.70 | <.0001 |
| **BsmtQual Fa** | 1 | 0.018125 | 0.020256 | 0.89 | 0.3711 |
| **BsmtQual Gd** | 1 | 0.012877 | 0.010252 | 1.26 | 0.2093 |
| **BsmtQual No** | 1 | 0.762193 | 0.080598 | 9.46 | <.0001 |
| **BsmtQual TA** | 0 | 0 | . | . | . |
| **BsmtExposure Av** | 1 | 0.017469 | 0.008995 | 1.94 | 0.0523 |
| **BsmtExposure Gd** | 1 | 0.070202 | 0.011616 | 6.04 | <.0001 |
| **BsmtExposure Mn** | 1 | 0.011229 | 0.011027 | 1.02 | 0.3087 |
| **BsmtExposure No** | 0 | 0 | . | . | . |
| **LogBsmtFinSF1** | 1 | 0.011027 | 0.001143 | 9.65 | <.0001 |
| **LogTotalBsmtSF** | 1 | 0.122092 | 0.011475 | 10.64 | <.0001 |
| **KitchenAbvGr** | 1 | -0.088173 | 0.015400 | -5.73 | <.0001 |
| **KitchenQual Ex** | 1 | 0.090587 | 0.016381 | 5.53 | <.0001 |
| **KitchenQual Fa** | 1 | -0.024620 | 0.019083 | -1.29 | 0.1972 |
| **KitchenQual Gd** | 1 | 0.014087 | 0.008821 | 1.60 | 0.1105 |
| **KitchenQual TA** | 0 | 0 | . | . | . |
| **Functional Maj1** | 1 | -0.097155 | 0.030515 | -3.18 | 0.0015 |
| **Functional Maj2** | 1 | -0.244696 | 0.050097 | -4.88 | <.0001 |
| **Functional Min1** | 1 | -0.067305 | 0.020296 | -3.32 | 0.0009 |
| **Functional Min2** | 1 | -0.044537 | 0.019816 | -2.25 | 0.0248 |
| **Functional Mod** | 1 | -0.106805 | 0.028957 | -3.69 | 0.0002 |
| **Functional Sev** | 1 | -0.447723 | 0.109425 | -4.09 | <.0001 |
| **Functional Typ** | 0 | 0 | . | . | . |
| **Fireplaces** | 1 | 0.027338 | 0.005652 | 4.84 | <.0001 |
| **GarageCars** | 1 | 0.050188 | 0.005601 | 8.96 | <.0001 |
| **SaleCondition Abnorml** | 1 | -0.100906 | 0.016555 | -6.10 | <.0001 |
| **SaleCondition AdjLand** | 1 | -0.043451 | 0.057135 | -0.76 | 0.4471 |
| **SaleCondition Alloca** | 1 | -0.043504 | 0.035280 | -1.23 | 0.2177 |
| **SaleCondition Family** | 1 | -0.103590 | 0.027363 | -3.79 | 0.0002 |
| **SaleCondition Normal** | 1 | -0.040172 | 0.012240 | -3.28 | 0.0011 |
| **SaleCondition Partial** | 0 | 0 | . | . | . |

1. **Output B of Backward Selection Process in Figure 2 below:** Backward does not provide many variables therefore we will not use the output from this model and not many factors using AIC have significant p-values. We perform a backward selection using AIC with a cross-validation stop.

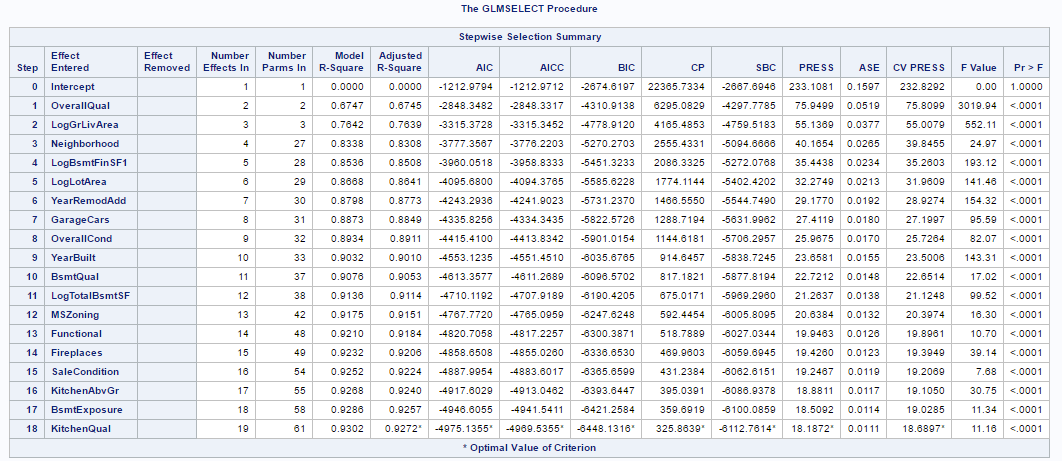
**Figure 2. Output of Backward Selection Process**

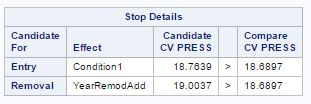


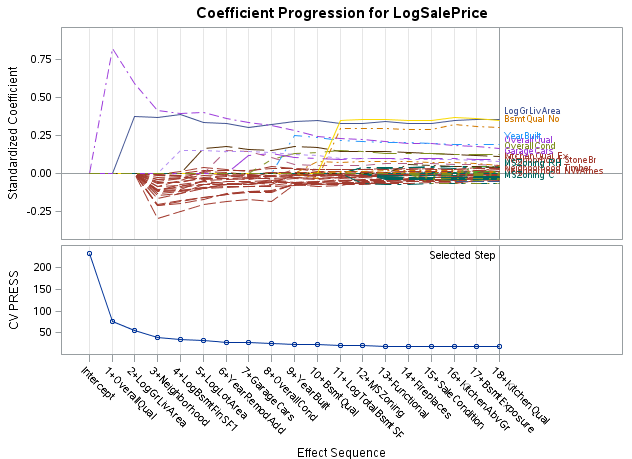
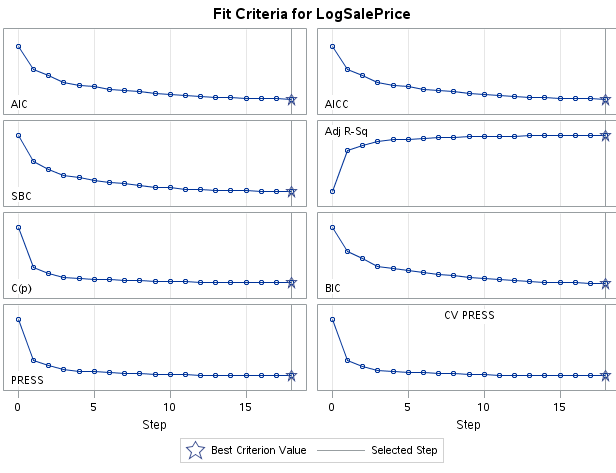


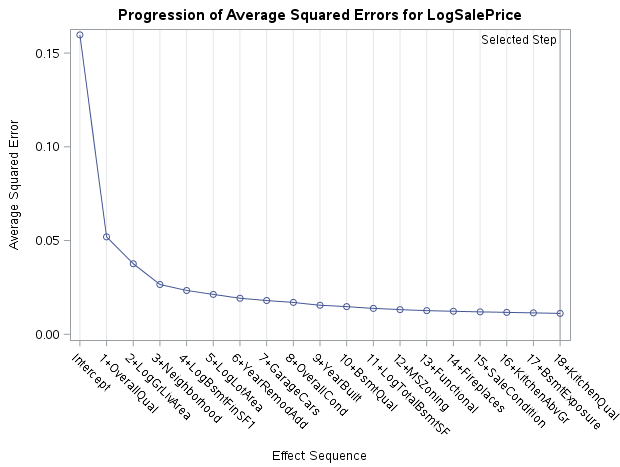
1. **Output of Stepwise Selection Process in Figure 3 below:** We perform a stepwise selection using AIC with a cross-validation stop.

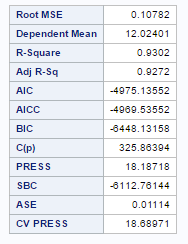
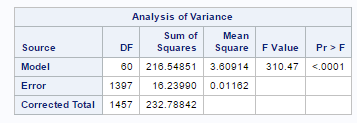
**Figure 3. Output of Stepwise Selection Process**





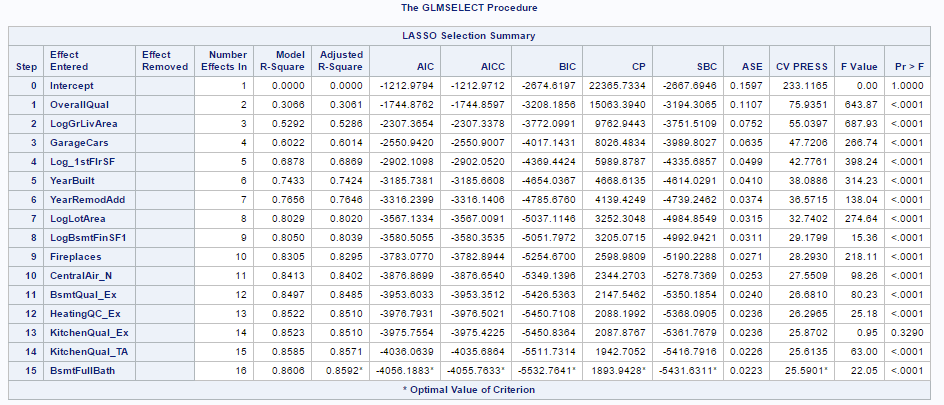


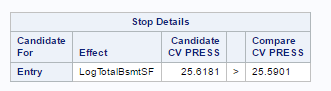


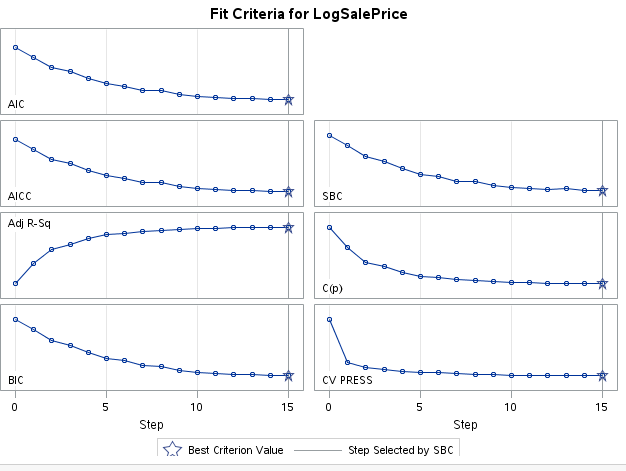
| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | -0.313522 | 0.621006 | -0.50 | 0.6137 |
| **LogGrLivArea** | 1 | 0.429445 | 0.014339 | 29.95 | <.0001 |
| **Neighborhood Blmngtn** | 1 | -0.023435 | 0.044116 | -0.53 | 0.5954 |
| **Neighborhood Blueste** | 1 | -0.014130 | 0.085893 | -0.16 | 0.8694 |
| **Neighborhood BrDale** | 1 | -0.021797 | 0.048141 | -0.45 | 0.6508 |
| **Neighborhood BrkSide** | 1 | 0.034340 | 0.039115 | 0.88 | 0.3801 |
| **Neighborhood ClearCr** | 1 | -0.016642 | 0.039023 | -0.43 | 0.6698 |
| **Neighborhood CollgCr** | 1 | -0.023492 | 0.034733 | -0.68 | 0.4989 |
| **Neighborhood Crawfor** | 1 | 0.095649 | 0.037478 | 2.55 | 0.0108 |
| **Neighborhood Edwards** | 1 | -0.072923 | 0.035940 | -2.03 | 0.0426 |
| **Neighborhood Gilbert** | 1 | -0.040870 | 0.035960 | -1.14 | 0.2559 |
| **Neighborhood IDOTRR** | 1 | -0.021575 | 0.044068 | -0.49 | 0.6245 |
| **Neighborhood MeadowV** | 1 | -0.057335 | 0.046702 | -1.23 | 0.2198 |
| **Neighborhood Mitchel** | 1 | -0.067385 | 0.036924 | -1.82 | 0.0682 |
| **Neighborhood NAmes** | 1 | -0.034215 | 0.034564 | -0.99 | 0.3224 |
| **Neighborhood NPkVill** | 1 | -0.033398 | 0.050463 | -0.66 | 0.5082 |
| **Neighborhood NWAmes** | 1 | -0.072289 | 0.035593 | -2.03 | 0.0424 |
| **Neighborhood NoRidge** | 1 | 0.066272 | 0.037698 | 1.76 | 0.0790 |
| **Neighborhood NridgHt** | 1 | 0.023534 | 0.036595 | 0.64 | 0.5203 |
| **Neighborhood OldTown** | 1 | -0.032066 | 0.039844 | -0.80 | 0.4211 |
| **Neighborhood SWISU** | 1 | -0.005428 | 0.042261 | -0.13 | 0.8978 |
| **Neighborhood Sawyer** | 1 | -0.047474 | 0.036150 | -1.31 | 0.1893 |
| **Neighborhood SawyerW** | 1 | -0.045520 | 0.036314 | -1.25 | 0.2102 |
| **Neighborhood Somerst** | 1 | -0.013161 | 0.041773 | -0.32 | 0.7528 |
| **Neighborhood StoneBr** | 1 | 0.100867 | 0.040344 | 2.50 | 0.0125 |
| **Neighborhood Timber** | 1 | -0.045838 | 0.037433 | -1.22 | 0.2210 |
| **Neighborhood Veenker** | 0 | 0 | . | . | . |
| **MSZoning C** | 1 | -0.313630 | 0.039663 | -7.91 | <.0001 |
| **MSZoning FV** | 1 | 0.094758 | 0.031588 | 3.00 | 0.0027 |
| **MSZoning RH** | 1 | 0.044960 | 0.032785 | 1.37 | 0.1705 |
| **MSZoning RL** | 1 | 0.039234 | 0.015330 | 2.56 | 0.0106 |
| **MSZoning RM** | 0 | 0 | . | . | . |
| **LogLotArea** | 1 | 0.089663 | 0.009083 | 9.87 | <.0001 |
| **OverallQual** | 1 | 0.047468 | 0.004167 | 11.39 | <.0001 |
| **OverallCond** | 1 | 0.044289 | 0.003415 | 12.97 | <.0001 |
| **YearBuilt** | 1 | 0.002502 | 0.000254 | 9.85 | <.0001 |
| **YearRemodAdd** | 1 | 0.001024 | 0.000225 | 4.55 | <.0001 |
| **BsmtQual Ex** | 1 | 0.080827 | 0.017184 | 4.70 | <.0001 |
| **BsmtQual Fa** | 1 | 0.018125 | 0.020256 | 0.89 | 0.3711 |
| **BsmtQual Gd** | 1 | 0.012877 | 0.010252 | 1.26 | 0.2093 |
| **BsmtQual No** | 1 | 0.762193 | 0.080598 | 9.46 | <.0001 |
| **BsmtQual TA** | 0 | 0 | . | . | . |
| **BsmtExposure Av** | 1 | 0.017469 | 0.008995 | 1.94 | 0.0523 |
| **BsmtExposure Gd** | 1 | 0.070202 | 0.011616 | 6.04 | <.0001 |
| **BsmtExposure Mn** | 1 | 0.011229 | 0.011027 | 1.02 | 0.3087 |
| **BsmtExposure No** | 0 | 0 | . | . | . |
| **LogBsmtFinSF1** | 1 | 0.011027 | 0.001143 | 9.65 | <.0001 |
| **LogTotalBsmtSF** | 1 | 0.122092 | 0.011475 | 10.64 | <.0001 |
| **KitchenAbvGr** | 1 | -0.088173 | 0.015400 | -5.73 | <.0001 |
| **KitchenQual Ex** | 1 | 0.090587 | 0.016381 | 5.53 | <.0001 |
| **KitchenQual Fa** | 1 | -0.024620 | 0.019083 | -1.29 | 0.1972 |
| **KitchenQual Gd** | 1 | 0.014087 | 0.008821 | 1.60 | 0.1105 |
| **KitchenQual TA** | 0 | 0 | . | . | . |
| **Functional Maj1** | 1 | -0.097155 | 0.030515 | -3.18 | 0.0015 |
| **Functional Maj2** | 1 | -0.244696 | 0.050097 | -4.88 | <.0001 |
| **Functional Min1** | 1 | -0.067305 | 0.020296 | -3.32 | 0.0009 |
| **Functional Min2** | 1 | -0.044537 | 0.019816 | -2.25 | 0.0248 |
| **Functional Mod** | 1 | -0.106805 | 0.028957 | -3.69 | 0.0002 |
| **Functional Sev** | 1 | -0.447723 | 0.109425 | -4.09 | <.0001 |
| **Functional Typ** | 0 | 0 | . | . | . |
| **Fireplaces** | 1 | 0.027338 | 0.005652 | 4.84 | <.0001 |
| **GarageCars** | 1 | 0.050188 | 0.005601 | 8.96 | <.0001 |
| **SaleCondition Abnorml** | 1 | -0.100906 | 0.016555 | -6.10 | <.0001 |
| **SaleCondition AdjLand** | 1 | -0.043451 | 0.057135 | -0.76 | 0.4471 |
| **SaleCondition Alloca** | 1 | -0.043504 | 0.035280 | -1.23 | 0.2177 |
| **SaleCondition Family** | 1 | -0.103590 | 0.027363 | -3.79 | 0.0002 |
| **SaleCondition Normal** | 1 | -0.040172 | 0.012240 | -3.28 | 0.0011 |
| **SaleCondition Partial** | 0 | 0 | . | . | . |

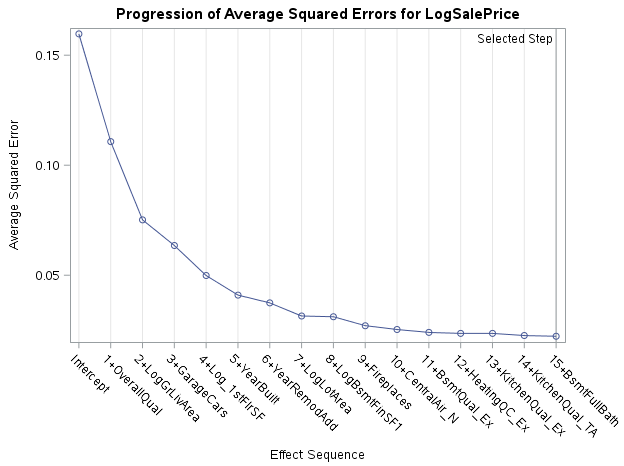
1. **Output D of LASSO Selection Process in Figure 3 below:** We perform a LASSO selection using SBC with a cross-validation stop.

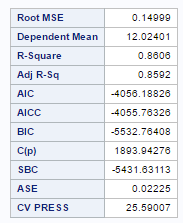
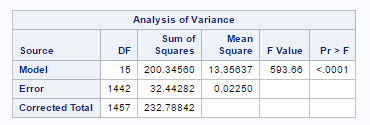
**Figure 4. Output of LASSO Selection Process**

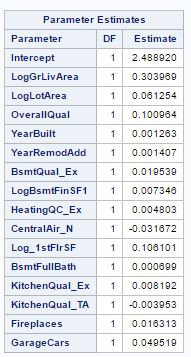




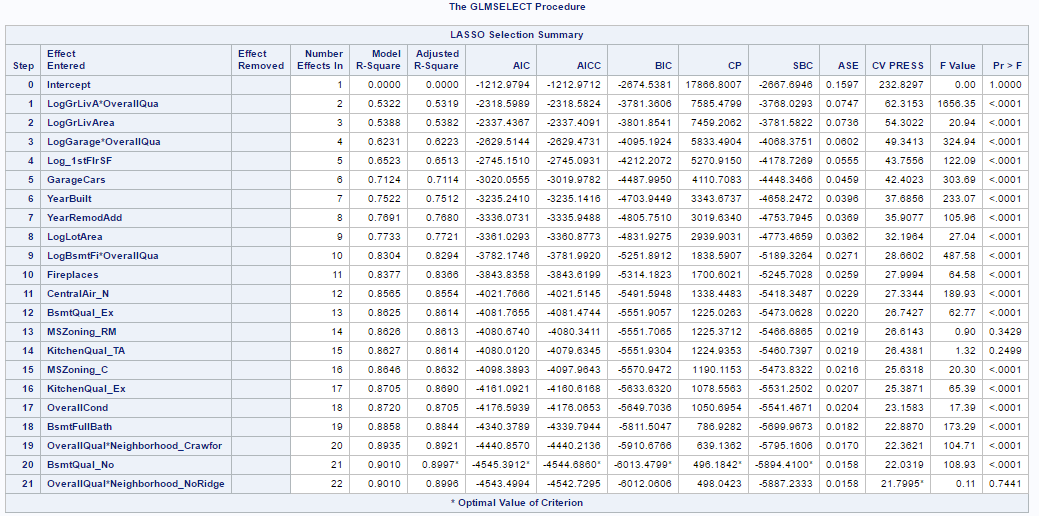
 







1. **Output E: We re-run LASSO checking for signficant interaction terms** among the terms chosen by the model. Those highlighted in green are signficatn interaction terms and those highlighted in red are not.



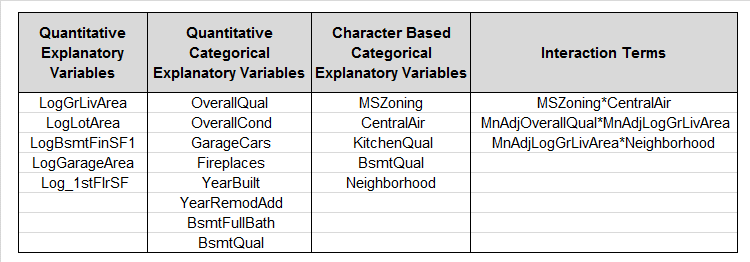
***For SAS code in this appendix, see*** [*Appendix XI*](#Appendix_XISAS) ***in this report and search for “APPENDIX on Model Variable-Selection Techniques”.***

**[APPENDIX VII: Final Model](#_top)**

Below is a summary of our final model output with all factors in detail in **Figure 1** and regression diagnostic plots in **Figure 2**. In **Table 1**, we provide a summary of the regression equation, the PROC GLM model statement and a table of all 21 variables in the model broken in quantitative, quantitative categorical and character categorical variables as well as interaction terms. For the regression equation, we do cannot list all the subcategories and coefficients for each categorical variable all so we enclose them in brackets [] and reference **Figure1** for all the coefficients and variables. Note the explanatory variables name below, a “Log “in a variable name refers to the natural log for our purposes.

As we can see all variables are statistically significant and model itself has p-value < 0.05. However, broken out into individual sub-categories of each categorical variable, some are not significant. Again, we keep the overall categorical variable if it is significant. In terms of VIFs, there are a few subcategories of categorical variables that show high VIFs but we do not remove them because other subcategories are still significant and have low VIFs (see table of Dependent Variable: LogSalePrice Tolerances in **Figure 1**). Finally, **Figure 3** provide our forecast (predicted SalePrice values) of our test data set based on our model. Our ***Kaggle score is 13.602%,*** which is the top 13.602%-tile of all submissions (see [Kaggle Website under Submission jhos25](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/submissions?sortBy=date&group=all&page=1)). In addition, we note that the intercept will include one subcategory of each categorical variable as the reference sub-category.

**Table 1. Regression Equation, PROC GLM Regression Equation and Variables in Final Model**



**PROC GLM Model:**

LogSalePrice = LogGrLivArea LogLotArea LogBsmtFinSF1 LogGarageArea   
Log\_1stFlrSF Neighborhood OverallQual OverallCond GarageCars Fireplaces YearBuilt YearRemodAdd MSZoning CentralAir KitchenQual BsmtFullBath BsmtQual   
MSZoning\*CentralAir MnAdjOverallQual\*MnAdjLogGrLivArea MnAdjLogGrLivArea\*Neighborhood

**Regression Equation:**

**LogSalePrice** = 2.645865289 + 0.627932446 \* LogGrLivArea - 0.100088503 \* LogLotArea

+ 0.010110255\* LogBsmtFinSF1 + 0.004155274 \* LogGarageArea + 0.107116329 \* Log\_1stFlrSF

+ [0.013093938 \*NeighborhoodBlmngtn +…… - 0.053878242 \* Neighborhood Timber]

+ 0.055859197 \* OverallQual + 0.047292374 \* OverallCond + 0.042101862 \* GarageCars

+ 0.035552140 \* Fireplaces + 0.003123955 \* YearBuilt + 0.000744956 \* YearRemodAdd

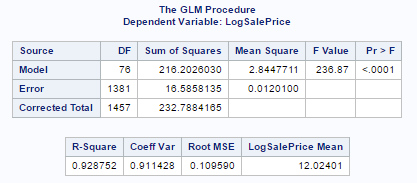
+ [-0.122906003 \* MSZoning\_C +…. + 0.027876581 \* MSZoning\_RL] - 0.085341481 \* CentralAir\_N + [0.066356275 \* KitchenQual\_Ex + …. + 0.014132672 \* KitchenQual Gd]

+ 0.031081582 \* BsmtFullBath + [0.087031022 \* BsmtQual\_Ex +…- 0.112433238 \* BsmtQual\_No]

+ [ -0.290463233 \* MSZoning\*CentralAir C\_N +….+0.000000\* MSZoning\*CentralAir\_RM\_N]

+ 0.034593053 \* MnAdjOverAllQual\*MnAdjLogGrLivArea + [-0.078464915 \* MnAdjLogGrLivArea\* NeighborhoodBlmngtn +… -0.159524676 \* MnAdjLogGrLivArea \* Neighborhood Timber]

**Figure 1. Summary of Table Outputs of Final Regression Model**

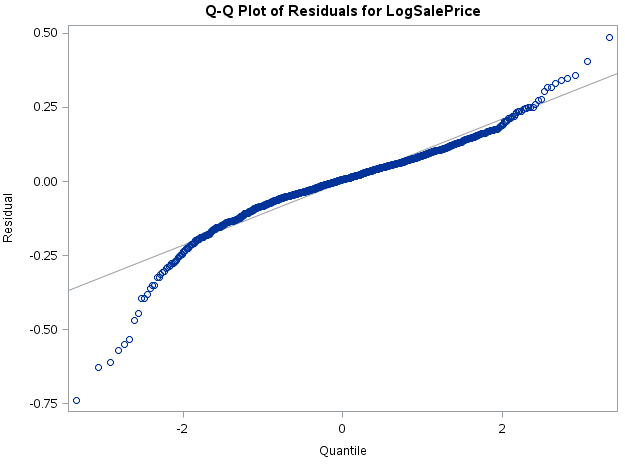
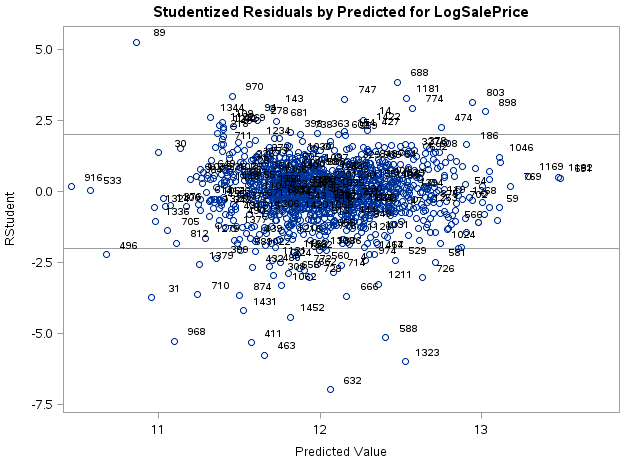


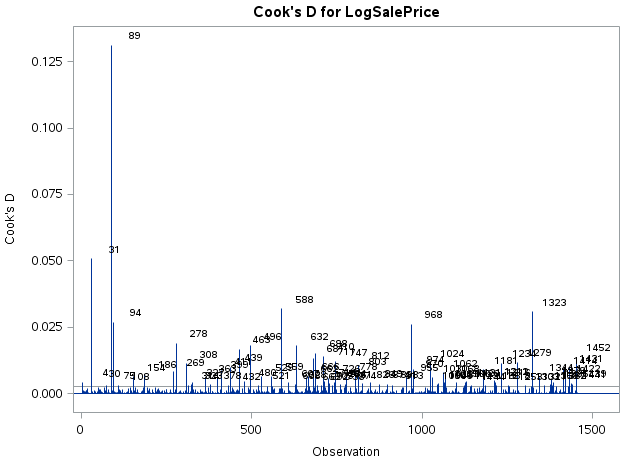
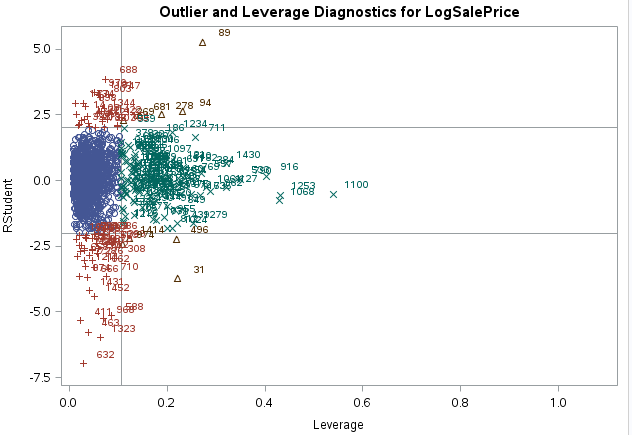
 

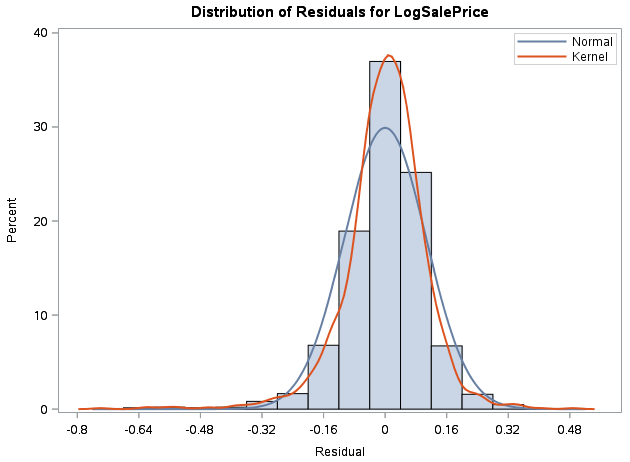
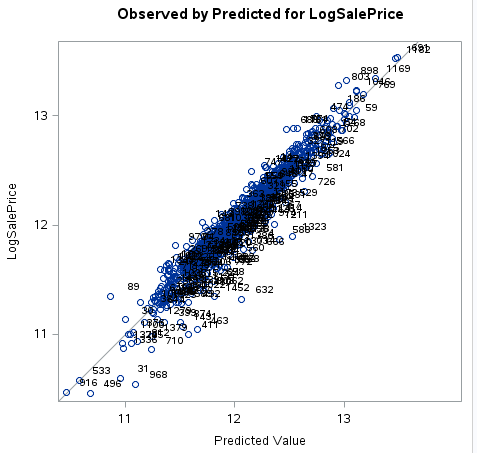
| **Parameter** | **Estimate** |  | **Standard Error** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Intercept** | -2.645865289 | B | 1.45406849 | -1.82 | 0.0690 | -5.498287109 | 0.206556531 |
| **LogGrLivArea** | 0.627932446 | B | 0.17977200 | 3.49 | 0.0005 | 0.275276732 | 0.980588160 |
| **LogLotArea** | 0.100088503 |  | 0.00948962 | 10.55 | <.0001 | 0.081472865 | 0.118704141 |
| **LogBsmtFinSF1** | 0.010110255 |  | 0.00134358 | 7.52 | <.0001 | 0.007474584 | 0.012745925 |
| **LogGarageArea** | 0.004155274 |  | 0.00328462 | 1.27 | 0.2061 | -0.002288110 | 0.010598659 |
| **Log\_1stFlrSF** | 0.107116329 |  | 0.01433482 | 7.47 | <.0001 | 0.078995952 | 0.135236705 |
| **Neighborhood Blmngtn** | 0.013093938 | B | 0.04596645 | 0.28 | 0.7758 | -0.077077675 | 0.103265551 |
| **Neighborhood Blueste** | -0.013073425 | B | 0.09073939 | -0.14 | 0.8855 | -0.191075358 | 0.164928509 |
| **Neighborhood BrDale** | 0.020115019 | B | 0.06996623 | 0.29 | 0.7738 | -0.117136555 | 0.157366594 |
| **Neighborhood BrkSide** | 0.062373829 | B | 0.04232573 | 1.47 | 0.1408 | -0.020655856 | 0.145403514 |
| **Neighborhood ClearCr** | 0.016844413 | B | 0.04346621 | 0.39 | 0.6984 | -0.068422528 | 0.102111354 |
| **Neighborhood CollgCr** | -0.018834071 | B | 0.03648486 | -0.52 | 0.6058 | -0.090405808 | 0.052737666 |
| **Neighborhood Crawfor** | 0.093238414 | B | 0.04048056 | 2.30 | 0.0214 | 0.013828381 | 0.172648447 |
| **Neighborhood Edwards** | -0.074042540 | B | 0.03820028 | -1.94 | 0.0528 | -0.148979391 | 0.000894311 |
| **Neighborhood Gilbert** | -0.043914823 | B | 0.03865527 | -1.14 | 0.2561 | -0.119744228 | 0.031914581 |
| **Neighborhood IDOTRR** | 0.055975036 | B | 0.04982425 | 1.12 | 0.2614 | -0.041764353 | 0.153714425 |
| **Neighborhood MeadowV** | -0.119136816 | B | 0.05537243 | -2.15 | 0.0316 | -0.227759987 | -0.010513645 |
| **Neighborhood Mitchel** | -0.092853609 | B | 0.03952649 | -2.35 | 0.0190 | -0.170392061 | -0.015315156 |
| **Neighborhood NAmes** | -0.067967162 | B | 0.03659567 | -1.86 | 0.0635 | -0.139756266 | 0.003821942 |
| **Neighborhood NPkVill** | -0.014529152 | B | 0.06069240 | -0.24 | 0.8108 | -0.133588423 | 0.104530119 |
| **Neighborhood NWAmes** | -0.084848763 | B | 0.03829060 | -2.22 | 0.0269 | -0.159962783 | -0.009734743 |
| **Neighborhood NoRidge** | -0.038187934 | B | 0.05994056 | -0.64 | 0.5242 | -0.155772328 | 0.079396460 |
| **Neighborhood NridgHt** | 0.026351764 | B | 0.04091209 | 0.64 | 0.5196 | -0.053904807 | 0.106608335 |
| **Neighborhood OldTown** | -0.024058217 | B | 0.04241100 | -0.57 | 0.5706 | -0.107255162 | 0.059138729 |
| **Neighborhood SWISU** | -0.018015898 | B | 0.04499101 | -0.40 | 0.6889 | -0.106274012 | 0.070242215 |
| **Neighborhood Sawyer** | -0.095256682 | B | 0.03919864 | -2.43 | 0.0152 | -0.172151998 | -0.018361366 |
| **Neighborhood SawyerW** | -0.067279141 | B | 0.03808854 | -1.77 | 0.0776 | -0.141996787 | 0.007438505 |
| **Neighborhood Somerst** | 0.015808020 | B | 0.04435726 | 0.36 | 0.7216 | -0.071206879 | 0.102822919 |
| **Neighborhood StoneBr** | 0.081723342 | B | 0.04595523 | 1.78 | 0.0756 | -0.008426269 | 0.171872953 |
| **Neighborhood Timber** | -0.053878242 | B | 0.04212482 | -1.28 | 0.2011 | -0.136513804 | 0.028757319 |
| **Neighborhood Veenker** | 0.000000000 | B | . | . | . | . | . |
| **OverallQual** | 0.055859197 |  | 0.00421720 | 13.25 | <.0001 | 0.047586380 | 0.064132015 |
| **OverallCond** | 0.047292374 |  | 0.00354085 | 13.36 | <.0001 | 0.040346339 | 0.054238410 |
| **GarageCars** | 0.042101862 |  | 0.00784990 | 5.36 | <.0001 | 0.026702840 | 0.057500885 |
| **Fireplaces** | 0.035552140 |  | 0.00584154 | 6.09 | <.0001 | 0.024092890 | 0.047011389 |
| **YearBuilt** | 0.003123955 |  | 0.00027760 | 11.25 | <.0001 | 0.002579386 | 0.003668524 |
| **YearRemodAdd** | 0.000744956 |  | 0.00023162 | 3.22 | 0.0013 | 0.000290588 | 0.001199324 |
| **MSZoning C** | -0.122906003 | B | 0.05987562 | -2.05 | 0.0403 | -0.240363011 | -0.005448995 |
| **MSZoning FV** | 0.063952419 | B | 0.03299561 | 1.94 | 0.0528 | -0.000774525 | 0.128679363 |
| **MSZoning RH** | 0.081780048 | B | 0.04045156 | 2.02 | 0.0434 | 0.002426907 | 0.161133189 |
| **MSZoning RL** | 0.027876581 | B | 0.01629785 | 1.71 | 0.0874 | -0.004094646 | 0.059847807 |
| **MSZoning RM** | 0.000000000 | B | . | . | . | . | . |
| **CentralAir N** | -0.085341481 | B | 0.02190041 | -3.90 | 0.0001 | -0.128303157 | -0.042379804 |
| **CentralAir Y** | 0.000000000 | B | . | . | . | . | . |
| **KitchenQual Ex** | 0.066356275 | B | 0.01702283 | 3.90 | 0.0001 | 0.032962877 | 0.099749673 |
| **KitchenQual Fa** | -0.011151455 | B | 0.02025198 | -0.55 | 0.5820 | -0.050879431 | 0.028576521 |
| **KitchenQual Gd** | 0.014132672 | B | 0.00901852 | 1.57 | 0.1173 | -0.003558800 | 0.031824145 |
| **KitchenQual TA** | 0.000000000 | B | . | . | . | . | . |
| **BsmtFullBath** | 0.031081582 |  | 0.00743400 | 4.18 | <.0001 | 0.016498433 | 0.045664732 |
| **BsmtQual Ex** | 0.087031022 | B | 0.01777553 | 4.90 | <.0001 | 0.052161064 | 0.121900979 |
| **BsmtQual Fa** | 0.001601600 | B | 0.02039855 | 0.08 | 0.9374 | -0.038413889 | 0.041617089 |
| **BsmtQual Gd** | 0.018867552 | B | 0.01058921 | 1.78 | 0.0750 | -0.001905130 | 0.039640233 |
| **BsmtQual No** | -0.112433238 | B | 0.02082010 | -5.40 | <.0001 | -0.153275679 | -0.071590798 |
| **BsmtQual TA** | 0.000000000 | B | . | . | . | . | . |
| **MSZoning\*CentralAir C N** | -0.290463233 | B | 0.07504538 | -3.87 | 0.0001 | -0.437678499 | -0.143247967 |
| **MSZoning\*CentralAir C Y** | 0.000000000 | B | . | . | . | . | . |
| **MSZoning\*CentralAir FV Y** | 0.000000000 | B | . | . | . | . | . |
| **MSZoning\*CentralAir RH N** | -0.063162100 | B | 0.06203833 | -1.02 | 0.3088 | -0.184861649 | 0.058537450 |
| **MSZoning\*CentralAir RH Y** | 0.000000000 | B | . | . | . | . | . |
| **MSZoning\*CentralAir RL N** | 0.105777687 | B | 0.02812867 | 3.76 | 0.0002 | 0.050598137 | 0.160957238 |
| **MSZoning\*CentralAir RL Y** | 0.000000000 | B | . | . | . | . | . |
| **MSZoning\*CentralAir RM N** | 0.000000000 | B | . | . | . | . | . |
| **MSZoning\*CentralAir RM Y** | 0.000000000 | B | . | . | . | . | . |
| **MnAdjOver\*MnAdjLogGr** | 0.034593053 |  | 0.00855122 | 4.05 | <.0001 | 0.017818268 | 0.051367838 |
| **MnAdjLogG\*Neighborho Blmngtn** | -0.078464915 | B | 0.32767434 | -0.24 | 0.8108 | -0.721258178 | 0.564328347 |
| **MnAdjLogG\*Neighborho Blueste** | -0.149979426 | B | 0.68136046 | -0.22 | 0.8258 | -1.486592834 | 1.186633983 |
| **MnAdjLogG\*Neighborho BrDale** | -0.018486812 | B | 0.27676205 | -0.07 | 0.9468 | -0.561406296 | 0.524432672 |
| **MnAdjLogG\*Neighborho BrkSide** | -0.084577575 | B | 0.18605847 | -0.45 | 0.6495 | -0.449565369 | 0.280410220 |
| **MnAdjLogG\*Neighborho ClearCr** | -0.272556063 | B | 0.19734798 | -1.38 | 0.1675 | -0.659690287 | 0.114578161 |
| **MnAdjLogG\*Neighborho CollgCr** | -0.242916234 | B | 0.18181661 | -1.34 | 0.1818 | -0.599582826 | 0.113750358 |
| **MnAdjLogG\*Neighborho Crawfor** | -0.268863457 | B | 0.18686339 | -1.44 | 0.1504 | -0.635430246 | 0.097703332 |
| **MnAdjLogG\*Neighborho Edwards** | -0.206047805 | B | 0.18488821 | -1.11 | 0.2653 | -0.568739911 | 0.156644301 |
| **MnAdjLogG\*Neighborho Gilbert** | -0.221335870 | B | 0.19201767 | -1.15 | 0.2492 | -0.598013711 | 0.155341972 |
| **MnAdjLogG\*Neighborho IDOTRR** | -0.009696668 | B | 0.19436505 | -0.05 | 0.9602 | -0.390979336 | 0.371585999 |
| **MnAdjLogG\*Neighborho MeadowV** | -0.314700774 | B | 0.19325584 | -1.63 | 0.1037 | -0.693807528 | 0.064405980 |
| **MnAdjLogG\*Neighborho Mitchel** | -0.309432185 | B | 0.18785011 | -1.65 | 0.0997 | -0.677934609 | 0.059070239 |
| **MnAdjLogG\*Neighborho NAmes** | -0.390035724 | B | 0.18173993 | -2.15 | 0.0320 | -0.746551905 | -0.033519544 |
| **MnAdjLogG\*Neighborho NPkVill** | -0.223761520 | B | 0.26829742 | -0.83 | 0.4044 | -0.750076071 | 0.302553030 |
| **MnAdjLogG\*Neighborho NWAmes** | -0.259326763 | B | 0.18751733 | -1.38 | 0.1669 | -0.627176375 | 0.108522849 |
| **MnAdjLogG\*Neighborho NoRidge** | -0.095540229 | B | 0.19913440 | -0.48 | 0.6315 | -0.486178848 | 0.295098389 |
| **MnAdjLogG\*Neighborho NridgHt** | -0.189825513 | B | 0.18992409 | -1.00 | 0.3177 | -0.562396428 | 0.182745402 |
| **MnAdjLogG\*Neighborho OldTown** | -0.166262916 | B | 0.18226714 | -0.91 | 0.3618 | -0.523813315 | 0.191287483 |
| **MnAdjLogG\*Neighborho SWISU** | -0.171911220 | B | 0.18855149 | -0.91 | 0.3621 | -0.541789516 | 0.197967075 |
| **MnAdjLogG\*Neighborho Sawyer** | -0.377141124 | B | 0.18714767 | -2.02 | 0.0441 | -0.744265578 | -0.010016670 |
| **MnAdjLogG\*Neighborho SawyerW** | -0.170748887 | B | 0.18496456 | -0.92 | 0.3561 | -0.533590774 | 0.192093001 |
| **MnAdjLogG\*Neighborho Somerst** | -0.234793192 | B | 0.19191329 | -1.22 | 0.2214 | -0.611266272 | 0.141679888 |
| **MnAdjLogG\*Neighborho StoneBr** | -0.147903130 | B | 0.19718071 | -0.75 | 0.4533 | -0.534709218 | 0.238902958 |
| **MnAdjLogG\*Neighborho Timber** | -0.159524676 | B | 0.19847677 | -0.80 | 0.4217 | -0.548873229 | 0.229823878 |
| **MnAdjLogG\*Neighborho Veenker** | 0.000000000 | B | . | . | . | . | . |

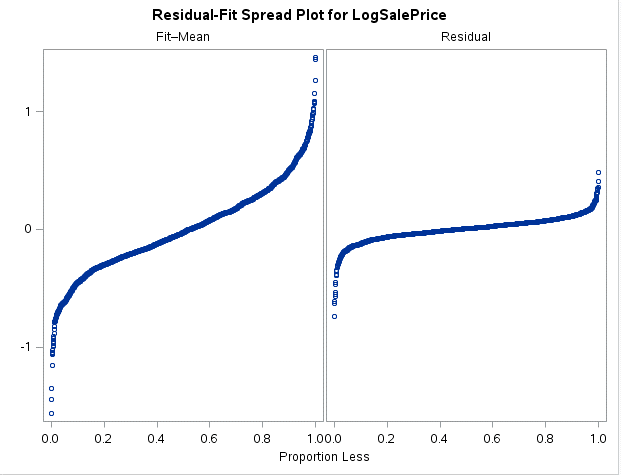
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **The GLM Procedure** | | **Dependent Variable: LogSalePrice Tolerances** | | |
| **Matrix Element Representation** | |
| **Dependent Variable: LogSalePrice** | |
| **Effects** | **Representation** | **Type I Tolerance** | **Type II Tolerance** | **VIF** |
| **Intercept** | Intercept | 1458.0 | 0.005680 | 0.0007 |
| **LogGrLivArea** | LogGrLivArea | 1.0 | 0.002337 | 1.0000 |
| **LogLotArea** | LogLotArea | 0.859766 | 0.346863 | 1.1631 |
| **LogBsmtFinSF1** | Dummy001 | 0.987948 | 0.510599 | 1.0122 |
| **LogGarageArea** | Dummy002 | 0.887798 | 0.360704 | 1.1264 |
| **Log\_1stFlrSF** | Log\_1stFlrSF | 0.597624 | 0.406356 | 1.6733 |
| **Neighborhood Blmngtn** | Dummy003 | 0.905517 | 0.338303 | 1.1043 |
| **Neighborhood Blueste** | Dummy004 | 0.978636 | 0.730327 | 1.0218 |
| **Neighborhood BrDale** | Dummy005 | 0.864087 | 0.155038 | 1.1573 |
| **Neighborhood BrkSide** | Dummy006 | 0.951503 | 0.120375 | 1.0510 |
| **Neighborhood ClearCr** | Dummy007 | 0.940386 | 0.231474 | 1.0634 |
| **Neighborhood CollgCr** | Dummy008 | 0.980095 | 0.067047 | 1.0203 |
| **Neighborhood Crawfor** | Dummy009 | 0.980557 | 0.148917 | 1.0198 |
| **Neighborhood Edwards** | Dummy010 | 0.921444 | 0.090033 | 1.0853 |
| **Neighborhood Gilbert** | Dummy011 | 0.885371 | 0.107570 | 1.1295 |
| **Neighborhood IDOTRR** | Dummy012 | 0.930222 | 0.134160 | 1.0750 |
| **Neighborhood MeadowV** | Dummy013 | 0.861970 | 0.233131 | 1.1601 |
| **Neighborhood Mitchel** | Dummy014 | 0.958146 | 0.162337 | 1.0437 |
| **Neighborhood NAmes** | Dummy015 | 0.793618 | 0.047130 | 1.2601 |
| **Neighborhood NPkVill** | Dummy016 | 0.943841 | 0.364520 | 1.0595 |
| **Neighborhood NWAmes** | Dummy017 | 0.912932 | 0.118126 | 1.0954 |
| **Neighborhood NoRidge** | Dummy018 | 0.873997 | 0.083889 | 1.1442 |
| **Neighborhood NridgHt** | Dummy019 | 0.831175 | 0.098381 | 1.2031 |
| **Neighborhood OldTown** | Dummy020 | 0.745331 | 0.064054 | 1.3417 |
| **Neighborhood SWISU** | Dummy021 | 0.899646 | 0.241470 | 1.1115 |
| **Neighborhood Sawyer** | Dummy022 | 0.737406 | 0.111273 | 1.3561 |
| **Neighborhood SawyerW** | Dummy023 | 0.754738 | 0.146232 | 1.3250 |
| **Neighborhood Somerst** | Dummy024 | 0.448197 | 0.075426 | 2.2312 |
| **Neighborhood StoneBr** | Dummy025 | 0.651065 | 0.231443 | 1.5359 |
| **Neighborhood Timber** | Dummy026 | 0.229999 | 0.182874 | 4.3478 |
| **Neighborhood Veenker** | Dummy027 | 0.000000 | 0.000000 | 5.0375E+13 |
| **OverallQual** | OverallQual | 0.349807 | 0.244661 | 2.8587 |
| **OverallCond** | OverallCond | 0.787468 | 0.530392 | 1.2699 |
| **GarageCars** | GarageCars | 0.271027 | 0.239658 | 3.6897 |
| **Fireplaces** | Fireplaces | 0.628320 | 0.586105 | 1.5915 |
| **YearBuilt** | YearBuilt | 0.175102 | 0.117328 | 5.7109 |
| **YearRemodAdd** | YearRemodAdd | 0.420724 | 0.360605 | 2.3769 |
| **MSZoning C** | MSZoning C | 0.768182 | 0.337312 | 1.3018 |
| **MSZoning FV** | MSZoning FV | 0.235561 | 0.177633 | 4.2452 |
| **MSZoning RH** | MSZoning RH | 0.891057 | 0.463815 | 1.1223 |
| **MSZoning RL** | MSZoning RL | 0.210742 | 0.185678 | 4.7451 |
| **MSZoning RM** | MSZoning RM | 0.000000 | 0.000000 | 1.3733E+15 |
| **CentralAir N** | CentralAir N | 0.695974 | 0.281953 | 1.4368 |
| **CentralAir Y** | CentralAir Y | 0.000000 | 0.000000 | 4.6292E+14 |
| **KitchenQual Ex** | Dummy028 | 0.689340 | 0.453390 | 1.4507 |
| **KitchenQual Fa** | Dummy029 | 0.827191 | 0.771469 | 1.2089 |
| **KitchenQual Gd** | Dummy030 | 0.440357 | 0.421324 | 2.2709 |
| **KitchenQual TA** | Dummy031 | 0.000000 | 0.000000 | 1.4249E+15 |
| **BsmtFullBath** | BsmtFullBath | 0.581999 | 0.557158 | 1.7182 |
| **BsmtQual Ex** | BsmtQual Ex | 0.563369 | 0.347799 | 1.7750 |
| **BsmtQual Fa** | BsmtQual Fa | 0.863600 | 0.844946 | 1.1579 |
| **BsmtQual Gd** | BsmtQual Gd | 0.317858 | 0.300819 | 3.1461 |
| **BsmtQual No** | BsmtQual No | 0.812207 | 0.768315 | 1.2312 |
| **BsmtQual TA** | BsmtQual TA | 0.000000 | 0.000000 | 5.6312E+15 |
| **MSZoning\*CentralAir C N** | Dummy032 | 0.378368 | 0.356890 | 2.6429 |
| **MSZoning\*CentralAir C Y** | Dummy033 | 0.000000 | 0.000000 | -4.4913E+15 |
| **MSZoning\*CentralAir FV Y** | Dummy034 | 0.000000 | 0.000000 | -4.7607E+31 |
| **MSZoning\*CentralAir RH N** | Dummy035 | 0.566823 | 0.522231 | 1.7642 |
| **MSZoning\*CentralAir RH Y** | Dummy036 | 0.000000 | 0.000000 | -6.2121E+14 |
| **MSZoning\*CentralAir RL N** | Dummy037 | 0.368122 | 0.348055 | 2.7165 |
| **MSZoning\*CentralAir RL Y** | Dummy038 | 0.000000 | 0.000000 | 1.3237E+15 |
| **MSZoning\*CentralAir RM N** | Dummy039 | 0.000000 | 0.000000 | 3.4724E+15 |
| **MSZoning\*CentralAir RM Y** | Dummy040 | 0.000000 | 0.000000 | 1.2003E+15 |
| **MnAdjOver\*MnAdjLogGr** | Dummy041 | 0.651861 | 0.379936 | 1.5341 |
| **MnAdjLogG\*Neighborho Blmngtn** | Dummy042 | 0.985962 | 0.687337 | 1.0142 |
| **MnAdjLogG\*Neighborho Blueste** | Dummy043 | 0.919635 | 0.858059 | 1.0874 |
| **MnAdjLogG\*Neighborho BrDale** | Dummy044 | 0.239009 | 0.137827 | 4.1839 |
| **MnAdjLogG\*Neighborho BrkSide** | Dummy045 | 0.563766 | 0.034149 | 1.7738 |
| **MnAdjLogG\*Neighborho ClearCr** | Dummy046 | 0.623661 | 0.107952 | 1.6034 |
| **MnAdjLogG\*Neighborho CollgCr** | Dummy047 | 0.764834 | 0.026042 | 1.3075 |
| **MnAdjLogG\*Neighborho Crawfor** | Dummy048 | 0.685252 | 0.052798 | 1.4593 |
| **MnAdjLogG\*Neighborho Edwards** | Dummy049 | 0.632121 | 0.035313 | 1.5820 |
| **MnAdjLogG\*Neighborho Gilbert** | Dummy050 | 0.662257 | 0.088761 | 1.5100 |
| **MnAdjLogG\*Neighborho IDOTRR** | Dummy051 | 0.434261 | 0.062909 | 2.3028 |
| **MnAdjLogG\*Neighborho MeadowV** | Dummy052 | 0.464854 | 0.065579 | 2.1512 |
| **MnAdjLogG\*Neighborho Mitchel** | Dummy053 | 0.734119 | 0.069178 | 1.3622 |
| **MnAdjLogG\*Neighborho NAmes** | Dummy054 | 0.563118 | 0.017194 | 1.7758 |
| **MnAdjLogG\*Neighborho NPkVill** | Dummy055 | 0.589240 | 0.331243 | 1.6971 |
| **MnAdjLogG\*Neighborho NWAmes** | Dummy056 | 0.604879 | 0.060042 | 1.6532 |
| **MnAdjLogG\*Neighborho NoRidge** | Dummy057 | 0.115648 | 0.022676 | 8.6469 |
| **MnAdjLogG\*Neighborho NridgHt** | Dummy058 | 0.307660 | 0.037460 | 3.2503 |
| **MnAdjLogG\*Neighborho OldTown** | Dummy059 | 0.514531 | 0.023665 | 1.9435 |
| **MnAdjLogG\*Neighborho SWISU** | Dummy060 | 0.620545 | 0.072486 | 1.6115 |
| **MnAdjLogG\*Neighborho Sawyer** | Dummy061 | 0.416750 | 0.044682 | 2.3995 |
| **MnAdjLogG\*Neighborho SawyerW** | Dummy062 | 0.452202 | 0.054931 | 2.2114 |
| **MnAdjLogG\*Neighborho Somerst** | Dummy063 | 0.460379 | 0.095136 | 2.1721 |
| **MnAdjLogG\*Neighborho StoneBr** | Dummy064 | 0.294317 | 0.094129 | 3.3977 |
| **MnAdjLogG\*Neighborho Timber** | Dummy065 | 0.106650 | 0.106650 | 9.3765 |
| **MnAdjLogG\*Neighborho Veenker** | Dummy066 | 0.000000 | 0.000000 | -5.0012E+09 |

**Figure 2. Diagnostic Plots for Final Model**

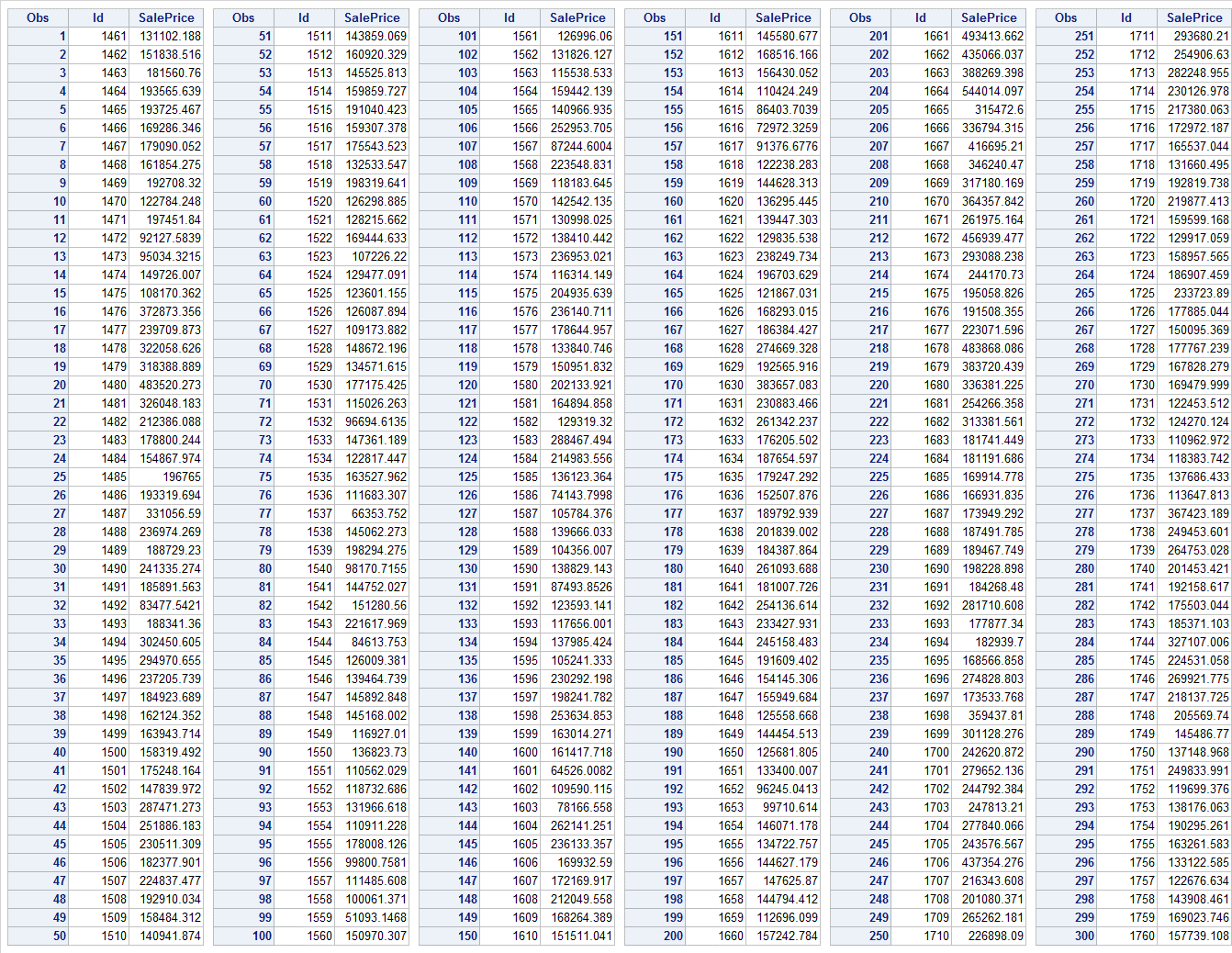


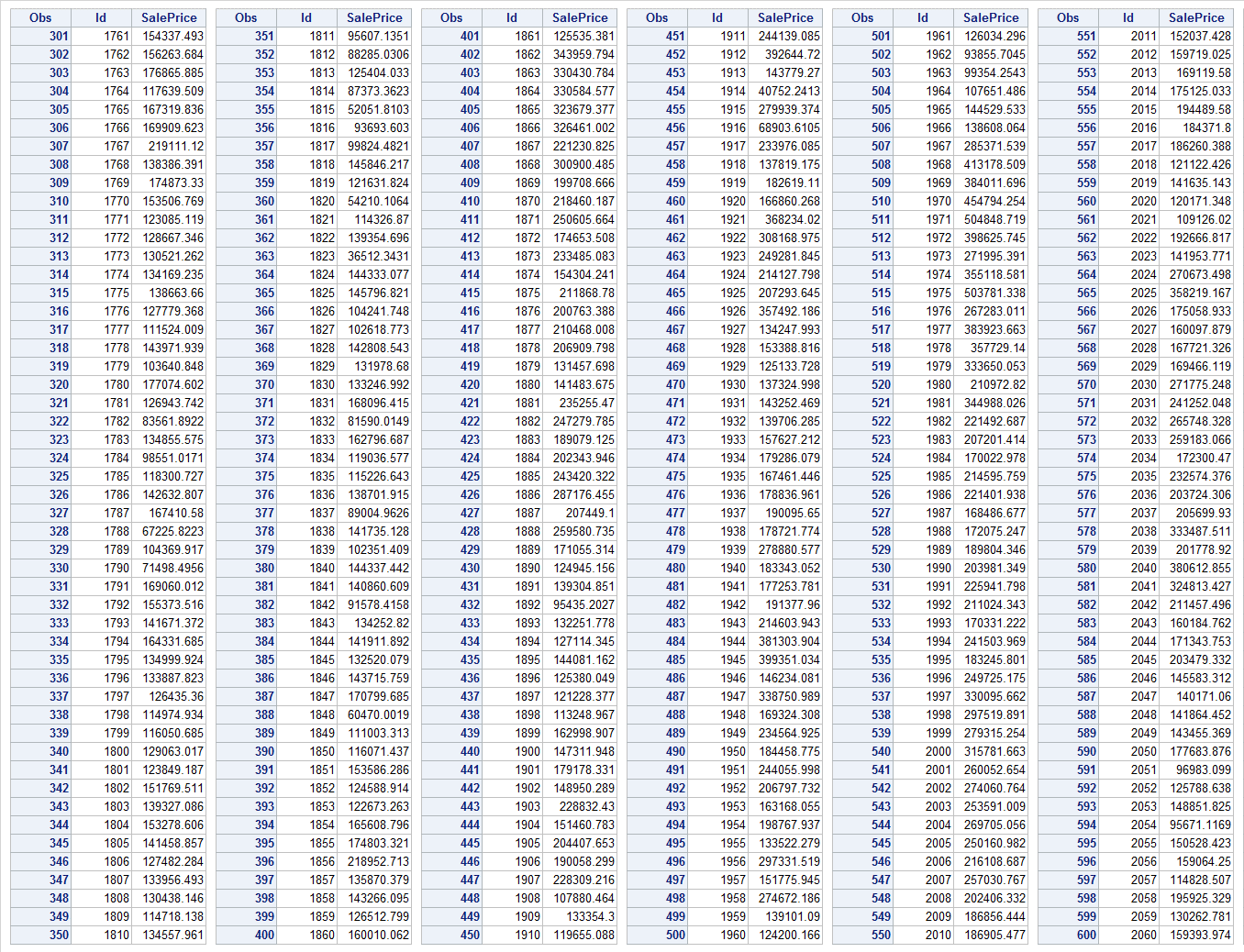


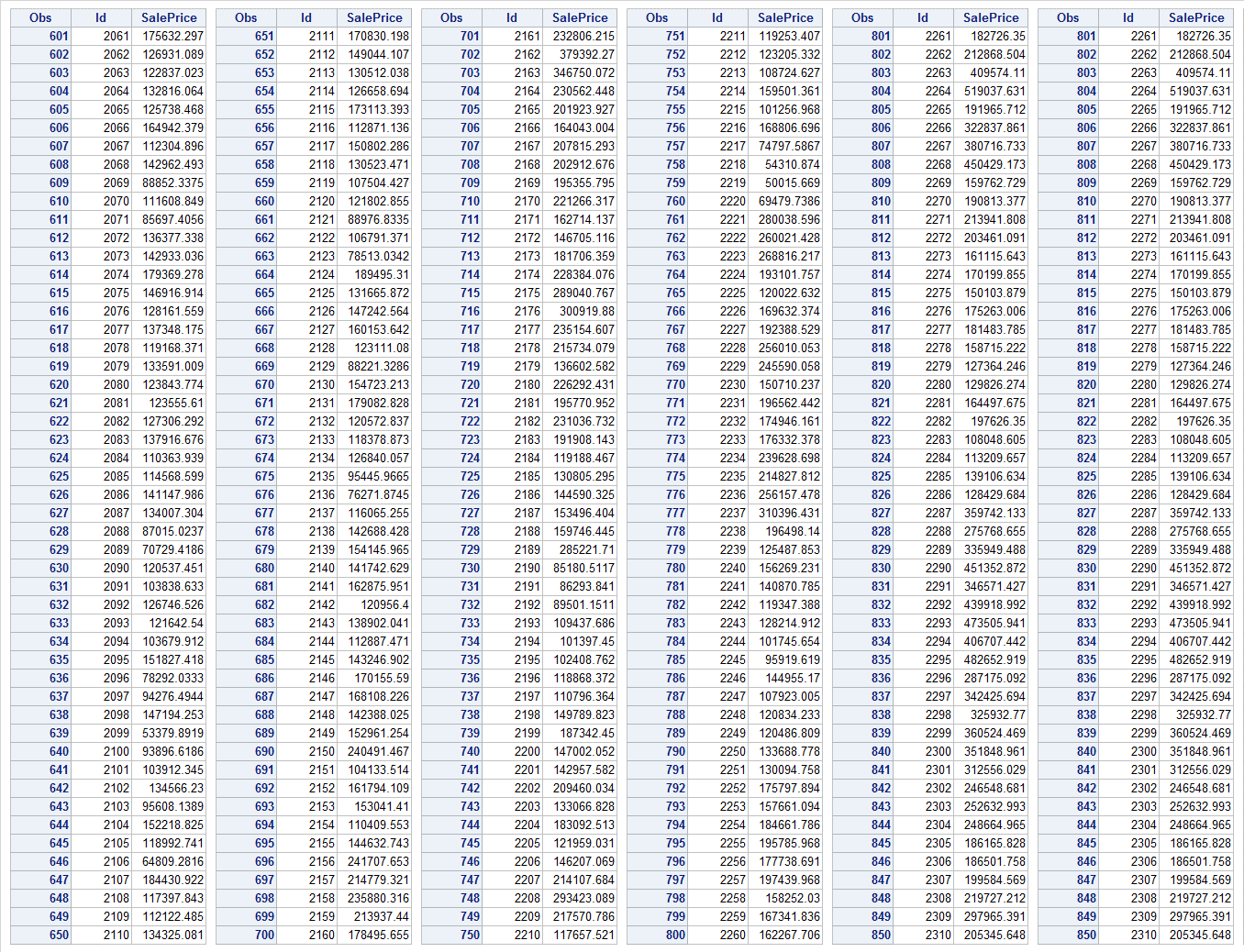


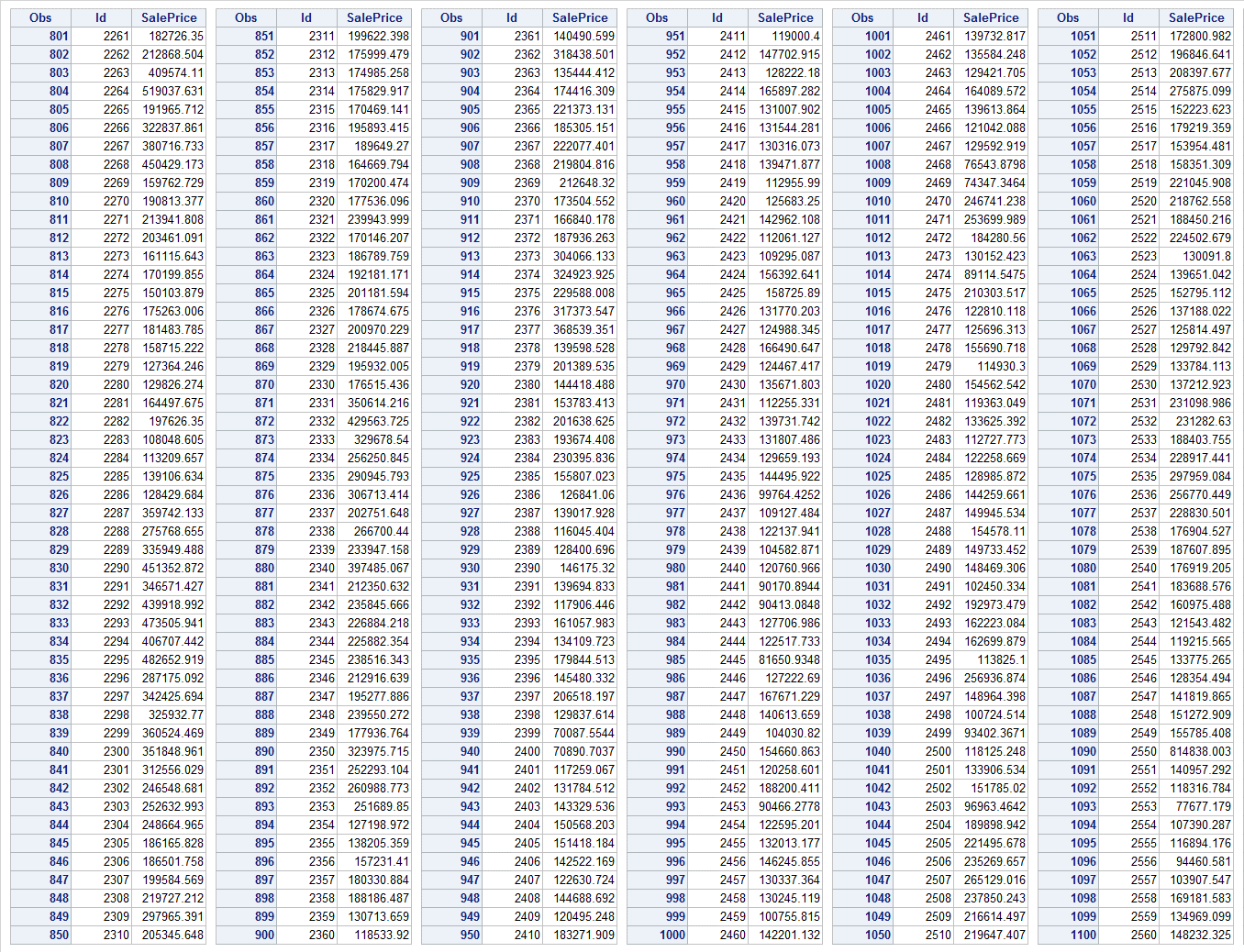


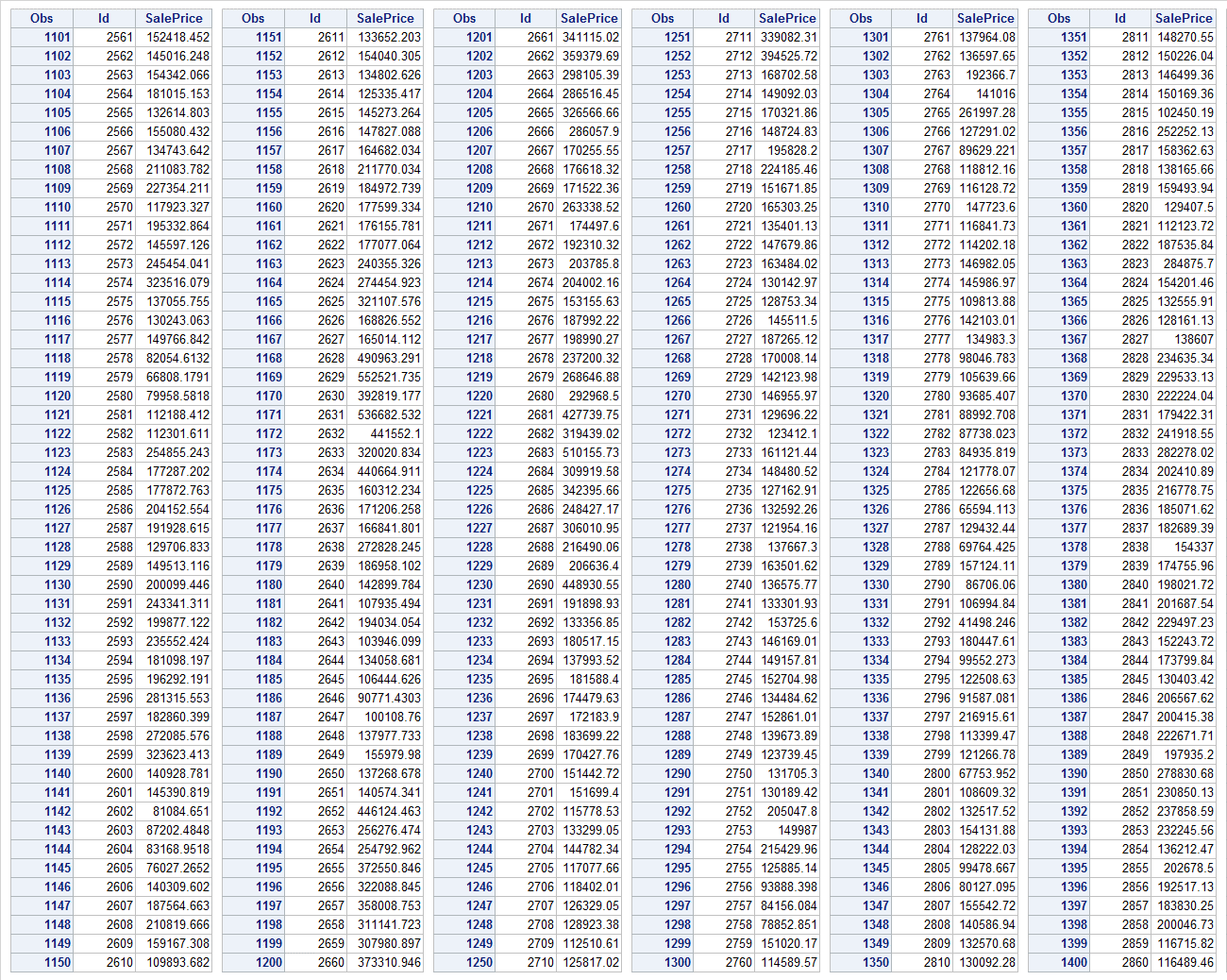
**Figure 3. Predicted Forecast of Test Data**

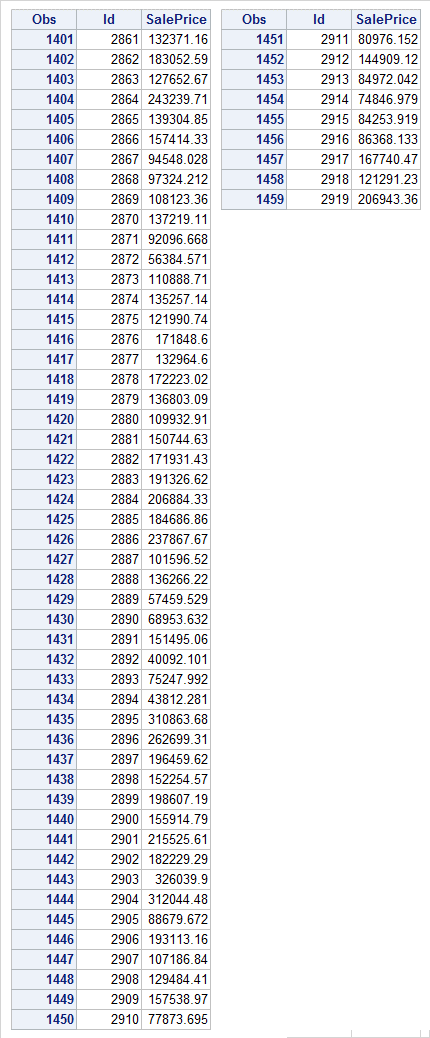










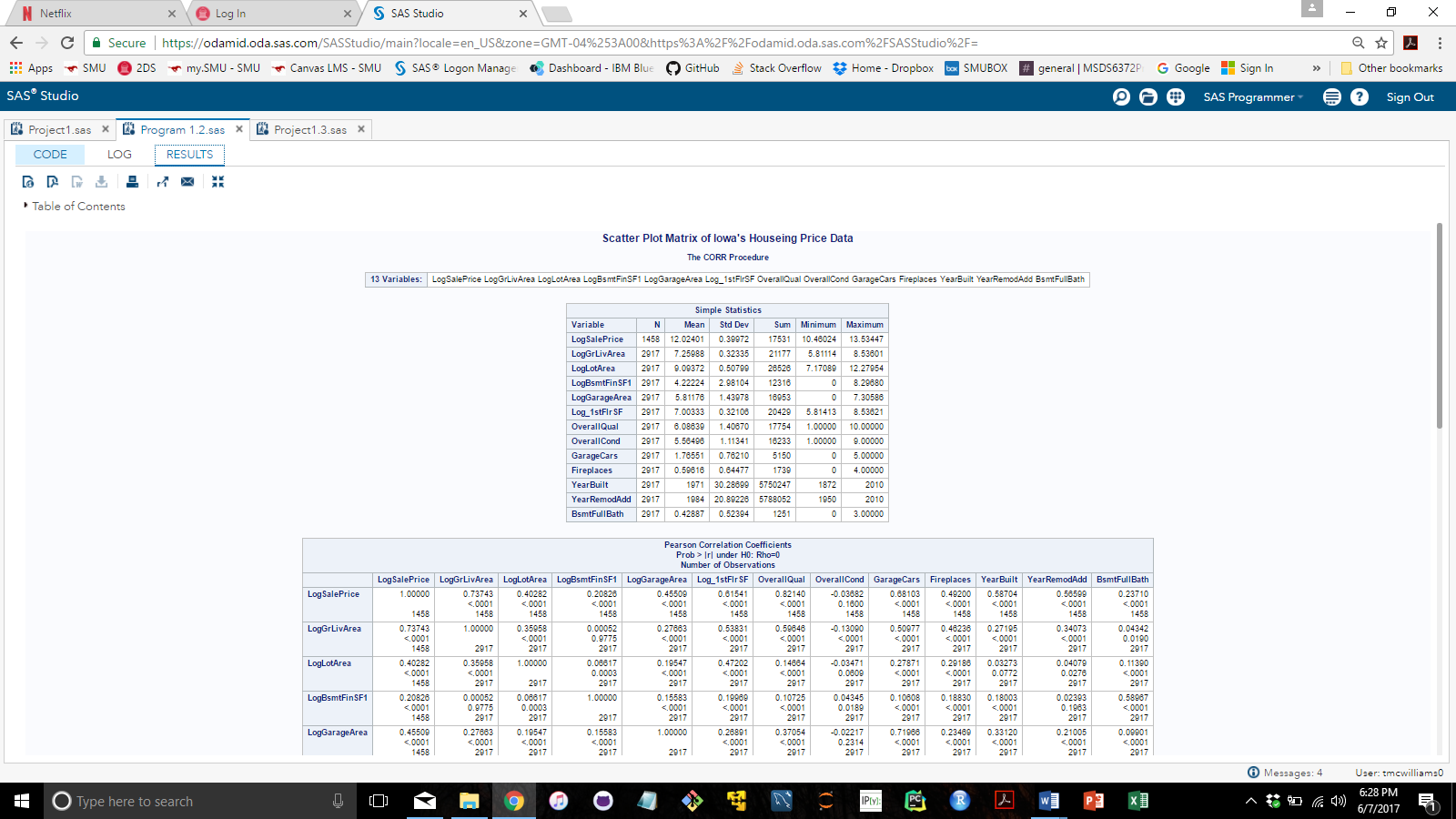


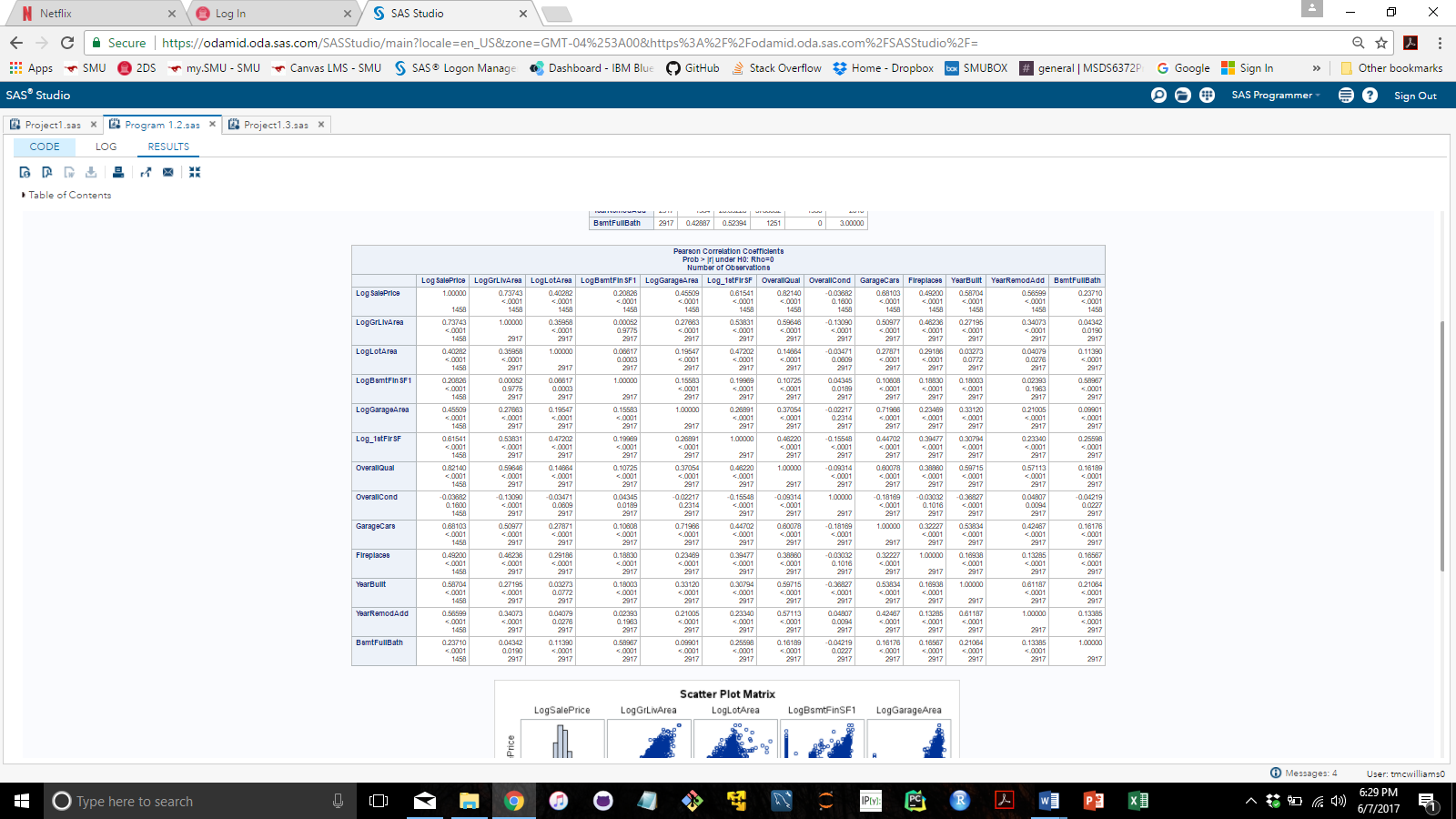
***For SAS code in this appendix, see*** [*Appendix XI*](#Appendix_XISAS) ***in this report and search for “APPENDIX on Final Model”.***

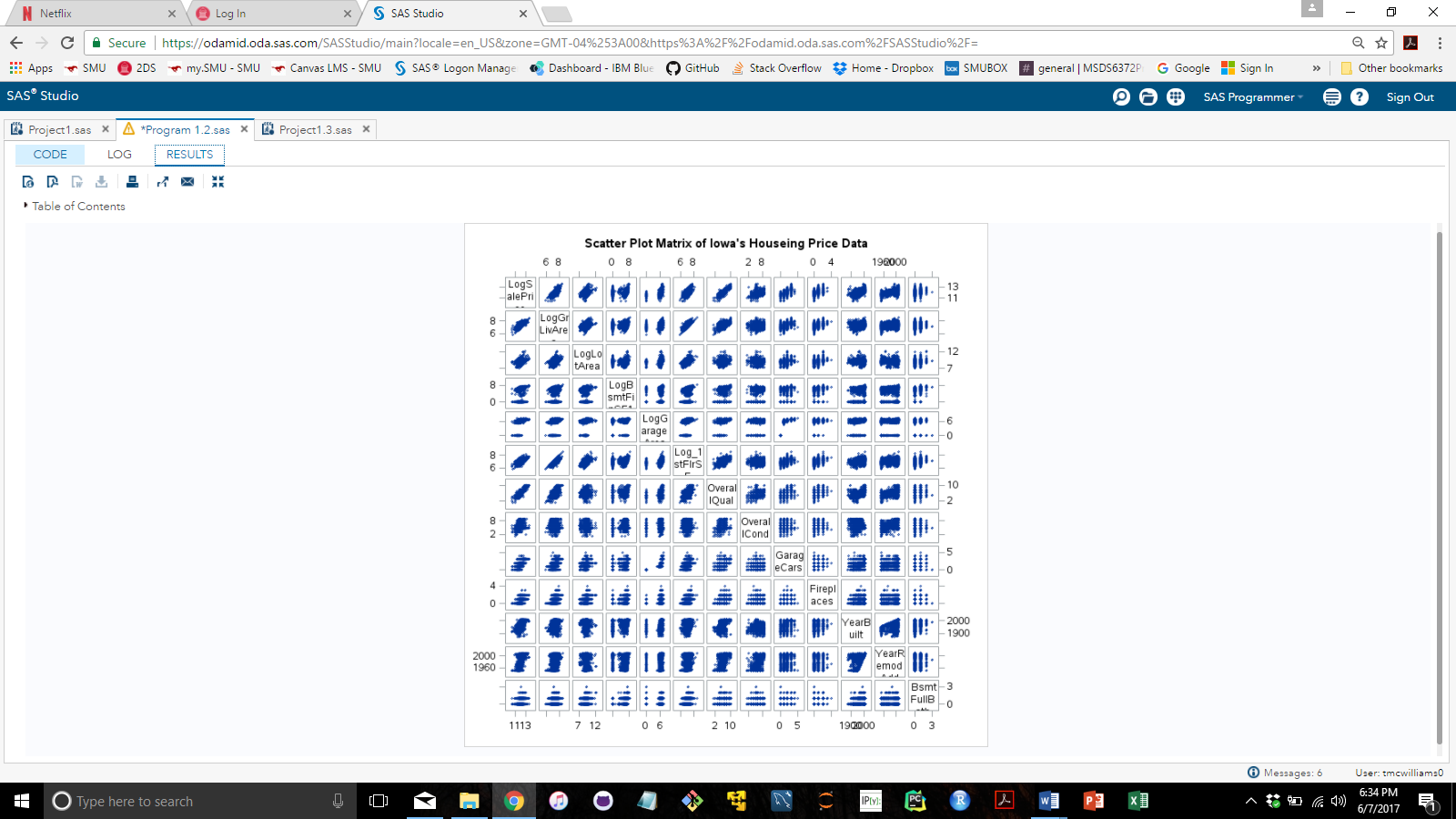
**APPENDIX VIII: Correlation Analysis**

This section of the appendix will focus on the correlation between the explanatory variables. Since there are two types of data in this analysis, numeric and categorical, interpreting the correlation between these two types can be extensive. First, we will assess the correlation among all the numeric variables: LogSalePrice, LogGrLivArea, LogLotArea, LogBsmtFinSF1, LogGarageArea, Log\_1stFlrSF, OverallQual, OverallCond, GarageCars Fireplaces, YearBuilt, YearRemodAdd, and BsmtFullBath. We are going to perform a Persons correlation and look at a matrix of scatterplots.

The Pearson correlation coefficients are numbers that measure the strength and direction of the linear relationship between the two variables. The correlation coefficient can range from -1 to +1, with -1 indicating a perfect negative correlation, +1 indicating a perfect positive correlation, and 0 indicating no correlation at all. Note, a variable correlated with itself will always have a correlation coefficient of 1. **Table 2** shows the Pearson’s correlation coefficients for all numeric explanatory variables. There are weak correlations between the following variables: LogSalePrice and OverallQual, LogSalePrice and GarageCars, LogSalePrice and LogGrLivArea, LogSalePrice and Log\_1stFlrSF, and OverallQual and GarageCars. Figure 1 displays the scatter plot matrix of all the numeric explanatory variables. Said scatter plots reinforce the correlations between the variables stated.

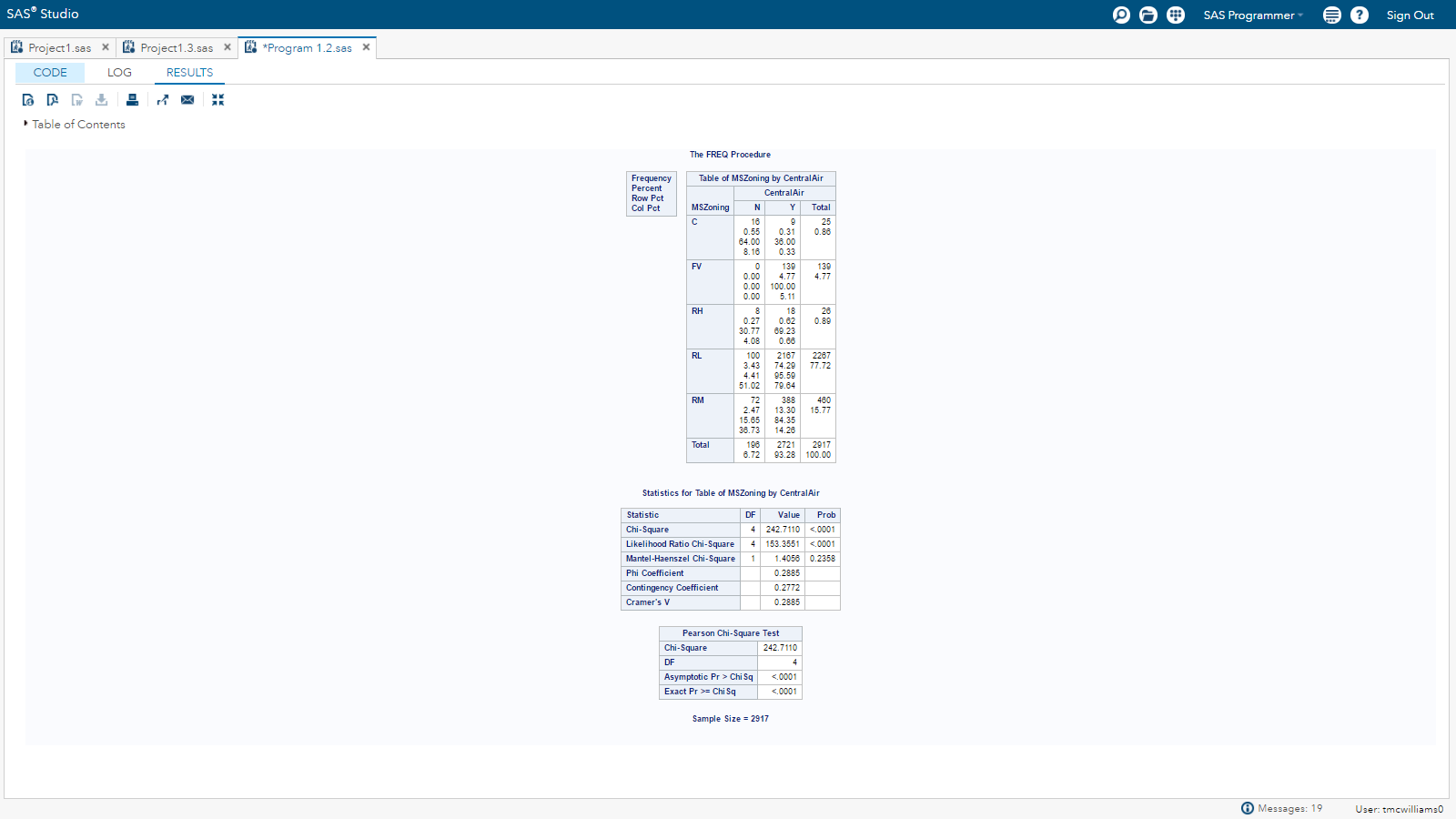
**Table 1: Descriptive statistics of the numeric explanatory variables**.

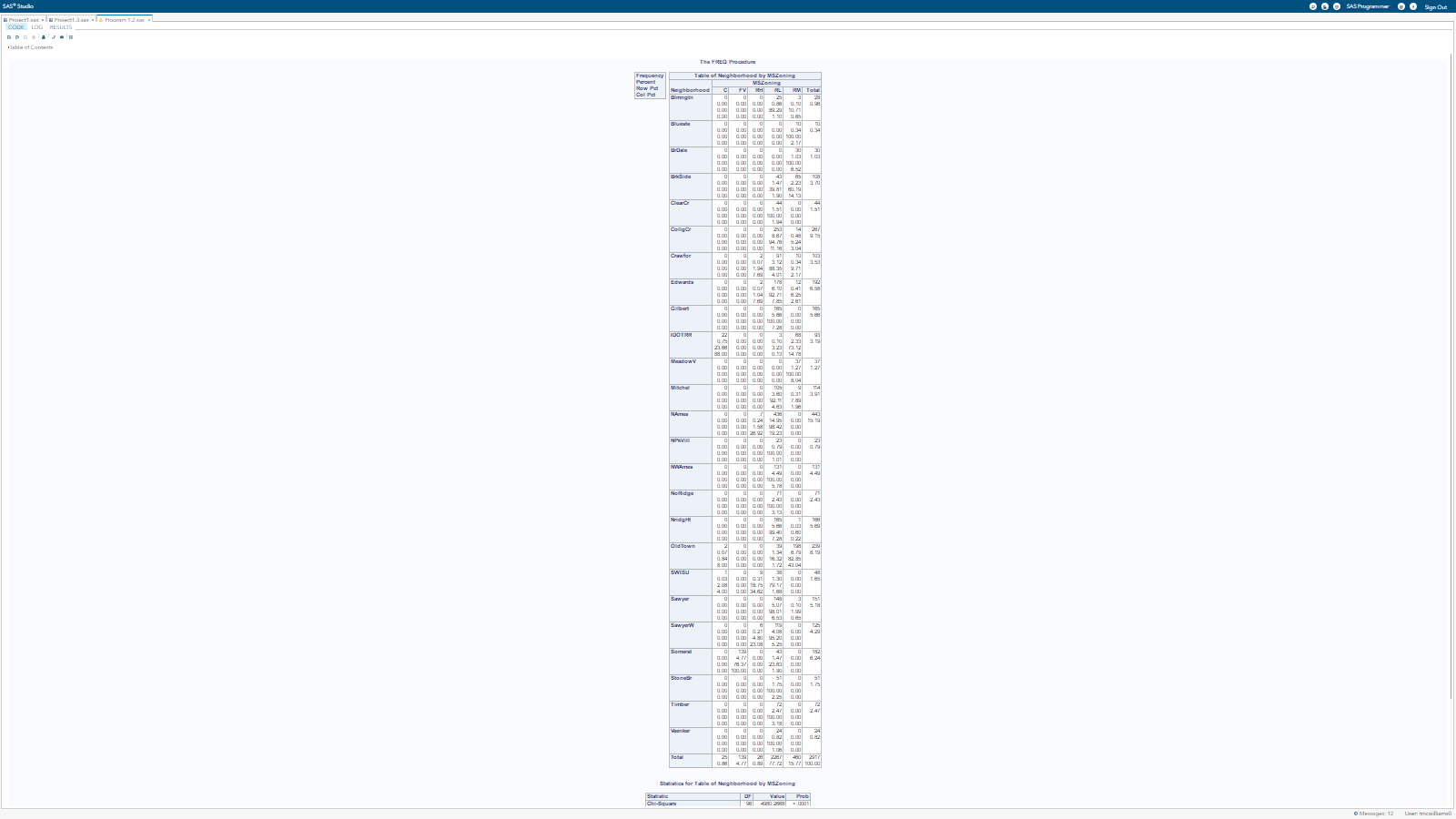
**Table 2: Pearson’s correlation coefficients.**

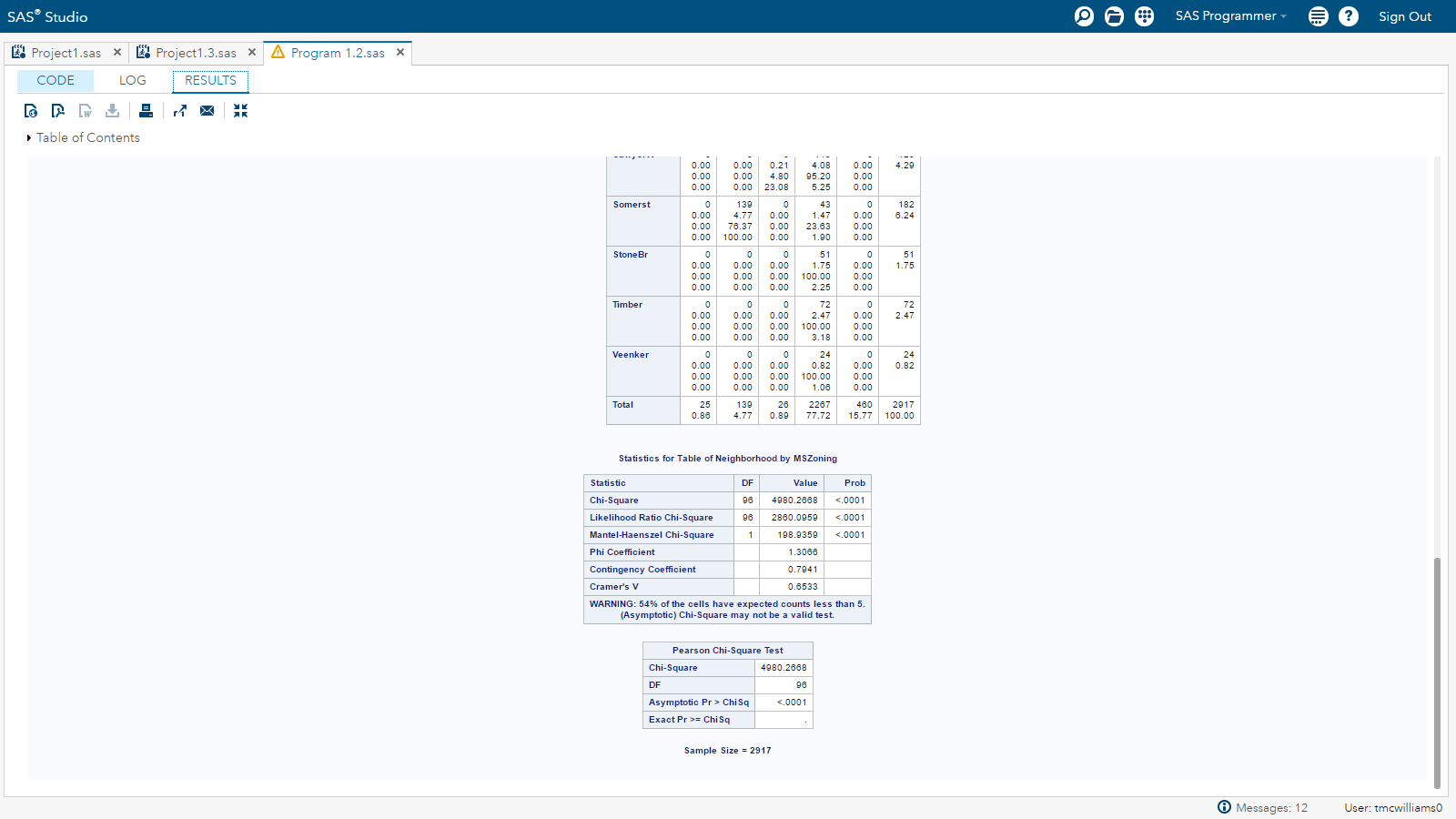
**Figure 1: Scatter plot matrix of all numeric explanatory variables.**

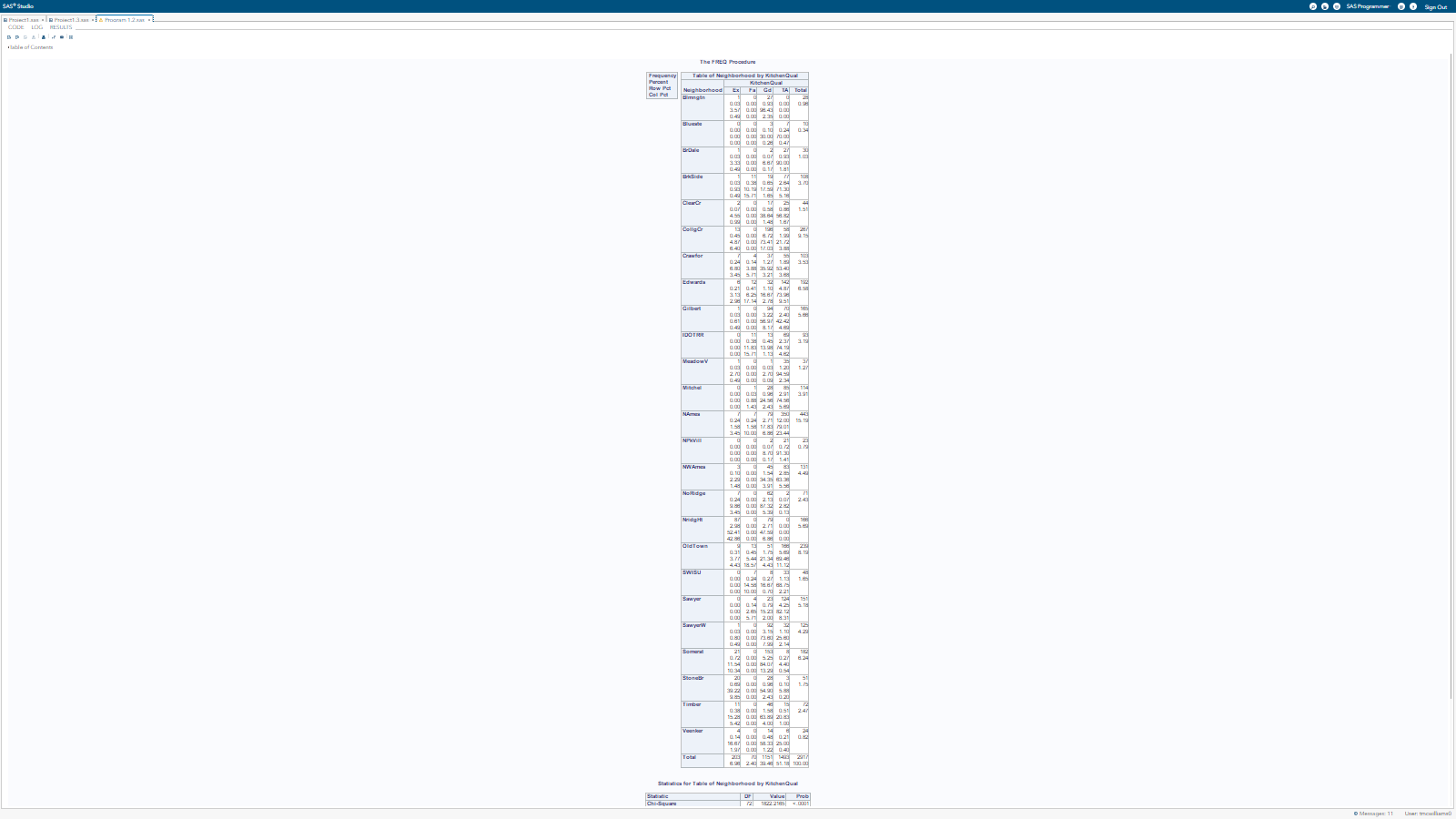
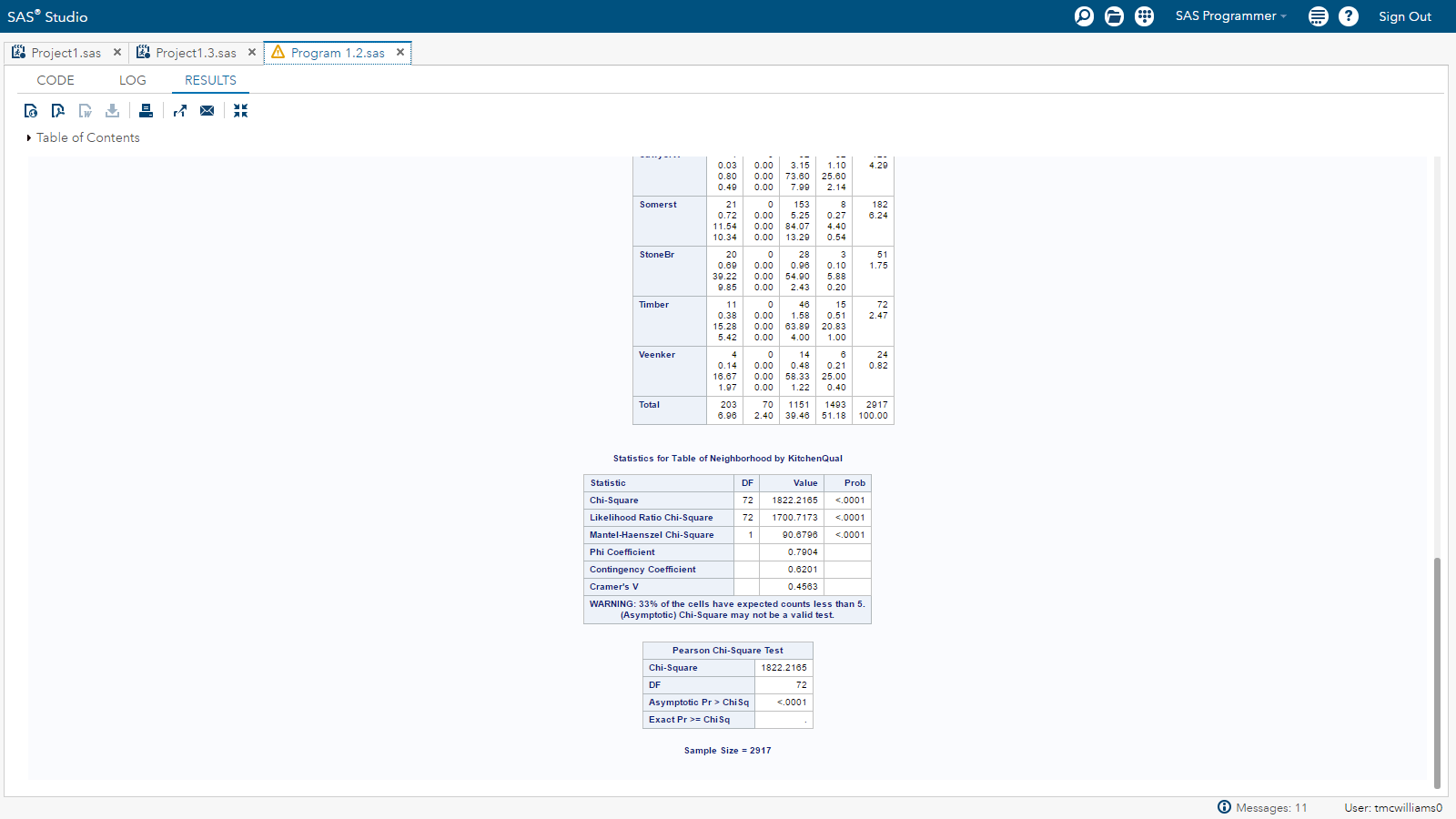
Next, we will be assessing the correlation between the five categorical variables: Neighborhood, CentralAir, KitchenQual, BsmtQual, and MSZoning. Using contingency tables and the Chi-Squared test we were able to find that there is no correlation between every possible association of the categorical variables. However, the correlation between MSZoning and CentralAir was not able to be assessed. There are too many possible associations of the subcategories for the machine to out put the results. We are assuming that these two categorical variables have no association or mild association with each other. The main thing we want to see is if the variables are redundant.

**Table 3: Contingency tables and the Chi-squared test on MSZoning and CenteralAir**

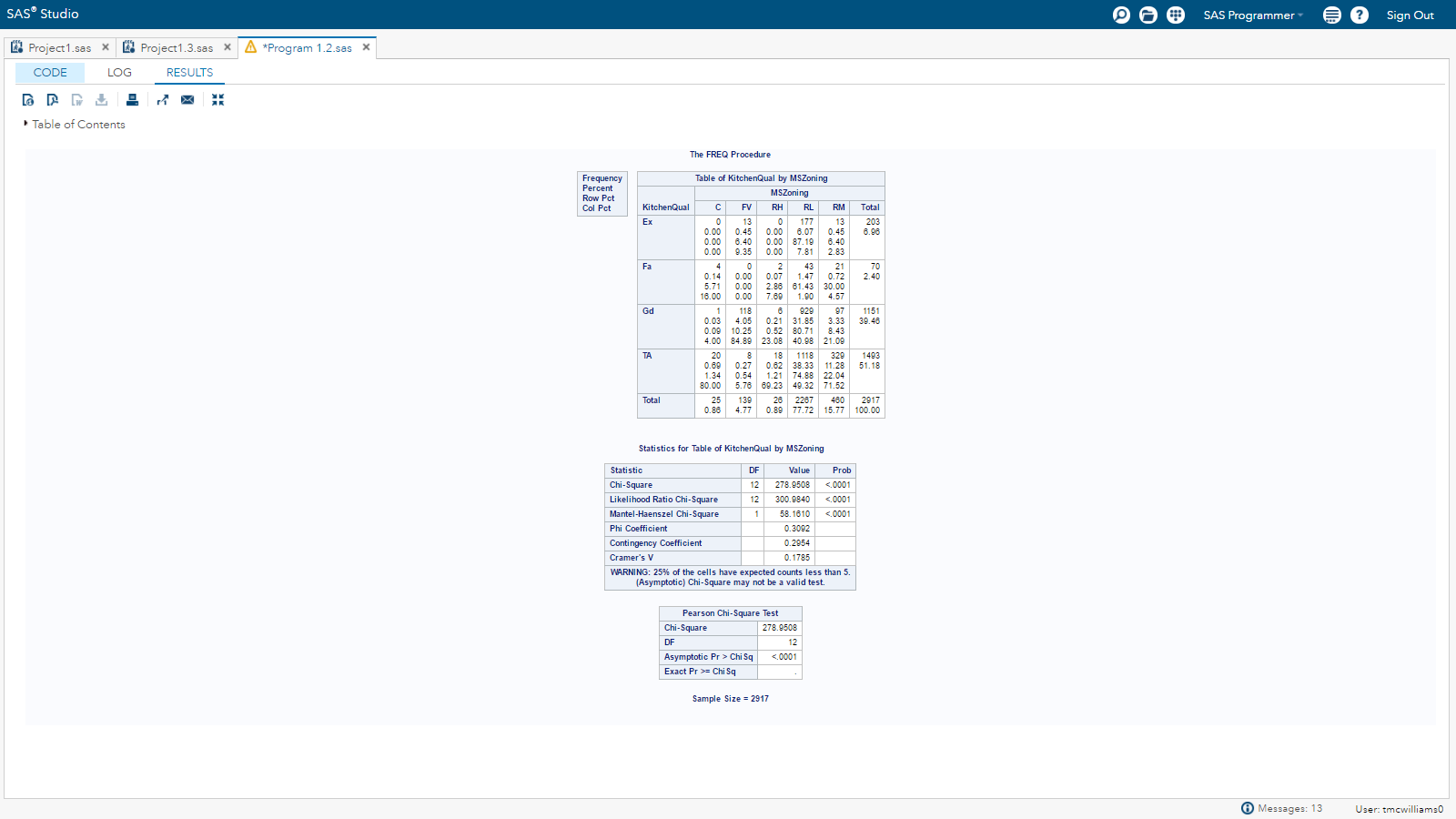


**Table 4: Contingency tables and the Chi-squared test on Neighborhood and MSZoning**

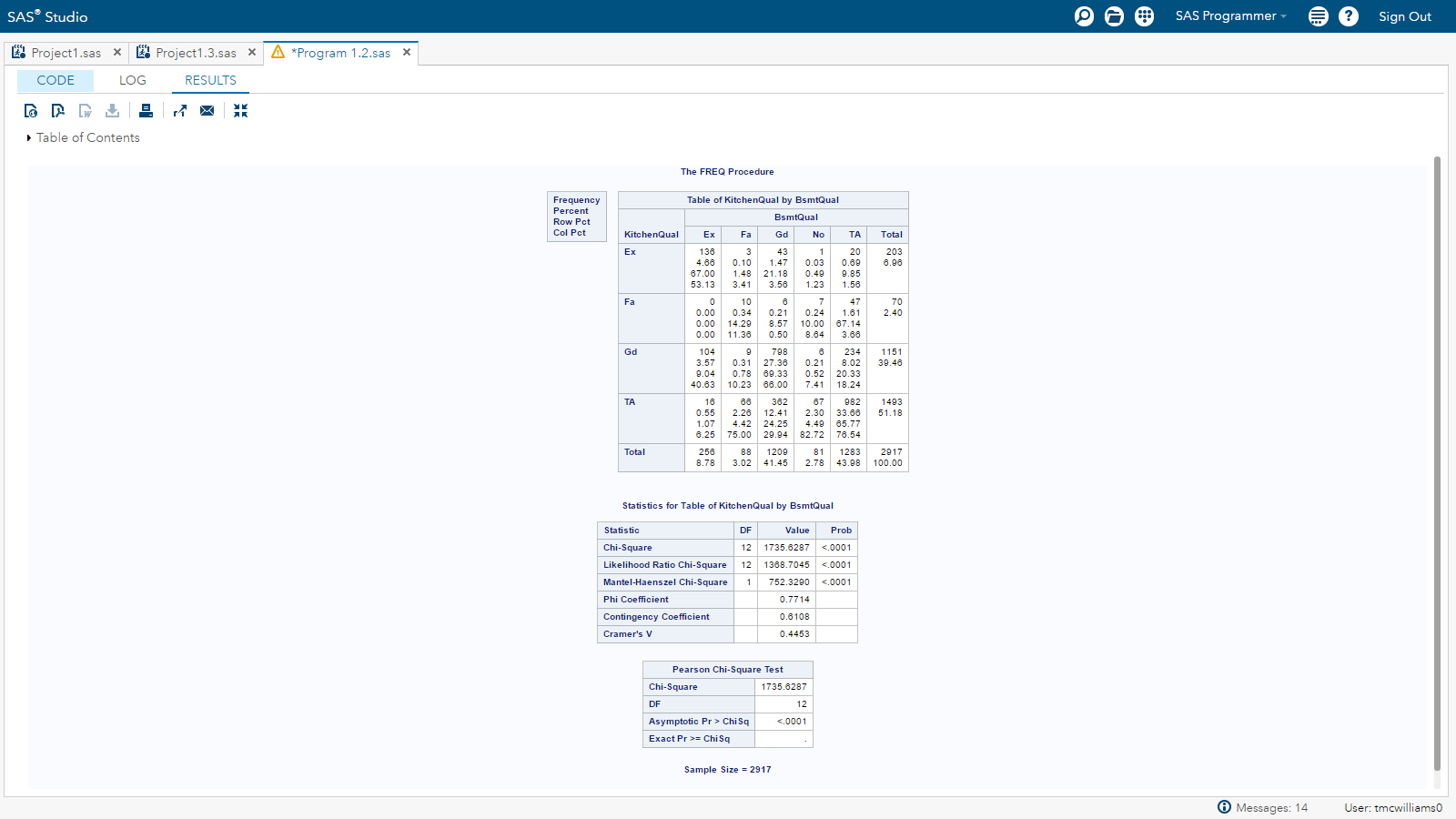


**Table 5: Contingency tables and the Chi-squared test on Neighborhood and** **KitchenQual**

**Table 6: Contingency tables and the Chi-squared test on KitchenQual and MSZoning**



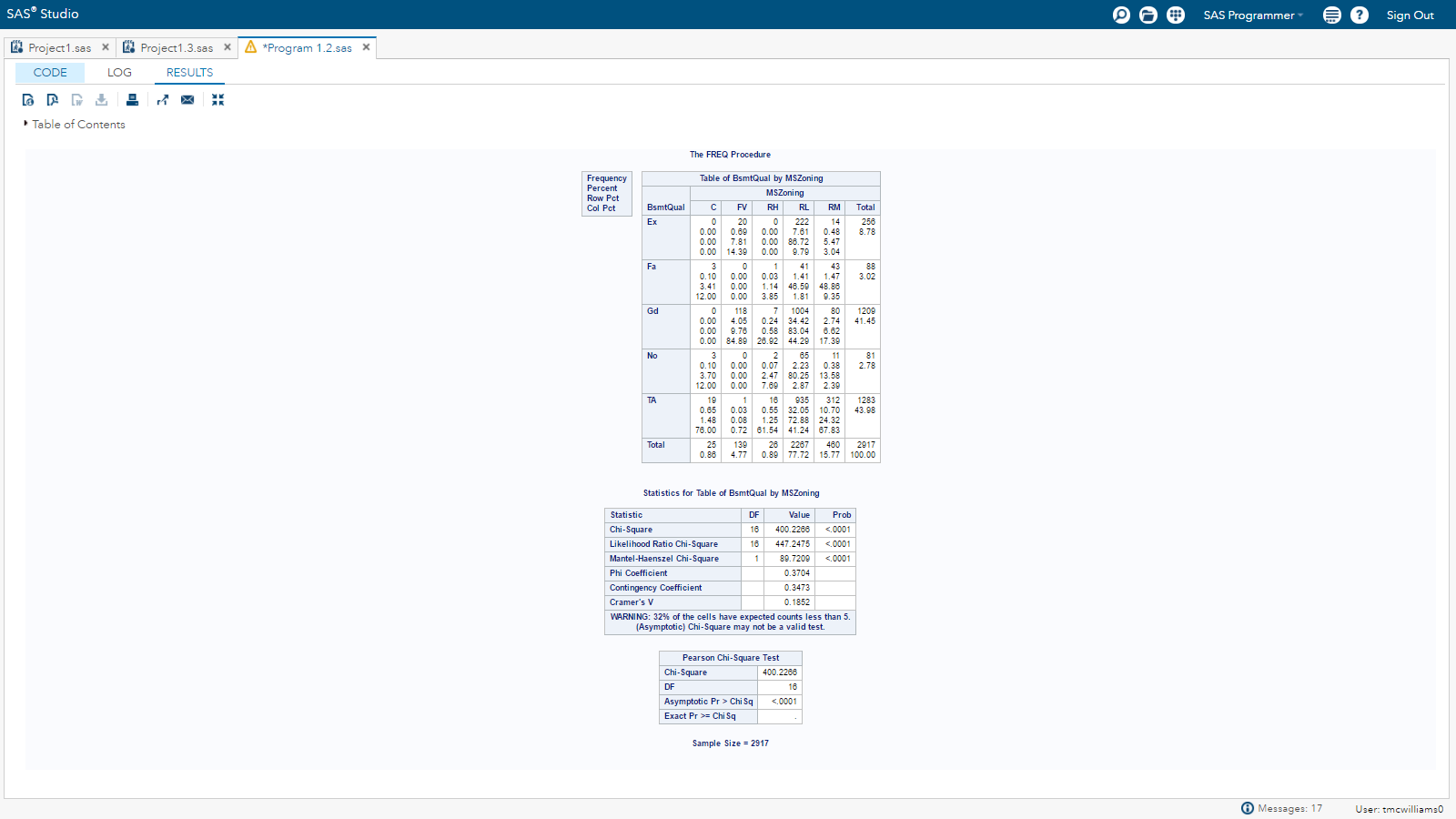
**Table 7: Contingency tables and the Chi-squared test on KitchenQual and MBsmtQual**

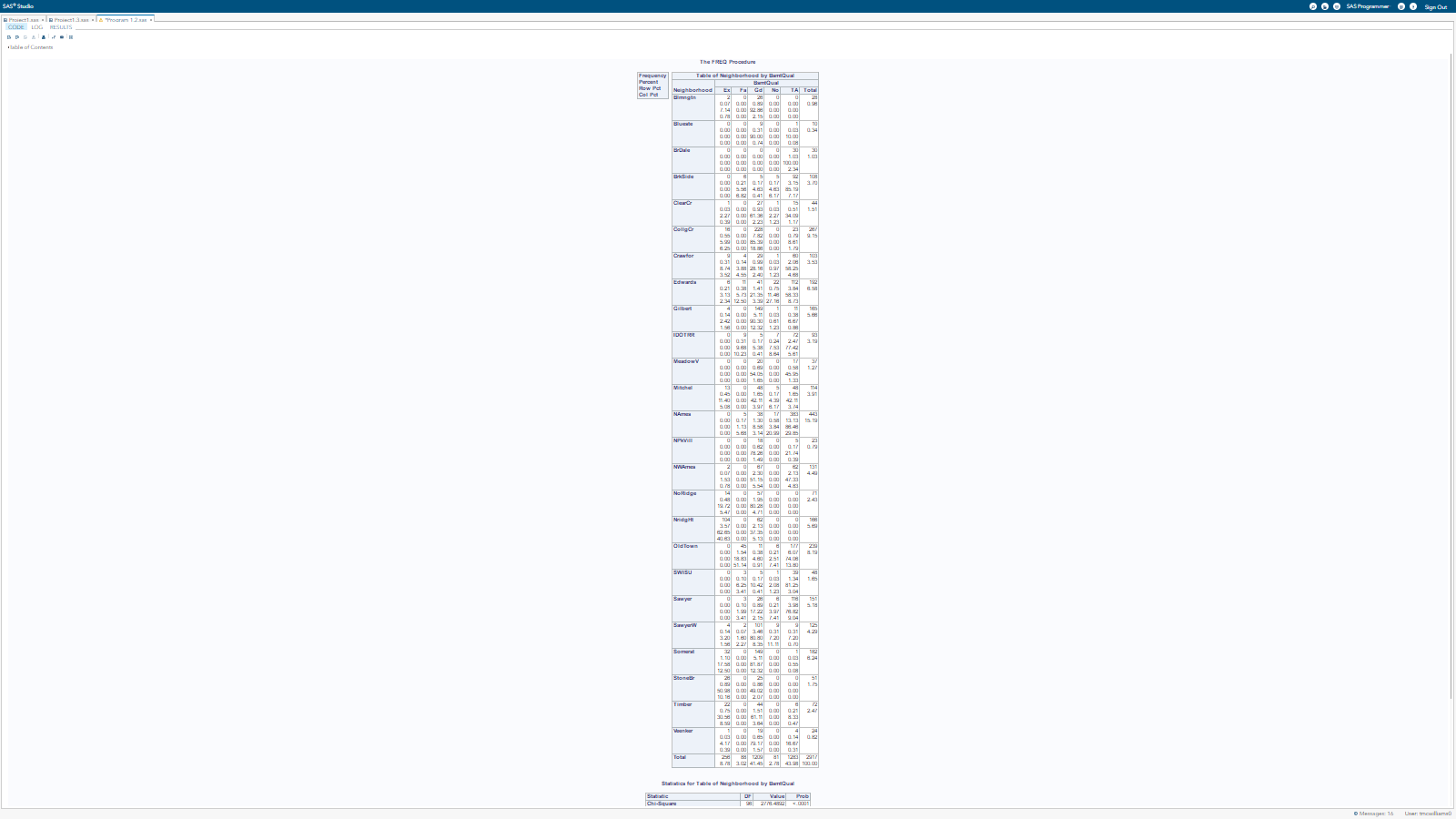


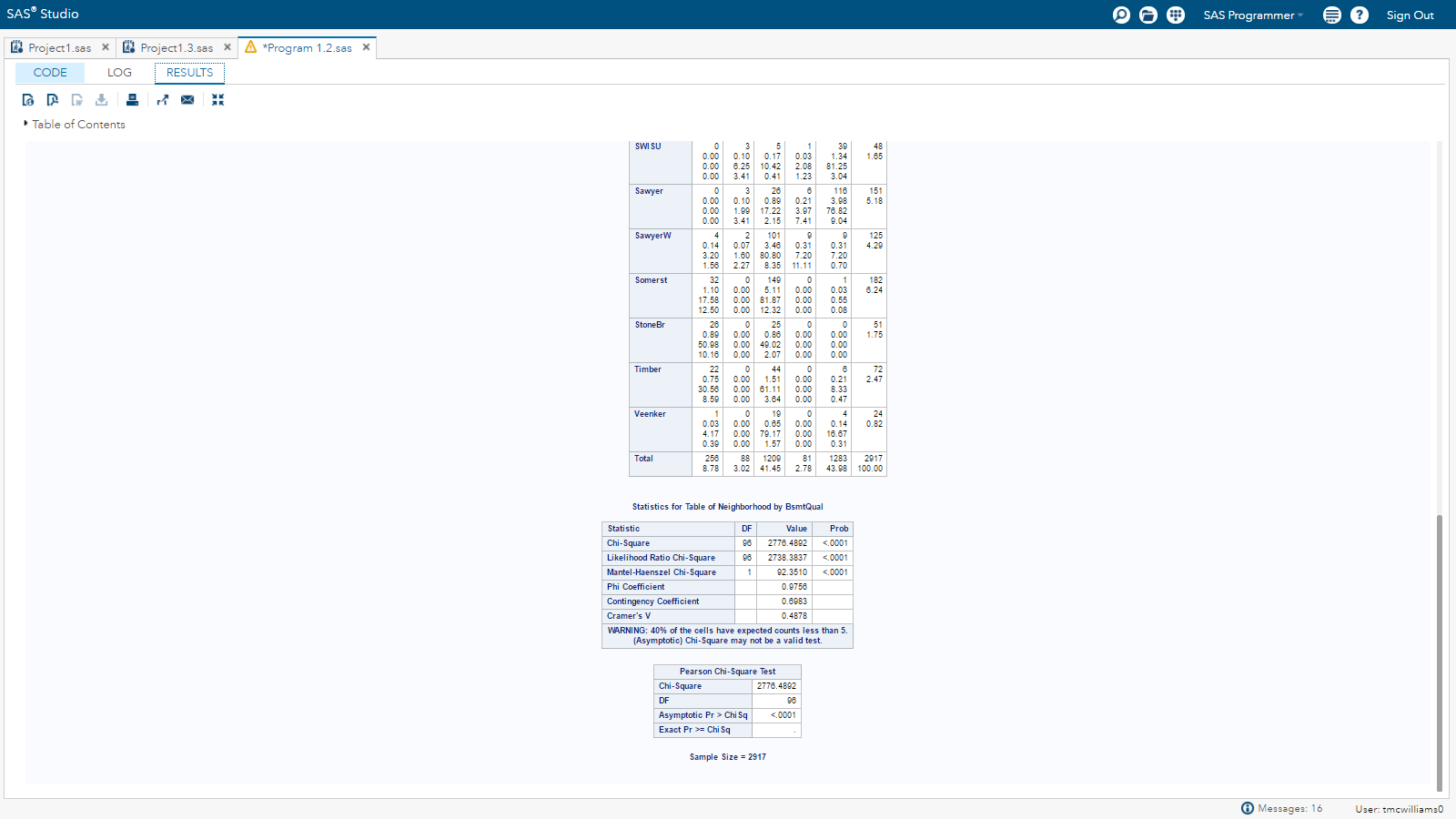
**Table 8: Contingency tables and the Chi-squared test on KitchenQual and CenteralAir**



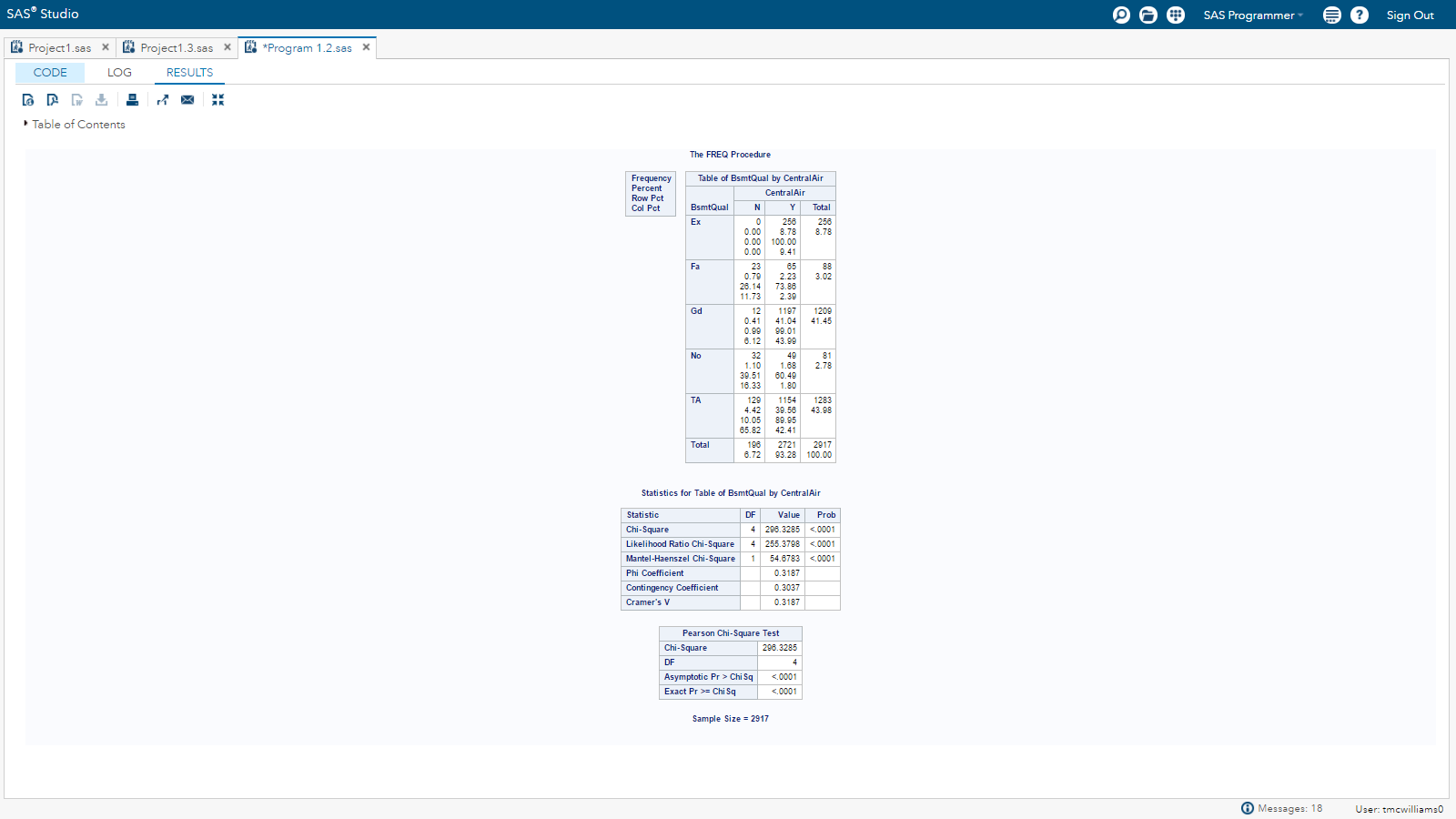
**Table 9: Contingency tables and the Chi-squared test on BsmtQual and MSZoning**



**Table 10: Contingency tables and the Chi-squared test on Neighborhood and BsmtQual**



**Table 11: Contingency tables and the Chi-squared test on BsmtQual and CentralAir**



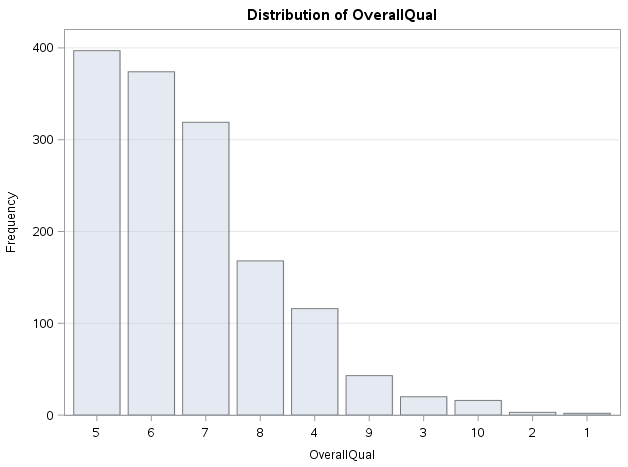
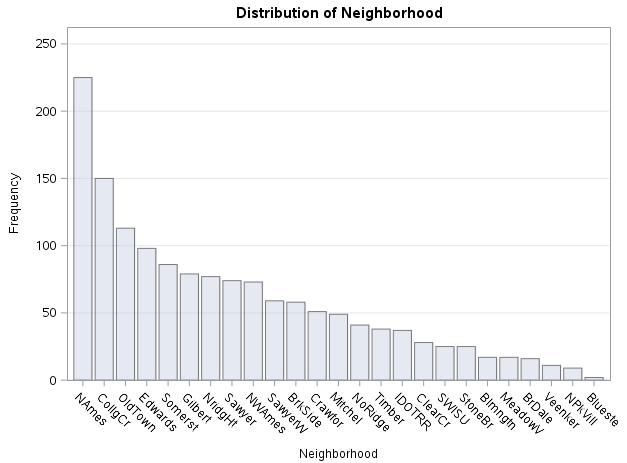
Finally, we want to assess the correlation between the categorical vs numeric variables. For this a one-way ANOVA should be performed on one categorical variable vs one numeric variable and the F-tests will be examined to help assess if there is correlation between the two. For example, a one-way ANOVA will be performed on Neighborhood(categorical) vs LogLotArea(numeric). If the assumptions of the one-way ANOVA are violated then a nonparametric alternative test, Kruskal-Wallis, should be performed. However, for this report we are not too concerned with the correlation between the categorical and numeric variables. One should keep in mind that there is a practical association versus a statistical association. Since this is a big data set the test may be extremely sensitive and say there are significant changes when there isn’t any.

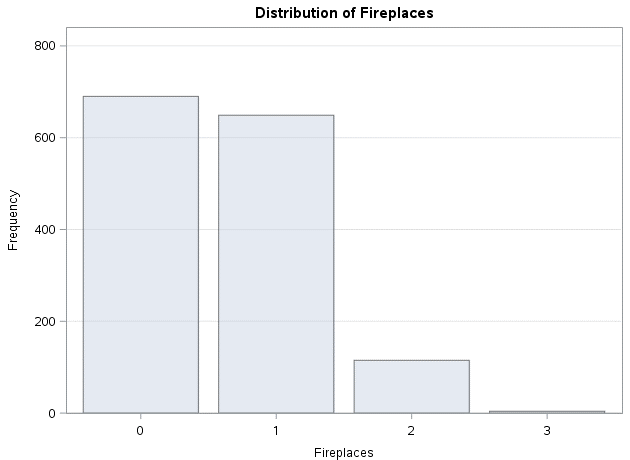
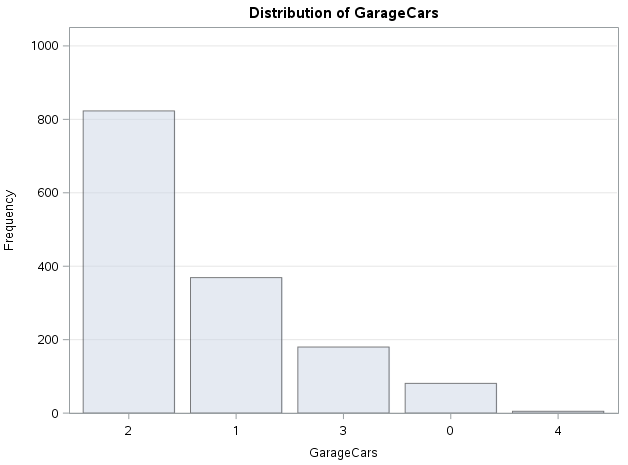
***For SAS code in this appendix, see*** [*Appendix XI*](#Appendix_XISAS) ***in this report and search for “APPENDIX on Correlation Analysis”.***

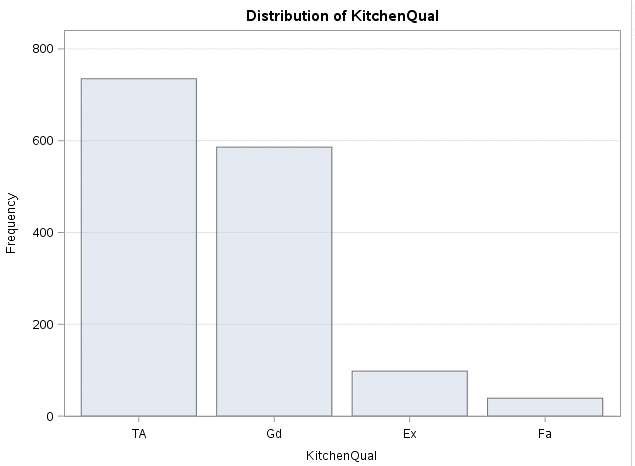
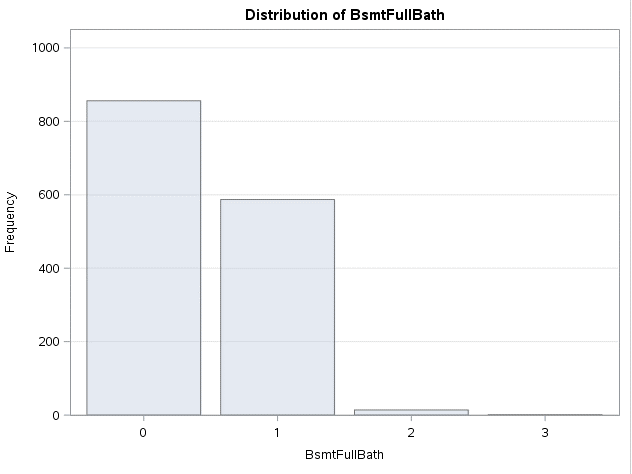
**APPENDIX IX: Distribution of Some Categorical Variables**

In **Figure 1**, we select a subset of categorical variables to display. We show the distribution of the number of houses sold for each Neighborhood showing Neighborhood Names with the large number of houses and Blueste with the least. We also show the distribution of the overall quality (OverallQual) of the houses (most are ranked 5); and the number of cars on can fit in the garage (2 cars seems to be most frequent). In addition, we show houses have 1 fireplace in general and that houses with basement full baths generally have 1. Finally, we provide the distribution of the kitchen quality.

**Figure 1: Distribution or Frequency of Sub-categories for Different Categorical Variables.**







***For SAS code in this appendix, see*** [*Appendix XI*](#Appendix_XISAS) ***in this report and search for “APPENDIX on Exploratory Data Analysis”.***

**APPENDIX X: Interpretation of Some Variables in Model**

We cannot interpret all the variables but below we interpret a few from the final regression equation, ***assuming all other variables remain unchanged when we make the change to increase or decrease the explanatory variable of interest***. We note before we begin that the in our regression equation we are takin the natural log of the SalePrice response variable.

**LogSalePrice** = …0.627932446 \* LogGrLivArea …..+ 0.010110255\* LogBsmtFinSF1 +

0.042101862 \* GarageCars....- 0.159524676 \* MnAdjLogGrLivArea \* Neighborhood Timber

1. **LogBsmtFinSF1**: Natural Log of type 1 finished square feet
   * SalePrice estimate will change by If x gets increased by 50%, then the SalesPrice estimate will increase by = 1.004108 or 0.41%.
2. **LogGrLivArea**: The natural log of above grade (ground) living area square feet

**NeighborhoodTimber:** Physical locations within Ames city limits

* + **LogGrLivArea** ; The variable can be interpreted in the following way If x gets increased by 20%, then the SalesPrice estimate will increase by = 1.121297.
  + **MnAdjLogGrLivArea \* Neighborhood Timber**: However, we also have an interaction term so we have to adjust for that assuming the neighborhood is Timber for the house location. For the interaction term, we take the natural log of GrLivArea (LogGrLivArea) normalized by the mean of the LogGrLivArea series to reduce the VIF. We are stubracting off a constant, and increasing GrLivArea by 20%. Therefore, we calculate the following:
    - [(ln(1.2\*GrLivArea)-Mean(LnGrLivArea)) - (ln(GrLivArea)-Mean(LnGrLivArea))]
    - = ln(1.2\*GrLivArea) – ln(GrLivArea)
    - =ln(1.2) + ln(GrLivArea) – ln(GrLivArea) = ln(1.2)
    - Therefore, = =
  + Overall increase is 8.915%, which is the multiplication of the two above.
    - = 1.08915 or an overall increase of 8.915%.
  + We are assuming all other variables are unchanged and are ignoring other terms that may have the Timber Neighborhood as part of them. Here are only looking at how SalePrice is changed byGrLivArea when the neighborhood is Timber.

1. **GarageCars:** Size of garage in car capacity (scale is 0 to 4)
   * SalePrice estimate will change by For a 1 car increase in the garage, the SalePrice estimate will increase by or increase by 4.3%. We are assuming all other variables are unchanged.

**APPENDIX XI: ALL SAS CODE**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\* Project #1: Analysis

\*\* Create by: James Hosker

\*\* Timothy McWilliams

\*\* Leticia Valadez

\*\* SMU Course: MSDS6372-4023

\*\* Date Created: 19-Apr-2017

\*\* Last Update: 27-May-2017

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\* Import the test and train data sets

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Generated Code (IMPORT) \*/

/\* Source File: test.csv \*/

/\* Source Path: /home/jhosker0/sasuser.v94/MSDS6372/Project1 \*/

/\* Code generated on: 4/15/17, 10:58 PM \*/

%web\_drop\_table(WORK.IMPORT);

FILENAME REFFILE '/home/jhosker0/sasuser.v94/MSDS6372/Project1/test.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.IMPORT;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

%web\_open\_table(WORK.IMPORT);

/\* Generated Code (IMPORT) \*/

/\* Source File: train.csv \*/

/\* Source Path: /home/jhosker0/sasuser.v94/MSDS6372/Project1 \*/

/\* Code generated on: 4/15/17, 10:59 PM \*/

%web\_drop\_table(WORK.IMPORT2);

FILENAME REFFILE '/home/jhosker0/sasuser.v94/MSDS6372/Project1/train.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.IMPORT2;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT2; RUN;

%web\_open\_table(WORK.IMPORT2);

data test;

set WORK.IMPORT;

run;

data train;

set WORK.IMPORT2;

run;

data test;

set test;

SalePrice = .;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Create Filtered Regression Set: TrainReg \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

data trainreg;

set WORK.IMPORT2;

/\* LotFrontage: there are 259 NAs which is ~18% of train.csv (1460)

and 227 NAs which is ~16% of test.csv (1459)

so repalced NAs in train set with mean \*/

if LotFrontage="NA" then LotFrontage=70.04996; /\* mean value of lotfrontage train data \*/

/\* Drop Alley: there are 1369 NAs in train.csv out of 1460

and 1352 NAs in test.csv out of 1459, so we remove it \*/

Drop Alley;

/\* Drop Utilities only one entries is different in train.csv and

no entries that are different in test.csv but has 2 NAs\*/

Drop Utilities;

/\* Only 4 NA in test.csv MSZoning, RL is most frequent \*/

if MSZoning="NA" then MSZoning="RL";

/\* Only 1 NA in test.csv Exteior1st Edwards neighborhood

which mostly had "Wd Sdng" \*/

if Exterior1st="NA" then Exterior1st="Wd Sdng";

/\* Only 1 NA in test.csv Exteior2st Edwards neighborhood

which mostly had "Wd Sdng" \*/

if Exterior2nd="NA" then Exterior2nd= "Wd Shng";

/\* In test.csv and train.csv, MasVnrType replace NA with None \*/

if MasVnrType="NA" then MasVnrType="None";

/\* In test.csv and train.csv, MasVnrArea Replace NA

with 0, since all MasVnrType None have 0 Area \*/

if MasVnrArea="NA" then MasVnrArea=0;

/\* In 44 test.csv and 37 train.csv, BsmtQual Replace NA with None \*/

if BsmtQual="NA" then BsmtQual="None";

/\* In 45 test.csv and 37 train.csv, BsmtCond Replace NA with None \*/

if BsmtCond="NA" then BsmtCond="None";

/\* In 44 test.csv and 38 train.csv, BsmtExposure Replace NA with None \*/

if BsmtExposure="NA" then BsmtExposure="None";

/\* In 42 test.csv and 37 train.csv, NA in test for BsmtFinType1 set NA to None\*/

if BsmtFinType1="NA" then BsmtFinType1="None";

/\* In test.csv, 1 NA in test for BsmtFinSF1 set NA to 0,

all NAs in BsmtFinType1 in train.csv have 0 for BsmtFinSF1 \*/

if BsmtFinSF1="NA" then BsmtFinSF1=0;

/\* In 42 test.csv and 38 train.csv, NA in test for BsmtFinType1 set NA to None\*/

if BsmtFinType2="NA" then BsmtFinType2="None";

/\* In test.csv, 1 NA in test for BsmtFinSF2 set NA to 0,

all NAs in BsmtFinType2 in train.csv have 0 for BsmtFinSF2 \*/

if BsmtFinSF2="NA" then BsmtFinSF2=0;

/\* In test.csv, 1 NA in test for BsmtUnfSF set NA to 0 \*/

if BsmtUnfSF="NA" then BsmtUnfSF=0;

/\* In train.csv, 1 NA in test for Electrical set NA

to SBrkr the most frequent response \*/

if Electrical="NA" then Electrical="SBrkr";

/\* In test.csv, 1 NA in test for TotalBsmtSF set NA to 0 \*/

if TotalBsmtSF="NA" then TotalBsmtSF=0;

/\* In test.csv, 2 NA in test for BsmtFullBath set NA to 0 \*/

if BsmtFullBath="NA" then BsmtFullBath=0;

/\* In test.csv, 2 NA in test for BsmtHalfBath set NA to 0 \*/

if BsmtHalfBath="NA" then BsmtHalfBath=0;

/\* In test.csv, 1 NA for KitchenQual set NA to TA, most frequent \*/

if KitchenQual="NA" then KitchenQual="TA";

/\* In test.csv, 2 NA in test for Functional set NA to Typ, most frequent \*/

if Functional="NA" then Functional="Typ";

/\* In 730 test.csv and 690 train.csv, FireplaceQu NA make category None \*/

if FireplaceQu="NA" then FireplaceQu="None";

/\* In 76 test.csv and 81 train.csv, GarageType NA make category None \*/

if GarageType="NA" then GarageType="None";

/\* In 78 test.csv and 81 train.csv, GarageYrBlt NA equal to mean 1978 in train.csv \*/

if GarageYrBlt="NA" then GarageYrBlt=1978; /\* mean year of dataset in train.csv \*/

/\* In 78 test.csv and 81 train.csv, GarageFinish NA equal to None \*/

if GarageFinish="NA" then GarageFinish="None";

/\* In test.csv, GarageCars 1 NA set equal to 0 \*/

if GarageCars="NA" then GarageCars=0;

/\* In test.csv, GarageArea 1 NA set equal to 0 \*/

if GarageArea="NA" then GarageArea=0;

/\* In 78 test.csv and 81 train.csv, GarageQual NA set equal to None \*/

if GarageQual="NA" then GarageQual="None";

/\* In 78 test.csv and 81 train.csv, GarageCond NA set equal to None \*/

if GarageCond="NA" then GarageCond="None";

/\* Drop PoolQC: there are 1453 NAs in train.csv out of 1460

and 1456 NAs in test.csv out of 1459, so we remove it \*/

Drop PoolQC;

/\* Drop Fence: there are 1179 NAs in train.csv out of 1460

and 1169 NAs in test.csv out of 1459, so we remove it \*/

Drop Fence;

/\* Drop MiscFeature: there are 1406 NAs in train.csv out of 1460

and 1408 NAs in test.csv out of 1459, so we remove it \*/

Drop MiscFeature;

/\* In test.csv, 1 NA for SaleType set NA to WD, most frequent \*/

if SaleType="NA" then SaleType="WD";

/\* Remove Data Points 524 and 1299 \*/

if id = 524 then delete;

if id = 1299 then delete;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Create Filtered Regression/Prediction \*/

/\* Train2 Set \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

data train2;

set train test;

/\* LotFrontage: there are 259 NAs which is ~18% of train.csv (1460)

and 227 NAs which is ~16% of test.csv (1459)

so repalced NAs in train set with mean \*/

if LotFrontage="NA" then LotFrontage=70.04996; /\* mean value of lotfrontage train data \*/

/\* Drop Alley: there are 1369 NAs in train.csv out of 1460

and 1352 NAs in test.csv out of 1459, so we remove it \*/

Drop Alley;

/\* Drop Utilities only one entries is different in train.csv and

no entries that are different in test.csv but has 2 NAs\*/

Drop Utilities;

/\* Only 4 NA in test.csv MSZoning, RL is most frequent \*/

if MSZoning="NA" then MSZoning="RL";

/\* Only 1 NA in test.csv Exteior1st Edwards neighborhood

which mostly had "Wd Sdng" \*/

if Exterior1st="NA" then Exterior1st="Wd Sdng";

/\* Only 1 NA in test.csv Exteior2st Edwards neighborhood

which mostly had "Wd Sdng" \*/

if Exterior2nd="NA" then Exterior2nd= "Wd Shng";

/\* In test.csv and train.csv, MasVnrType replace NA with None \*/

if MasVnrType="NA" then MasVnrType="None";

/\* In test.csv and train.csv, MasVnrArea Replace NA

with 0, since all MasVnrType None have 0 Area \*/

if MasVnrArea="NA" then MasVnrArea=0;

/\* In 44 test.csv and 37 train.csv, BsmtQual Replace NA with None \*/

if BsmtQual="NA" then BsmtQual="None";

/\* In 45 test.csv and 37 train.csv, BsmtCond Replace NA with None \*/

if BsmtCond="NA" then BsmtCond="None";

/\* In 44 test.csv and 38 train.csv, BsmtExposure Replace NA with None \*/

if BsmtExposure="NA" then BsmtExposure="None";

/\* In 42 test.csv and 37 train.csv, NA in test for BsmtFinType1 set NA to None\*/

if BsmtFinType1="NA" then BsmtFinType1="None";

/\* In test.csv, 1 NA in test for BsmtFinSF1 set NA to 0,

all NAs in BsmtFinType1 in train.csv have 0 for BsmtFinSF1 \*/

if BsmtFinSF1="NA" then BsmtFinSF1=0;

/\* In 42 test.csv and 38 train.csv, NA in test for BsmtFinType1 set NA to None\*/

if BsmtFinType2="NA" then BsmtFinType2="None";

/\* In test.csv, 1 NA in test for BsmtFinSF2 set NA to 0,

all NAs in BsmtFinType2 in train.csv have 0 for BsmtFinSF2 \*/

if BsmtFinSF2="NA" then BsmtFinSF2=0;

/\* In test.csv, 1 NA in test for BsmtUnfSF set NA to 0 \*/

if BsmtUnfSF="NA" then BsmtUnfSF=0;

/\* In train.csv, 1 NA in test for Electrical set NA

to SBrkr the most frequent response \*/

if Electrical="NA" then Electrical="SBrkr";

/\* In test.csv, 1 NA in test for TotalBsmtSF set NA to 0 \*/

if TotalBsmtSF="NA" then TotalBsmtSF=0;

/\* In test.csv, 2 NA in test for BsmtFullBath set NA to 0 \*/

if BsmtFullBath="NA" then BsmtFullBath=0;

/\* In test.csv, 2 NA in test for BsmtHalfBath set NA to 0 \*/

if BsmtHalfBath="NA" then BsmtHalfBath=0;

/\* In test.csv, 1 NA for KitchenQual set NA to TA, most frequent \*/

if KitchenQual="NA" then KitchenQual="TA";

/\* In test.csv, 2 NA in test for Functional set NA to Typ, most frequent \*/

if Functional="NA" then Functional="Typ";

/\* In 730 test.csv and 690 train.csv, FireplaceQu NA make category None \*/

if FireplaceQu="NA" then FireplaceQu="None";

/\* In 76 test.csv and 81 train.csv, GarageType NA make category None \*/

if GarageType="NA" then GarageType="None";

/\* In 78 test.csv and 81 train.csv, GarageYrBlt NA equal to mean 1978 in train.csv \*/

if GarageYrBlt="NA" then GarageYrBlt=1978; /\* mean year of dataset in train.csv \*/

/\* In 78 test.csv and 81 train.csv, GarageFinish NA equal to None \*/

if GarageFinish="NA" then GarageFinish="None";

/\* In test.csv, GarageCars 1 NA set equal to 0 \*/

if GarageCars="NA" then GarageCars=0;

/\* In test.csv, GarageArea 1 NA set equal to 0 \*/

if GarageArea="NA" then GarageArea=0;

/\* In 78 test.csv and 81 train.csv, GarageQual NA set equal to None \*/

if GarageQual="NA" then GarageQual="None";

/\* In 78 test.csv and 81 train.csv, GarageCond NA set equal to None \*/

if GarageCond="NA" then GarageCond="None";

/\* Drop PoolQC: there are 1453 NAs in train.csv out of 1460

and 1456 NAs in test.csv out of 1459, so we remove it \*/

Drop PoolQC;

/\* Drop Fence: there are 1179 NAs in train.csv out of 1460

and 1169 NAs in test.csv out of 1459, so we remove it \*/

Drop Fence;

/\* Drop MiscFeature: there are 1406 NAs in train.csv out of 1460

and 1408 NAs in test.csv out of 1459, so we remove it \*/

Drop MiscFeature;

/\* In test.csv, 1 NA for SaleType set NA to WD, most frequent \*/

if SaleType="NA" then SaleType="WD";

/\* Remove Data Points 524 and 1299 see Appendix in Report \*/

if id = 524 then delete;

if id = 1299 then delete;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* 1. Fix Issues with LotFrontage Data characters instead of number \*/

/\* 2. Add LogSalPrice and LogGrLivArea to both trainreg and train2 \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

data train2;

set train2 (rename=(LotFrontage = LotFrontageChar)) ;

LotFrontage = input(LotFrontageChar,best.);

LogSalePrice = log(SalePrice);

LogGrLivArea = log(GrLivArea);

LogTotalBsmtSF = log(TotalBsmtSF+1);

LogLotArea = log(LotArea+1);

LogBsmtFinSF1 = log(BsmtFinSF1+1);

Log\_1stFlrSF = log(\_1stFlrSF+1);

Log\_2ndFlrSF = log(\_2ndFlrSF+1);

LogGarageArea = log(GarageArea + 1);

LogBsmtFinSF2 = log(BsmtFinSF2 + 1);

LogBsmtUnfSF = log(BsmtUnfSF + 1);

LogLotFrontage = log(LotFrontage+1);

LogWoodDeckSF = log(WoodDeckSF+1);

LogOpenPorchSF = log(OpenPorchSF+1);

LogMasVnrArea = log(MasVnrArea+1);

LogEnclosedPorch = log(EnclosedPorch+1);

Log\_3SsnPorch = log(\_3SsnPorch+1);

LogScreenPorch = log(ScreenPorch+1);

LogPoolArea = log(PoolArea+1);

Drop LotFrontageChar;

run;

data trainreg;

set trainreg (rename=(LotFrontage = LotFrontageChar)) ;

LotFrontage = input(LotFrontageChar,best.);

LogSalePrice = log(SalePrice);

LogGrLivArea = log(GrLivArea);

LogTotalBsmtSF = log(TotalBsmtSF+1);

LogLotArea = log(LotArea+1);

LogBsmtFinSF1 = log(BsmtFinSF1+1);

Log\_1stFlrSF = log(\_1stFlrSF+1);

Log\_2ndFlrSF = log(\_2ndFlrSF+1);

LogGarageArea = log(GarageArea + 1);

LogBsmtFinSF2 = log(BsmtFinSF2 + 1);

LogBsmtUnfSF = log(BsmtUnfSF + 1);

LogLotFrontage = log(LotFrontage+1);

LogWoodDeckSF = log(WoodDeckSF+1);

LogOpenPorchSF = log(OpenPorchSF+1);

LogMasVnrArea = log(MasVnrArea+1);

LogEnclosedPorch = log(EnclosedPorch+1);

Log\_3SsnPorch = log(\_3SsnPorch+1);

LogScreenPorch = log(ScreenPorch+1);

LogPoolArea = log(PoolArea+1);

Group = 1;

Drop LotFrontageChar;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* FINAL MODEL using PROC GLM Output \*

\* Output Regression \*

\* Get plots and graphs requested \*

\* unpack diagnostics with labels \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc means maxdec=4 N mean stddev min Q1 median Q3 max data=trainreg;

var LogSalePrice OverallQual OverAllCond LogGrLivArea GarageCars LogTotalBsmtSF LogGarageArea LogBsmtFinSF1

LogGrLivArea LogTotalBsmtSF LogBsmtFinSF1 LogLotArea Log\_1stFlrSF;

run;

data train3; set train2;

MnAdjOverallQual = OverallQual-6.094;

MnAdjOverallCond = OverallCond - 5.5761;

MnAdjLogGrLivArea = LogGrLivArea - 7.266;

MnAdjGarageCars = GarageCars - 1.7661;

MnAdjLogTotalBsmtSF = LogTotalBsmtSF - 6.7483;

MnAdjLogGarageArea = LogGarageArea - 5.8065;

MnAdjLogBsmtFinSF1 = LogBsmtFinSF1 - 4.2243;

MnAdjLogLotArea = LogLotArea - 9.1086;

MnAdjLogBsmtFinSF1 = LogBsmtFinSF1 - 4.2243;

MnAdjLog\_1stFlrSF = Log\_1stFlrSF - 7.0067;

run;

ODS GRAPHICS ON / ATTRPRIORITY=NONE;

proc glm data=train3 plots(unpack)=diagnostic(label); /\* plot=all \*/

class Neighborhood MSZoning CentralAir KitchenQual BsmtQual Foundation;

model LogSalePrice = LogGrLivArea LogLotArea LogBsmtFinSF1 LogGarageArea

Log\_1stFlrSF Neighborhood OverallQual OverallCond GarageCars Fireplaces YearBuilt

YearRemodAdd MSZoning CentralAir KitchenQual BsmtFullBath BsmtQual

MSZoning\*CentralAir MnAdjOverallQual\*MnAdjLogGrLivArea MnAdjLogGrLivArea\*Neighborhood

/ CLI CLPARM CLM SOLUTION TOLERANCE CLM;

output out=CustData p=predict PRESS=CVPress;

run;quit;

ODS GRAPHICS OFF;

/\* Filter out and calculate Prediction \*/

data resultsCust;

set CustData;

if Predict > 0 then LogSalePrice = Predict;

if Predict < 0 then LogSalePrice = log(180932.92); /\* mean of training data \*/

if Predict='.' then LogSalePrice = log(180932.92); /\* mean of training data \*/

SalePrice2 = SalePrice;

SalePrice = exp(LogSalePrice);

keep id SalePrice;

where id > 1460;

run;

proc print data = resultsCust;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* APPENDIX on Data Cleaning (Additonal Code) \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc freq data=WORK.IMPORT2;

table Utilities PoolQC Fence MiscFeature;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* APPENDIX on Data Removal (Additonal Code) \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Please note to run this analysis you may need comment

out the section of SAS code that removed data points

1299 and 524.

/\* Breakout of all Quantitative variables by

Neighborhood and OverallQual \*/

proc means maxdec=2 N mean stddev min Q1 median Q3 max data=trainreg;

class Neighborhood;

var SalePrice OverallQual OverAllCond GrLivArea GarageCars GarageArea

BsmtFinSF1 GrLivArea TotalBsmtSF BsmtFinSF1 LotArea \_1stFlrSF;

run;

proc means maxdec=2 N mean stddev min Q1 median Q3 max data=trainreg;

class OverallQual;

var SalePrice OverallQual OverAllCond GrLivArea GarageCars GarageArea

BsmtFinSF1 GrLivArea TotalBsmtSF BsmtFinSF1 LotArea \_1stFlrSF;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* APPENDIX Exploratory Data Analysis \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Exploratory Data Analysis \*/

proc means maxdec=4 N mean stddev min Q1 median Q3 max data=trainreg;

var LogSalePrice LogGrLivArea LogGarageArea LogBsmtFinSF1

LogLotArea Log\_1stFlrSF

LogLotFrontage LogMasVnrArea LogBsmtFinSF2

LogBsmtUnfSF LogTotalBsmtSF Log\_2ndFlrSF LowQualFinSF

LogWoodDeckSF LogOpenPorchSF LogEnclosedPorch Log\_3SsnPorch

LogScreenPorch LogPoolArea ;

run;

data sortdat; set trainreg;

run;

proc sort data = sortdat;

by LogSalePrice;

run;

proc univariate data = trainreg;

var LogSaleprice SalePrice;

histogram LogSaleprice SalePrice;

run;

/\* Scatterplot matrix \*/

proc sgscatter data=trainreg;

matrix SalePrice GrLivArea LotArea BsmtFinSF1 GarageArea / diagonal=(kernel histogram);

run;

proc sgscatter data=trainreg;

matrix LogSalePrice LogGrLivArea LogLotArea LogBsmtFinSF1 LogGarageArea / diagonal=(kernel histogram);

run;

/\* Frequency Plot \*/

proc freq data=trainreg order=freq;

tables Neighborhood OverallQual OverallCond GarageCars Fireplaces

YearBuilt YearRemodAdd BsmtFullBath BsmtQual

MSZoning CentralAir KitchenQual BsmtQual Neighborhood

/plots=freqplot;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* APPENDIX on Data Transformation to Natural \*

\* Log (Additonal Code) \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\* Linear regression of orginal response \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\* and some explanatory variables \*\*\*\*/

proc glm data=trainreg plots(unpack)=diagnostic(label); /\* plot=all \*/

class MSZoning Street LotShape LandContour LotConfig LandSlope Neighborhood

Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu

GarageType GarageFinish GarageQual GarageCond PavedDrive

SaleType SaleCondition;

model SalePrice = GrLivArea Neighborhood MSSubClass MSZoning LotFrontage LotArea Street LotShape LandContour

LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle OverallQual OverallCond YearBuilt

YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation

BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF

TotalBsmtSF Heating HeatingQC CentralAir Electrical \_1stFlrSF \_2ndFlrSF LowQualFinSF

BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd

Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea

GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch

PoolArea MiscVal MoSold YrSold SaleType SaleCondition / CLI CLPARM CLM SOLUTION TOLERANCE CLM;

output out=CustData p=predict PRESS=CVPress;

run;quit;

/\*\*\*\*\*\*\*\*\*\*\*\*\* Linear regression of transformed response \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\* and some explanatory variables by the natural log \*\*\*\*/

proc glm data=trainreg plots(unpack)=diagnostic(label); /\* plot=all \*/

class MSZoning Street LotShape LandContour LotConfig LandSlope Neighborhood

Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu

GarageType GarageFinish GarageQual GarageCond PavedDrive

SaleType SaleCondition;

model LogSalePrice = LogGrLivArea Neighborhood MSSubClass MSZoning LogLotFrontage LogLotArea

Street LotShape LandContour LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle

OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType LogMasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure

BsmtFinType1 LogBsmtFinSF1 BsmtFinType2 LogBsmtFinSF2 LogBsmtUnfSF LogTotalBsmtSF Heating

HeatingQC CentralAir Electrical Log\_1stFlrSF Log\_2ndFlrSF LowQualFinSF BsmtFullBath BsmtHalfBath

FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd

Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars LogGarageArea

GarageQual GarageCond PavedDrive LogWoodDeckSF LogOpenPorchSF LogEnclosedPorch Log\_3SsnPorch

LogScreenPorch LogPoolArea MiscVal MoSold YrSold SaleType SaleCondition / CLI CLPARM CLM SOLUTION TOLERANCE CLM;

output out=CustData p=predict PRESS=CVPress;

run;quit;

ODS GRAPHICS OFF;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* APPENDIX for Model \*/

/\* Variable-Selection Techniques \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* 1. FORWARD SELECTION \*/

/\* Train2 Data vs. LogSalePrice \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

ODS GRAPHICS ON / ATTRPRIORITY=NONE;

proc glmselect data = train2 plots = all;

class MSZoning Street LotShape LandContour LotConfig LandSlope Neighborhood

Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu

GarageType GarageFinish GarageQual GarageCond PavedDrive

SaleType SaleCondition;

model LogSalePrice = LogGrLivArea Neighborhood MSSubClass MSZoning LogLotFrontage LogLotArea

Street LotShape LandContour LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle

OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType LogMasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure

BsmtFinType1 LogBsmtFinSF1 BsmtFinType2 LogBsmtFinSF2 LogBsmtUnfSF LogTotalBsmtSF Heating

HeatingQC CentralAir Electrical Log\_1stFlrSF Log\_2ndFlrSF LowQualFinSF BsmtFullBath BsmtHalfBath

FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd

Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars LogGarageArea

GarageQual GarageCond PavedDrive LogWoodDeckSF LogOpenPorchSF LogEnclosedPorch Log\_3SsnPorch

LogScreenPorch LogPoolArea MiscVal MoSold YrSold SaleType SaleCondition

/ selection=forward(select=AIC stop=CV) cvmethod=random(2) hierarchy=single showpvalues stat=all;

output out=FwdData p=predict r=resid;

run;quit;

ODS GRAPHICS OFF;

/\* Filter out and calculate Prediction \*/

data resultsFwd;

set FwdData;

if Predict > 0 then LogSalePrice = Predict;

if Predict < 0 then LogSalePrice = log(180932.92); /\* mean of training data \*/

SalePrice2 = SalePrice;

SalePrice = exp(LogSalePrice);

keep id SalePrice;

where id > 1460;

run;

proc print data = resultsFwd;

run;

/\* Calculated stats around SalePrice \*/

proc means maxdec=2 N mean stddev min Q1 median Q3 max data=resultsFwd;

var SalePrice;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* 2. BACKWARD SELECTION \*/

/\* Train2 Data vs. LogSalePrice \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

ODS GRAPHICS ON / ATTRPRIORITY=NONE;

proc glmselect data = train2 plots = all;

class MSZoning Street LotShape LandContour LotConfig LandSlope Neighborhood

Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu

GarageType GarageFinish GarageQual GarageCond PavedDrive

SaleType SaleCondition;

model LogSalePrice = LogGrLivArea Neighborhood MSSubClass MSZoning LogLotFrontage LogLotArea

Street LotShape LandContour LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle

OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType LogMasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure

BsmtFinType1 LogBsmtFinSF1 BsmtFinType2 LogBsmtFinSF2 LogBsmtUnfSF LogTotalBsmtSF Heating

HeatingQC CentralAir Electrical Log\_1stFlrSF Log\_2ndFlrSF LowQualFinSF BsmtFullBath BsmtHalfBath

FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd

Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars LogGarageArea

GarageQual GarageCond PavedDrive LogWoodDeckSF LogOpenPorchSF LogEnclosedPorch Log\_3SsnPorch

LogScreenPorch LogPoolArea MiscVal MoSold YrSold SaleType SaleCondition

/ selection=backward(select=AIC stop=CV) cvmethod=random(2) hierarchy=single showpvalues stat=all;

output out=BckData p=predict r=resid;

run;quit;

ODS GRAPHICS OFF;

/\* Filter out and calculate Prediction \*/

data resultsBck;

set BckData;

if Predict > 0 then LogSalePrice = Predict;

if Predict < 0 then LogSalePrice = log(180932.92); /\* mean of training data \*/

SalePrice2 = SalePrice;

SalePrice = exp(LogSalePrice);

keep id SalePrice;

where id > 1460;

run;

proc print data = resultsFwd;

run;

/\* Calculated stats around SalePrice \*/

proc means maxdec=2 N mean stddev min Q1 median Q3 max data=resultsBck;

var SalePrice;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* 3. STEPWISE SELECTION \*/

/\* Train2 Data vs. LogSalePrice \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

ODS GRAPHICS ON / ATTRPRIORITY=NONE;

proc glmselect data = train2 plots = all;

class MSZoning Street LotShape LandContour LotConfig LandSlope Neighborhood

Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu

GarageType GarageFinish GarageQual GarageCond PavedDrive

SaleType SaleCondition;

model LogSalePrice = LogGrLivArea Neighborhood MSSubClass MSZoning LogLotFrontage LogLotArea

Street LotShape LandContour LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle

OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType LogMasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure

BsmtFinType1 LogBsmtFinSF1 BsmtFinType2 LogBsmtFinSF2 LogBsmtUnfSF LogTotalBsmtSF Heating

HeatingQC CentralAir Electrical Log\_1stFlrSF Log\_2ndFlrSF LowQualFinSF BsmtFullBath BsmtHalfBath

FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd

Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars LogGarageArea

GarageQual GarageCond PavedDrive LogWoodDeckSF LogOpenPorchSF LogEnclosedPorch Log\_3SsnPorch

LogScreenPorch LogPoolArea MiscVal MoSold YrSold SaleType SaleCondition

/ selection=stepwise(select=AIC stop=cv) cvmethod=random(2) hierarchy=single showpvalues stat=all;

output out=StepData p=predict r=resid;

run;quit;

ODS GRAPHICS OFF;

/\* Filter out and calculate Prediction \*/

data resultsStep;

set StepData;

if Predict > 0 then LogSalePrice = Predict;

if Predict < 0 then LogSalePrice = log(180932.92); /\* mean of training data \*/

if Predict='.' then LogSalePrice = log(180932.92); /\* mean of training data \*/

SalePrice2 = SalePrice;

SalePrice = exp(LogSalePrice);

keep id SalePrice;

where id > 1460;

run;

proc print data = resultsStep;

run;

/\* Calculated stats around SalePrice \*/

proc means maxdec=2 N mean stddev min Q1 median Q3 max data=resultsStep;

var SalePrice;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* 4. LASSO SELECTION \*/

/\* Train2 Data vs. LogSalePrice \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

ODS GRAPHICS ON / ATTRPRIORITY=NONE;

proc glmselect data = train2 plots = all;

class MSZoning Street LotShape LandContour LotConfig LandSlope Neighborhood

Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu

GarageType GarageFinish GarageQual GarageCond PavedDrive

SaleType SaleCondition;

model LogSalePrice = LogGrLivArea Neighborhood MSSubClass MSZoning LogLotFrontage LogLotArea

Street LotShape LandContour LotConfig LandSlope Condition1 Condition2 BldgType HouseStyle

OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType LogMasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure

BsmtFinType1 LogBsmtFinSF1 BsmtFinType2 LogBsmtFinSF2 LogBsmtUnfSF LogTotalBsmtSF Heating

HeatingQC CentralAir Electrical Log\_1stFlrSF Log\_2ndFlrSF LowQualFinSF BsmtFullBath BsmtHalfBath

FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd

Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars LogGarageArea

GarageQual GarageCond PavedDrive LogWoodDeckSF LogOpenPorchSF LogEnclosedPorch Log\_3SsnPorch

LogScreenPorch LogPoolArea MiscVal MoSold YrSold SaleType SaleCondition

/ selection=LASSO(choose=SBC stop=cv) cvmethod=random(2) hierarchy=single showpvalues stat=all;

output out=LassoData p=predict r=resid;

run;quit;

ODS GRAPHICS OFF;

/\* Filter out and calculate Prediction \*/

data resultsLasso;

set LassoData;

if Predict > 0 then LogSalePrice = Predict;

if Predict < 0 then LogSalePrice = log(180932.92); /\* mean of training data \*/

if Predict='.' then LogSalePrice = log(180932.92); /\* mean of training data \*/

SalePrice2 = SalePrice;

SalePrice = exp(LogSalePrice);

keep id SalePrice;

where id > 1460;

run;

proc print data = resultsLasso;

run;

/\* Calculated stats around SalePrice \*/

proc means maxdec=2 N mean stddev min Q1 median Q3 max data=resultsLasso;

var SalePrice;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* 5. LASSO interaction term example \*

\* regression \*

\* Get plots and graphs requested \*

\* unpack diagnostics with labels \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc means maxdec=4 N mean stddev min Q1 median Q3 max data=trainreg;

var LogSalePrice OverallQual OverAllCond LogGrLivArea GarageCars LogTotalBsmtSF LogGarageArea LogBsmtFinSF1

LogGrLivArea LogTotalBsmtSF LogBsmtFinSF1 LogLotArea Log\_1stFlrSF;

run;

data train3; set train2;

MnAdjOverallQual = OverallQual-6.094;

MnAdjOverallCond = OverallCond - 5.5761;

MnAdjLogGrLivArea = LogGrLivArea - 7.266;

MnAdjGarageCars = GarageCars - 1.7661;

MnAdjLogTotalBsmtSF = LogTotalBsmtSF - 6.7483;

MnAdjLogGarageArea = LogGarageArea - 5.8065;

MnAdjLogBsmtFinSF1 = LogBsmtFinSF1 - 4.2243;

MnAdjLogLotArea = LogLotArea - 9.1086;

MnAdjLogBsmtFinSF1 = LogBsmtFinSF1 - 4.2243;

MnAdjLog\_1stFlrSF = Log\_1stFlrSF - 7.0067;

run;

ODS GRAPHICS ON / ATTRPRIORITY=NONE;

proc glmselect data = train2 plots = all;

class Neighborhood MSZoning CentralAir KitchenQual BsmtQual Foundation;

model LogSalePrice = GarageCars Fireplaces YearBuilt

YearRemodAdd MSZoning CentralAir KitchenQual BsmtFullBath BsmtQual LogLotArea

Log\_1stFlrSF OverallCond

LogGrLivArea | LogBsmtFinSF1 | LogGarageArea | OverallQual | Neighborhood

@2 / selection=LASSO(choose=SBC stop=cv) cvmethod=random(2) hierarchy=single showpvalues stat=all;

output out=LassoData p=predict r=resid;

/\* LogTotalBsmtSF LogBsmtFinSF1 MnAdjOverallQual\*Foundation Fireplaces\*CentralAir \*/

run;quit;

ODS GRAPHICS OFF;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* APPENDIX Correlation Analysis \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Matrix scatter plot and Correlation anaylsis \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc corr data=train3;

var LogSalePrice LogGrLivArea LogLotArea LogBsmtFinSF1 LogGarageArea Log\_1stFlrSF OverallQual OverallCond GarageCars Fireplaces YearBuilt YearRemodAdd BsmtFullBath;

run;

proc sgscatter data=train3;

matrix LogSalePrice LogGrLivArea LogLotArea LogBsmtFinSF1 LogGarageArea Log\_1stFlrSF OverallQual OverallCond GarageCars Fireplaces YearBuilt YearRemodAdd BsmtFullBath;

title "Scatter Plot Matrix of Iowa's Houseing Price Data";

run;

/\* Correlation on Neighborhood, MSZoning, KitchenQual, BsmtQual; & CentralAir:

(categorical explanitory variables using ChiSq \*/

proc freq data=train3;

tables Neighborhood\*MSZoning / chisq;

exact pchi;

run;

proc freq data=train3;

tables Neighborhood\*KitchenQual / chisq;

exact pchi;

run;

proc freq data=train3;

tables KitchenQual\*MSZoning / chisq;

exact pchi;

run;

proc freq data=train3;

tables KitchenQual\*BsmtQual / chisq;

exact pchi;

run;

proc freq data=train3;

tables KitchenQual\*CentralAir / chisq;

exact pchi;

run;

proc freq data=train3;

tables Neighborhood\*BsmtQual/ chisq;

exact pchi;

run;

proc freq data=train3;

tables BsmtQual\*MSZoning / chisq;

exact pchi;

run;

proc freq data=train3;

tables BsmtQual\*CentralAir / chisq;

exact pchi;

run;

/\* not enough memory \*/

/\*proc freq data=train3;

tables Neighborhood\*CentralAir / chisq;

exact pchi;

run;

\*/

proc freq data=train3;

tables MSZoning\*CentralAir / chisq;

exact pchi;

run;