MSDS 6371 Project Description (Weeks 13 and 14)

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this Kaggle competition's dataset proves that much more influences price negotiations than the number of bedrooms or the presence of a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

**Data and Description:**

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

**HOW TO KAGGLE VIDEO:**

**QOIs:** [**https://www.youtube.com/watch?v=0QJtczDPxZQ**](https://www.youtube.com/watch?v=0QJtczDPxZQ)

Read everything BEFORE you begin.

* Use SAS for this project.

Your team’s objective is to conduct 2 analyses:

1. ANALYSIS 1: Assume that Century 21 Ames (a real estate company) in Ames Iowa has commissioned you to answer a very important question with respect to their business. Century 21 Ames only sells houses in the NAmes, Edwards and BrkSide neighborhoods and would like to simply get an estimate of how the SalePrice of the house is related to the square footage of the living area of the house (GrLIvArea) and if the SalesPrice (and its relationship to square footage) depends on which neighborhood the house is located in. Build and fit a model that will answer this question, keeping in mind that realtors prefer to talk about living area in increments of 100 sq. ft. Provide your client with the estimate (or estimates if it varies by neighborhood) as well as confidence intervals for any estimate(s) you provide. It turns out that Century 21’s leadership team has a member that has some statistical background. Therefore, make sure and provide evidence that the model assumptions are met and that any suspicious observations (outliers / influential observations) have been identified and addressed. Finally, of course, provide your client with a well written conclusion that quantifies the relationship between living area and sale price with respect to these three neighborhoods. Remember that the company is only concerned with the three neighborhoods they sell in.

1. ANALYSIS 2: Build the most predictive model for sales prices of homes in all of Ames Iowa. This includes all neighborhoods. Your group is limited to only the techniques we have learned in 6371 (no random forests or other methods we have not yet covered). Specifically, you should produce 4 models: one from forward selection, one from backwards elimination, one from stepwise selection, and one that you build custom. The custom model could be one of the three preceding models or one that you build by adding or subtracting variables at your will. Generate an adjusted R2, CV Press and Kaggle Score for each of these models and clearly describe which model you feel is the best in terms of being able to predict future sale prices of homes in Ames, Iowa. In your paper, please include a table similar to the one below. The group with the lowest public Kaggle score will receive an extra 3 bonus points on the final exam!

Quick note on Kaggle completion: We only have one course under our belts so far (almost), but you can compete in this competition with the tools you have now (top 40th percentile or better!). After your next course (6372), you will really be able to do well (top 25th percentile or better!). With these skills as well as the skills you pick up in Data Mining and Quantifying the World, you will be able to compete with anyone!

NOTE 1: ALL ANALYSES MUST BE DONE IN SAS and all code must be placed in the appendix. Part of the grading process will be to run the code and verify the Kaggle score for each group.

Note 2: An extra 3 points on the final exam will be awarded to the team with the model with the lowest (best) Kaggle Score. In the unlikely event of a tie will split these points.

**Deliverables:**

Your group is to turn in a paper should be no more than 7 pages long (without the appendix). Please put your code in the appendix. If you are concerned with staying within the allotted 7 pages, put more screenshots and such in the appendix.

**Sample Format**

Required deliverables in the complete report:

The format of your paper (headers, sections, etc.) is flexible, although it should contain the following information.

Introduction

Data Description

(Where did the data come from? How big is it? How many observations? Where can we find out more? What are the specific variables that we need to know with respect to your analysis?)

Analysis Question 2

Restatement of Problem

Build the most predictive model for sales prices of homes in all of Ames Iowa including all neighborhoods.

Based on instructions all variables must be included in all models with exception to the custom model.

Model Selection

Type of Selection

Forward (using log transform of sale price)

|  |  |
| --- | --- |
| **Effects:** | Intercept Neighborhood OverallQual BsmtFinSF1 GrLivArea |

Backward (using log transform of sale price)

|  |  |
| --- | --- |
| **Effects:** | Intercept 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 |

Stepwise (using log transform of sale price)

|  |  |
| --- | --- |
| **Effects:** | Intercept Neighborhood BldgType OverallQual OverallCond YearBuilt RoofMatl BsmtFinSF1 TotalBsmtSF GrLivArea |

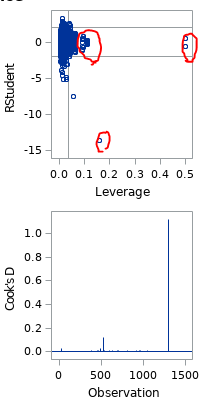
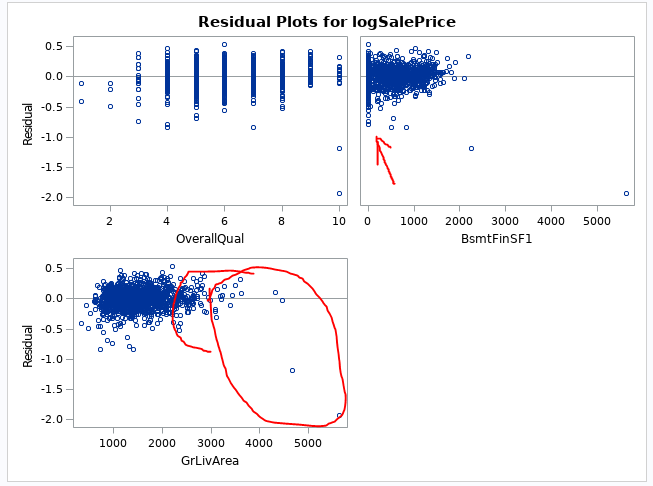
Custom

|  |  |
| --- | --- |
| **Effects:** | Intercept Neighborhood BldgType OverallQual OverallCond YearBuilt RoofMatl BsmtFinSF1 TotalBsmtSF GrLivArea |

**Checking Assumptions**

Residual PlotsInfluential point analysis (Cook’s D and Leverage)

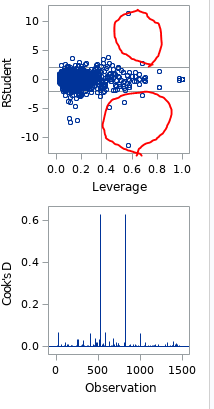
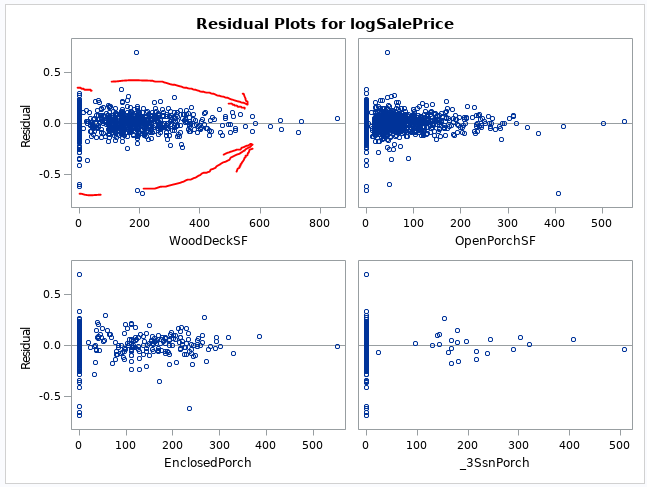
Forward Selected



Backward Selected

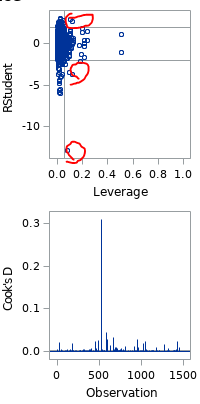
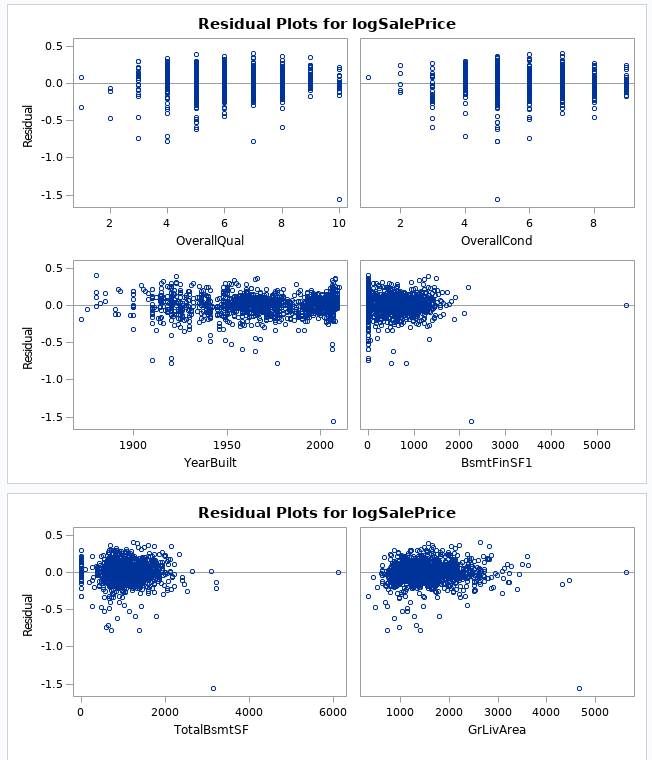
There are a lot of variables with a lot of non-constant variance. The common theme is illustrated with WoodDeckSF below

Cooks D is a train wreck



Stepwise Selected

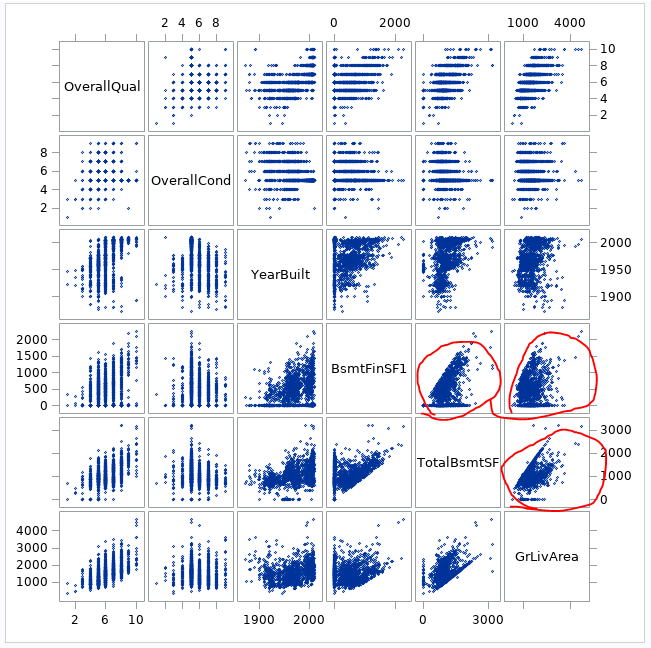
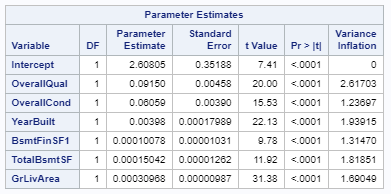
Residuals look better than prior 2 models however there is still some uneven spread in variance of YearBuilt and Bsmt variables



Custom

Here we attempt to improve the stepwise selected model:

There’s somewhat of a pattern for collinearity in Ground Level Living Area, Finished Basement SF and TotalBsmtSF though variance inflation appears acceptable.

Using intuition; Total Basement SF is likely correlated to GRLivArea so left it out.

Comparing Competing Models

Adj R2

Internal CV Press

Kaggle Score

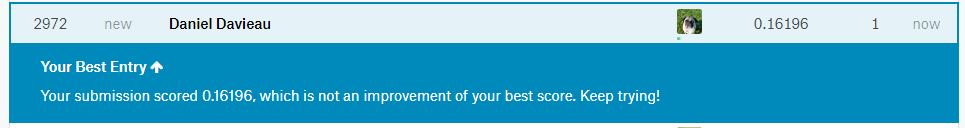
|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Models | Adjusted R2 | CV PRESS | Kaggle Score |
| Forward | 0.8895 | 32.13272 | .16196 |
| Backward | 0.9447 | 32.28760 | 5.77017 |
| Stepwise | 0.8895 | 30.27482 | .13303 |
| CUSTOM | 0.8886 | 26.97725 | .13313 |

Conclusion: A short summary of the analysis.

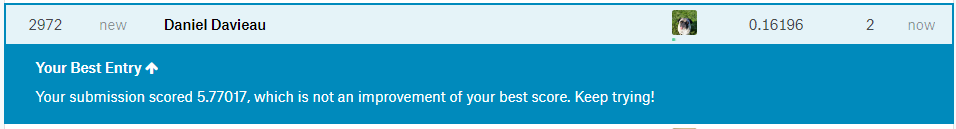
Appendix

Well commented SAS Code for Analysis 1 and 2

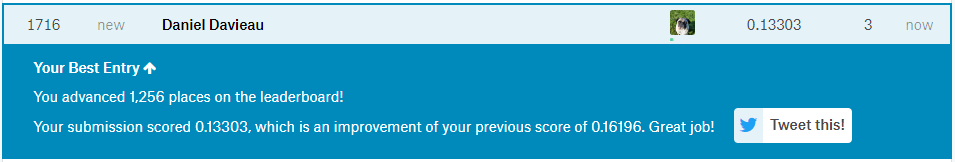
Forward (Had to use EXP() Function in Excel to undo natural Log)



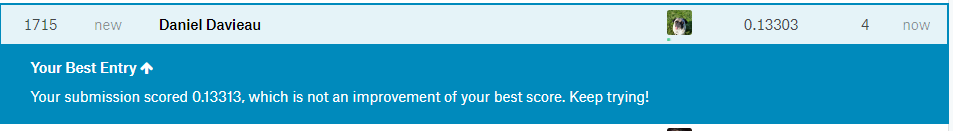
Backward



Stepwise



Custom



/\* 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;

**Rubric:**

Presentation (30%):

Organized paper with title, headings, subheadings, etc.

Labeled plots, figures, tables and charts.

Every plot, figure, table and chart included is referenced in the paper and vice versa.

No spelling or grammatical errors.

Analysis Question 1: (35%)

Analysis Question 2: (35 %)