Introduction

This report demonstrates an advanced regression model to predict the sale prices of home in Ames Iowa based on 79 explanatory variables provided by Kaggle House Price Competition. The report consists of two main analyses. The first analysis is to simply get an estimate of how the sale prices of the house are based on the square footage of the living area of the house and which neighborhood the house is located in (NAmes, Edwards and BrkSide). The second analysis of the report focuses on building the most predictive model for sale prices of home in all of Ames Iowa. In this analysis, we provide in-depth explanations on how variables are selected for the final model. Our model is evaluated based on Root-Mean-Squared-Error (RMSE) between the predicted value and the observed sale price.

Data

House sale data and various house attributes covering houses in multiple Ames, Iowa neightborhoods were gathered from <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>.

Description

The sales data contains a considerable number of specific attributes (i.e. 80 attributes + 1 id) for each house contained in the data set. Contained in the training data set, there are 1460 observations with a detailed description of each attribute in the data\_description.txt file found at the link previously provided.

Each analysis will state and cover the specific data points considered within that work.

Cleaning NA Values

There were many places in both the training and test datasets where NA values were recorded for quantitative fields. These data fields required some cleaning in preparation for use in the statistical package used for analyses, SAS. Both the training and test data sets were pre-processed prior to use using Excel. In these instances, these data fields were set to zero. Some variables could be considered either as a quantitative data field or a classification field (e.g. YearRemodAdd). In these instances, these fields were considered to be quantitative and thus were set to zero for NA values.

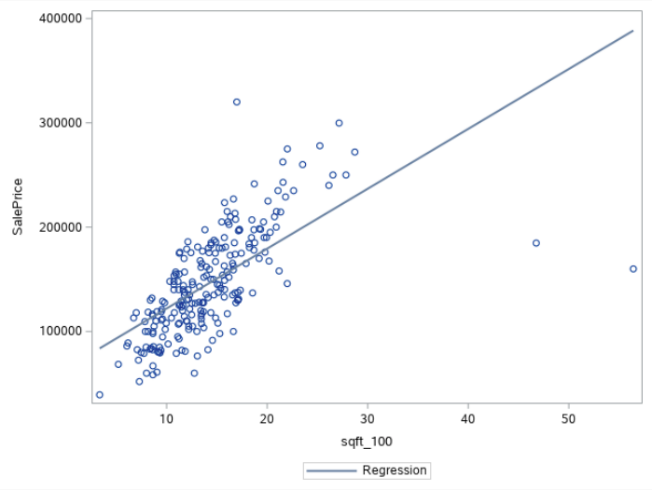
Analysis Question 1:

Problem

Analyze the relationship between Greater Living Area and Sales Price for houses sold in Ames, Iowa for select neighborhoods (NAmes, Edwards and BrkSide). Also to analyze if there is any dependency on the neighborhood and Sales Price.

Build and Fit the Model

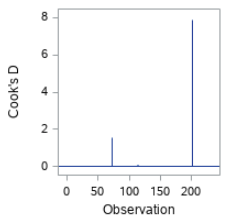
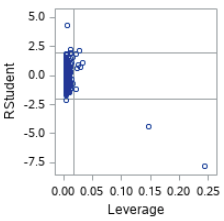
First, we will look at the data (just for the three neighborhoods as a collective data set.

Scatter Plot****

Let’s start with analyzing the relationship between Greater Living Area and Sales Price for these neighborhoods.

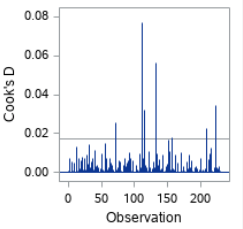
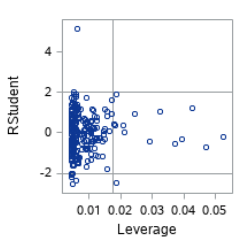
Here is a scatter plot for Sales Price vs Greater Living Area (In increments of 100 square feet):

At first glance, two outliers are visible which influence the fit line to tilt downwards.

Residual Plots****

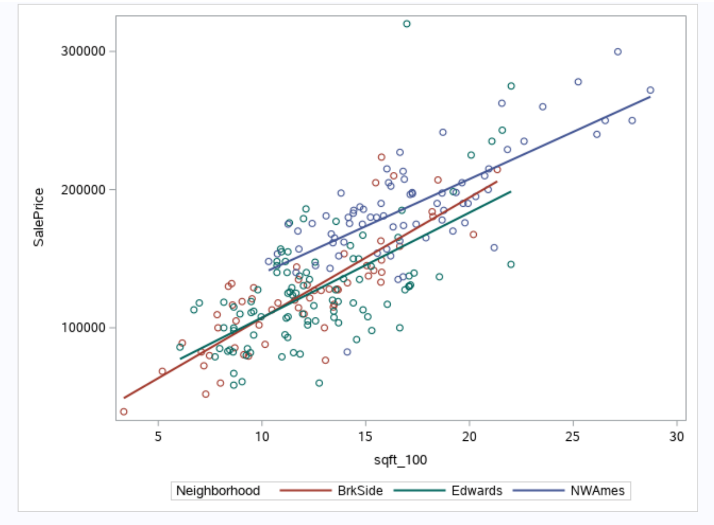
Influential point analysis (Cook’s D and Leverage)

The two charts to the right show the leverage and Cook’s D statistic (including outliers)

After taking a closer look at the outliers in the training data, it appears these transactions were made on model Homes having overall quality of the home rating of 10. 

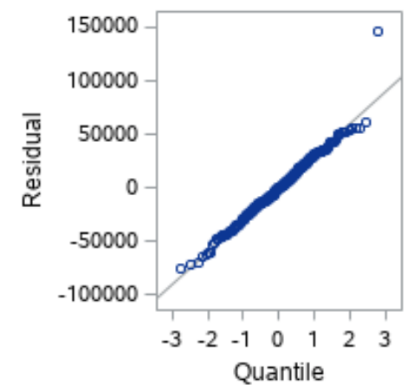
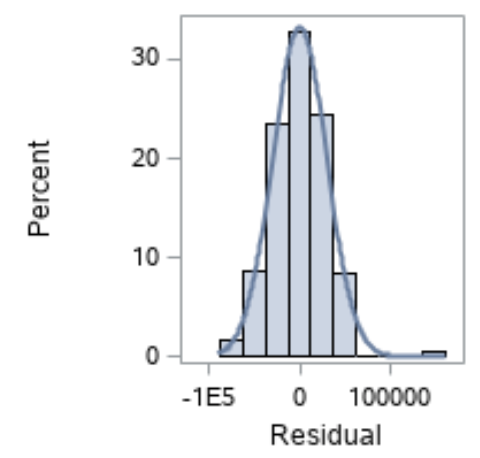
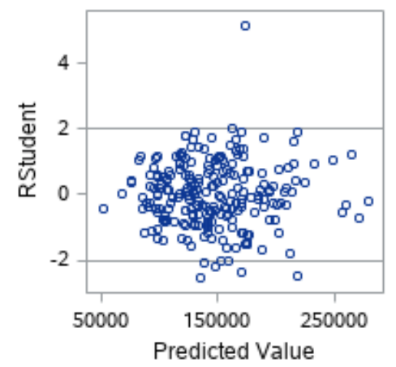
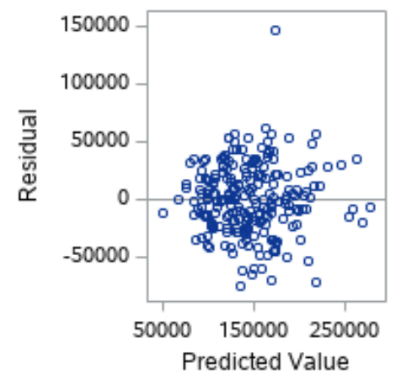
Deleting the outliers help better fit the model. The Leverage and Cook’s D after deleting the outliers are now within reasonable limits.

As only the two outliers were encountered, and they were similar in nature, only one assumption for their removal is required.

Grouped by Neighborhood****

After grouping the data by neighborhood, it is evident through visual analysis, that the regression lines differ by neighborhood.

Assumptions Continued



Residual Plot: The residual plot resembles a random scatter of points around the 0 line, suggesting that variance is constant throughout the dataset.

Studentized Residual Plot: This plot is very similar to the residual plot, although this plot identifies potential outlying observations. This plot identifies a potentially very outlying point with a predicted t-value of 4.5, however the Cook’s Distance value for this point is within the acceptable range.

Histogram of Residuals: The histogram of residuals looks normally distributed with little skew, providing evidence that the residuals are normally distributed.

Q-Q Plot of Residuals: Since the points fall along a straight diagonal line, the Q-Q Plot of residuals provides no evidence against the residuals being normally distributed.

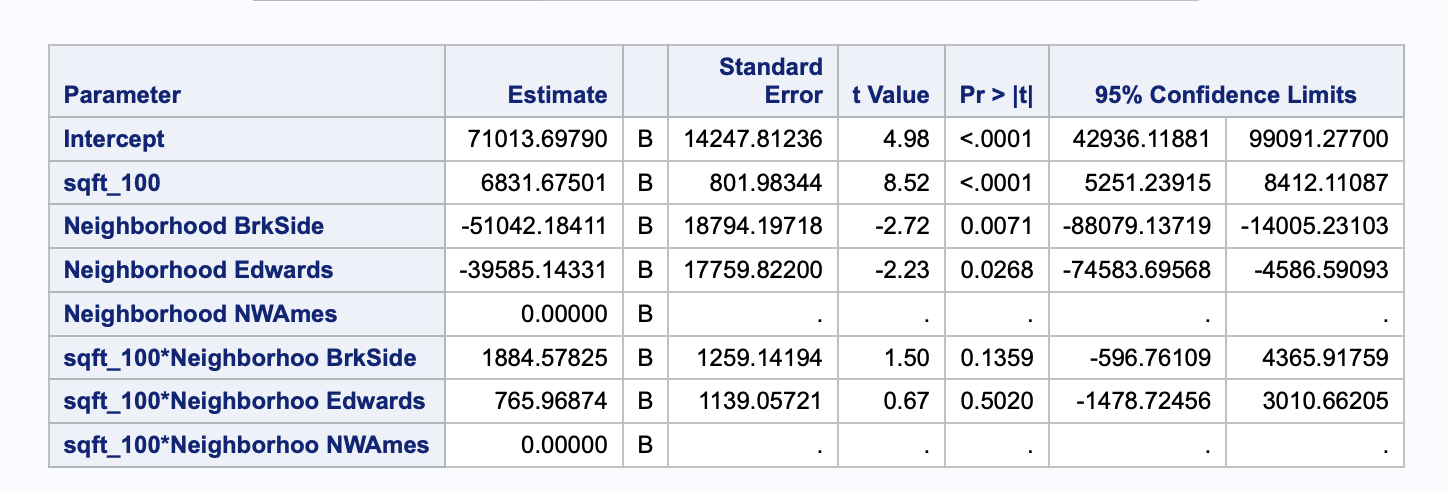
The model is a reasonable fit without transformations, although transformations may be investigated to handle the possible problem with equal standard deviations.

Model

Model Metrics: Adj R2 : 0.685

Parameters: As the parameters were provided, no further analysis on reasoning behind their selection is provided.

Estimates



Overall:    Sales\_price = $71013.70 + ($6831.68 \* SqFt\_100) - ($51042.18 \* BrkSide) - ($39585.14 \* Edwards) + ($1884.58 \* SqFt\_100 \* BrkSide) + ($765.97 \* SqFt\_100 \* Edwards)

NWAmes:     Sales\_price = $71013.70 + $6831.68 \* SqFt\_100

BrkSide:     Sales\_price = $19971.52 + $8716.26 \* SqFt\_100

Edwards:     Sales\_price = $31428.56 + $7597.65 \* SqFt\_100

Interpretation with Confidence Intervals

Northwest Ames neighborhood was selected to be the reference neighborhood for this model.

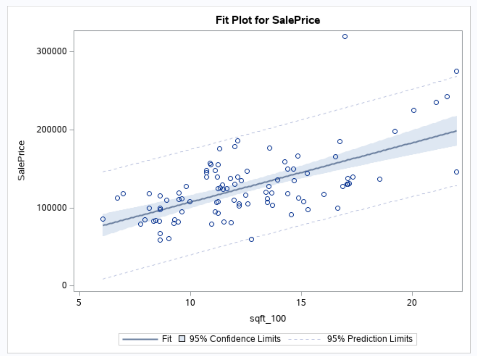
For NWAmes: Every 100 Sqft increase in Greater Living Area increases the Sales Price for the home by $6831.68 (p-value < 0.0001) with a 95% Confidence interval of ($5251.24, $8412.11). If the Greater Living Area has zero sqft then the Sales price for the is $71013.70, but it may not have any practical significance.

For BrkSide: Every 100 sq ft increase in Greater Living Area increases the Sales Price for the home by $8716.26 (p-value < 0.0001) with a 95% Confidence interval of ($7179.73, $10252.78). If the Greater Living Area has zero sqft then the Sales price for the is $19971.52, but it may not have any practical significance.

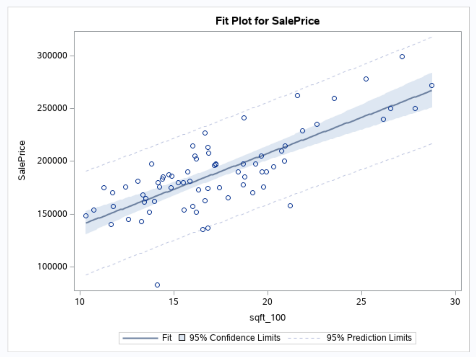
For Edwards:100 sq ft increase in Greater Living Area increases the Sales Price for the home by $7597.65 (p-value < 0.0001) with a 95% Confidence interval of ($5681.88, $9513.40). If the Greater Living Area has zero sqft then the Sales price for the is $31428.56, but it may not have any practical significance.

Resulting p-values suggest that the neighborhood impacts the intercept or base price of a given house. This is statistically significant with p-values of 0.0071 for Brookside and 0.0268 for Edwards. However the impact of square footage on sales price does not fluctuate from neighborhood to neighborhood; square footage in Brookside has a p-value of 0.1359 and square footage in Edwards has a p-value of 0.502.

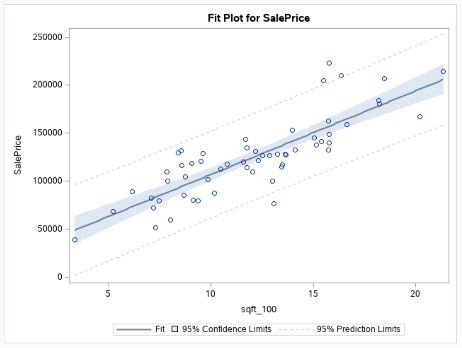
Edwards



NWAmes



Brkside



Conclusion

The sale price of the home differs on which neighborhood the house is located and its square footage. Whereas the general equation for estimated price trends similarly between neighborhoods, the price is offset depending on which neighborhood is under equation. In add cases, home price is positively correlated with the square footage.

Analysis Question 2

Problem

Build the most predictive model for home sales prices in Ames Iowa using these variables: 

* GrLivArea
* KitchenQual
* Neighborhood
* ExterQual
* OverallQual
* BsmtQual
* GarageCars
* GarageQual
* SaleCondition
* YearBuilt

Data

Influential Data Points

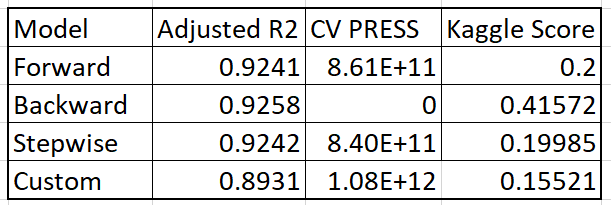
In order to find influential points, a regression model was run for SalePrice versus GrLivArea and an influence table was generated for each observation. The list of DFBETAS was then organized from largest to smallest in order to find the most influential points. According to Belsley, Kuh, and Welsch (1980), the recommended size-adjusted cutoff value is 2/√(n), or in this case 2/√ (1460) = 0.0523. As a result, 47 points were found to overly influence the dataset, and those points are displayed below according to their influence ordered from largest to smallest. These points are present in the Cook’s D graph also presented below, and appear as vertical lines indicating unusually large values. With such a large number of influential data points it is hard to justify removal from the dataset, especially when those points comprise 3.2% of the data.

However, there are two data points that are significant outliers and are able to be removed. When comparing residual graphs from before and after, two points have t-values below -5, indicating they are statistically unlikely to occur. The two points in question are 524 and 1299. Upon removal, we can see that the range of the Cook’s Distances for the data drops from 1.5 to 0.25. We can conclude from this that these two points were causing undue influence to the dataset.

Variable Significance

Insignificant variables: ExterCond (0.06), BsmtFullBath (0.3356), BsmtHalfBath (0.6537), FullBath (0.5), HalfBath (0.1), TotRmsAbvGrd (0.175), FireplaceQu (0.1), GarageType (0.19), GarageFinish (0.5778), PavedDrive (0.73), OpenPorchSF (0.8449), EnclosedPorch (0.6523), X3SsnPorch (0.21), Fence (0.11), MiscFeature (0.9), MiscVal (0.93), MoSold (0.2), SaleCondition (0.20), Heating (0.86), CentralAir (0.17)

Model Selection

Type of Selection****

In total, four models were generated using forward, backward, and stepwise selection methods. The forward model began with zero variables, and added variables until the CVPRESS value for entry candidates was greater than the CVPRESS value for the previous candidate entered. The backwards selection model was generated by starting with all the variables in the model and subtracting variables until the CVPRESS value for the final exit candidate was greater than the CVPRESS value for the final variable removed. Stepwise selection did both steps at the same time for two different variables.

Results from each individual model can be seen above. The CVPRESS for backwards selection is zero due to a beginning CVPRESS of zero; once the computer detected a CVPRESS greater than zero it halted the process. Though backward selection has the best R-squared value, it performed the worst on kaggle. This is likely due to overfitting the model to the training data through the inclusion of too many junk variables. The forward and stepwise models did surprisingly well with kaggle scores of 0.2 and 0.1999 respectively.

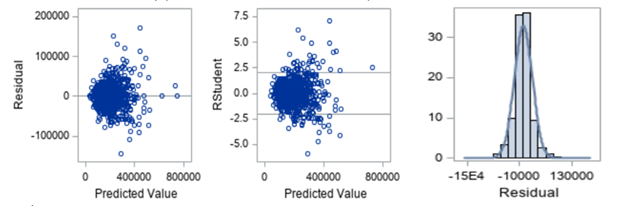
A custom model was generated and performed the best out of all models. Despite having the lowest R-square value of 0.89 and the highest CVPRESS value of 1.11E+12, it resulted in a kaggle score of 0.1556. The resulting regression equation included 178 parameters from 18 effects. The full list of parameters and coefficient estimates can be found in the Appendix under A5: Parameters and estimated coefficients for custom model Question 2. The variables chosen for the final model were: GrLivArea\*KitchenQua, GrLivArea\*Neighborhood, ExterQual, GrLivArea\*OverallQual, BsmtQual, GarageCars, SaleCondition, YearBuilt, and YearRemodAdd.

In order to construct the model, the variables were first screened based on their adjusted R-square values when compared one-on-one to Sale Price. From these models, the computer performed a stepwise selection to choose the most significant variables. Next the model was refined to eliminate variables that described the same qualities of a house, to group together variables that had effects on each other, and to include variables that described leftover variance in the model. The following examples reflect each scenario respectively:

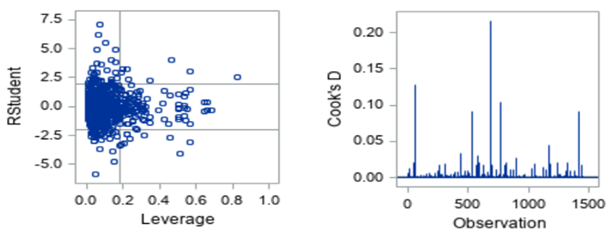
* It was found that GrLivArea, X1stFlrSF, and TotalBsmtSF were closely related, as the size of the basement and the size of the general living area both rely on the first floor square footage. In order to avoid redundancy, X1stFlrSF and TotalBsmtSF were removed from the model.
* It was found that the neighborhood and overall quality have an impact on sales price that is related to the general living area; home purchasers care about the quality of the living area and the neighborhood that the house is present in. As a result the computer chose to group these variables together.
* Some variables that were screened out due to having low initial R-squared values explained certain aspects of pricing information that were not captured by other variables; an example of this is YearBuilt, the year the house was built.

Once the model was run on the test set, there were SalePrice values that were found to be negative. To counteract this, the mean of the data from the results was taken and any negative predicted value was replace with the value of 180,932. It came to our attention that these negative predictions constituted opportunity for improvement through a second fit model. It also came to our attention that the negative values were included in figuring the mean; however when the negative numbers were removed from the mean calculation it was found that this decreased the model’s kaggle score.

Checking Assumptions



By looking at the residual plot, the residuals are distributed evenly on both sides of the zero line. Although there is a dense cloud around the predicted value of 250,000-300,000 that might suggest some evidence of non-constant variance. The studentized residual plot shares a similar resemblance to the residual plot. It identifies potential outlying observations. There are a few outlying data points around the predicted value of 400,000 that might provide some evidence against normality and require further examination. On the other hand, the histogram of residuals does not provide strong evidence against normal distribution of the residuals.



Based on the displays, there are a few observations with high leverage. However, there is no visual evidence that there is any observation with both high leverage and high residual.

Conclusion

It was found that a custom model did the best at predicting housing prices in the test dataset. The variables that had the most effect on the model were GrLivArea\*KitchenQua, GrLivArea\*Neighborhood, ExterQual, GrLivArea\*OverallQual, BsmtQual, GarageCars, SaleCondition, YearBuilt, and YearRemodAdd.

Many of the variables that were initially influential were ultimately removed from the model as other variables did a better job of explaining SalePrice. Though there were some points that had higher than normal leverage, it was found that they were not acting as outliers for the dataset.

Ultimately human interaction was needed to optimize the computer-generated models.

Appendix

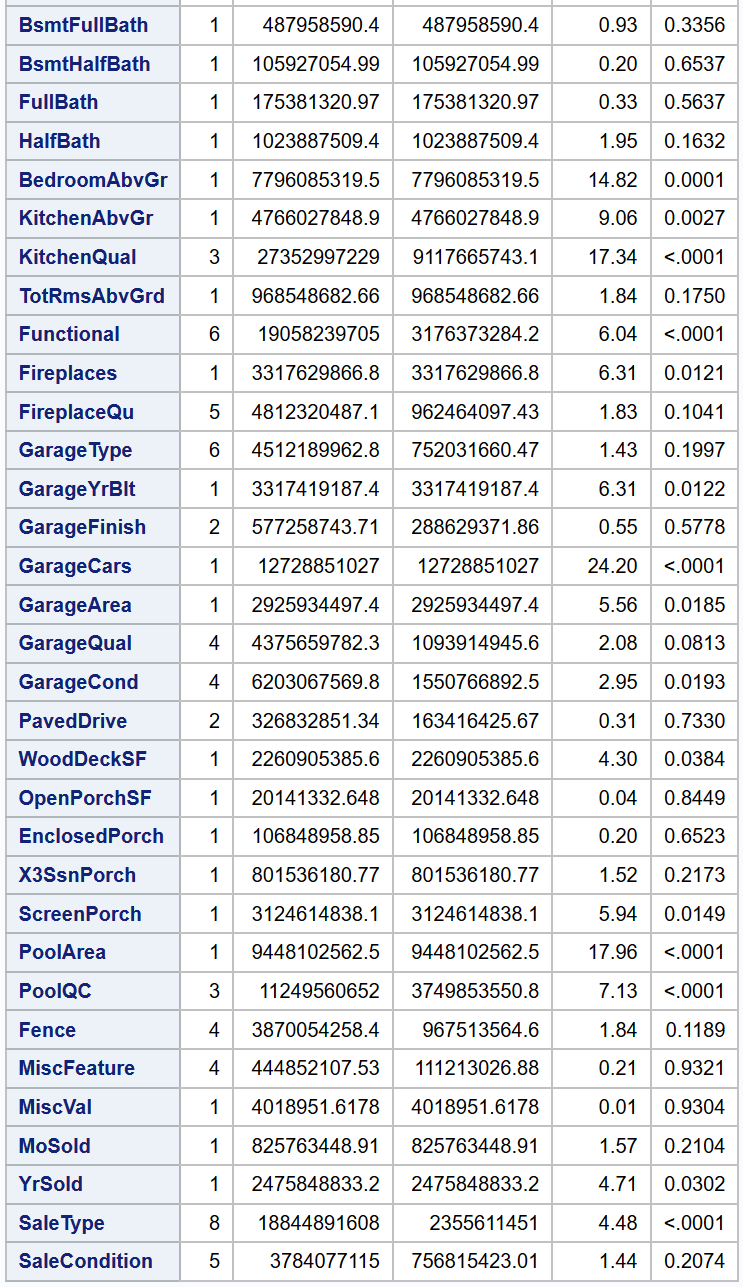
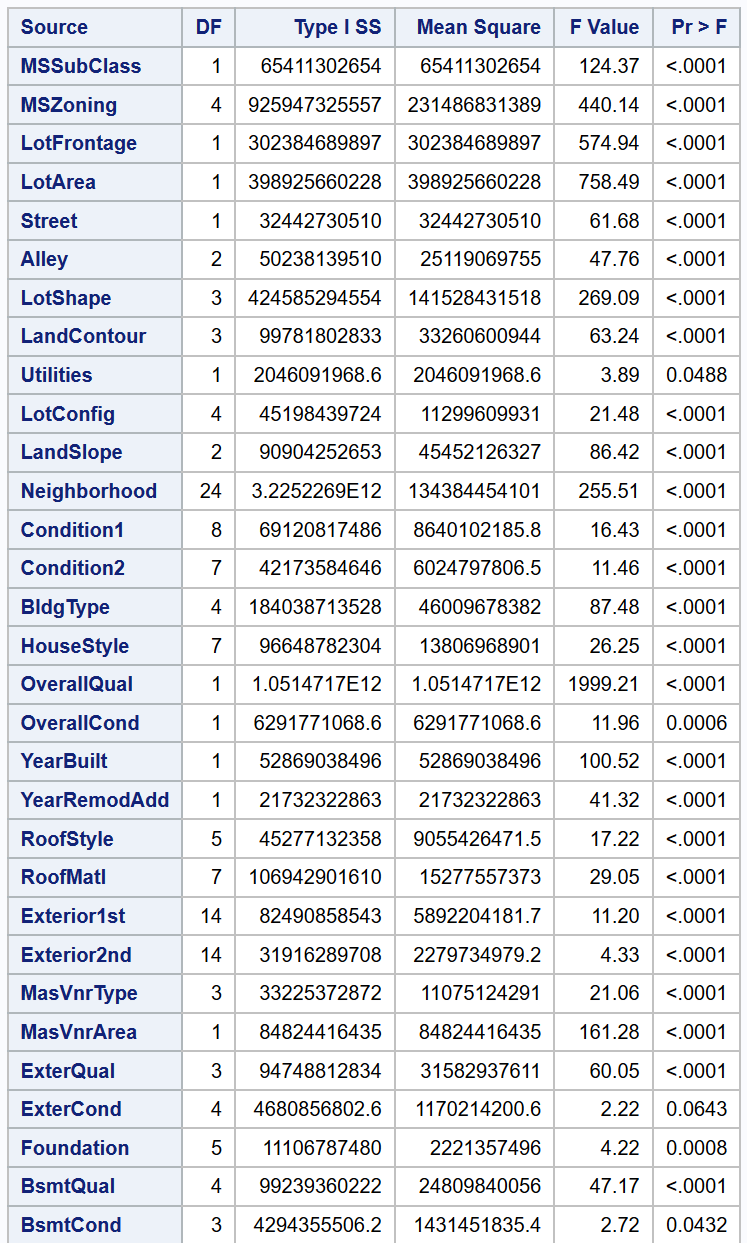
Github Repository

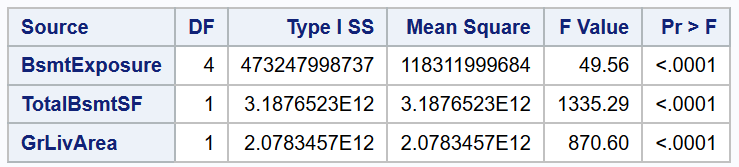
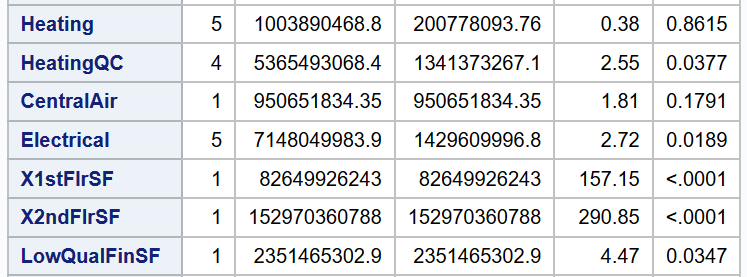
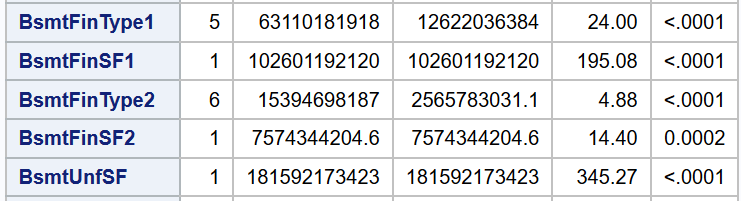
A1: Github Repository

<https://github.com/jules-stacy/StatsDS6371_proj1.git>

Data: Variable Significance Outputs

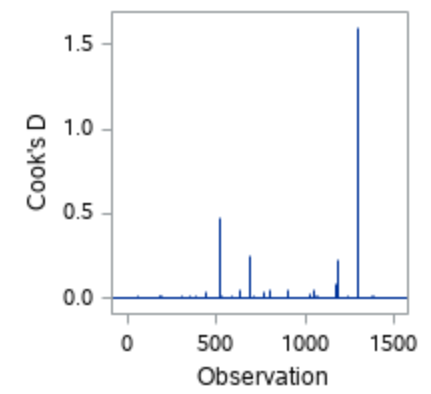
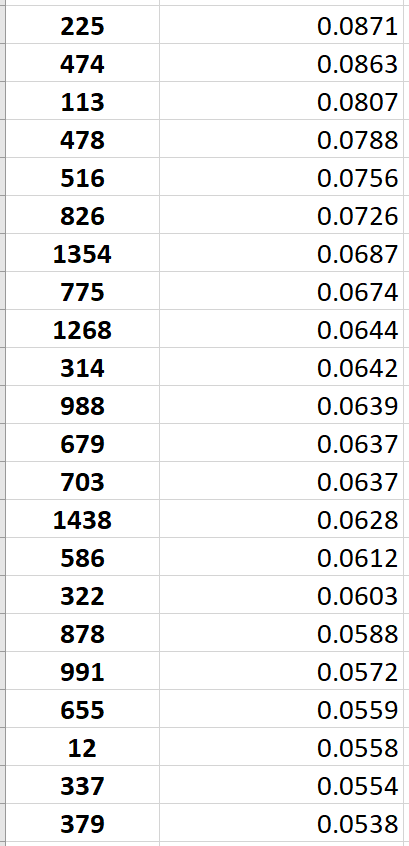
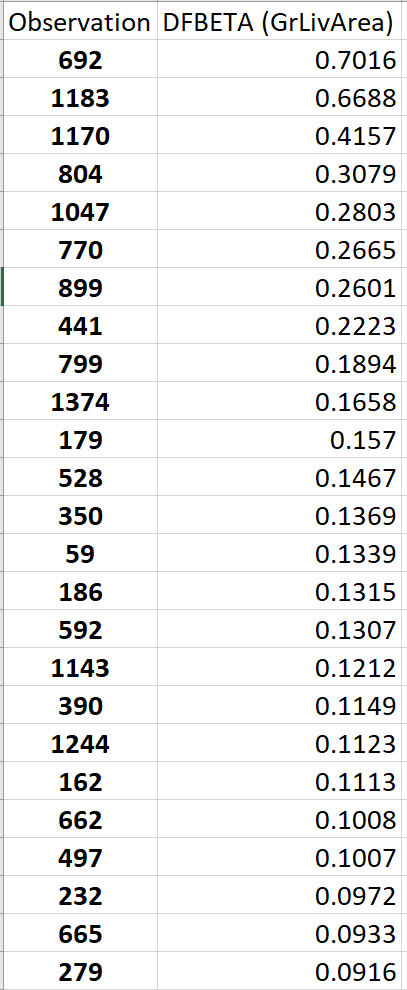
A2: P-values for each variable regressed against SalePrice

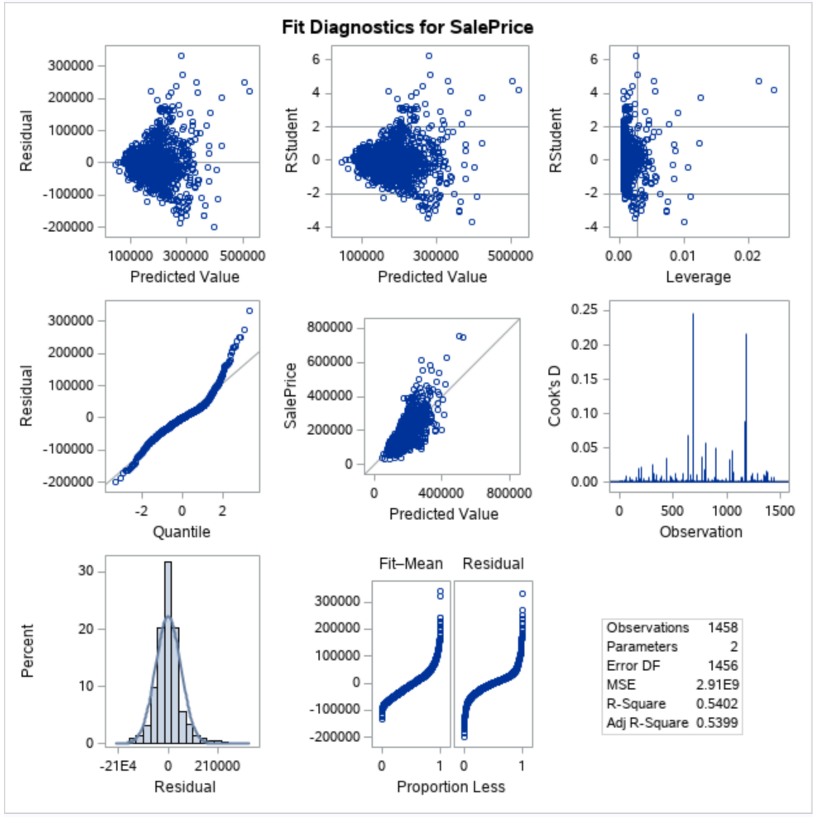
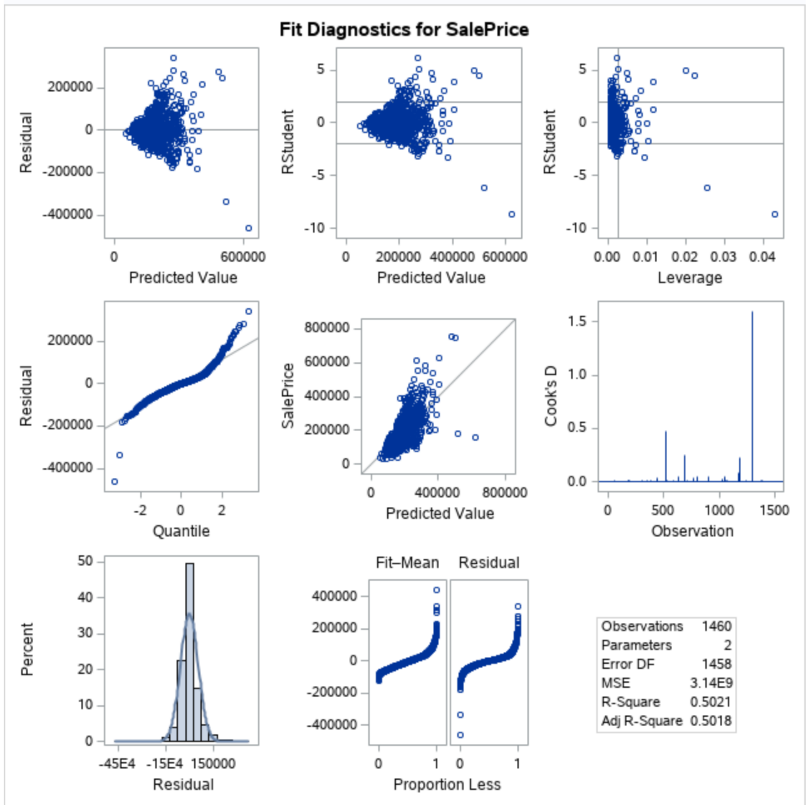




Data: Influential Data Point Graphs

A3: Influential data points analyses





Analysis 1

A4: Question 1 Code

/\* Loading Data \*/

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

FILENAME REFFILE '/folders/myshortcuts/StatsSAScode/Project/train\_clean.csv';

PROC IMPORT DATAFILE=REFFILE

    DBMS=CSV

    OUT=WORK.train;

    GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.train; RUN;

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

data train2;

set train;

log\_price = log(SalePrice);

sqft\_100 = GrLivArea/100;

if (id = 524 or id = 1299) then delete;

run;

/\* Data Filtering and Preparation \*/

proc glm data = sub\_house;

class neighborhood (ref= "NWAmes");

model SalePrice = sqft\_100 | neighborhood / solution clparm;

run;

/\* Initial Scatterplot (non-grouped) \*/

proc sgplot data=train;

reg x=sqft\_100 y=SalePrice;

run;

/\* Initial Cook's D and Leverage (non-grouped) (with outliers) \*/

proc reg data=train;

model SalePrice = sqft\_100  / clm cli;

run;

/\* Remove the two outliers (i.e. model homes) \*/

data train;

set train;

if (id = 524 or id = 1299) then delete;

run;

/\* Cook's D and Leverage Again (non-grouped) (without outliers) \*/

proc reg data=train;

model SalePrice = sqft\_100  / clm cli;

run;

/\* Calculate the regression line \*/

proc sgplot data=train;

reg x=sqft\_100 y=SalePrice /group=neighborhood;

run;

/\* Calculate all estimate and confidence intervals \*/

/\* This also generates the graphs for assumption checking \*/

proc glm data = train plots=all;

class neighborhood (ref= "NWAmes");

model SalePrice = sqft\_100 | neighborhood / solution clparm;

run;

/\*CI by Neighborhood\*/

proc glm data = train ;

where neighborhood eq 'Edwards';

model SalePrice = sqft\_100 / solution clparm;

run;

proc glm data = train ;

where neighborhood eq 'NWAmes';

model SalePrice = sqft\_100 / solution clparm;

run;

proc glm data = train ;

where neighborhood eq 'BrkSide';

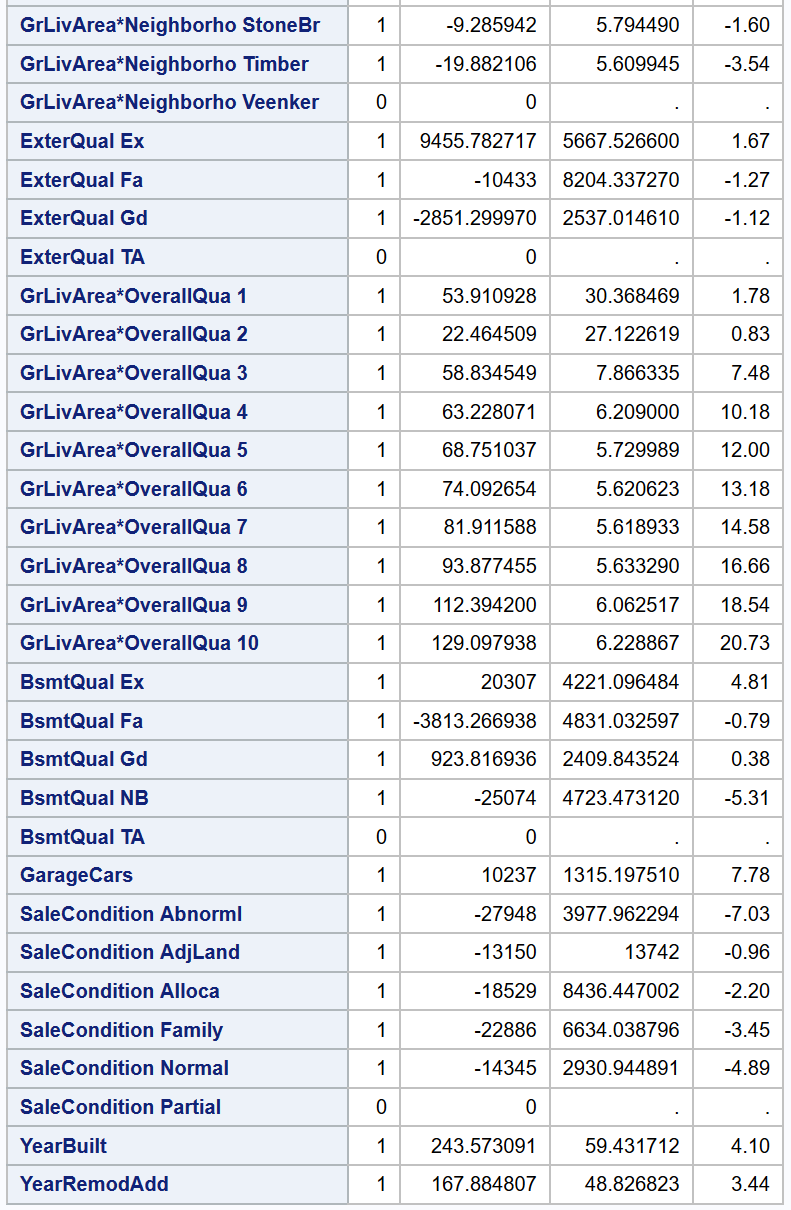
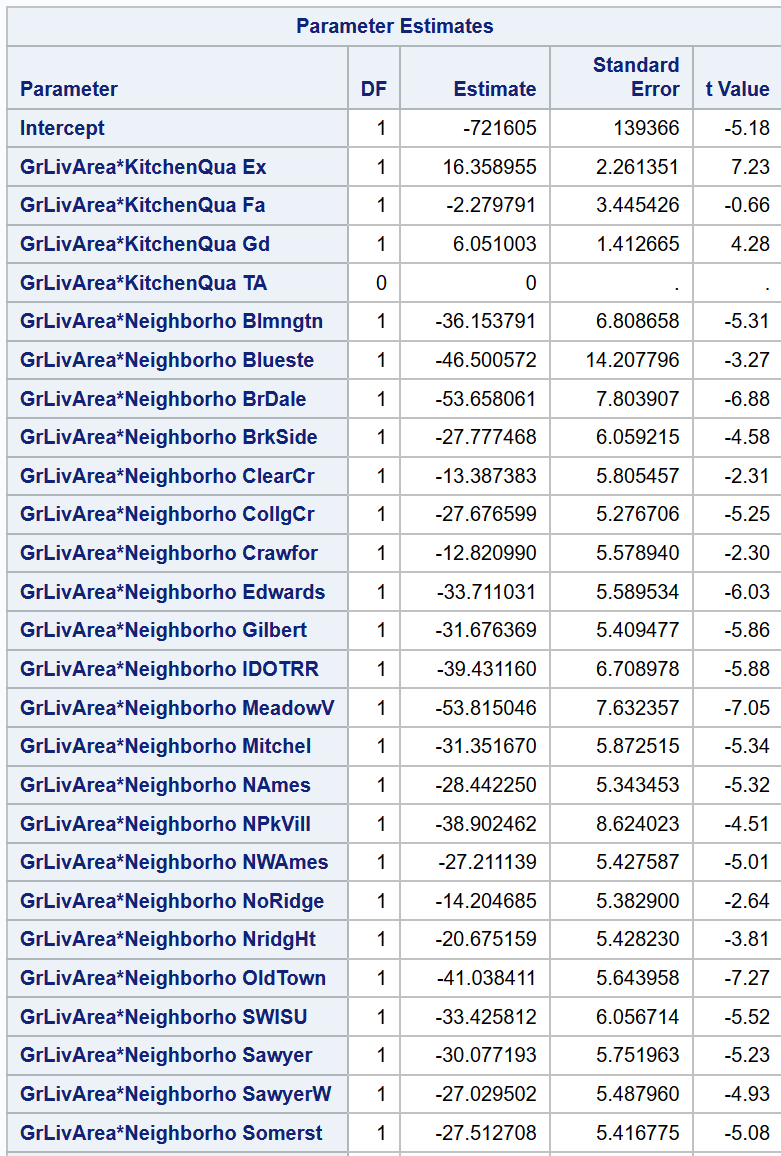
model SalePrice = sqft\_100 / solution clparm;

run;

Analysis 2

Custom Model

A5: Parameters and estimated coefficients for custom model Question 2



Code

A6: Code used for Question 2

/\*prediction data import\*/

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

FILENAME REFFILE '/folders/myshortcuts/StatsSAScode/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);

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

FILENAME REFFILE '/folders/myshortcuts/StatsSAScode/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);

data combo;

   set train test;

run;

data combo;

set combo;

if MSZoning = "NA" then MSZoning = "NZ";

if LotFrontage = "NA" then LotFrontage = 0;

if Alley = "NA" then Alley = "NL";

if Utilities = "NA" then Utilities = "NU";

if Exterior1st = "NA" then Exterior1st = "Other";

if Exterior2nd = "NA" then Exterior2nd = "Other";

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

if MasVnrArea = "NA" then MasVnrArea = 0;

if BsmtQual = "NA" then BsmtQual = "NB";

if BsmtCond = "NA" then BsmtCond = "NB";

if BsmtExposure = "NA" then BsmtExposure = "NB";

if BsmtFinType1 = "NA" then BsmtFinType1 = "NB";

if BsmtFinType2 = "NA" then BsmtFinType2 = "NB";

if BsmtFinSF1 = "NA" then BsmtFinSF1 = 0;

if BsmtFinSF2 = "NA" then BsmtFinSF2 = 0;

if BsmtUnfSF = "NA" then BsmtUnfSF = 0;

if TotalBsmtSF = "NA" then TotalBsmtSF = 0;

if BsmtFullBath = "NA" then BsmtFullBath = 0;

if BsmtHalfBath = "NA" then BsmtHalfBath = 0;

if KitchenQual = "NA" then KitchenQual = "NK";

if Functional = "NA" then Functional = "NH";

if Electrical = "NA" then Electrical = "NE";

if FireplaceQu = "NA" then FireplaceQu = "NF";

if GarageType = "NA" then GarageType = "NG";

if GarageYrBlt = "NA" then GarageYrBlt = 0;

if GarageFinish = "NA" then GarageFinish = "NG";

if GarageCars = "NA" then GarageCars = 0;

if GarageArea = "NA" then GarageArea = 0;

if GarageQual = "NA" then GarageQual = "NG";

if GarageCond = "NA" then GarageCond = "NG";

if PoolQC = "NA" then PoolQC = "NP";

if Fence = "NA" then Fence = "NF";

if MiscFeature = "NA" then MiscFeature = "NM";

if SaleType = "NA" then SaleType = "Other";

if (id = 524 or id = 1299) then delete;

run;

/\* Computer Selection~~~~\*/

/\*forward selection\*/

proc glmselect data = combo plots(stepaxis = number) = (criterionpanel ASEplot);

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

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            X1stFlrSF              X2ndFlrSF           LowQualFinSF    GrLivArea              BsmtFullBath     BsmtHalfBath    FullBath              HalfBath             BedroomAbvGr  KitchenAbvGr             KitchenQual       TotRmsAbvGrd  Functional           Fireplaces           FireplaceQu       GarageType              GarageYrBlt       GarageFinish      GarageCars              GarageArea        GarageQual       GarageCond              PavedDrive        WoodDeckSF     OpenPorchSF              EnclosedPorch   X3SsnPorch     ScreenPorch              PoolArea            PoolQC Fence    MiscFeature              MiscVal MoSold YrSold   SaleType     SaleCondition

/selection = Forwards (select=cv) cvmethod=random(5) stats=adjrsq;

output out = resultsf p = predict;

run;

data results\_f;

set resultsf;

if Predict > 0 then SalePrice = Predict;

if Predict < 0 then SalePrice = 180932;

keep id SalePrice;

where id>1460;

;

proc export data=WORK.results\_f

outfile='/folders/myshortcuts/StatsSAScode/Project/results\_f.csv'

dbms=csv

replace;

run;

/\*backward selection\*/

proc glmselect data = combo plots(stepaxis = number) = (criterionpanel ASEplot);

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

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            X1stFlrSF              X2ndFlrSF           LowQualFinSF    GrLivArea              BsmtFullBath     BsmtHalfBath    FullBath              HalfBath             BedroomAbvGr  KitchenAbvGr             KitchenQual       TotRmsAbvGrd  Functional           Fireplaces           FireplaceQu       GarageType              GarageYrBlt       GarageFinish      GarageCars              GarageArea        GarageQual       GarageCond              PavedDrive        WoodDeckSF     OpenPorchSF              EnclosedPorch   X3SsnPorch     ScreenPorch              PoolArea            PoolQC Fence    MiscFeature              MiscVal MoSold YrSold   SaleType     SaleCondition

/selection = Backward (select=CV) cvmethod=random(5) stats=adjrsq;

output out = resultsb p = predict;

run;

data results\_b;

set resultsb;

if Predict > 0 then SalePrice = Predict;

if Predict < 0 then SalePrice = 180932;

keep id SalePrice;

where id>1460;

;

proc export data=WORK.results\_b

outfile='/folders/myshortcuts/StatsSAScode/Project/results\_b.csv'

dbms=csv

replace;

run;

/\*stepwise selection\*/

proc glmselect data = combo plots(stepaxis = number) = (criterionpanel ASEplot);

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

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            X1stFlrSF              X2ndFlrSF           LowQualFinSF    GrLivArea              BsmtFullBath     BsmtHalfBath    FullBath              HalfBath             BedroomAbvGr  KitchenAbvGr             KitchenQual       TotRmsAbvGrd  Functional           Fireplaces           FireplaceQu       GarageType              GarageYrBlt       GarageFinish      GarageCars              GarageArea        GarageQual       GarageCond              PavedDrive        WoodDeckSF     OpenPorchSF              EnclosedPorch   X3SsnPorch     ScreenPorch              PoolArea            PoolQC Fence    MiscFeature              MiscVal MoSold YrSold   SaleType     SaleCondition

/selection = Stepwise (select=cv) cvmethod=random(5) stats=adjrsq;

output out = resultss p = predict;

run;

data results\_s;

set resultss;

if Predict > 0 then SalePrice = Predict;

if Predict < 0 then SalePrice = 180932;

keep id SalePrice;

where id>1460;

;

proc export data=WORK.results\_s

outfile='/folders/myshortcuts/StatsSAScode/Project/results\_s.csv'

dbms=csv

replace;

run;

/\*custom model, hybrid computer human selection~~~~~\*/

/\*best code so far: final model \*/

proc glmselect data = combo plots(stepaxis = number) = (criterionpanel ASEplot);

class Neighborhood OverallQual ExterQual KitchenQual BsmtQual SaleCondition GarageQual;

model SalePrice =   GrLivArea | KitchenQual GrLivArea |Neighborhood ExterQual | OverallQual GrLivArea | OverallQual BsmtQual GarageCars | GarageQual  SaleCondition | OverallQual YearBuilt YearRemodAdd

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

output out = results p = predict;

run;

proc means data=results;

var SalePrice;

run;

data results2;

set results;

if Predict > 0 then SalePrice = Predict;

if Predict < 0 then SalePrice = 180932;

keep id SalePrice;

where id>1460;

;

proc export data=WORK.results2

outfile='/folders/myshortcuts/StatsSAScode/Project/results2.csv'

dbms=csv

replace;

run;

Sources

1. Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: John Wiley & Sons.