Housing Prices Project

MSDS 6371

July 31st, 2021

Randy Kim

**Introduction**

The sale price of a home can vary depending on a multitude of variables. This is the case for virtually all residential areas, including the residences in Ames, Iowa. While current market conditions may be influential, the physical aspects of the homes can help predict what the sale price will be. This analysis dives into the physical aspects of homes to find a feasible model that can predict the sale price of homes in Ames, Iowa.

**Data Description**

A data set, Train, obtained from Kaggle contains 1,460 observations of homes in the Ames, Iowa area and a total of 81 columns; a link to the data set on Kaggle can be viewed in the appendix. Of these 81 columns, 79 are potential predictor variables and Sale Price is the response variable. The predictor variables of interest to this analysis are presented in a table in the appendix along with a brief explanation of the variable meaning.

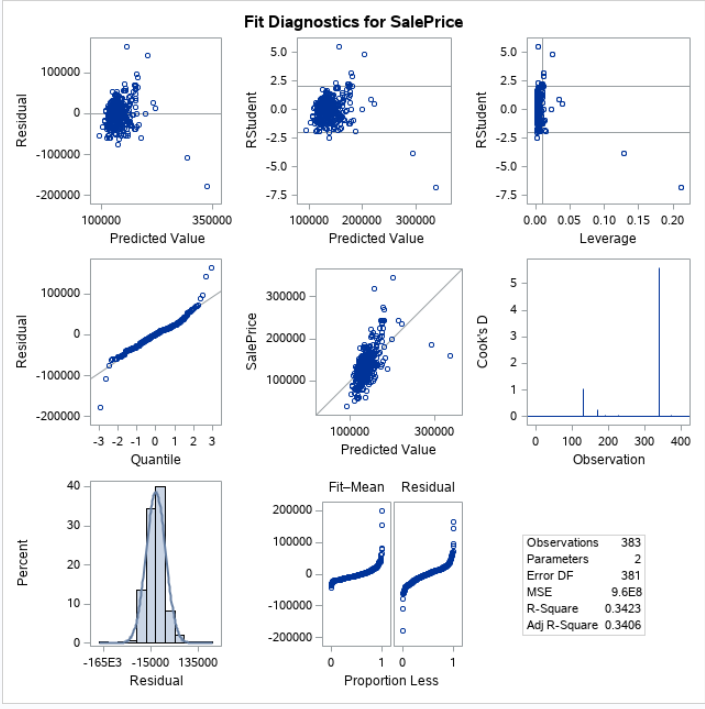
A testing data set, Test, was also provided from Kaggle which contains 1,459 observations and 80 columns; a link to this data set can be found in the appendix. The columns of the Test data set are the same as those from the Train data set, only the Sale Price of these data are missing. The goal is to predict what the sale price of the homes in the Test data set will be using a model built from the Train data set.

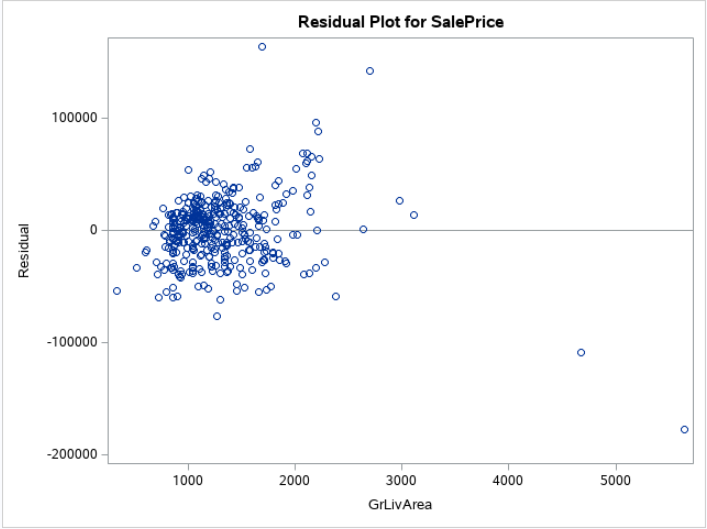
**Analysis Question 1**

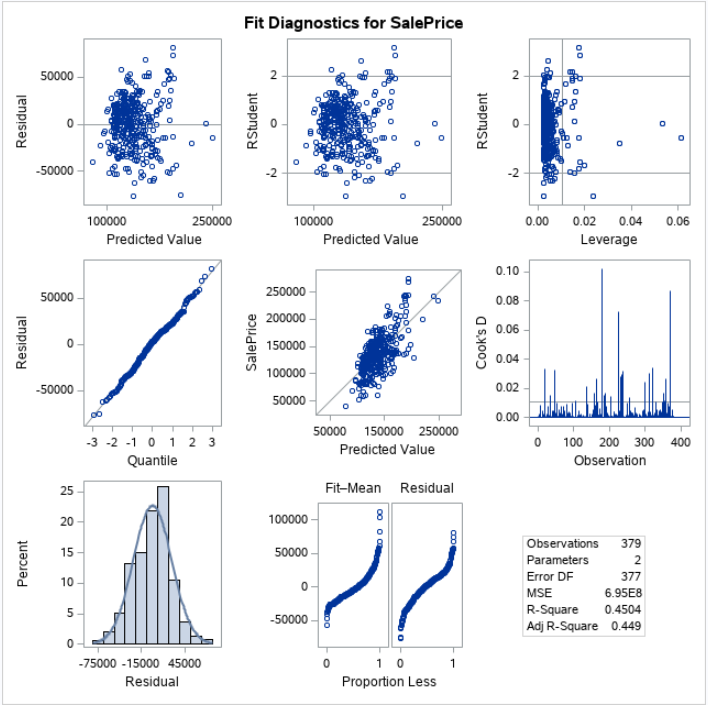
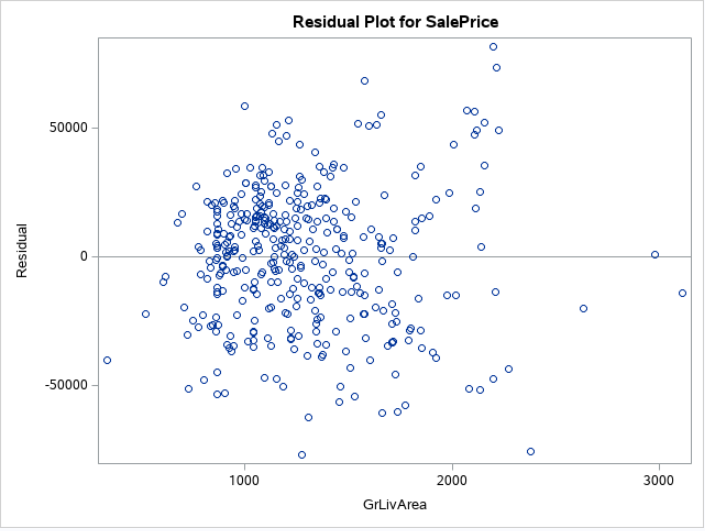
**Problem**

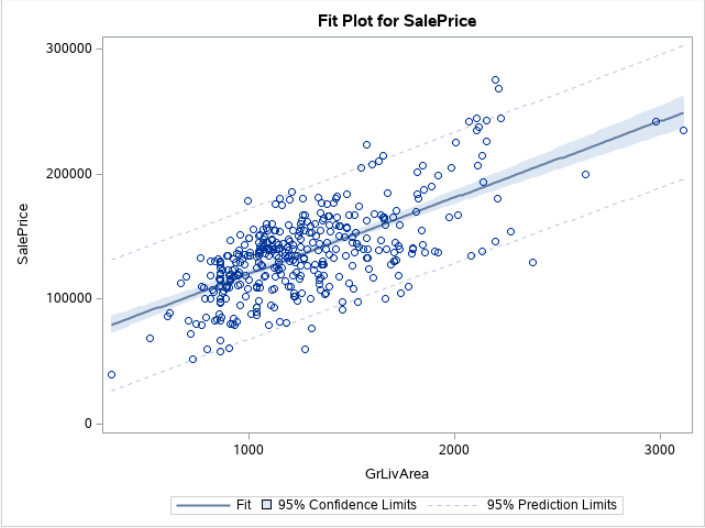
Century 21 Ames, a real estate company in Ames, Iowa, only sells homes in the NAmes, BrkSide, and Edwards neighborhoods. They would like to get an estimate of the relationship between a home’s sale price and its square footage of the living area, GrLivArea. They would also like to know if this relationship varies by the neighborhoods that they sell in.

**Assumptions**

To check the assumptions, the QQ plots, residual plots, and scatter plots were assessed for abnormal behavior on the data from the neighborhoods Names, BrkSide, and Edwards.



 From these plots, four outliers were found within the data that have large studentized residuals, ± 2.5, and one of the outliers has high Cook’s D, >5. The outliers are influential to the assumptions of equal standard deviations and linearity. These outliers were removed and the assumptions were tested again.



With the outliers removed, the assumption of linearity and equal standard deviations are met. Based on the scatterplot, the assumption of normality appears to be met and indicates randomly distributed residuals. We will assume that the data are independent.

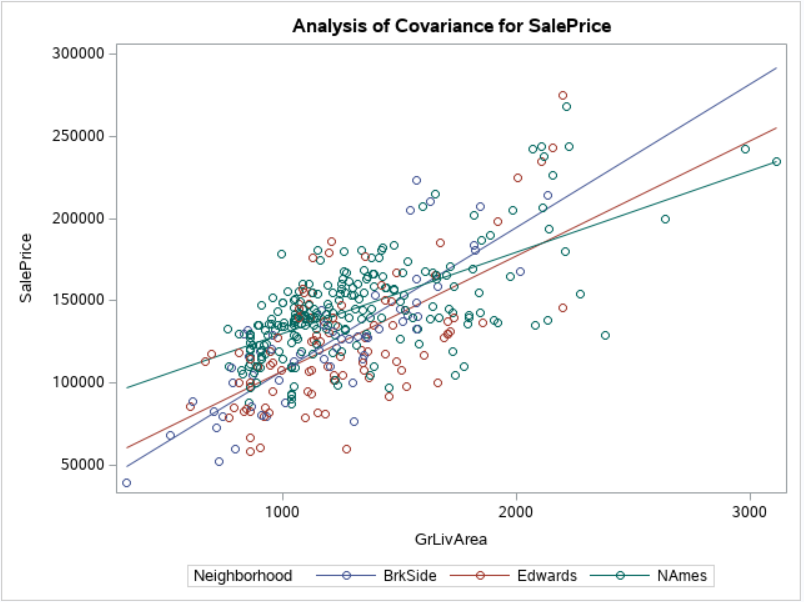
While the assumptions of the model are better met excluding the outliers, there was not a difference in significance of the model with or without the outliers. However, since the assumptions were better met without the outliers, we proceeded using the data that had the outliers removed. This resulted in better adjusted R-squared values for the models we compared.

**Comparing Models**

The models that we compared are the simple linear model with no outliers, the full model with no outliers and the reduced model with no outliers. Each of these models and their estimates, along with their associated adjusted R-squared values can be found below, the full regression estimation tables can be found in the plots 1.2 to 1.4 of the appendix.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Estimated Equation** | **Adjusted R-Squared Value** | **CV PRESS Statistic** |
| Simple Linear Regression | Sale Price = 58785.07 + 61.15(GrLivArea) | 0.449 | 265778619334 |
| Full Model | Sale Price = 67292.53 + 59.56(GrLivArea) - 14107.65(BrkSide Neighborhood) - 16857.65(Edwards Neighborhood) | 0.4948 | 245495825073 |
| Reduced Model | Sale Price = 80325.71 + 49.56(GrLivArea) - 60354.20( BrkSide Neighborhood) - 43225.29(Edwards Neighborhood) + 37.60(GrLivArea\*BrkSide Neighborhood) + 20.60(GrLivArea\*Edwards Neighborhood) | 0.5165 | 237769407660 |

**Build and Fit**

Due to the reduced model with no outliers having the highest adjusted R-squared value and the lowest CV PRESS statistic, we will choose the reduced model as the best fit for this data. Below is the regression estimated scatterplot for this selected model. 

**Parameters**

The estimated best fitting regression line is

**Sale Price = 80325.71 + 49.56(GrLivArea) - 60354.20( BrkSide Neighborhood) - 43225.29(Edwards Neighborhood) + 37.60(GrLivArea\*BrkSide Neighborhood) + 20.60(GrLivArea\*Edwards Neighborhood)**

The confidence intervals for these estimates can be found in the chart in plot 1.4 of the appendix showing the estimates table.

**Conclusion**

This means that for every 100 additional square feet in the living space, assuming that the house is in the reference neighborhood of NAmes, the sale price of the house will increase by approximately $4,956. We are 95% confident that this increase will be between $4,151 and $5,762.

Given that the estimates for the BrkSide and Edwards neighborhoods are negative, we assume that homes in the reference neighborhood, NAmes, on average have a higher sale price. Assuming that the square footage of the living space is the same, a house in the BrkSide neighborhood will have a sale price on average of $60,354.20 less than a home in the NAmes neighborhood. We are 95% confident that this decrease in price is between $36,640.02 to $84,068.38 less. Assuming that the square footage of the living space is the same, a house in the Edwards neighborhood will have a sale price on average of $43,225.29 less than a home in the NAmes neighborhood. We are 95% confident that this decrease in price is between $21,914.41 to $64,536.17 less.

Within the BrkSide neighborhood, each additional 100 square feet of living space will result in an increase of sale price on average of $3,760. We are 95% confident that this increase in price from an increase of 100 square feet of living space in a BrkSide home will be between $1,911 to $5,609. Within the Edwards neighborhood, each additional 100 square feet of living space will result in an increase of sale price on average of $2,060. We are 95% confident that this increase of 100 square feet of living space in an Edwards home will be between $447 to $3,673.

This is an observation study, therefore, we cannot make any causal inference. However, there is a correlation between the three neighborhoods, square feet of living space, and sale prices.

**Analysis Question 2**

**Problem**

Century 21 Ames would like to create a predictive model for sale prices of homes in all neighborhoods in Ames, Iowa. They would like to produce 4 models; forward selection, backward elimination, stepwise selection and a custom model. Comparing the parameters of the 4 models over adjusted R2, CV Press and Kaggle Score, the best model in terms of predicting future sale prices of homes in Ames, Iowa is chosen.

**Models**

**Forward Selection**

LotArea+

The model was selected after 14 steps with Adjusted R-Square of 0.9165. The residual plot indicates that there is a random distribution around sale price of $300,000, however, there are some outliers. There are a few outliers with high leverage and low and high residuals, but based on Cook’s D, which is under 0.10, the outliers are not influential enough to change the model. The histogram and qq plot show that the model is fairly normal. There is some evidence of a right skew in the qq plot but due to the number of observations, no action needed. The full summary table of the forward selection and the fit can be viewed in plot 2.1 in the appendix.

**Backward Elimination**

The model was selected after one elimination with Adjusted R-Square of 0,9385. The residual plot indicates that there is a random distribution around sale price of $200,000 and there are a number of outliers that can be observed. Studentzied vs. leverage plot shows that the outliers aren’t influential enough to be concerned. However, the Cook’s D plot indicates that there are two observations that need to be examined. The histogram and qq plot show that the model is normally distributed. There is some evidence of a right skew in the qq plot but due to the number of observations, no action needed The full summary table of the backward elimination and the fit can be viewed in plot 2.2 in the appendix.

**Stepwise Selection**

LotArea+

Similar to Forward selection, the model was selected after 13 steps with Adjusted R-Square of 0.9165. The residual plot shows that there is a random distribution around sale price of $150,000 and there are a number of outliers that can be observed. Studenized vs. leverage plot shows that the outliers aren’t influential and the Cook’s D, which is under 0.8, indicates the outliers are not to be concerned. The histogram and qq plot show that the model is fairly normally distributed. Again, there is some evidence of a right skew in the qq plot but due to the number of observations, no action needed. The full summary table of the stepwise selection and the fit can be viewed in plot 2.3 of the appendix.

**Custom Model**

The customized model contains 3 variables, Neighborhood, Overall Quality, and Living area by square feet. The Adjusted R-Square came out to be 0.8397. The residual plot indicates that there is a random distribution around sale price of $125,000 and a number of outliers can be observed. Studentized vs. leverage plot and Cook’s D plots, <0.1, show that the outliers are not influential enough to change the model. The histogram and qq plot indicate that the model is fairly normally distributed as well. The fit of the Custom model can be viewed in plot 2.4 of the appendix.

**Conclusion**

Comparing the Forward Selection, Backward Elimination, Stepwise Selection, and Custom Model, the Forward Selection model had the highest Adjusted R-Square and the lowest CV Press and Kaggle Score. Therefore, the forward Selection model with cross validation will be a final predictive model for sale prices of homes in all neighborhoods in Ames, Iowa.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Model** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Forward Selection | 0.9165 | 7.900801E11 | 0.15010 |
| Backward Elimination | 0.9324 | 9.220347E11 | 0.28859 |
| Stepwise Selection | 0.9165 | 8.39466E11 | 0.15045 |
| Custom Model | 0.8397 | 1.526254E12 | 0.17701 |

**Appendix**

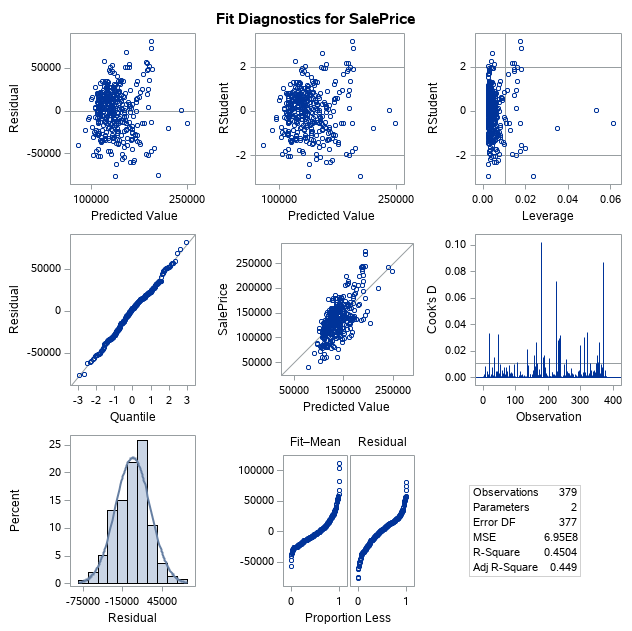
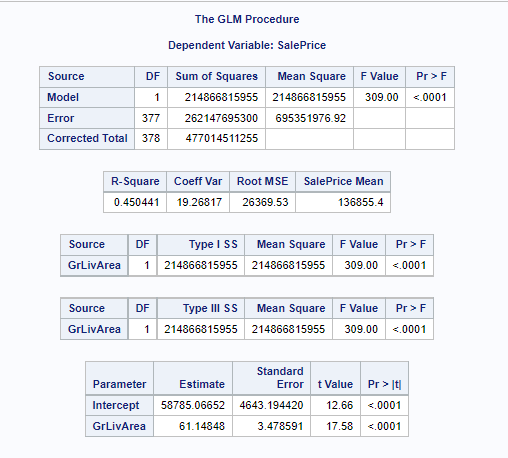
Link to Kaggle Datasets (Both Train and Test):

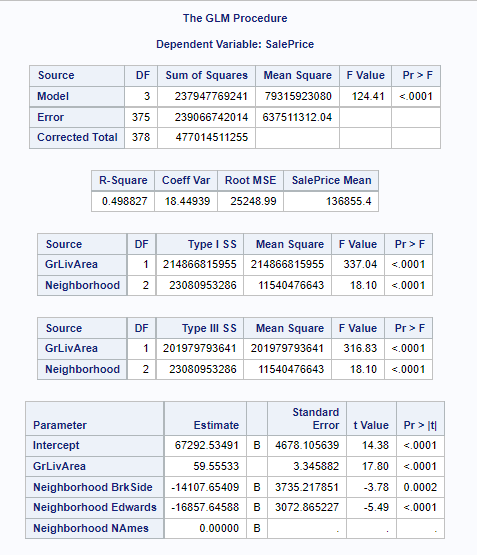
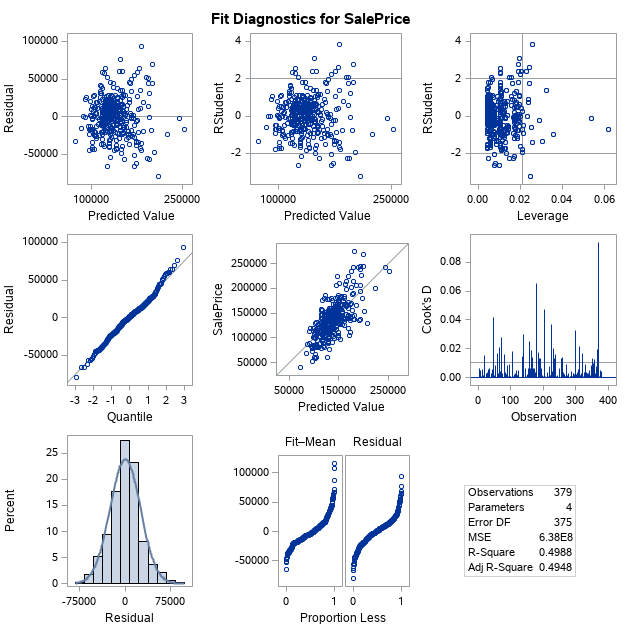
<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview>

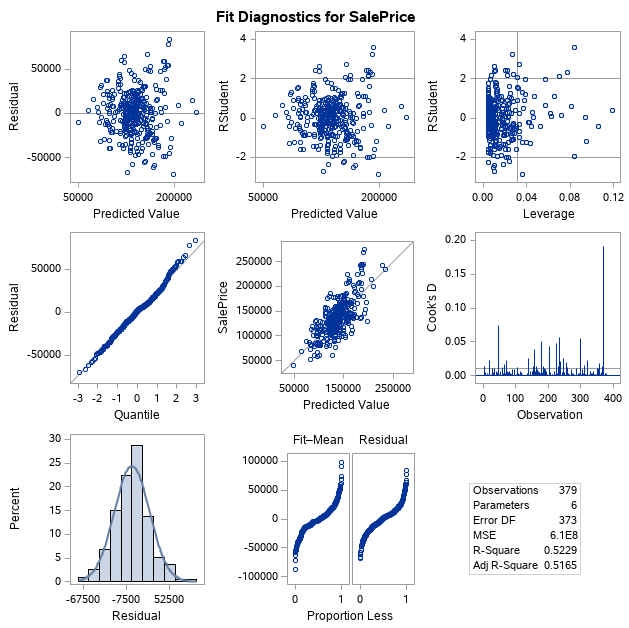
Table of Response variables used in the analysis (Alphabetical Order)

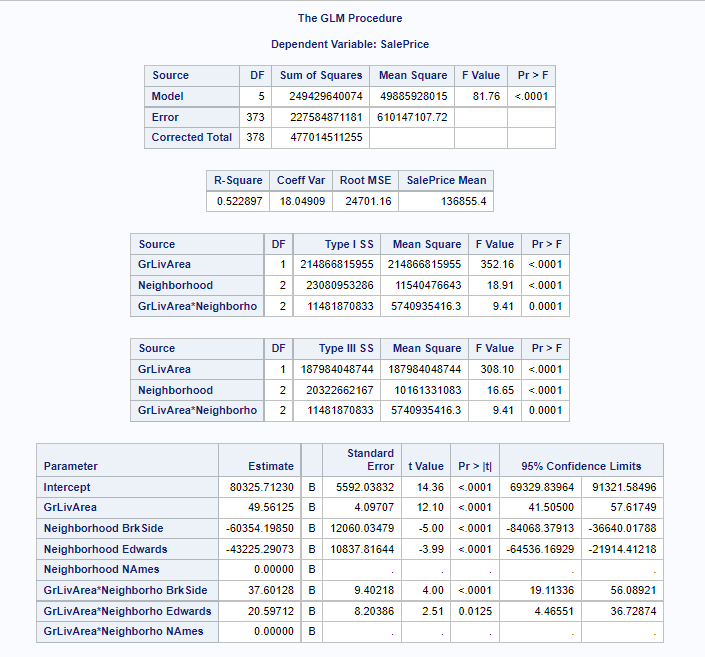
|  |  |
| --- | --- |
| **Variable Name** | **Brief Description** |
| Alley | Type of alley access to property |
| Bedroom | Bedrooms above grade (does NOT include basement bedrooms) |
| BldgType | Type of dwelling |
| BsmtCond | Evaluates the general condition of the basement |
| BsmtExposure | Refers to walkout or garden level walls |
| BsmtFinSF1 | Type 1 finished square feet |
| BsmtFinSF2 | Type 2 finished square feet |
| BsmtFinType1 | Rating of basement finished area |
| BsmtFinType2 | Rating of basement finished area (if multiple types) |
| BsmtFullBath | Basement full bathrooms |
| BsmtHalfBath | Basement half bathrooms |
| BsmtQual | Evaluates the height of the basement |
| BsmtUnfSF | Unfinished square feet of basement area |
| CentralAir | Central air conditioning |
| Condition1 | Proximity to various conditions |
| Condition2 | Proximity to various conditions (if more than one is present) |
| Electrical | Electrical system |
| EnclosedPorch | Enclosed porch area in square feet |
| ExterCond | Evaluates the present condition of the material on the exterior |
| Exterior1st | Exterior covering on house |
| Exterior2nd | Exterior covering on house (if more than one material) |
| ExterQual | Evaluates the quality of the material on the exterior |
| Fence | Fence quality |
| FireplaceQu | Fireplace quality |
| Fireplaces | Number of fireplaces |
| Foundation | Type of foundation |
| FullBath | Full bathrooms above grade |
| Functional | Home functionality (Assume typical unless deductions are warranted) |
| GarageArea | Size of garage in square feet |
| GarageCars | Size of garage in car capacity |
| GarageCond | Garage condition |
| GarageFinish | Interior finish of the garage |
| GarageQual | Garage quality |
| GarageType | Garage location |
| GarageYrBlt | Year garage was built |
| GrLivArea | Above Grade Living Area in Square Feet |
| HalfBath | Half baths above grade |
| Heating | Type of heating |
| HeatingQC | Heating quality and condition |
| HouseStyle | Style of dwelling |
| Kitchen | Kitchens above grade |

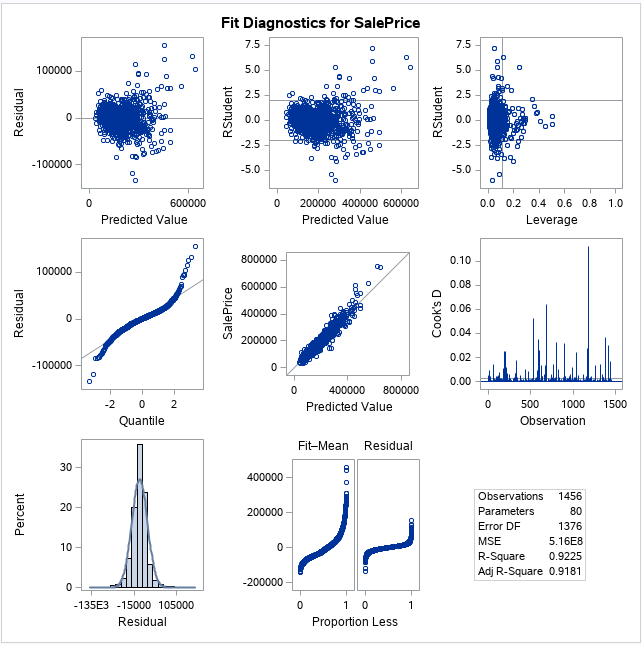
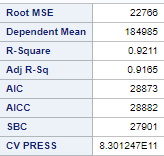
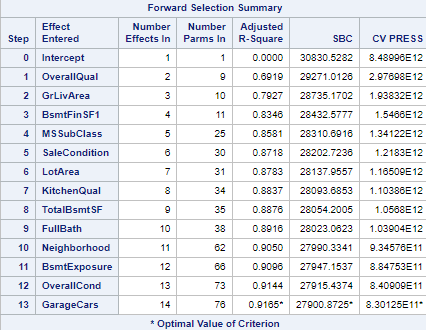
|  |  |
| --- | --- |
| KitchenQu | Kitchen quality |
| LandContour | Flatness of the property |
| LandSlope | Slope of property |
| LotArea | Lot size in square feet |
| LotConfig | Lot configuration |
| LotFrontage | Linear feet of street connected to property |
| LotShape | General shape of property |
| LowQualFinSF | Low quality finished square feet (all floors) |
| MasVnrArea | Masonry veneer area in square feet |
| MasVnrType | Masonry veneer type |
| MiscFeature | Miscellaneous feature not covered in other categories |
| MiscVal | $Value of miscellaneous feature |
| MoSold | Month Sold (MM) |
| MSSubClass | Identifies the type of dwelling involved in the sale |
| MSZoning | Identifies the general zoning classification of the sale |
| Neighborhood | Physical locations within Ames city limits |
| OpenPorchSF | Open porch area in square feet |
| OverallCond | Rates the overall condition of the house |
| OverallQual | Rates the overall material and finish of the house |
| PavedDrive | Paved driveway |
| PoolArea | Pool area in square feet |
| PoolQC | Pool quality |
| RoofMatl | Roof material |
| RoofStyle | Type of roof |
| SaleCondition | Condition of sale |
| SaleType | Type of sale |
| ScreenPorch | Screen porch area in square feet |
| Street | Type of road access to property |
| TotalBsmtSF | Total square feet of basement area |
| TotRmsAbvGrd | Total rooms above grade (does not include bathrooms) |
| Utilities | Type of utilities available |
| WoodDeckSF | Wood deck area in square feet |
| YearBuilt | Original construction date |
| YearRemodAdd | Remodel date (same as construction date if no remodeling or additions) |
| YrSold | Year Sold (YYYY) |

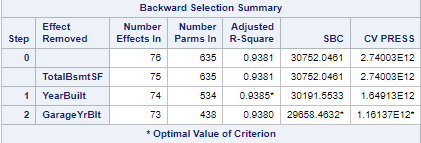
Plot 1.2 - Simple Model 

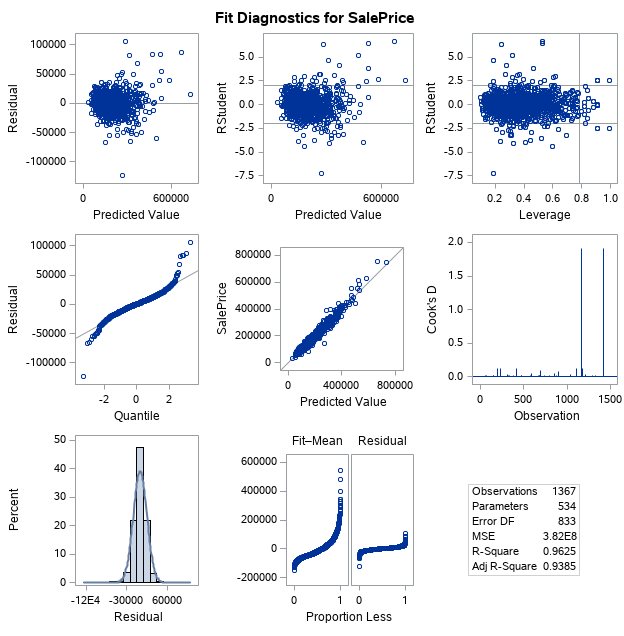
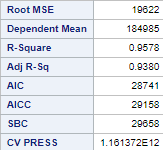
Plot 1.3 - Full Model

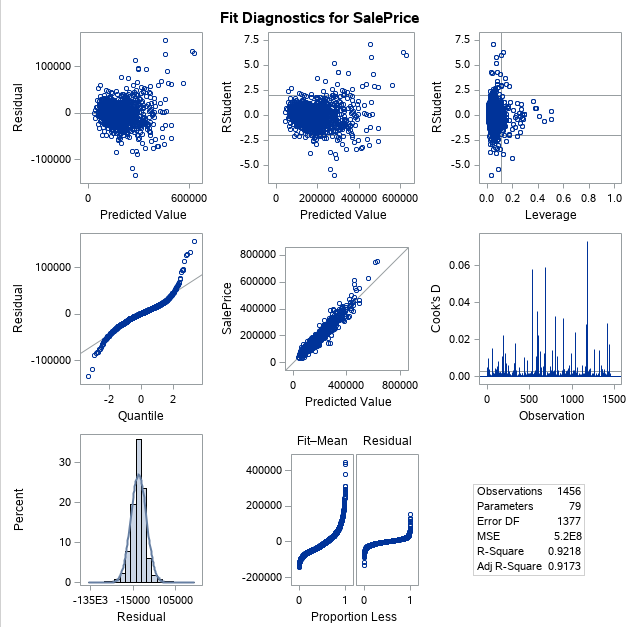
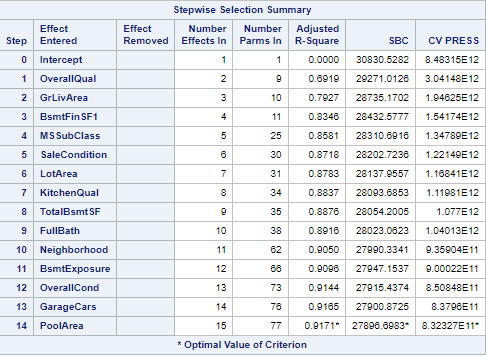
Plot 1.4 - Reduced Model

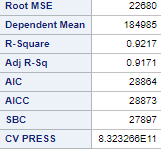


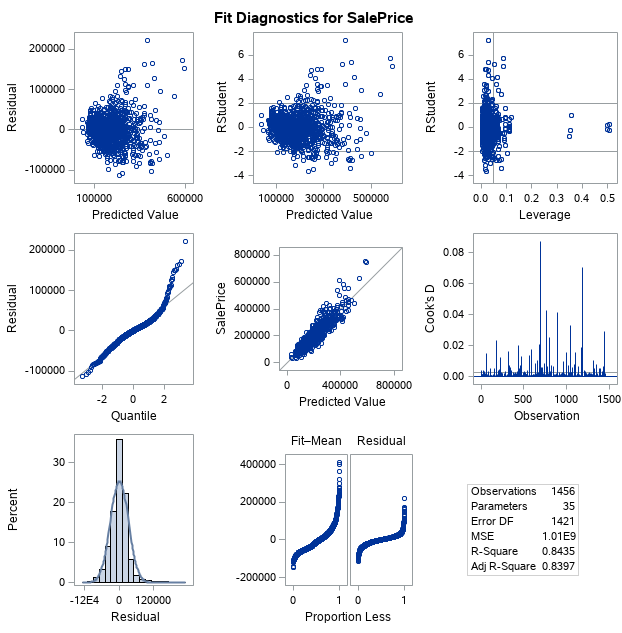
Plot 2.1 - Forward selection

Plot 2.2 - Backward Elimination



Plot 2.3 - Stepwise Selection



Plot 2.4 - Custom Model

Full SAS Code

/\* DS6371 - Project \*/

/\* Randy Kim \*/

/\* Importing the Training Data Set from Kaggle \*/

Proc Import Datafile='/home/u42892787/DS 6371 Project/Train.csv'

Out=Train Replace;

Run;

Proc Print Data=Train;

Run;

/\* Importing the Testing Data Set from Kaggle \*/

Proc Import Datafile='/home/u42892787/DS 6371 Project/Test.csv'

Out=Test Replace;

Run;

Proc Print Data=Test;

Run;

/\* Start Analysis Question 1 \*/

/\* Shortening the data to the 3 Neighborhoods of interest; NAmes, BrkSide, Edwards\*/

Data TrainNeighborhood;

Set Train;

Where Neighborhood = 'NAmes' OR

Neighborhood = 'BrkSide' OR

Neighborhood = 'Edwards';

Run;

Proc Print Data=TrainNeighborhood;

Run;

/\* Proc GLM to check the Assumptions \*/

/\* Simple Model on Full Data\*/

Proc GLM Data=TrainNeighborhood Plots=All;

Model SalePrice = GrLivArea/ solution;

Run;

/\* It appears that there are 4 outliers, 2 with large GrLivArea and 2 with large SalePrice \*/

/\* The outliers are observation 643, 725, 1299, and 524 \*/

/\* Removing the Outliers \*/

Data TrainNoOutlier;

Set TrainNeighborhood;

Where Id ~= 643 AND Id ~= 725 AND Id ~= 1299 AND Id ~= 524;

Run;

Proc Print Data = TrainNoOutlier;

Run;

/\* Proc GLM to check Assumptions with No Ouliers \*/

/\* Also Shows the simple linear model on No Outlier Data \*/

Proc GLM Data=TrainNoOutlier Plots=All;

Model SalePrice = GrLivArea/ Solution clm;

Run;

/\* Simple Model NO INTERCEPT on All Data \*/

Proc GLM Data=TrainNeighborhood Plots=All;

Model SalePrice = GrLivArea/ noint solution;

Run;

/\* Simple Model NO INTERCEPT on No Outlier Data \*/

Proc GLM Data=TrainNoOutlier Plots=All;

Model SalePrice = GrLivArea/ noint solution;

Run;

/\* Proc GLM showing the Full Model on All Data \*/

Proc GLM Data=TrainNeighborhood Plots=ALL;

Class Neighborhood (ref = 'NAmes');

Model SalePrice = GrLivArea Neighborhood/ Solution;

Run;

/\* Proc GLM showing Full Model on No Outlier Data \*/

Proc GLM Data=TrainNoOutlier Plots=All;

Class Neighborhood (ref = 'NAmes');

Model SalePrice = GrLivArea Neighborhood/ Solution clm;

Run;

/\* Pric GLM Showing Full Model on All Data with NO INTERCEPT \*/

Proc GLM Data=TrainNeighborhood Plots=All;

Class Neighborhood (ref = 'NAmes');

Model SalePrice = GrLivArea Neighborhood/ noint solution;

Run;

/\* Proc GLM showing Full Model on No Outlier Data with NO INTERCEPT \*/

Proc GLM Data=TrainNoOutlier Plots=ALL;

Class Neighborhood (ref = 'NAmes');

Model SalePrice = GrLivArea Neighborhood/ noint solution;

Run;

/\* Proc GLM showing Reduced Model on All Data \*/

Proc GLM Data=TrainNeighborhood Plots=ALL;

Class Neighborhood (ref = 'NAmes');

Model SalePrice = GrLivArea | Neighborhood/ Solution;

Run;

/\* Proc GLM showing Reduced Model on No Outlier Data \*/

/\* Chosen Model! \*/

Proc GLM Data=TrainNoOutlier Plots=All;

Class Neighborhood (ref = 'NAmes');

Model SalePrice = GrLivArea | Neighborhood/ Solution clparm clm;

Run;

/\* Proc GLM showing Reduced Model on All Data with NO INTERCEPT \*/

Proc GLM Data=TrainNeighborhood Plots=ALL;

Class Neighborhood (ref = 'NAmes');

Model SalePrice = GrLivArea | Neighborhood/ noint solution;

Run;

/\* Proc GLM showing Reduced Model on No Outlier Data with NO INTERCEPT \*/

Proc GLM Data=TrainNoOutlier PLots=All;

Class Neighborhood (ref = 'NAmes');

Model SalePrice = GrLivArea | Neighborhood/ noint solution;

Run;

/\* End Analysis Question 1 \*/

/\* Start Analysis Question 2 \*/

/\* Adding the variable SalePrice to the Test data set \*/

Data Test;

Set Test;

SalePrice = .;

Run;

/\* Adding the Train and Test set together to get the estimates \*/

/\* No Outliers \*/

Data TrainNoOutlier2;

Set Train;

Where Id ~= 643 AND Id ~= 725 AND Id ~= 1299 AND Id ~= 524;

Run;

Proc Print Data = TrainNoOutlier2;

Run;

/\* Adding the Train and Test set together to get the estimates \*/

Data TestEst2;

Set TrainNoOutlier2 Test;

Run;

/\* Forward Selection \*/

Proc GLMSelect Data = TrainNoOutlier2;

Class MSSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope

Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual

OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond

BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical

BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual

TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish

GarageCars GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature MoSold YrSold

SaleType SaleCondition LotFrontage;

model SalePrice = 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 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath

HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces

FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond

PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea PoolQC Fence

MiscFeature MoSold YrSold SaleType SaleCondition/ selection=FORWARD(stop=cv)

cvmethod = random(5) stats=ADJRSQ;

Run;

/\* Getting estimates using the forward selection equation \*/

Proc GLM Data= TestEst2 plots=ALL;

Class Neighborhood OverallQual MSSubClass SaleCondition KitchenQual FullBath BsmtExposure

OverallCond GarageCars;

Model SalePrice = OverallQual GrLivArea BsmtFinSF1 MSSubClass SaleCondition LotArea

KitchenQual TotalBsmtSF FullBath Neighborhood BsmtExposure OverallCond GarageCars

PoolArea/ cli solution;

Output out = resultsF2 p = Predict;

Run;

/\* Creating the data set with just the predictions from forward selection\*/

Data resultsforward2;

Set resultsF2;

If Predict > 0 then SalePrice = Predict;

If Predict < 0 then SalePrice = 163000;

Keep Id SalePrice;

Where Id > 1460;

Run;

Proc Print Data = resultsforward2;

Run;

/\* Exporting the created estimates of Forward Selection \*/

proc export data = resultsforward2

outfile='/home/u42892787/DS 6371 Project/resultsforward2.csv'

dbms=csv;

run;

/\* Backward Elimination \*/

Proc GLMSelect Data = TrainNoOutlier2;

Class MSSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope

Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual

OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond

BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical

BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual

TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish

GarageCars GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature MoSold YrSold

SaleType SaleCondition LotFrontage;

model SalePrice = 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 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath

HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces

FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond

PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea PoolQC Fence

MiscFeature MoSold YrSold SaleType SaleCondition/ selection=BACKWARD(stop=cv)

cvmethod = random(5) stats=ADJRSQ;

Run;

/\* Getting estimates using the backward elimination equation \*/

Proc GLM Data = TestEst2 Plots=ALL;

Class MSSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope

Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual

OverallCond YearRemodAdd RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond

BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical

BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual

TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish

GarageCars GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature MoSold YrSold

SaleType SaleCondition LotFrontage;

model SalePrice = MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour

Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle

OverallQual OverallCond YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd

MasVnrType MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure

BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF Heating HeatingQC

CentralAir Electrical LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath

HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces

FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond

PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea PoolQC Fence

MiscFeature MoSold YrSold SaleType SaleCondition/ cli solution;

Output out = resultsB p = Predict;

Run;

/\* Creating the data set with just the predictions from backward elimination\*/

Data resultsbackward;

Set resultsB;

If Predict > 0 then SalePrice = Predict;

If Predict < 0 then SalePrice = 163000;

Keep Id SalePrice;

Where Id > 1460;

Run;

Proc Print Data = resultsbackward;

Run;

/\* Exporting the created estimates of Backward Elimination \*/

proc export data = resultsbackward

outfile='/home/u42892787/DS 6371 Project/resultsbackward.csv'

dbms=csv;

run;

/\* Stepwise Selection \*/

Proc GLMSelect Data = TrainNoOutlier2;

Class MSSubClass MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope

Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual

OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl

Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond

BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical

BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual

TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish

GarageCars GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature MoSold YrSold

SaleType SaleCondition LotFrontage;

model SalePrice = 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 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath

HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces

FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond

PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea PoolQC Fence

MiscFeature MoSold YrSold SaleType SaleCondition/ selection=Stepwise(stop=cv)

cvmethod = random(5) stats=ADJRSQ;

Run;

/\* Getting estimates using the Stepwise selection equation \*/

Proc GLM Data = TestEst2 plots=all;

Class MSSubClass Neighborhood OverallQual OverallCond BsmtExposure FullBath KitchenQual

GarageCars SaleCondition;

model SalePrice = OverallQual GrLivArea BsmtFinSF1 MSSubClass SaleCondition LotArea

KitchenQual TotalBsmtSF FullBath Neighborhood BsmtExposure OverallCond

GarageCars/ cli solution;

Output out = resultsS2 p = Predict;

Run;

/\* Creating the data set with just the predictions from stepwise selection\*/

Data resultsstepwise2;

Set resultsS2;

If Predict > 0 then SalePrice = Predict;

If Predict < 0 then SalePrice = 163000;

Keep Id SalePrice;

Where Id > 1460;

Run;

Proc Print Data = resultsstepwise2;

Run;

/\* Exporting the created estimates of Stewise Selection \*/

proc export data = resultsstepwise2

outfile='/home/u42892787/DS 6371 Project/resultsstepwise2.csv'

dbms=csv;

run;

/\* Custom Model \*/

Proc GLM Data= TestEst2 Plots=All;

Class Neighborhood (ref = 'BrDale') OverallQual (ref = '2');

Model SalePrice = Neighborhood OverallQual GrLivArea/ cli solution;

Output out = resultsC p = Predict;

Run;

/\* Creating the data set with just the predictions from stepwise selection\*/

Data resultscustom;

Set resultsC;

If Predict > 0 then SalePrice = Predict;

If Predict < 0 then SalePrice = 163000;

Keep Id SalePrice;

Where Id > 1460;

Run;

Proc Print Data = resultscustom;

Run;

/\* Exporting the created estimates of Custom Model \*/

proc export data = resultscustom

outfile='/home/u42892787/DS 6371 Project/resultscustom.csv'

dbms=csv;

run;

/\* End Analysis Question 2 \*/