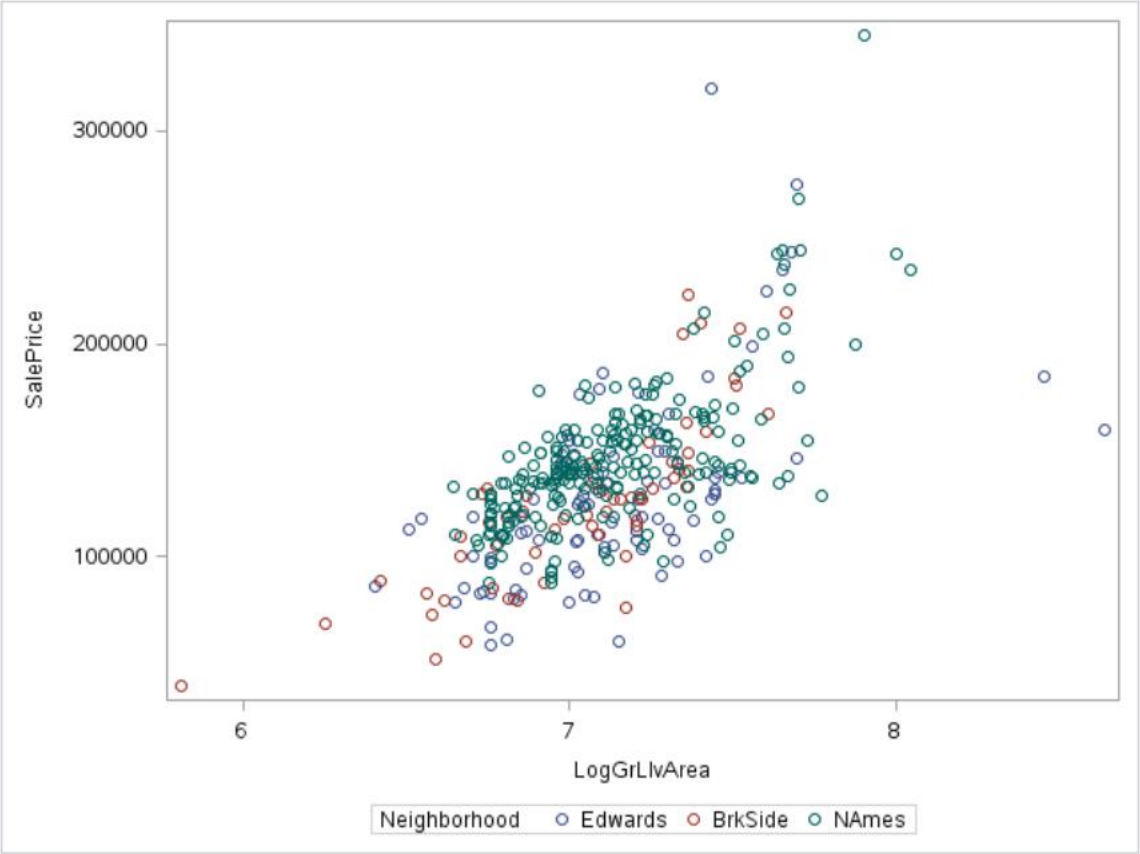
David Tran
Evangelos Giakoumakis
Amber Whelpley
MSDS 6371-403 HW 13/14

The Boston Housing Data Set is a collection of data from approximately 500 US census reports in the Boston area. In order to attempt to predict home sale prices in certain neighborhoods, multiple linear regression will be used determine which home features are related to sale price and whether the square footage of the home is related to the sale price.

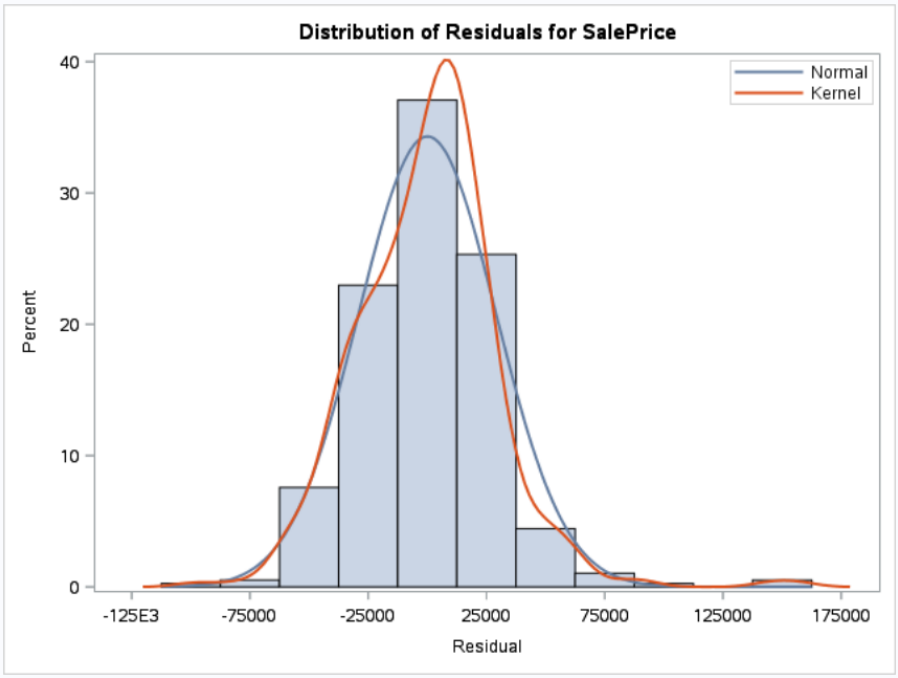
The data used for this analysis is the Ames Housing dataset, compiled by Dean De Cock as an alternative to, and subset of, the standard Boston Housing Data Set. The Ames Housing dataset includes only residential sales from Ames between 2006 and 2010. Information collected about these residential sales includes dimensions of various parts of the home, lot size and dwelling square footage, and quantifying variables about the number of specific rooms in the homes.

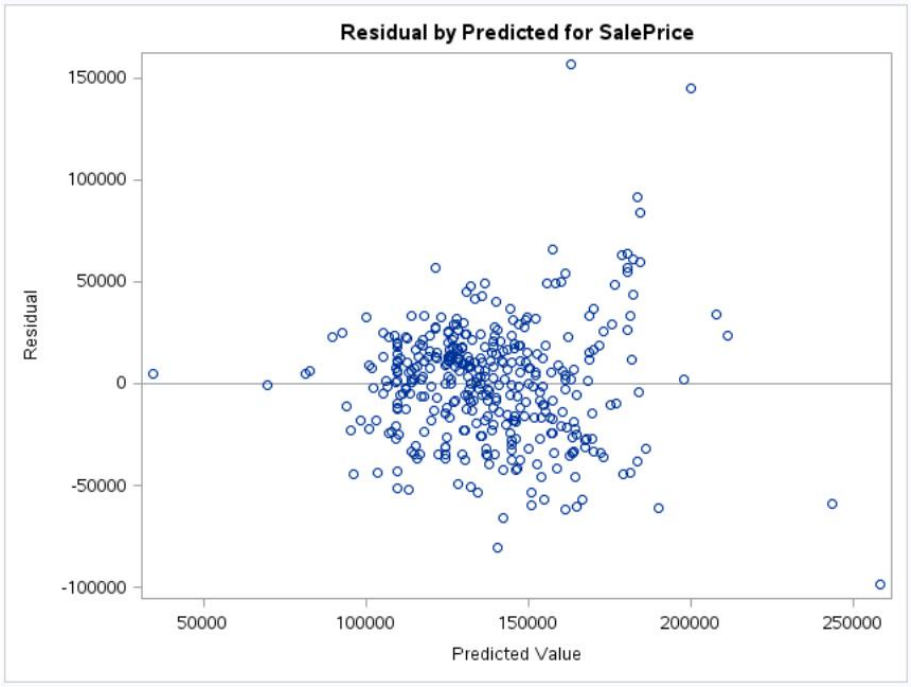
The Century 21 office in Ames, IA, would like to determine whether the final sale price of homes in their sales area (namely, the Names, Edwards, and BrkSide neighborhoods) is related to the size of the living area of the home. In order to analyze the information, home sales in the three neighborhoods of interest from 2006 to 2010 were gathered and a model was created using forward selection.

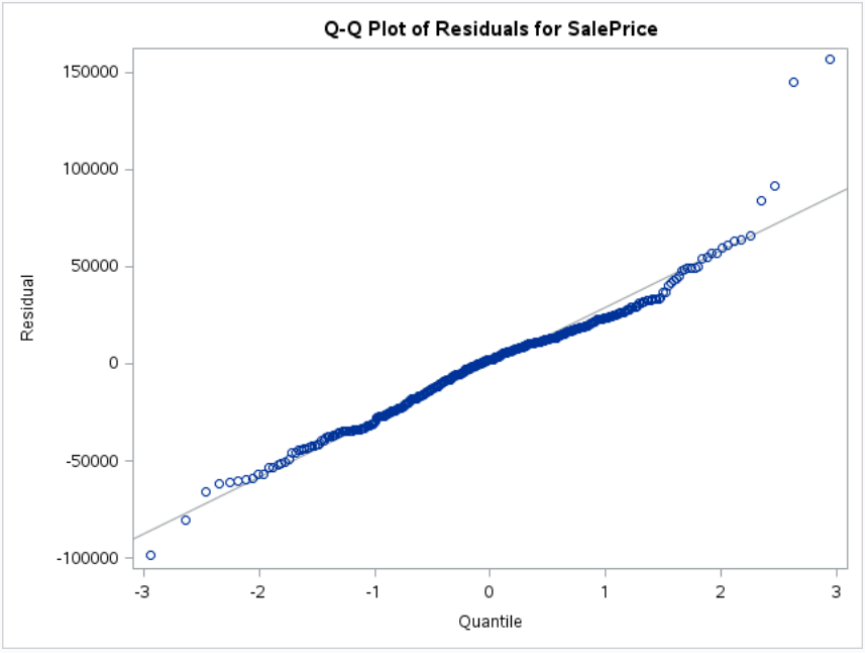
Before creating the model, the assumptions for multiple linear regression were checked. Based on the REG procedure in SAS, it was decided that a log transformation of the sale price and living area would be appropriate for the data. Seen below is the scatterplot of the transformed values.



The histogram seems to be normally distributed, and the scatterplot of residuals shows a normal variance with minimal clustering. The Q-Q plot shows that the data appears linear and fits the line well. From the Crooks D chart there appear to be a couple of outliers, however, since this data has already been filtered down to only residential sales and only the neighborhoods of interest the decision was made to keep these outliers and proceed.


Distribution of Residuals for SalePrice
Normal
Kernel
Percent
40
30
20
10
0
-125E3
-75000
-25000
25000
75000
125000
175000
Residual


Residual by Predicted for SalePrice
Residual
150000
100000
50000
0
-50000
-100000
50000
100000
150000
200000
250000
Predicted Value



The R^2 for the forward selection regression model is 0.4196 and the adjusted R^2 is 0.4181. These are relatively low values for the comparison of sale price and the log-transformed living area.

Root MSE	29106	**R-Square**	0.4196
Dependent Mean	138063	**Adj R-Sq**	0.4181
Coeff Var	21.08163		

For the final model using forward selection, we have an intercept of -425,922 for the sale price and parameter estimate of 79,219 for the log-transformed square footage of the homes:

$$sale\ price = -425{,}922 + 79{,}219\ln(square\ footage)$$

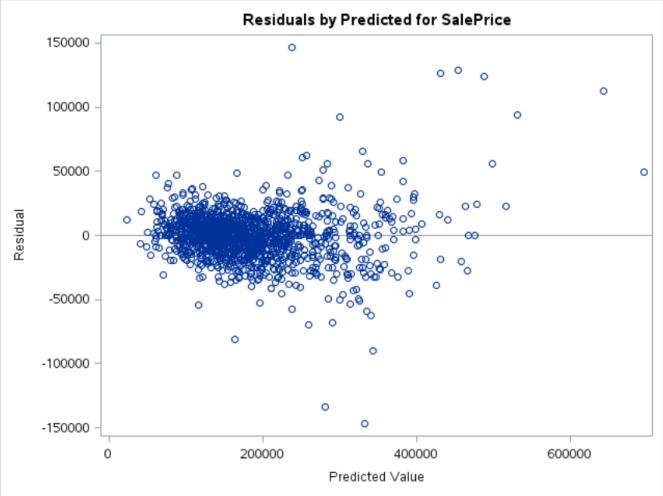
When the square footage is 1, the sale price would be -$425,922. Multiplying the square footage by e will increase the sale price by $79,219. More practically, a home with 1,000 square feet of living area is expected to have a sale price of $121,303. The price would increase by approximately $79,219 when the living area is multiplied by e (approximately 2.718).

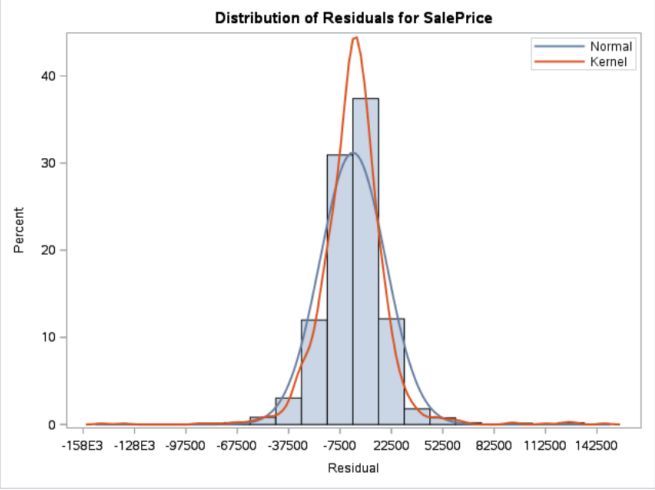
In conclusion, it does appear that the living area of the home in square feet has a correlation to the ultimate sale price of the home. The larger area available for living space, the greater the sale price is expected to be.

To build a predictive model for sales prices of homes in *all* neighborhoods of Ames, IA, three regression models (forward selection, backwards elimination, and stepwise selection) were created. After reviewing these models, a fourth custom model was created. A brief summary of the four models can be seen below:

Predictive Model	Adjusted R^2	CV Press	Kaggle Score
Forward selection	0.9472	5.41e11	0.15423
Backward elimination	0.9114	1.182e12	0.17135
Stepwise selection	0.9411	5.796e11	0.16432
Custom model	0.951	5.075e11	0.15032

As with the first analysis, the diagnostic plots were reviewed. The scatterplot showed randomly dispersed data, the Q-Q plot shows that the points relatively closely follow a straight line, and the histogram showed normal data. As before, Cook's D plot showed some outliers, but the decision was made to continue the analysis with the outliers.





The forward selection and stepwise selection models were created using all variables, two interactions maximum, single hierarchy, and random cvmethod. The backward elimination model used no interactions but still single hierarchy and random cvmethod.

Root MSE	18217
Dependent Mean	180615
R-Square	0.9515
Adj R-Sq	0.9472
AIC	30057
AICC	30080
PRESS	5.410381E11
SBC	29237

Root MSE	23597
Dependent Mean	180615
R-Square	0.9163
Adj R-Sq	0.9114
AIC	30773
AICC	30783
PRESS	1.182807E12
SBC	29746

Root MSE	19250
Dependent Mean	180615
R-Square	0.9445
Adj R-Sq	0.9411
AIC	30187
AICC	30198
PRESS	5.796288E11
SBC	29192

Forward selection, Backward elimination, and Stepwise selection

Reviewing the first three models revealed that the forward selection model seemed to be the most appropriate of the bunch, as it has the highest adjusted R^2 and the lowest CV press. Because of this, the forward selection model was used as the base for the custom model. Certain variables were removed from the model that were found to be irrelevant to the sale price of the homes: Alley, Building Style, Exterior Quality, Garage Year Built, Screen Porch, and Misc Feature.

Root MSE	17554
Dependent Mean	180615
R-Square	0.9544
Adj R-Sq	0.9510
AIC	29933
AICC	29949
PRESS	5.075127E11
SBC	29018

Overall, the most predictive model of the four described above is the custom model. This model can be used to predict future sale prices of homes in Ames, Iowa, based on the many varied features and sizes of the home.

Appendix

Analysis 1

```
/* Import Test Data*/;
FILENAME REFFILE '/home/davidtran0/Statistical Foundations/Week
13/test.csv';
PROC IMPORT DATAFILE=REFFILE
      DBMS=CSV
      OUT=STAT1.TEST1;
      GETNAMES=YES;
RUN;

PROC PRINT DATA=STAT1.TEST1; RUN;

/* Import Train Data*/;
FILENAME REFFILE '/home/davidtran0/Statistical Foundations/Week
13/train.csv';
PROC IMPORT DATAFILE=REFFILE
      DBMS=CSV
      OUT=STAT1.TRAIN1;
      GETNAMES=YES;
RUN;
PROC Print DATA=STAT1.TRAIN1; RUN;

/* Separating the Edward Neighborhood to its own Dataset*/;
DATA TRAIN2;
      SET STAT1.TRAIN1;
      IF Neighborhood EQ 'Edwards';
run;
PROC PRINT Data=TRAIN2;
run;

/* Separating the BrkSide Neighborhood to its own Dataset*/;
DATA TRAIN3;
      SET STAT1.TRAIN1;
      IF Neighborhood EQ 'BrkSide';
run;
PROC PRINT Data=TRAIN3;
run;

/* Separating the NAmes Neighborhood to its own Dataset*/;
DATA TRAIN4;
      SET STAT1.TRAIN1;
      IF Neighborhood EQ 'NAmes';
run;
PROC PRINT Data=TRAIN4;
run;

/* Combining the 3 Neighborhood Data*/;
Data FULL;
      SET TRAIN2 TRAIN3 TRAIN4;
run;
PROC PRINT Data=FULL;
```

```
run;

/* Log Transform Data */
Data LogFull;
      SET Full;
      LogGrLIvArea = log(GrLIvArea);
      LogSalePrice = log(SalePrice);
run;
PROC PRINT Data=LogFull;
run;

/* Plotting the Data */
PROC SGPLOT Data = Full;
STYLEATTRS DATASYMBOLS=(CIRCLE TRIANGLE
Asterisk);
SCATTER X=GrLIvArea Y=SalePrice / GROUP=Neighborhood;
       axis1 ORDER=(0 to 6000 by 100);
run;

/* Plotting the transformed data */
PROC SGPLOT Data = LogFull;
STYLEATTRS DATASYMBOLS=(CIRCLE TRIANGLE
Asterisk);
SCATTER X=LogGrLIvArea Y=SalePrice / GROUP=Neighborhood;
       axis1 ORDER=(0 to 6000 by 100);
run;

/* Building the model between SalePrice = logGrLIvArea & Neighborhood*/;
PROC REG data=LogFull PLOTS(unpack) = diagnostics;
      MODEL SalePrice = LogGrLIvArea
      /VIF CLM CLI;
run;
```

Analysis 2

```
/* Combine train and test data sets and clean up*/
data final_train;
set STAT1.train STAT1.test;
if LotFrontage EQ 'NA' THEN LotFrontage = 0;
if GarageYrBlt EQ 'NA' THEN GarageYrBlt = 0;
run;

proc print data=final_train;
run;

/* proc glm to review assumptions */
proc glm data = STAT_KAG.train plots=diagnostics;
class MSZoning LotFrontage Street Alley LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1
Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd
MasVnrType ExterQual ExterCond Foundation
BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC
CentralAir Electrical KitchenQual Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC
Fence MiscFeature SaleType SaleCondition;
model SalePrice = MSSubClass MSZoning LotFrontage LotArea Street LotShape
LandContour Utilities LotConfig
LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle OverallQual
OverallCond YearBuilt YearRemodAdd
RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual
ExterCond Foundation BsmtQual BsmtCond
BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF
TotalBsmtSF Heating HeatingQC CentralAir
Electrical "1stFlrSF"n "2ndFlrSF"n LowQualFinSF GrLivArea BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces FireplaceQu
GarageType GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF
OpenPorchSF EnclosedPorch "3SsnPorch"n ScreenPorch
PoolArea PoolQC Fence MiscVal MoSold YrSold SaleType SaleCondition;
run;


/* Stepwise selection using all variables */
proc glmselect data = final_train;
class MSZoning LotFrontage Street Alley LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1
Condition2 HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType
ExterCond Foundation
BsmtQual BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir
Electrical KitchenQual Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC
Fence MiscFeature SaleType SaleCondition;
model SalePrice = MSSubClass | MSZoning | LotFrontage | LotArea | Street |
Alley | LotShape | LandContour | Utilities | LotConfig
| LandSlope | Neighborhood | Condition1 | Condition2 | HouseStyle |
OverallQual | OverallCond | YearBuilt | YearRemodAdd
| RoofStyle | RoofMatl | Exterior1st | Exterior2nd | MasVnrType | MasVnrArea
| ExterCond | Foundation | BsmtQual
```

```
| BsmtExposure | BsmtFinType1 | BsmtFinSF1 | BsmtFinType2 | BsmtFinSF2 |
BsmtUnfSF | TotalBsmtSF | Heating | HeatingQC | CentralAir
| Electrical | "1stFlrSF"n | "2ndFlrSF"n | LowQualFinSF | GrLivArea |
BsmtFullBath | BsmtHalfBath | HalfBath | BedroomAbvGr
| KitchenAbvGr | KitchenQual | TotRmsAbvGrd | Functional | Fireplaces |
FireplaceQu | GarageType | GarageYrBlt | GarageFinish
| GarageCars | GarageArea | GarageQual | GarageCond | PavedDrive |
WoodDeckSF | OpenPorchSF | EnclosedPorch | "3SsnPorch"n | ScreenPorch
| PoolArea | PoolQC | Fence | MiscFeature | MiscVal | MoSold | YrSold |
SaleType | SaleCondition @2
/ selection=stepwise(choose=PRESS) hierarchy=single showpvalues
cvmethod=random(2);
output out = results_step p = Predict;
run;

/* Make results Kaggle friendly and fix prediction issues */
data step_results_final;
set results_step;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 160000; /* Mean = 180000 */
keep Id SalePrice;
where Id > 1460;
;

proc print data=step_results_final;
run;

/* Forward selection using all variables */
proc glmselect data = final_train;
class MSZoning LotFrontage Street Alley LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1
Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd
MasVnrType ExterQual ExterCond Foundation
BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC
CentralAir Electrical KitchenQual Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC
Fence MiscFeature SaleType SaleCondition;
model 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 | "1stFlrSF"n | "2ndFlrSF"n | LowQualFinSF | GrLivArea |
BsmtFullBath | BsmtHalfBath | FullBath | HalfBath | BedroomAbvGr
| KitchenAbvGr | KitchenQual | TotRmsAbvGrd | Functional | Fireplaces |
FireplaceQu | GarageType | GarageYrBlt | GarageFinish
| GarageCars | GarageArea | GarageQual | GarageCond | PavedDrive |
WoodDeckSF | OpenPorchSF | EnclosedPorch | "3SsnPorch"n | ScreenPorch
| PoolArea | PoolQC | Fence | MiscFeature | MiscVal | MoSold | YrSold |
SaleType | SaleCondition @2
/ selection=forward(choose=PRESS) hierarchy=single showpvalues
cvmethod=random(2);
output out = results_for p = Predict;
run;
```

```
/* Make results Kaggle friendly and fix prediction issues */
data for_results_final;
set results_for;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 160000; /* Mean = 180000 */
keep Id SalePrice;
where Id > 1460;
;

proc print data=for_results_final;
run;

/* Backward selection using all variables */
proc glmselect data = final_train;
class MSZoning LotFrontage Street Alley LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1
Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd
MasVnrType ExterQual ExterCond Foundation
BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC
CentralAir Electrical KitchenQual Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC
Fence MiscFeature SaleType SaleCondition;
model 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 "1stFlrSF"n "2ndFlrSF"n LowQualFinSF GrLivArea BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces FireplaceQu
GarageType GarageYrBlt GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF
OpenPorchSF EnclosedPorch "3SsnPorch"n ScreenPorch
PoolArea PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType
SaleCondition
/ selection=backward(choose=PRESS) hierarchy=single showpvalues
cvmethod=random(2);
output out = results_back p = Predict;
run;

/* Make results Kaggle friendly and fix prediction issues */
data back_results_final;
set results_back;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 160000; /* Mean = 180000 */
keep Id SalePrice;
where Id > 1460;
;

proc print data=back_results_final;
run;

/* Custom selection */
```

```
proc glmselect data = final_train;
class MSZoning LotFrontage Street Alley LotShape LandContour Utilities
LotConfig LandSlope Neighborhood Condition1
Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd
MasVnrType ExterQual ExterCond Foundation
BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC
CentralAir Electrical KitchenQual Functional
FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC
Fence MiscFeature SaleType SaleCondition;
model SalePrice = MSSubClass | MSZoning | LotFrontage | LotArea | Street |
LotShape | LandContour | Utilities | LotConfig
| LandSlope | Neighborhood | Condition1 | Condition2 | BldgType | HouseStyle
| OverallQual | OverallCond | YearBuilt | YearRemodAdd
| RoofStyle | RoofMatl | Exterior1st | Exterior2nd | MasVnrType | MasVnrArea
| ExterQual | ExterCond | Foundation | BsmtQual | BsmtCond
| BsmtExposure | BsmtFinType1 | BsmtFinSF1 | BsmtFinType2 | BsmtFinSF2 |
BsmtUnfSF | TotalBsmtSF | Heating | HeatingQC | CentralAir
| Electrical | "1stFlrSF"n | "2ndFlrSF"n | LowQualFinSF | GrLivArea |
BsmtFullBath | BsmtHalfBath | FullBath | HalfBath | BedroomAbvGr
| KitchenAbvGr | KitchenQual | TotRmsAbvGrd | Functional | Fireplaces |
FireplaceQu | GarageType | GarageFinish
| GarageCars | GarageArea | GarageQual | GarageCond | PavedDrive |
WoodDeckSF | OpenPorchSF | EnclosedPorch | "3SsnPorch"n | ScreenPorch
| PoolArea | PoolQC | Fence | MiscVal | MoSold | YrSold | SaleType |
SaleCondition @2
/ selection=forward(choose=PRESS) showpvalues;
output out = results_cus p = Predict;
run;

/* Make results Kaggle friendly and fix prediction issues */
data cus_results_final;
set results_cus;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 160000; /* Mean = 180000 */
keep Id SalePrice;
where Id > 1460;
;

proc print data=cus_results_final;
run;

/* Go to library, export data in SAS student folder and then download to pc
before uploading to kaggle */
/* Current best kaggle attempt rank# 1086 score 0.15032 */
```