Table of Contents

[Data Cleansing 1](#_Toc480732098)

[Question 1 2](#_Toc480732099)

[Problem Statement 2](#_Toc480732100)

[Specify the model 2](#_Toc480732101)

[Check Assumptions 2](#_Toc480732102)

[Compare Competing Models 2](#_Toc480732103)

[Parameter Interpretation 2](#_Toc480732104)

[Conclusion 2](#_Toc480732105)

[Question 2 2](#_Toc480732106)

[Problem Statement 2](#_Toc480732107)

[Model Selection 2](#_Toc480732108)

[Forward Selection 2](#_Toc480732109)

[Assumption Checking 2](#_Toc480732110)

[Backward Elimination & Custom 2](#_Toc480732111)

[Assumption Checking 2](#_Toc480732112)

[Stepwise Selection 2](#_Toc480732113)

[Assumption Checking 2](#_Toc480732114)

[Comparing Competing Models 2](#_Toc480732115)

[Conclusion 2](#_Toc480732116)

[Appendix 2](#_Toc480732117)

# **No table of figures entries found.**

# Data Cleansing

The initial training data contained 'NA' for data items that were missing. Numeric fields that contain NA were treated as character by SAS import. Consequently, we had to convert the NAs to the missing value symbol, ".", in SAS, and then we exported the data and reimported the data. The same procedure had to be applied to the test data.

# Question 1

# Problem Statement

Century 21 has requested analysis of home sales data in regards to the region where they sell. We are to provide them with estimate of how the sale price of the house is related to the square footage of the living area of the house and if it depends on which neighborhood the house is located in.

## Specify the model

Problem Statement

The model used to represent the analysis is represented as follows.

Sales Price = β0 + β1\*(Square footage in units of 100 sqft) + β 2\*(Edwards) +β3\*(Brookside)+β4\*(Edwards Square footage interaction term)+β5\*(Brookside Square footage interaction term)

|  |  |
| --- | --- |
| Parameter | Estimate |
| β0 | 60860 |
| β1 | 6492.91199 |
| β2 | -15050 |
| β3 | -11945 |
| β4 | 2166.05748 |
| β5 | 3284.66699[[1]](#footnote-1) |

Table See attached Parameter estimates table in appendix

Here, North Ames Neighborhood is used as the reference neighborhood. The interaction terms are also centered to reduce variation inflation.

## Check Assumptions

**Linear relationship:**

A plot of the sales price vs the square footage reveals a possible linear relationship. (Plot show in appendix)

**Patterns in the residual plots:**

The Corresponding residual plots show no distinguishable patterns in the residuals. Patterns would suggest a lack of fit for the linear model and unequal variance. (Residual plots in appendix)

**Normality:**

The corresponding Q-Q plot and histogram show a normal distribution and little deviation from normality. Hence, the normality assumption is met. (Plot in Appendix)

**Multicollinearity:**

The corresponding correlation matrix and variation inflation factors show little collinearity between the variables exists in the model. (Plot in appendix)

## Compare Competing Models

**Model 1**

The simplest model for analysis includes terms for the continuous variable(square footage) and the levels of the categorical variable(Neighborhood). This model has an adj. R-sq value of 0.3917. All terms are statistically significant.

**Model 2**

Logically, one might think that there could be a relationship between square footage and the neighborhood in which a house sits. This prompts the test for statistically significant interaction terms in the model. Upon doing this, the model found has a higher R-sq value of 0.44, an 11% increase. However, in this model the Variance Inflation terms are significantly higher for the categorical and interaction terms. Given that the interaction terms are based in part from the categorical terms, this could be expected. The solution is to center the interaction terms. Additionally, the coefficient for one of the categorical terms is now no longer statistically significant. This could be due to the apparent collinearity.

**Model 3**

Centering the interaction terms results in a model with low values for variation inflation and statistically significant coefficients.

**Model 4**

Upon further investigation, there are 2 observations that are highly influential, Cook’s D well above 1. After removing these 2 observations, we arrive at a new model with an adj R-sq at 0.5060 with all statistically significant coefficients. This is the final model that we use for our analysis.

## Parameter Interpretation

The final model consists of 6 coefficients, their interpretations are listed below.

|  |  |
| --- | --- |
| Coefficient | Interpretation |
| β0 | This is the intercept of the linear regression model. For all variables held constant, if there is a zero square footage, this would represent the value of the property in the reference neighborhood (Names). Physically, this might represent the value of the plot in which a property sits. |
| β1 | This is the rate of change of the sales price given the square footage with all things held constant. |
| β2 | This term represents the difference in mean sale price between Edwards and N Ames with all things held constant. |
| β3 | This term represents the difference in mean sale price between Brookside and N Ames with all things held constant. |
| β4 | This is the coefficient for the interaction term between the Edwards neighborhood and the square footage. This term accounts for the relationship that Edwards has on the square footage of a home. |
| β­5 | This is the coefficient for the interaction term between the Brookside neighborhood and the square footage. This term accounts for the relationship that Edwards has on the square footage of a home. |

## Conclusion

After investigation, we have created a model that effectively describes the relationship between the sales price and square footage of a home in the three neighborhoods, North Ames, Edwards, and Brookside. This model has a statistically significant dependence on the neighborhood of the property. The relationship takes the following form with the coefficient values shown below.

Sales Price = β0 + β1\*(Square footage in units of 100 sqft) + β 2\*(Edwards) +β3\*(Brookside)+β4\*(Edwards Square footage interaction term)+β5\*(Brookside Square footage interaction term)

|  |  |
| --- | --- |
| Parameter | Estimate |
| β0 | 60860 |
| β1 | 6492.91199 |
| β2 | -15050 |
| β3 | -11945 |
| β4 | 2166.05748 |
| β5 | 3284.66699 |

This model assumes that properties are within the 3 listed neighborhoods. The effectiveness of this model is shown in the attached analysis of variance F-test.

# Question 2

Problem Statement  
Century 21 has asked us to construct the most predictive model for sales prices in all of Ames Iowa. They have provided 1,460 instances of home sales. Each home sale comprises 80 explanatory variables. Not all variables were collect, and missing values exist throughout the data set. In order to create the most predictive model, we will three forms of automatic variable selections: forward selection, backwards eliminations, and stepwise selection.

## Model Selection

## Forward Selection[[2]](#footnote-2)

The forward selection began with the following explantory variables: MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold.

After running the model, the following explanatory variables were selected by SAS: OverallQual GrLivArea BsmtFinSF1 GarageCars MSSubClass YearRemodAdd MasVnrArea LotArea KitchenAbvGr BedroomAbvGr TotRmsAbvGrd YearBuilt OverallCond BsmtFullBath ScreenPorch.

### Model Adjustment

The VIF for GrLivArea and TotRmsAbvGrd showed collinearity. Consequently TotRmsAbvGround was removed from the model, which decreased R-Squared and Adjusted R-Squared by .0021 but resolved the collinearity issue. Having resolved for Collinearity, it was observed that KitchenAbvGr was no statistically significant, so it was removed from the model, which decreased R-Square by .0001 without affecting Adjusted R-Square.

### Assumption Checking

Having looked as the residual plots, we noticed that the residuals were distributed around zero. The QQ plot and histogram show evidence of normality. However, there was a highly influential and leverage point with a Cook's D of a 3.2902. We identified this point as id=1299. This point was excluded from the analysis inasmuch as it denoted as 12 bedroom home, which represent a different population from the under consideration. The removal of the data point increase RSquared by .0317 and increased Adj-RSquared by .0320.

### Forward Selection Final Model and Adjusted R-Squared

SalePrice = -1168370 + 18049OveralQual + 70.83GrLivArea + 24.80 BsmtFinSF1 + 8231GarageCars + -209.74 MSSubClass + 201.877 YearRemodAdd + 32.36MasVnrArea + .59LotArea - 9977.25BedroomAbvGr + 361.09YearBuilt + 4112.28 OverallCond + 4035BsmtFullBath + 40.35ScreenPorch

The adjusted R2 is .8360.

## Backward Elimination (a.k.a. Custom)

The backward elimination began with the same variables as the forward selection. After running the model, the following explanatory variables were selected by SAS: MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 \_1stFlrSF \_2ndFlrSF BsmtFullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd GarageCars ScreenPorch.

### Model Adjustment

In the initial SAS model, TotRmsAbvGrd has a VIF greater than 4. We removed the instance that we found in the forward selection model adjust (I.e., id-1299). Having done so, increased R Squred by .0321 and increased Adjusted R Squae by .0324, but it did not resolve the collinearity problem. We removed TotRmsAbvGrd from the model and this resolved the collinearity; however, R-Squared and Adjusted R-Squared decreased by .002.

## Assumption Checking

Having looked as the residual plots, we noticed that the residuals were distributed around zero. The QQ plot and histogram show evidence of normality. Having already taken care of the high leverage point, we didn't find any other points with a Cook's D greater than 1.

### Backward Elimination Final Model and Adjusted R-Squared

SalePrice = -868542 + -154.19MSSubClass+.53LotArea+17910OverallQual OverallCond+407.65YearBuilt+28.25MasVnrArea+20.89BsmtFinSF1+83.27\_1stFlrSF+69.11\_2ndFlrSF+4466.87BsmtFullBath+-9174.28BedroomAbvGr+-11719KitchenAbvGr+7848.91GarageCars+33.80ScreenPorch

The adjusted RSquared is .8362.

## Stepwise Selection

The stepwise elimination began with the same variables as the backward elimination. After running the mode, the following explanatory variables were selected by SAS: MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold

### Model Adjustment

In the initial regression, all of the parameter estimates were significant; hwoever, TotRmsAbvGrd and GrLivArea appeared to be collinear. The instance with the high cook's d . (I.e., id=1299) was removed and the model was rerun. This increased the R-Square and Adjusted R-Square, but did not resolve the collinearity. We then removed TotRmsAbvGrd from the model. Having removed TotRmsAbvGrd resolved the collinearity, but KitchenAbvGrd became statistically insignificant so it was removed from the model, which results in a drop of .0001 for R Squared and Adjusted R-Squared

## Assumption Checking

Having looked as the residual plots, we noticed that the residuals were distributed around zero. The QQ plot and histogram show evidence of normality. Having already taken care of the high leverage point, we didn't find any other points with a Cook's D greater than 1.

## Comparing Competing Models

|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Models | Adjusted R2 | CV Press | Kaggle Score |
| Forward | .8360 | 1.26E12 | .30167 |
| Backward | .8362 | 1.28E12 | .21992 |
| Stepwise | .8348 | 1.33E12 | .21992 |
| Custom[[3]](#footnote-3) | .8362 | 1.28E12 | .21992 |

## Looking athe aformentioned models, we ar choosing the back elimination model because it has the highest Adjusted R2 and tied for the lowest Kaggle score among our models.

## Conclusion

The best predictive model is the one that we found using backward elimination after we have adjusted for collinearity and high Cook's D.

SalePrice = -868542 + -154.19MSSubClass+.53LotArea+17910OverallQual OverallCond+407.65YearBuilt+28.25MasVnrArea+20.89BsmtFinSF1+83.27\_1stFlrSF+69.11\_2ndFlrSF+4466.87BsmtFullBath+-9174.28BedroomAbvGr+-11719KitchenAbvGr+7848.91GarageCars+33.80ScreenPorch

# Appendix

Figures

# Figures

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Q1 Analysis |  |  |  |  |
| Correlation matrix | Residual plots | Fit Diagnostics | Parameter Estimates | Analysis of variance |
|  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Forward Selection |  |  |  |  |  |
| Initial Selection | Remove TotRmsAbvGround | Remove KitchenAbvGrd | Residuals | Cook's D | Final Model |
|  |  |  | Inserting image... |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Backward Elimination |  |  |  |  |  |
| Initial Selection | Initial Regression | Remove Leverage Pt | Remove TotRmsAbvGrd | Residuals | Final Model |
|  |  |  |  | Inserting image... | Inserting image... |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stepwise Selection |  |  |  |  |  |
| Initial Selection | Initial Regression | Remove Leverage Pt | Removal of TotRmsAbvGrd | Removal of KitchenAbvGrd | Final Model |
| Inserting image... | Inserting image... |  | Inserting image... |  |  |
|  |  |  |  |  |  |
| Residuals |  |  |  |  |  |
|  |  |  |  |  |  |

Question 1 Code

|  |
| --- |
| \*Import data from csv file;  **proc** **import** datafile='c:\SASDATA\ExpStat\train1.csv' dbms=dlm out=RawTrain replace;  delimiter=",";  getnames=yes;  guessingrows=**1461**;  **run**;  \*get preliminary information of the type of data present in the file;  **proc** **contents** data=RawTrain;  **run**;  \*view formated data ensure correct import;  **proc** **print** data=train;  **run**;  \*slice data set to obtain narrow data set specific to current analysis;  **data** TrainAmes;  set RawTrain;  keep SalePrice GrLivArea Neighborhood Id;  **run**;  \*ensure proper slicing;  **proc** **contents** data=TrainAmes;  **run**;  \*further slice data set to only interested neighborhoods;  **data** TrainAmes2;  set TrainAmes;  if (Neighborhood = 'NAmes') or (Neighborhood = 'Edwards') or (Neighborhood = 'BrkSide');  **run**;  **proc** **print** data = TrainAmes2;  **run**;  \*give categorical variables numerical values for easier management;  **data** TrainAmes3;  set TrainAmes2;  number = **.**;  if (Neighborhood = 'NAmes') then number = **1**;  if (Neighborhood = 'Edwards') then number = **2**;  if (Neighborhood = 'BrkSide') then number = **3**;  **run**;  **proc** **print** data = TrainAmes3;  **run**;  \*sort by neighborhood;  **proc** **sort** data=TrainAmes3;  by number;  **proc** **print** data = TrainAmes3;  **run**;  \*investigate for linear relationship;  **proc** **gplot** data = TrainAmes3;  plot SalePrice\*GrLivArea=number;  **run**;  \*/format data to be in correct units;  **data** TrainAmes4;  set TrainAmes3;  GrLivArea = GrLivArea/**100**;  **run**;  **proc** **print** data=TrainAmes4;  **run**;  \*create dummy variables for anaysis of categorical levels;  **data** TrainAmes5;  set TrainAmes4;  if number = **2** then d1 = **1**; else d1 = **0**;  if number = **3** then d2 = **1**; else d2 = **0**;  int1 = d1\*GrLivArea; int2 = d2\*GrLivArea;  **run**;  **proc** **print** data=TrainAmes5;  **run**;  \*run preliminary regression model;  **proc** **reg** data=TrainAmes5;  model SalePrice = GrLivArea d1 d2 / R CLB;  title 'Regression of Sale Price on Dummy Coded Therapy';  **run**;  \*run regression model with added interaction terms;  **proc** **reg** data=TrainAmes5;  model SalePrice = GrLivArea d1 d2 int1 int2/ R CLB VIF;  title 'Regression of Sale Price on Dummy Coded Therapy';  **run**;  \*calculate means for centerig terms;  **proc** **means** data=TrainAmes5;  var GrLivArea d1 d2;  **run**;  \*center interaction terms;  **data** TrainAmes6;  set TrainAmes5;  cent1 = (GrLivArea - **13.018**)\*(d1-**0.261**);  cent2 = (GrLivArea - **13.018**)\*(d2-**0.151**);  **run**;  \*run regression with centered interaction terms;  **proc** **reg** data=TrainAmes6;  model SalePrice = GrLivArea d1 d2 cent1 cent2/ R CLB VIF;  **run**;  \*check correlation of variables;  **proc** **corr** data=TrainAmes6;  var GrLivArea d1 d2 cent1 cent2;  **run**;  \*remove highly influencial observations;  **data** TrainAmes7;  set TrainAmes6;  if Id~=**1299**;  if Id~=**524**;  **run**;  \*run regression after removing influencial observations;  **proc** **reg** data=TrainAmes7;  model SalePrice = GrLivArea d1 d2 cent1 cent2/ R CLB VIF;  **run**; |

## Question 2 Code

/\* import initial data \*/  
 proc import datafile='/home/herreram1/sasuser.v94/WILEYDATA/train1.csv' dbms=dlm out=RawTrain replace;  
 delimiter=",";  
 getnames=yes;  
 guessingrows=1461;  
 run;  
   
/\* clean up numerica data with NA and make them "." so that SAS can interpret the datatypes correctly \*/  
 /\* The aforementioned import brings some numeric fields as character if NA is in the column \*/  
 /\* Need to change all NA's to . then export to csv and reimport \*/  
 data noNA;  
 set RawTrain;  
   
 /\* Clean Numeric Values \*/  
 if LotFrontage = 'NA' then LotFrontage= .;  
 if MasVnrArea = 'NA' then MasVnrArea=.;  
 if GarageYrBlt = 'NA' then GarageYrBlt=.;  
 if Alley = 'NA' then Alley= .;  
 if FireplaceQu = 'NA' then FireplaceQu = .;  
   
 /\* Clean Alpha Values \*/  
 if MasVnrType = 'NA' then MasVnrType = .;  
 if BsmtQual = 'NA' then BsmtQual = .;  
 if BsmtCond = 'NA' then BsmtCond = .;  
 if BsmtExposure = 'NA' then BsmtExposure = .;  
 if BsmtFinType1 = 'NA' then BsmtFinType1 = .;  
 if BsmtFinType2 = 'NA' then BsmtFinType2 = .;  
 if GarageType = 'NA' then GarageType = .;  
 if GarageFinish = 'NA' then GarageFinish = .;  
 if GarageQual = 'NA' then GarageQual = .;  
 if GarageCond = 'NA' then GarageCond = .;  
 if PoolQC = 'NA' then PoolQC = .;  
 if Fence = 'NA' then Fence = .;  
 if MiscFeature = 'NA' then MiscFeature = .;  
 run;  
 /\* export the data that has been cleaned up with missing values understandable to SAS \*/  
   
proc export data=noNA outfile= '/home/herreram1/sasuser.v94/WILEYDATA/train2.csv' dbms='csv' replace;  
 run;  
   
/\* Report the data: This time LotFrontage MasVnrAreaGarageYrBlt will be numeric \*/  
 proc import datafile='/home/herreram1/sasuser.v94/WILEYDATA/train2.csv' dbms=dlm out=RawTrain2 replace;  
 delimiter=",";  
 getnames=yes;  
 guessingrows=1461;  
 run;  
   
  
data All\_Data;  
 set RawTrain2;  
 lnSalePrice = log(SalePrice);  
 lnGrLivArea = log(GrLivArea);  
 lnOverallQual = log(OverallQual);  
 lnBsmtFinSF1 = log(BsmtFinSF1);  
 lnGarageYrBlt = log(GarageYrBlt);  
 lnOverallCond = log(OverallCond);  
 lnLotArea = log(LotArea);  
 lnYearBuilt = log(YearBuilt);  
 lnBsmtFinSF1 = log(BsmtFinSF1);  
 lnBsmtFinSF2 = log(BsmtFinSF2);  
 lnFireplaces = log(Fireplaces);  
 lnGarageCars = log(GarageCars);  
 lnMSSubClass = log(MSSubClass);  
 lnYearRemodAdd = log(YearRemodAdd);  
 lnMasVnrArea = log(MasVnrArea);  
 lnKitchenAbvGr = log(lnKitchenAbvGr);  
 lnBedroomAbvGr = log(BedroomAbvGr);  
 lnTotRmsAbvGrd = log(TotRmsAbvGrd);  
 lnBsmtFullBath = log(BsmtFullBath);  
 lnScreenPorch = log(ScreenPorch);  
 lnLotFrontage = log(LotFrontage);  
 lnBsmtFullBath = log(BsmtFullBath);  
 lnBsmtUnfSF = log(BsmtUnfSF);  
 lnTotalBsmtSF = log(TotalBsmtSF);  
 ln\_1stFlrSF = log(ln\_1stFlrSF);  
 ln\_2ndFlrSF = log(ln\_2ndFlrSF);  
 lnLowQualFinSF = log(lnLowQualFinSF);  
 lnBsmtHalfBath = log(BsmtHalfBath);  
 lnFullBath = log(FullBath);  
 lnLowQualFinSF = log(LowQualFinSF);  
 lnHalfBath = log(HalfBath);  
 lnLowQualFinSF = log(lnLowQualFinSF);  
 lnGarageArea = log(GarageArea);  
 lnWoodDeckSF = log(WoodDeckSF);  
 lnOpenPorchSF = log(OpenPorchSF);  
 lnEnclosedPorch = log(EnclosedPorch);  
 ln\_3SsnPorch = log(\_3SsnPorch);  
 lnScreenPorch = log(lnScreenPorch);  
 lnPoolArea = log(PoolArea);  
 lnMiscVal = log(MiscVal);  
 lnMoSold = log(MoSold);  
 lnYrSold = log(YrSold);  
 GrlivAreaSquared = GrLivArea\*\*2;  
 run;  
   
  
  
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\* Forward Selection \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
   
/\* Plot Data: Step 1 from forward regression\*/  
 proc sgscatter data = All\_Data;  
 matrix SalePrice GrLivArea OverallQual BsmtFinSF1 GarageYrBlt OverallCond LotArea YearBuilt BsmtFinSF2 Fireplaces ;  
 run;  
 proc sgscatter data = All\_Data;  
 matrix lnSalePrice lnGrLivArea lnOverallQual lnBsmtFinSF1 lnGarageYrBlt lnOverallCond lnLotArea lnYearBuilt lnBsmtFinSF2 lnFireplaces ;  
 run;  
   
  
/\* Develop Tentative Model: Step 2 from forward regression \*/  
 /\* Step 2 Develop a Tentative Model \*/  
 /\* Forward selection \*/  
 proc glmselect data = All\_Data;  
 model SalePrice = MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold / selection=forward ;  
 run;  
   
/\* Account for Confounders and look at residuals. Consider Consider remvoing TotRmsAbvGrd\*/  
 proc reg data=All\_Data;  
 model SalePrice = OverallQual GrLivArea BsmtFinSF1 GarageCars MSSubClass YearRemodAdd MasVnrArea LotArea KitchenAbvGr BedroomAbvGr TotRmsAbvGrd YearBuilt OverallCond BsmtFullBath ScreenPorch / tol VIF CLB;  
 run;  
   
/\* Account for Confounders and look at residuals. Without TotRmsAbvGrd\*/  
 proc reg data=All\_Data;  
 model SalePrice = OverallQual GrLivArea BsmtFinSF1 GarageCars MSSubClass YearRemodAdd MasVnrArea LotArea KitchenAbvGr BedroomAbvGr YearBuilt OverallCond BsmtFullBath ScreenPorch / tol VIF CLB;  
 run;  
   
/\* Account for Confounders and look at residuals. Without KitchenAbvGr\*/  
 proc reg data=All\_Data;  
 model SalePrice = OverallQual GrLivArea BsmtFinSF1 GarageCars MSSubClass YearRemodAdd MasVnrArea LotArea BedroomAbvGr YearBuilt OverallCond BsmtFullBath ScreenPorch / tol VIF CLB;  
 output out = t student=res cookd=cookd h=lev p= yhat;  
 run;  
   
/\* find point with high cook D \*/  
 proc sgplot data= t;   
 scatter y = res x=cookd /datalabel = cookd;  
 run;  
   
/\* find Id for point with high cook D \*/  
 proc sgplot data= t;   
 scatter y = res x=cookd /datalabel = id;  
 run;  
   
/\* Account for Confounders and look at residuals. Without KitchenAbvGr\*/  
 proc reg data=All\_Data;  
 where id <> 1299; /\* exclude commercial property \*/  
 model SalePrice = OverallQual GrLivArea BsmtFinSF1 GarageCars MSSubClass YearRemodAdd MasVnrArea LotArea BedroomAbvGr YearBuilt OverallCond BsmtFullBath ScreenPorch / tol VIF CLB;  
 output out = t student=res cookd=cookd h=lev p= yhat;  
 run;  
   
  
  
  
  
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\*. Backward Elimination \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
   
proc sgscatter data = All\_Data;  
 matrix SalePrice MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 \_1stFlrSF \_2ndFlrSF BsmtFullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd GarageCars ScreenPorch ;  
 run;  
   
/\* Backward Selection \*/  
 proc glmselect data = All\_Data;  
 model SalePrice = MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold / selection=backward ;  
 run;  
   
  
  
/\* Account for Confounders look at VIFs - Determine if collinearity is caused by outlier values \*/  
 proc reg data=All\_Data;  
 model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 \_1stFlrSF \_2ndFlrSF BsmtFullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd GarageCars ScreenPorch / tol VIF CLB;  
 output out = t student=res cookd=cookd h=lev p= yhat;  
 run;  
 proc sgplot data= t;   
 scatter y = res x=cookd /datalabel = cookd;  
 run;  
   
/\* find Id for point with high cook D \*/  
 proc sgplot data= t;   
 scatter y = res x=cookd /datalabel = id;  
 run;  
   
/\* Account for Confounders look at VIFs - remove id 1299 \*/  
 proc reg data=All\_Data;  
 where id <> 1299; /\* exclude commercial property \*/  
 model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 \_1stFlrSF \_2ndFlrSF BsmtFullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd GarageCars ScreenPorch / tol VIF CLB;  
 output out = t student=res cookd=cookd h=lev p= yhat;  
 run;  
   
/\* exclude commercial property and Remove TotRmsAbvGrd\*/  
 proc reg data=All\_Data;  
 where id <> 1299;   
model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 \_1stFlrSF \_2ndFlrSF BsmtFullBath BedroomAbvGr KitchenAbvGr GarageCars ScreenPorch / tol VIF CLB;  
 output out = t student=res cookd=cookd h=lev p= yhat;  
 run;  
   
  
  
  
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\*. Stepwise Selection \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
   
proc sgscatter data = All\_Data;  
 matrix SalePrice MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 \_1stFlrSF \_2ndFlrSF BsmtFullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd GarageCars ScreenPorch ;  
 run;  
   
/\* Stepwise Selection \*/  
 proc glmselect data = All\_Data;  
 model SalePrice = MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold / selection=stepwise ;  
 run;  
   
  
  
/\* Account for Confounders look at VIFs - Determine if collinearity is caused by outlier values \*/  
 proc reg data=All\_Data;  
 model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 GrLivArea BsmtFullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd GarageCars ScreenPorch / tol VIF CLB;  
 output out = t student=res cookd=cookd h=lev p= yhat;  
 run;  
 proc sgplot data= t;   
 scatter y = res x=cookd /datalabel = cookd;  
 run;  
   
/\* find Id for point with high cook D \*/  
 proc sgplot data= t;   
 scatter y = res x=cookd /datalabel = id;  
 run;  
   
/\* Account for Confounders look at VIFs - remove id 1299 \*/  
 proc reg data=All\_Data;  
 where id <> 1299;  
 model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 GrLivArea BsmtFullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd GarageCars ScreenPorch / tol VIF CLB;  
 output out = t student=res cookd=cookd h=lev p= yhat;  
 run;  
   
/\* Account for Confounders look at VIFs - remove id 1299 Remove TotRmsAbvGrd \*/  
 proc reg data=All\_Data;  
 where id <> 1299;  
 model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 GrLivArea BsmtFullBath BedroomAbvGr KitchenAbvGr GarageCars ScreenPorch / tol VIF CLB;  
 output out = t student=res cookd=cookd h=lev p= yhat;  
 run;  
   
  
/\* Account for Confounders look at VIFs - remove id 1299 Remove KitchenAbvGr \*/  
 proc reg data=All\_Data;  
 where id <> 1299;  
 model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 GrLivArea BsmtFullBath BedroomAbvGr GarageCars ScreenPorch / tol VIF CLB;  
 output out = t student=res cookd=cookd h=lev p= yhat;  
 run;  
   
  
  
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\*. CROSS VALIDATION \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
   
  
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\* Forward Selection \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
 proc glmselect data = All\_Data;  
 where id<>1299;  
 model SalePrice = MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr  
 KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold / selection=forward(stop=CV) cvmethod=random(5) stats=adjrsq;  
 run;  
   
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\*. Backward Elimination \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
 proc glmselect data = All\_Data;  
 where id<>1299;  
 model SalePrice = MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr  
 KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold / selection=backward(stop=CV) cvmethod=random(5) stats=adjrsq; ;  
 run;  
   
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\*. Stepwise Selection \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
 proc glmselect data = All\_Data;  
 where id<>1299;  
 model SalePrice = MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF \_1stFlrSF \_2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr  
 KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch \_3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold / selection=stepwise(stop=CV) cvmethod=random(5) stats=adjrsq; ;  
 run;  
   
  
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\*. KAGGLE. \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
   
/\* import test data \*/  
 proc import datafile='/home/herreram1/sasuser.v94/WILEYDATA/testCleaned.csv' dbms=dlm out=DirtyTestData replace;  
 delimiter=",";  
 getnames=yes;  
 guessingrows=1461;  
 run;  
   
/\* clean up numerica data with NA and make them "." so that SAS can interpret the datatypes correctly \*/  
 /\* The aforementioned import brings some numeric fields as character if NA is in the column \*/  
 /\* Need to change all NA's to . then export to csv and reimport \*/  
 data CleanTestData;  
 set DirtyTestData;  
   
 /\* Clean Numeric Values \*/  
 if LotFrontage = 'NA' then LotFrontage= .;  
 if MasVnrArea = 'NA' then MasVnrArea=.;  
 if GarageYrBlt = 'NA' then GarageYrBlt=.;  
 if Alley = 'NA' then Alley= .;  
 if FireplaceQu = 'NA' then FireplaceQu = .;  
 if BsmtFinSF1 = 'NA' then BsmtFinSF1 = .;  
 if BsmtFinSF2 = 'NA' then BsmtFinSF2 = .;  
 if Utilities = 'NA' then Utilities = .;  
 if TotalBsmtSF = 'NA' then TotalBsmtSF = .;  
 if BsmtFullBath = 'NA' then BsmtFullBath = .;  
 if BsmtHalfBath = 'NA' then BsmtHalfBath = .;  
 if GarageCars = 'NA' then GarageCars = .;  
 if GarageArea = 'NA' then GarageArea = .;  
 if BsmtUnfSF = 'NA' then BsmtUnfSF = .;  
   
  
  
  
 /\* Clean Alpha Values \*/  
 if MasVnrType = 'NA' then MasVnrType = .;  
 if BsmtQual = 'NA' then BsmtQual = .;  
 if BsmtCond = 'NA' then BsmtCond = .;  
 if BsmtExposure = 'NA' then BsmtExposure = .;  
 if BsmtFinType1 = 'NA' then BsmtFinType1 = .;  
 if BsmtFinType2 = 'NA' then BsmtFinType2 = .;  
 if GarageType = 'NA' then GarageType = .;  
 if GarageFinish = 'NA' then GarageFinish = .;  
 if GarageQual = 'NA' then GarageQual = .;  
 if GarageCond = 'NA' then GarageCond = .;  
 if PoolQC = 'NA' then PoolQC = .;  
 if Fence = 'NA' then Fence = .;  
 if MiscFeature = 'NA' then MiscFeature = .;  
 run;  
 /\* export the data that has been cleaned up with missing values understandable to SAS \*/  
   
proc export data=CleanTestData outfile= '/home/herreram1/sasuser.v94/WILEYDATA/TestCleanedData.csv' dbms='csv' replace;  
 run;  
   
/\* Report the data: This time LotFrontage MasVnrAreaGarageYrBlt will be numeric \*/  
 proc import datafile='/home/herreram1/sasuser.v94/WILEYDATA/TestCleanedData.csv' dbms=dlm out=TestDataReady replace;  
 delimiter=",";  
 getnames=yes;  
 guessingrows=1461;  
 run;  
   
  
/\* add column sale price to testdata \*/  
 Data TestDataReady;  
 set TestDataReady;  
 SalePrice = .;  
 run;  
   
/\* append data sets \*/  
 Data MergedTestData;  
 set All\_Data TestDataReady;  
 run;  
   
/\* Use models and predict values \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\* Forward Selection - Kaggle \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
 proc glm data=MergedTestData;  
 where id <> 1299;   
model SalePrice = OverallQual GrLivArea BsmtFinSF1 GarageCars MSSubClass YearRemodAdd MasVnrArea LotArea BedroomAbvGr YearBuilt OverallCond BsmtFullBath ScreenPorch;  
 output out=results p=Predict;  
 run;  
   
Data KaggleForwardSelection;  
 set results;  
 if Predict > 0 then SalePrice=Predict;  
 if Predict < 0 then SalePrice= 180921; /\*use the average price of the training data prices \*/  
 keep id SalePrice;  
 where id > 1460;  
   
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\* Backward Selection - Kaggle \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
 proc glm data=MergedTestData;  
 where id <> 1299;   
model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 \_1stFlrSF \_2ndFlrSF BsmtFullBath BedroomAbvGr KitchenAbvGr GarageCars ScreenPorch;  
 output out=results p=Predict;  
 run;  
   
Data KaggleBackwardElimination;  
 set results;  
 if Predict > 0 then SalePrice=Predict;  
 if Predict < 0 then SalePrice= 180921; /\*use the average price of the training data prices \*/  
 keep id SalePrice;  
 where id > 1460;  
   
/\* ----------------------------------------------------------------------------------------------------------------\*/  
 /\*. Stepwise Selection \*/  
 /\* ----------------------------------------------------------------------------------------------------------------\*/  
 proc glm data=MergedTestData;  
 where id <> 1299;   
model SalePrice = MSSubClass LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1 GrLivArea BsmtFullBath BedroomAbvGr GarageCars ScreenPorch;  
 output out=results p=Predict;  
 run;  
   
Data KaggleStepwiseSelection;  
 set results;  
 if Predict > 0 then SalePrice=Predict;  
 if Predict < 0 then SalePrice= 180921; /\*use the average price of the training data prices \*/  
 keep id SalePrice;  
 where id > 1460;

1. [↑](#footnote-ref-1)
2. The SAS Output and figures for these findings can be found in the appendix. [↑](#footnote-ref-2)
3. Choosing the backward model as the custom model because it has the highest AdjustedR Squared and tied for lowest score in kaggle. [↑](#footnote-ref-3)