Grade: 94+3(best Kaggle score)=97

This a complete analysis with well written and organized report. I would suggest you to always write out the regression model with estimated coefficients when the number of predictors isn’t large.

Ames Housing Analysis

## Introduction

This research is conducted on behalf of Century 21 Ames. We will be analyzing housing data for Ames, Iowa collected between 2006 and 2010.

The original data set, collected by Dean De Cock, contains 2930 observations and a large variety of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing property values.

In the first analysis, we will be looking at the relationship between sale price and total square footage for each home in three specific neighborhoods: North Ames, Eastwood, and Brookside neighborhoods.

For the second analysis, we will be looking at all neighborhoods in Ames to predict the sale price of homes.

## Question 1

For this analysis we will be assessing the relationship between the square footage (sqft) of a property and the sale price in the North Ames, Eastwood and Brookside neighborhoods. Specifically, we will be examining a subset from the larger data pool, using 381 observations from the 3 neighborhoods of interest. Our model, taking in to account the neighborhoods as the interaction term, looks like this:

After initial assessment of the data (Figures 1 & 2), we deemed it best to perform a log of both sale price (log[sale]) and property square footage (log[sqft]). The results improved the normality and linearity of the data (Figures 1 & 3): specifically, we went from an initial adjusted r2 of .3917 to .5056.

Even after log transformation, however, we found that there were still a small number of potential outliers, specifically in the Edwards neighborhood. We omitted these properties based on their studentized residual leverage score and Cook’s D and found that the linearity and normality of the data improved even more (Figure 4); we determined that the degree of improvement in the model were worth the omissions, and further improved our r2 from .5056 to .5216. We ran the regression, using North Ames as reference, and found interaction between all terms to be significant at the p = 0.05 level (Table 1 and Figure 5).

Using these results, our model now looks like this:

This model further gives us three equations for sale price, one for each neighborhood:

{logsale | Brookside} = 5.91 + 0.82 logsqft

{logsale | Edwards} = 6.92 + 0.67 logsqft

{logsale | North Ames} = 8.49 + 0.47 logsqft

We will address our findings and the implications of these formulae for each of the neighborhoods in question, in turn.

### Brookside

For Brookside, the model indicates that a doubling of the median square footage of a property in Brookside is associated with a 2.82 = 1.765 multiplicative change in the median sale price. The predicted median for the sale price of a property in this location increases by 76.5% for each doubling of square footage, with a 95% confidence interval for this relationship being in the range of (2.65 = 1.56, 2.98 = 1.97), or 56% and 97%.

Using our formula, a baseline home with 100 square feet would have a median sale price of $16,092; each doubling of the square footage would multiply this by 1.765 (thus, 200 square feet would be $28,403; 400, $50,131; and so forth).

### Edwards

For Edwards, the model indicates that a doubling of the median square footage of a property in Edwards is associated with a 2.67 = 1.59 multiplicative change in the median sale price. The predicted median for the sale price of a property in this location increases by 59% for each doubling of square footage, with a 95% confidence interval for this relationship being in the range of (2.508 = 1.42, 2.83 = 1.78), or 42% and 78%.

Using our formula, a baseline home with 100 square feet would have a median sale price of $22,144; each doubling of the square footage would multiply this by 1.59 (thus, 200 square feet would be $35,209; 400, $55,983; and so forth).

### North Ames

For North Ames, the model indicates that a doubling of the median square footage of a property in North Ames is associated with a 2.47 = 1.38 multiplicative change in the median sale price. The predicted median for the sale price of a property in this location increases by 38% for each doubling of square footage, with a 95% confidence interval for this relationship being in the range of (2.38 = 1.3, 2.56 = 1.47), or 30% and 47%.

Using our formula, a baseline home with 100 square feet would have a median sale price of $42,376; each doubling of the square footage would multiply this by 1.38 (thus, 200 square feet would be $58,479; 400, $80,701; and so forth).

### Conclusions

This was an observational study, thus we cannot make inferences to causation. There could be many other variables that contribute to the relationship between property square footage and sale price, and given the number of variables, most likely are.

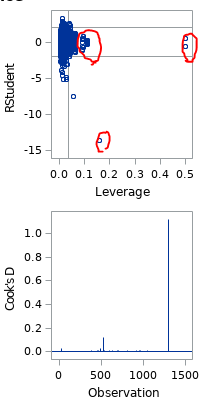
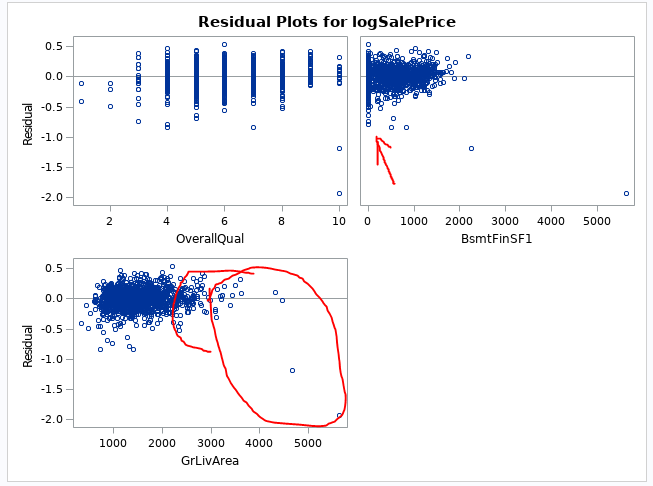
Using our models to determine sales prices of some of the homes within the data provided, these formulae are rough estimates: this should be understood especially given the range of the confidence intervals, provided earlier. In addition, the omission of all factors other than square footage would suggest that the results from these estimates should be taken as precisely that – estimates – with further analysis required to produce more accurate results.

## Question 2

For this analysis, we will be utilizing four different models to attempt to find the most accurate predictive model for homes in the Ames, Iowa area.

Because of the nature of the data, we have opted to utilize the log transform of the sales price, rather than the raw sales price, for each of our models.

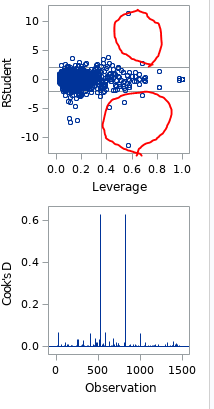
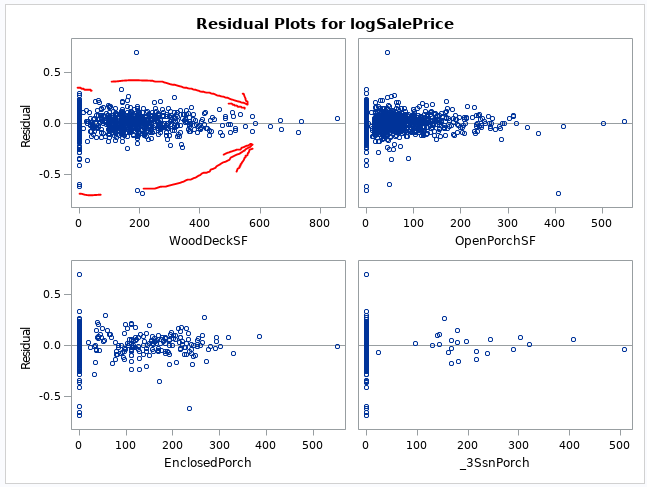
### Forward Selection

The best forward selection model we built relies on the neighborhood of the home, its overall quality, square footage of the first finished basement (if any), and the ground floor living area.

While there are some outliers (see the figure to right), the overall quality of the residual plots are acceptable: if the relatively low quantity of outliers were not present, these would be random clouds, which is what we are looking for.

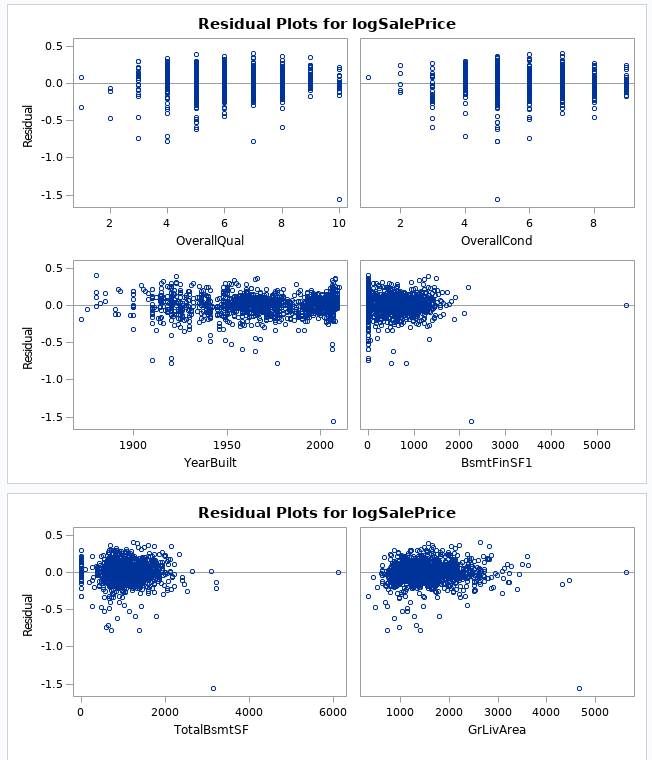
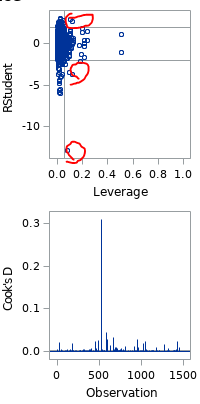
Similarly, the studentized residuals and Cook’s D appear to be acceptable, again with the exception of a small number of outliers.

### Backward Selection

Our backward selection model found that 76 of the 78 potential explanatory variables available to us within the data had a statistically significant role to play in determining sales price. Specifically, the total basement square footage and ground floor living area were omitted: given that these variables were redundant with others, this is not surprising.

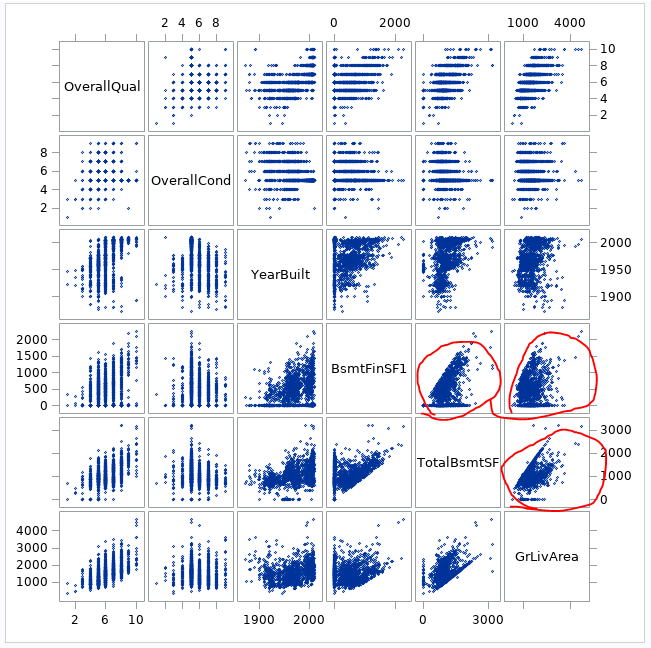
Within this model were a large number of variables with notable non-constant variance. The common theme is presented in the figure, above, with the square footage of wood decking. While the studentized residuals are mostly clumped into an acceptable range, the Cook’s D has several observations that are concerning, and indicate potential outliers, but should not impact the overall model strongly.

### Stepwise Selection

Our stepwise selection model selected nine variables: neighborhood; building type; overall quality; overall condition; year built; roof material; square footage of the first finished basement; total basement square footage; and ground floor living area.

Our findings indicate that the residuals looked better than the previous two models, however there was still some uneven spread in the variance for the year built and basement square footage variables. While there were still a few outliers in the studentized residuals, the Cook’s D overall was significantly better than both prior models, with the worst outlier coming out to roughly .3.

### Custom Selection

Here we attempted to improve on the quality of the stepwise selection, given that it was our best model, by reducing the number of variables we used to reduce problems potentially introduced by colineraity: specifically, we decided to remove the total basement square footage from the model, per our examination of colinearity in the table shown.

In testing, however, this model fared roughly comparably to our existing stepwise selection.

We then decided that removing outliers may have a positive effect on the quality of the model. We returned the removed variable, then removed observations with a lot area of larger than 100000, and a ground living area of greater than 5641 square feet. This model is effectively identical to the stepwise selection, only with extreme outliers removed.

### Model Comparisons

|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Models | Adjusted r2 | CV Press | Kaggle Score |
| Forward | .8895 | 32.13272 | .16196 |
| Backward | .9447 | 32.2876 | 5.77017 |
| Stepwise | .8895 | 30.27482 | .13303 |
| Custom | .8886 | 26.97725 | .13313 |

It would appear that our forward, stepwise, and custom models have roughly the same predictive power, given the almost-identical adjusted r2 values. Where they differ, however, is in the CV Press; this would indicate that there is most likely two variables in the forward model that are having a minor negative impact, while the stepwise most likely just has one – compared to our custom method, which has the lowest CV Press value. Kaggle scores are presented for completion’s sake, and given the opacity of their testing methodology, we offer no comment beyond an observation that the stepwise method narrowly won out over our custom: this indicates that there may be potentially some very minor overfitting occurring in a variable contributing little to the overall model.

That our custom model did not outperform the stepwise model indicates that the removed outliers were useful in establishing predictions: the lower CV Press value indicates that the custom model has a higher probability of overfitting, which bears out in the (slightly higher) Kaggle score.

What sounds out, however, is the backward model. It has a significantly higher adjusted r2, and a CV Press comparable with our forward model, but has an abysmal Kaggle score. The combination of the inflated adjusted r2 and the Kaggle score indicates that the backward selection most likely has a notable amount of overfitting occurring, which would result in test sets being – potentially – wildly inaccurate.

### Conclusions

Using this data, it appears to be entirely possible to construct a reasonably accurate model to assess the sales price of a home in Ames, Iowa. Of course, as with all statistical models, it is possible that any number of other variations can have an impact on the values used to inform these models and render them moot for usage in a context other than the one in which it was constructed.

Thus, it may be better to state that given the data, it is possible to construct a reasonably accurate model to assess the sales price of a home in Ames, Iowa, for the given time period from which these data were taken, under the economic conditions of the time.

## References

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

<https://ww2.amstat.org/publications/jse/v19n3/decock.pdf>

## Appendix A – Question 1

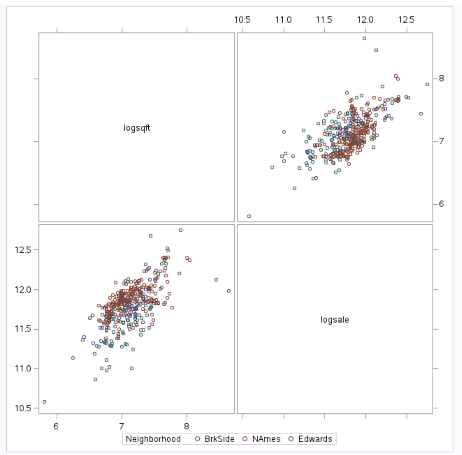
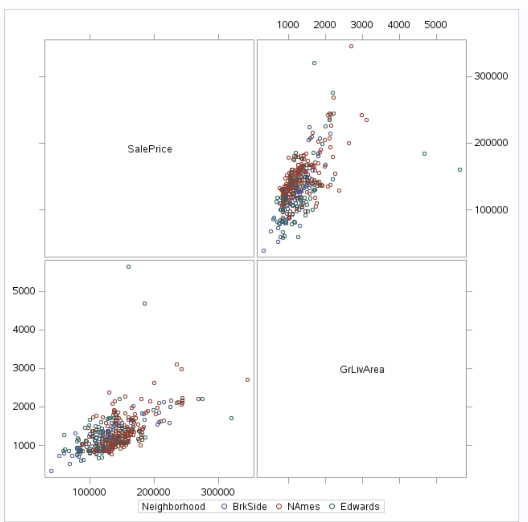


Figure 1 - Analysis 1 - EDA Pre and Post Log Log Transformation

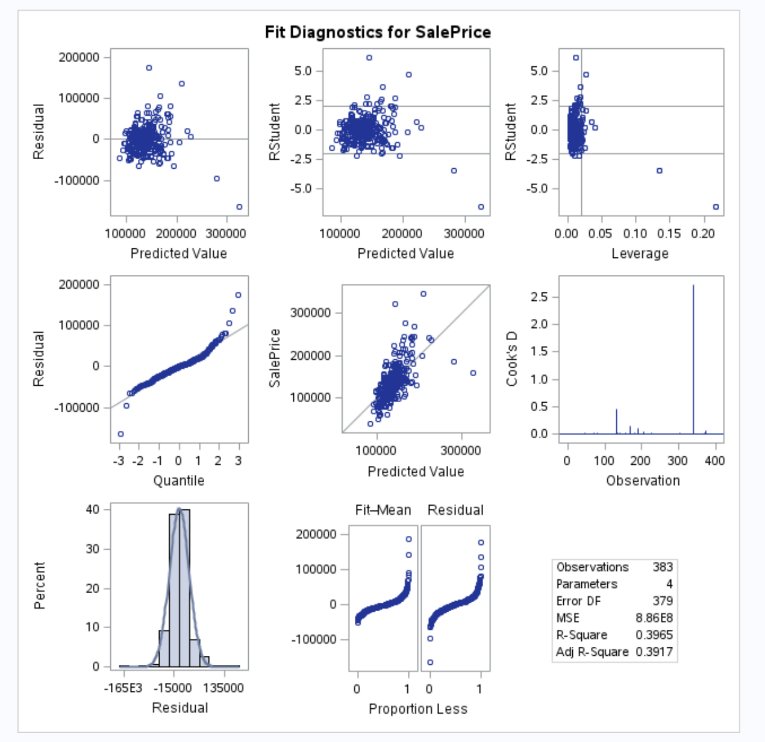


Figure 2 - Analysis 1 - Initial Fit Diagnostics

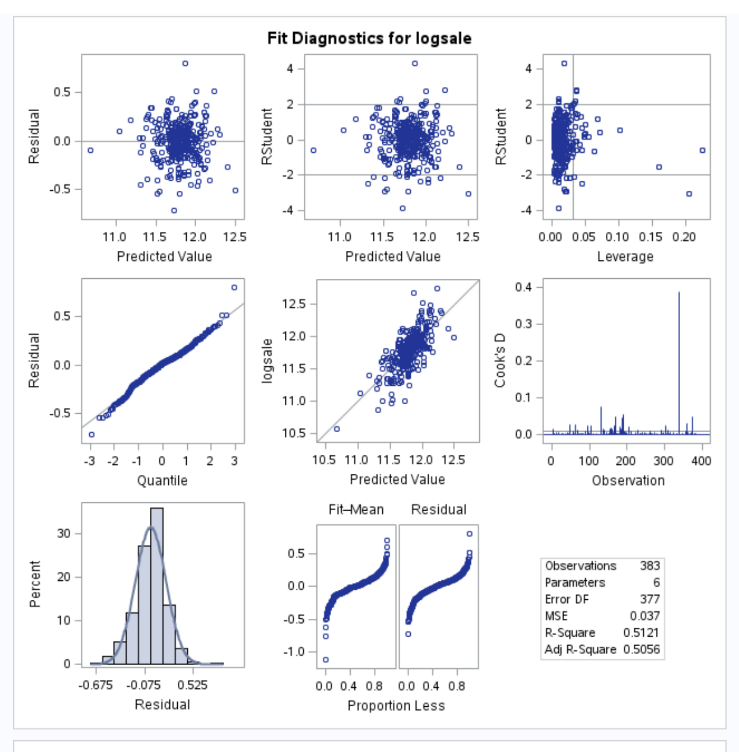


Figure 3- Analysis 1 - Post Log Log Fit Diagnostics

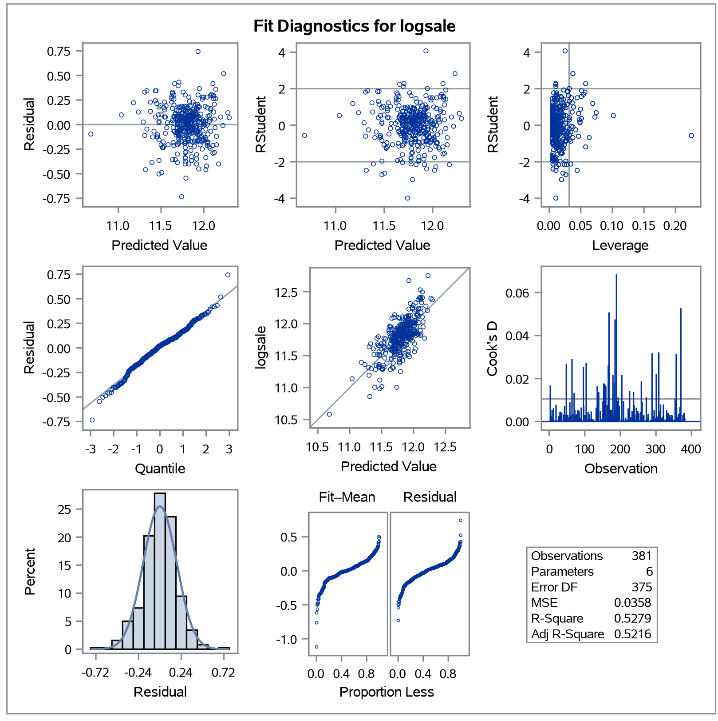


Figure 4 – Analysis 1 – Post Removal of Outliers

Table 1 - Question 1 - Result of analysis

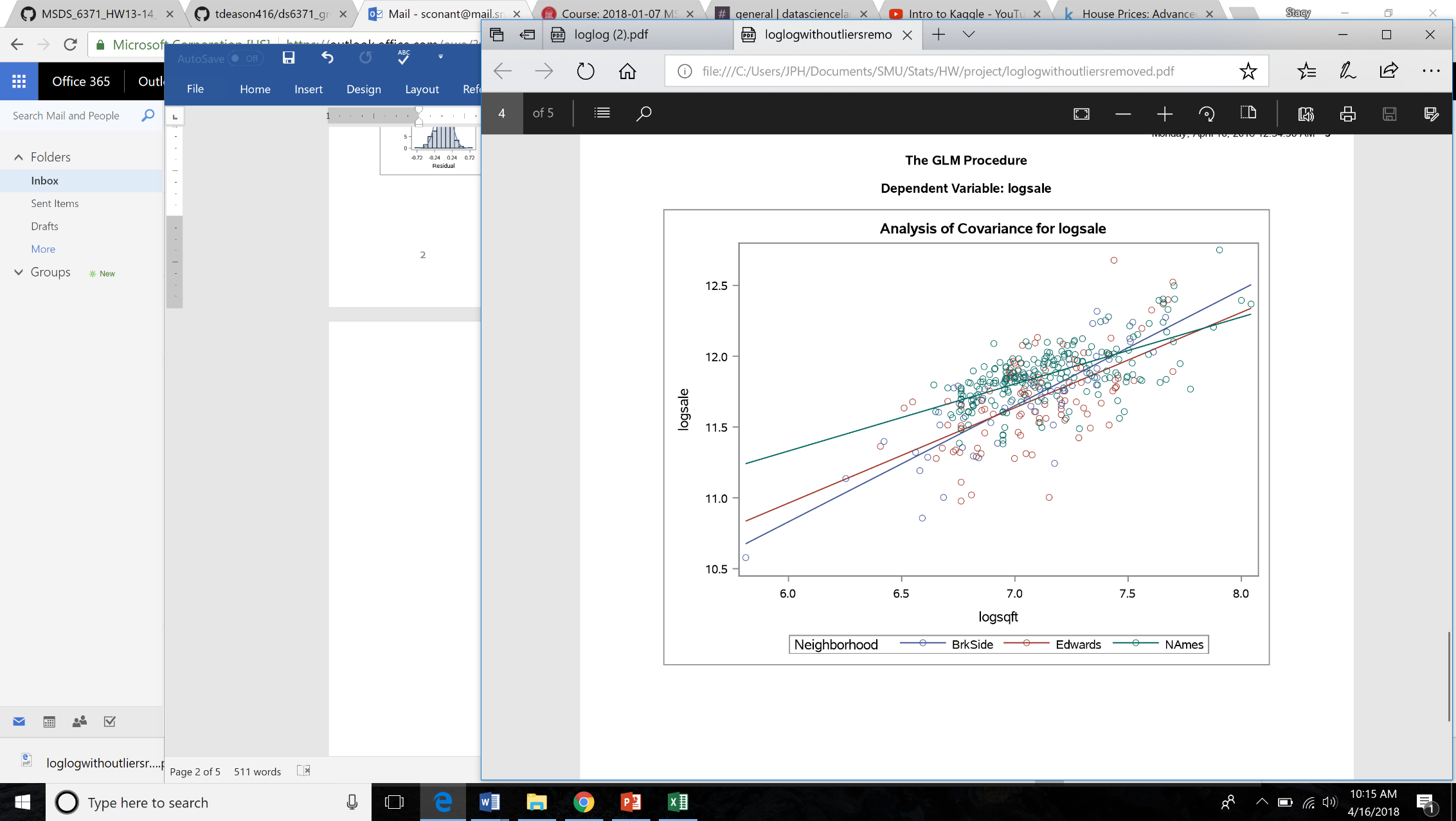
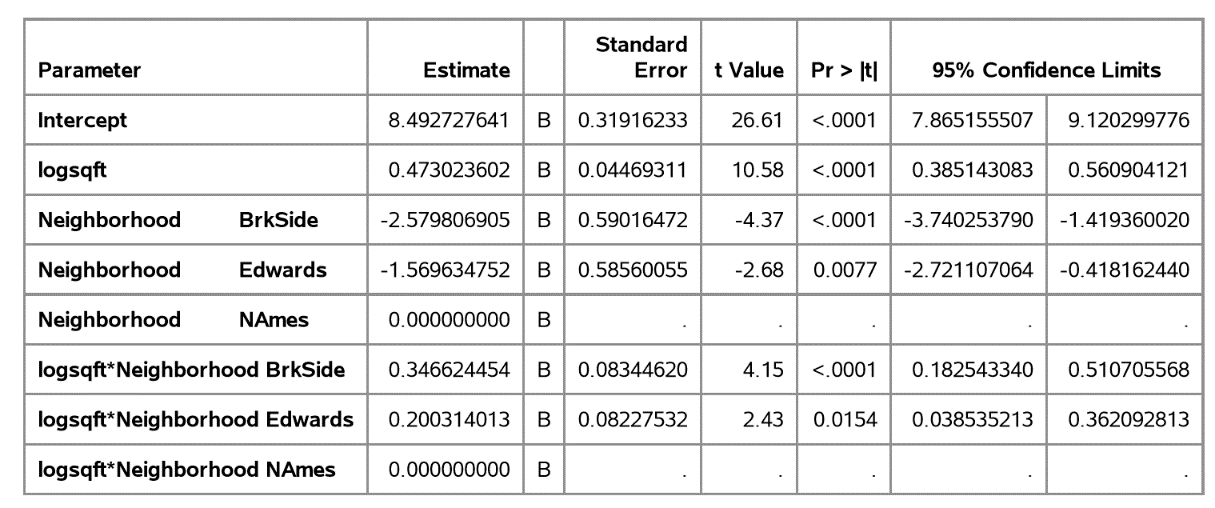


Figure 5 - Question 1 - Analysis Results

Code for Question 1:

\*uploading data;

data train1;

infile '/home/sconant0/sasuser.v94/train.csv' firstobs= 2 dlm=",";

input Id MSSubClass MSZoning$ LotFrontage LotArea Street$ Alley$ LotShape$ LandContour$ Utilities$ LotConfig$ LandSlope$ Neighborhood$ 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$ FirstFlrSF SecondFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual$ TotRmsAbvGrd Functional$ Fireplaces FireplaceQu$ GarageType$ GarageYrBlt GarageFinish$ GarageCars GarageArea GarageQual$ GarageCond$ PavedDrive$ WoodDeckSF OpenPorchSF EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea PoolQC$ Fence$ MiscFeature$ MiscVal MoSold YrSold SaleType$ SaleCondition$ SalePrice;

run;

\*uploading data;

data test1;

infile '/home/sconant0/sasuser.v94/test.csv' firstobs = 2 dlm=",";

input Id MSSubClass MSZoning$ LotFrontage LotArea Street$ Alley$ LotShape$ LandContour$ Utilities$ LotConfig$ LandSlope$ Neighborhood$ 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$ FirstFlrSF SecondFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual$ TotRmsAbvGrd Functional$ Fireplaces FireplaceQu$ GarageType$ GarageYrBlt GarageFinish$ GarageCars GarageArea GarageQual$ GarageCond$ PavedDrive$ WoodDeckSF OpenPorchSF EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea PoolQC$ Fence$ MiscFeature$ MiscVal MoSold YrSold SaleType$ SaleCondition$;

run;

\*adding sale price to test set;

data test2;

set test1;

SalePrice = .;

;

\*combining data set;

data test3;

set train1 test2;

run;

proc print data=test3; run;

\*subsetting data for question 1 with only neighborhoods of interest;

data test4;

set test3;

if Neighborhood = "NAmes" or Neighborhood = "Edwards" or Neighborhood = "BrkSide"; run;

proc print data=test4; run;

\*exporting cvs;

proc export data=test4

outfile='/home/sconant0/sasuser.v94/test4.csv'

dbms=csv; run;

\*EDA of data set for sale price and sqft;

proc sgscatter data = test4;

matrix SalePrice GrLivArea / group=Neighborhood; run;

\*run glm on raw data;

proc glm data=test4 plots=ALL;

class Neighborhood;

model SalePrice = GrLivArea | Neighborhood / solution clparm; run;

\*log transform both variables;

data test4log;

set test4;

logsale = log(SalePrice);

logsqft = log(GrLivArea); run;

\*re-run EDA on log log;

proc sgscatter data = test4log;

matrix logsqft logsale / group=neighborhood; run;

\*remove 2 outliers in Edwards: Id 524 and 1299;

data test4log;

set test4;

logsale = log(SalePrice);

logsqft = log(GrLivArea);

if \_n\_ = 131 or \_n\_ = 339 then delete; run;

proc print data=test4log; run;

\*re-run glm on log log data with outliers removed;

proc glm data=test4log plots=ALL;

class Neighborhood;

model logsale = logsqft | Neighborhood / solution clparm;

## Appendix B – Question 2

Code for Question 2:

/\* Import the train data \*/

/\* SAS Encountered errors with MasVnrArea and GarageYrBlt due to "NA" in Numeric Fields. SAS set the "NA" values to "." Therefore no issues for now \*/

**%web\_drop\_table**(WORK.TRAIN);

FILENAME REFFILE '/folders/myshortcuts/StatisticalFoundations/Group Project/train.csv';

**PROC** **IMPORT** DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.TRAIN;

GETNAMES=YES;

**RUN**;

**PROC** **CONTENTS** DATA=WORK.TRAIN; **RUN**;

**%web\_open\_table**(WORK.TRAIN);

/\* Analysis Question 2 \*/

/\* Correct Skew in Saleprice (log) looking much more normal and consistent with what stacy did for Q1\*/

**data** Q2TRAIN;

set WORK.TRAIN;

logSalePrice = log(SalePrice);

**run**;

/\* Variable Selection Models \*/

/\* Looking for Large R^2 and small CV Press \*/

/\* Forward \*/

**proc** **glmselect** data =Q2TRAIN plots=all;

class MSZoning Street Alley LotShape LandContour Utilities 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 PoolQC Fence MiscFeature SaleType SaleCondition;

model logSalePrice = MSSubClass MSZoning LotArea Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood 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 GrLivArea 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 PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition

/selection =Forward (stop=CV) cvmethod=random(**5**) stats=adjrsq;

/\* Backward \*/

**proc** **glmselect** data =Q2TRAIN plots=all;

class MSZoning Street Alley LotShape LandContour Utilities 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 PoolQC Fence MiscFeature SaleType SaleCondition;

model logSalePrice = MSSubClass MSZoning LotArea Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood 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 GrLivArea 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 PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition

/selection =backward (stop=CV) cvmethod=random(**5**) stats=adjrsq;

/\* Stepwise \*/

**proc** **glmselect** data =Q2TRAIN plots=all;

class MSZoning Street Alley LotShape LandContour Utilities 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 PoolQC Fence MiscFeature SaleType SaleCondition;

model logSalePrice = MSSubClass MSZoning LotArea Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood 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 GrLivArea 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 PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition

/selection =stepwise (stop=CV) cvmethod=random(**5**) stats=adjrsq;

/\* Assumptions \*/

/\* Forward Selected \*/

**proc** **glm** data=Q2TRAIN plots=all;

class Neighborhood;

model logSalePrice = Neighborhood OverallQual BsmtFinSF1 GrLivArea /solution;

/\* Backward Selected \*/

**proc** **glm** data=Q2TRAIN plots=all;

class MSZoning Street Alley LotShape LandContour Utilities 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 PoolQC Fence MiscFeature SaleType SaleCondition;

model logSalePrice = MSSubClass MSZoning LotArea Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood 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 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 PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition /solution;

/\* Stepwise Selected \*/

**proc** **glm** data=Q2TRAIN plots=all;

class Neighborhood BldgType RoofMatl;

model logSalePrice = Neighborhood BldgType OverallQual OverallCond YearBuilt RoofMatl BsmtFinSF1 TotalBsmtSF GrLivArea/ solution ;

/\* Custom \*/

/\* See if we can improve stepwise \*/

**data** Q2CUSTOMTRAIN;

set Q2TRAIN;

logSalePrice = log(SalePrice);

if LotArea > **100000** then delete;

if GrLivArea > **5641** then delete;

/\* Check for colinearity in variables \*/

**proc** **SGSCATTER** data=Q2CUSTOMTRAIN;

matrix Neighborhood BldgType OverallQual OverallCond YearBuilt RoofMatl BsmtFinSF1 TotalBsmtSF GrLivArea;

/\* Check for variable Inflation \*/

**proc** **reg** data=Q2CUSTOMTRAIN;

model logSalePrice = OverallQual OverallCond YearBuilt BsmtFinSF1 TotalBsmtSF GrLivArea / selection=cp VIF;

**proc** **glm** data=Q2CUSTOMTRAIN plots=all;

class Neighborhood BldgType RoofMatl;

model logSalePrice = Neighborhood BldgType OverallQual OverallCond YearBuilt RoofMatl BsmtFinSF1 TotalBsmtSF GrLivArea/ solution;

/\* Try Custom Model \*/

**proc** **glmselect** data =Q2CUSTOMTRAIN plots=all;

class Neighborhood BldgType RoofMatl;

model logSalePrice = Neighborhood BldgType OverallQual OverallCond YearBuilt RoofMatl BsmtFinSF1 TotalBsmtSF GrLivArea

/selection =stepwise (stop=CV) cvmethod=random(**5**) stats=adjrsq;

/\* Predictions \*/

/\* Import the test data \*/

**%web\_drop\_table**(WORK.TEST);

FILENAME REFFILE '/folders/myshortcuts/StatisticalFoundations/Group Project/test.csv';

**PROC** **IMPORT** DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.TEST;

GETNAMES=YES;

**RUN**;

**PROC** **CONTENTS** DATA=WORK.TEST; **RUN**;

**%web\_open\_table**(WORK.TEST);

/\* Add Saleprice column to test data \*/

**data** Q2TEST;

set WORK.TEST;

SalePrice = .;

**run**;

/\* Combine the train and test data \*/

**data** Q2PREDICT;

set WORK.Q2TRAIN WORK.Q2TEST;

**run**;

/\* Forward \*/

**proc** **glm** data = Q2PREDICT plots=all;

class Neighborhood;

model logSalePrice = Neighborhood OverallQual BsmtFinSF1 GrLivArea /cli solution;

output out = ForwardSelectedresults p = Predict;

**run**;

**data** ForwardSelectedresults;

set ForwardSelectedresults;

predictedSalePrice = logsaleprice;

keep id Predict saleprice logsaleprice predictedSalePrice;

**proc** **print** data=ForwardSelectedresults;

/\* Backward \*/

**proc** **glm** data = Q2PREDICT plots=all;

class MSZoning Street Alley LotShape LandContour Utilities 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 PoolQC Fence MiscFeature SaleType SaleCondition;

model logSalePrice = MSSubClass MSZoning LotArea Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood 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 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 PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition /cli solution;

output out = BackwardSelectedresults p = Predict;

**run**;

**data** BackwardSelectedresults;

set BackwardSelectedresults;

predictedSalePrice = logsaleprice;

keep id Predict saleprice logsaleprice predictedSalePrice;

**proc** **print** data=BackwardSelectedresults;

/\* Stepwise \*/

**proc** **glm** data = Q2PREDICT plots=all;

class Neighborhood BldgType RoofMatl;

model logSalePrice = Neighborhood BldgType OverallQual OverallCond YearBuilt RoofMatl BsmtFinSF1 TotalBsmtSF GrLivArea

/cli solution;

output out = StepwiseSelectedresults p = Predict;

**run**;

**data** StepwiseSelectedresults;

set StepwiseSelectedresults;

predictedSalePrice = logsaleprice;

keep id Predict saleprice logsaleprice predictedSalePrice;

**proc** **print** data=StepwiseSelectedresults;

/\* Custom \*/

**data** Q2PREDICT;

set WORK.Q2CUSTOMTRAIN WORK.Q2TEST;

**run**;

**proc** **glm** data =Q2PREDICT plots=all;

class Neighborhood BldgType RoofMatl;

model logSalePrice = Neighborhood BldgType OverallQual OverallCond YearBuilt RoofMatl BsmtFinSF1 TotalBsmtSF GrLivArea

/cli solution;

output out = CustomSelectedresults p = Predict;

**run**;

**data** CustomSelectedresults;

set CustomSelectedresults;

predictedSalePrice = logsaleprice;

keep id Predict saleprice logsaleprice predictedSalePrice;

**proc** **print** data=CustomSelectedresults;