DS 6371
8/06/23

Predictive Data Modeling for Sale Prices of Homes Using Linear Regression

Team members:
Sakava Kiv - skiv@smu.edu, Banumathi Pullaiahnaidu - bpullaiahnaidu@mail.smu.edu
Github: https://github.com/SakavaKiv/HousePricePredictions, https://banupullaiahnaidu.github.io/

Analysis I
Introduction
Brief introduction to the questions of interest and the setting of the problem

In this analysis, we delve deep into the intriguing connection between house sale prices and their living area square footage (GrLivArea) in distinct neighborhoods (NAmes, Edwards and BrkSide). Realtors frequently converse about living area in 100-square-feet increments, propelling us to craft a succinct and easy-to-understand estimate (or estimates) of this association for the company. Moreover, we embark on an exciting journey to uncover potential variations in this relationship across the three neighborhoods, ensuring our model incorporates the neighborhood as a categorical variable for a comprehensive exploration.

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

We gathered Kaggle data. For the Training Set we have 383 records out 1460 records if we filter out for neighborhoods in (NAmes, Edwards and BrkSide). And for the Test Set we have 362 out 1459 records respectively. Data dictionary can be found in the Appendix of this document. Finally for Analysis I, we will be concerned with the columns SalePrice, GrLivArea and the categorical variable Neighborhood in which we will do our analysis on.

Analysis Question 1:
Restatement of Problem
The main objectives of the analysis are as follows:
Develop a linear regression model to predict SalePrice based on GrLivArea, taking into account the different neighborhoods.
Provide estimates for the relationships between SalePrice and GrLivArea for each neighborhood.
Determine the confidence intervals for the estimated coefficients to understand the uncertainty in the estimates.
Ensure that the assumptions of linear regression, such as linearity, independence, constant variance, and normality of residuals, are met.
Identify and address any suspicious observations, including outliers and influential observations.
Present a well-written conclusion that quantifies the relationship between living area and SalePrice for the three specific neighborhoods.
By conducting this analysis, Century 21 Ames aims to gain valuable insights to support their business decisions and better understand the housing market in the selected neighborhoods.

Build and Fit the Model
SAS Code Display 1 (see appendix for code and outputs)
SAS Code Display 2 (see appendix for code and outputs)
Model for full data set
SAS Code Display 3 (see appendix for code)

$$Predicted\ Price = \beta_0 + \beta_1 GrLivArea + \beta_2 NeiNAmes + \beta_3 NeiEdwards + \beta_4 NeiBrkSide$$

Model for living area < 4500 sft

$$Predicted\ Price = \beta_0 + \beta_1 GrLivArea + \beta_2 NeiNAmes + \beta_3 NeiEdwards + \beta_4 NeiBrkSide$$

SAS Code Display 4 (see appendix for code)

Model for living area < 4500 sft and with variable interactions

$$\Pr e\,dicted\,Sale\,\Pr i\,ce = \beta_0 + \beta_1 GLA + \beta_2 NN + \beta_3 NE + \beta_4 NB + \beta_5 NN \cdot GLA + \beta_6 E \cdot GLA + \beta_7 NB \cdot GLA$$

SAS Code Display 5 (see appendix for code)

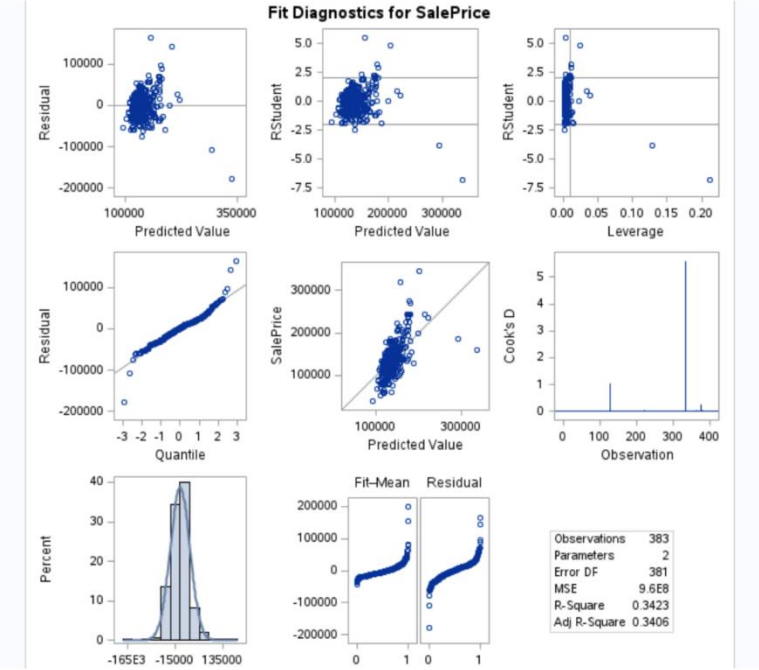
Checking Assumptions

Plot the Data

SAS Code Display 6 (see appendix for code)

Fit the Model

SAS Code Display 7 (see appendix for code)



Evaluate Assumptions

Linearity – Looking at the Pearson Correlation Coefficient below, we do believe a linear correlation exists between the variables.

SAS Code Display 8 (see appendix for code)

The CORR Procedure

2 Variables:	SalePrice GrLivArea

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SalePrice	383	138063	38156	52877938	39300	345000
GrLivArea	383	1302	485.51836	498602	334.00000	5642

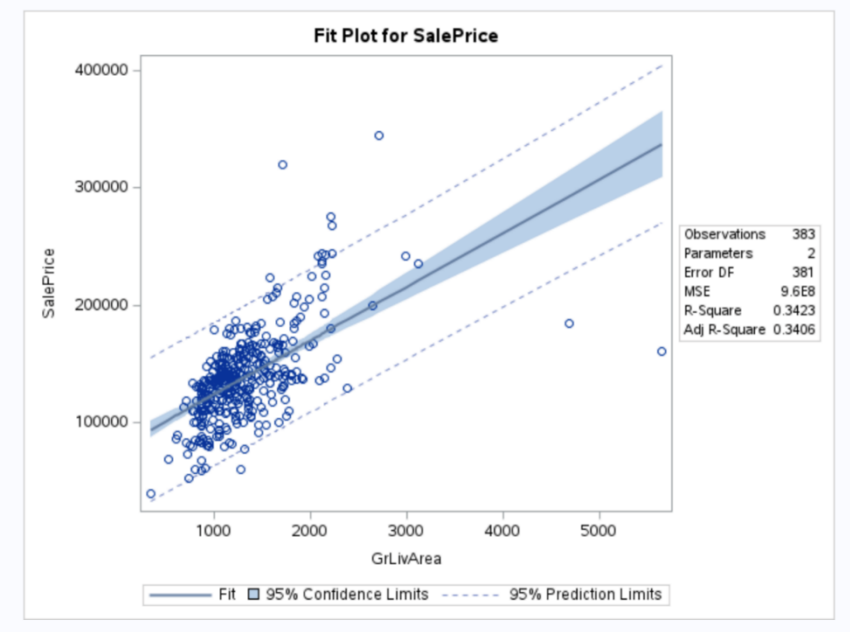
Pearson Correlation Coefficients, N = 383
Prob > |r| under H0: Rho=0

	SalePrice	GrLivArea
SalePrice	1.00000	0.58506 <.0001
GrLivArea	0.58506 <.0001	1.00000

Independence – We will assume the observations in the data set provided are independent of each other.

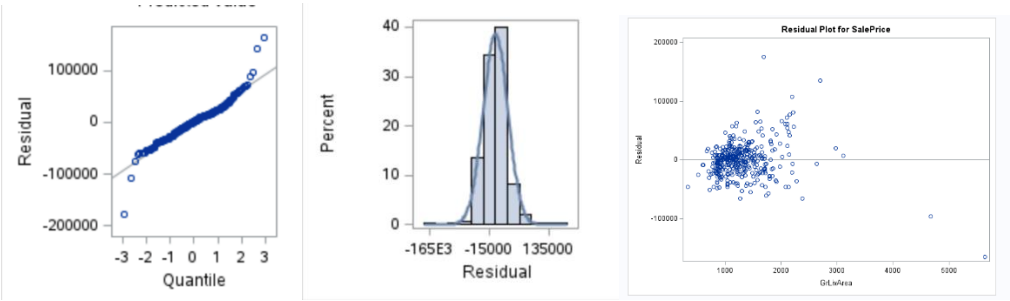
Constant Variance – From the below plot we see some evidence of increased standard deviation.

SAS Code Display 9 (see appendix for code)



Normality of Residuals – Looking at the qq-plot and histogram below there is evidence that the data is normally distributed. The Residual plot also suggest that, with some outliers

SAS Code Display 10 (see appendix for code)

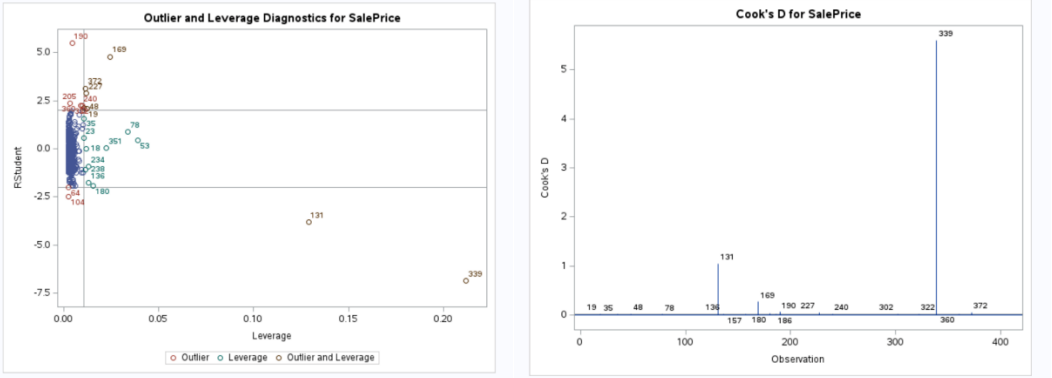


Identify and address any suspicious observations, including outliers and influential observations.

Influential point analysis (Cook's D and Leverage)

From the below plots we see evidence of high Cook's D for observations 131 and 339 and they also show up as having high leverage and as outliers

SAS Code Display 11 (see appendix for code)



CODE LOG RESULTS OUTPUT DATA

Table: WORK.OUTPUTDATASET | View: Column names | Filter: COOKD > 1.02

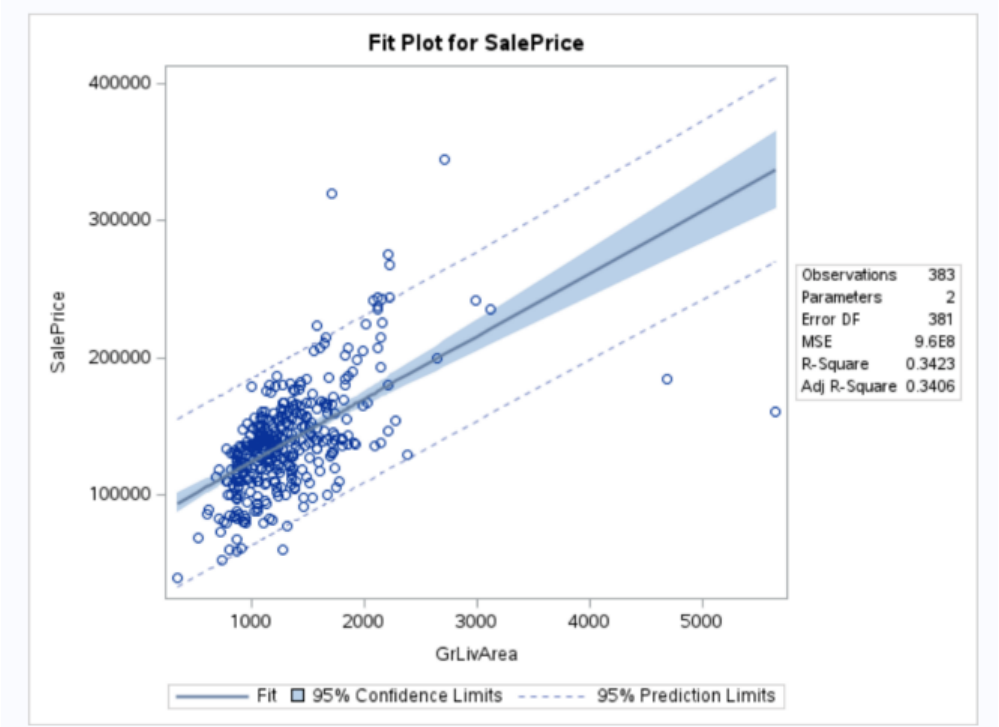
Total rows: 383 Total columns: 7 Filtered rows: 2 | Rows 1-2

	Id	GrLivArea	Neighborho...	SalePrice	COOKD	H	DFFITS
1	524	4676	Edwards	184750	1.0420704737	0.1290433969	-1.46913496
2	1299	5642	Edwards	160000	5.6014423708	0.2117996637	-3.542084972

Looking at all the above plots two observations (observation 131 and 339) show up as having a high leverage are outliers. We can remove these and reduce the range of our analysis to houses with GrLivArea less than 4500 sf. By doing this we will be improving the assumptions for constant variance and get better confidence interval estimates. The R-Square and Adjusted R-Square also goes up.

With Observations 131 and 139

SAS Code Display 12 (see appendix for code)



The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	383
Number of Observations Used	383

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.903676E11	1.903676E11	198.29	<.0001
Error	381	3.657846E11	960064442		
Corrected Total	382	5.561521E11			

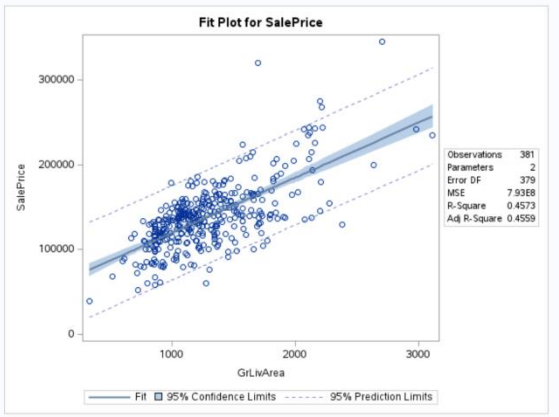
Root MSE	30985	R-Square	0.3423
Dependent Mean	138063	Adj R-Sq	0.3406
Coeff Var	22.44267		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Semi-partial Corr Type I	Squared Semi-partial Corr Type II	Variance Inflation	95% Confidence Limits	
Intercept	1	78206	4536.05353	17.24	<.0001	0	.	.	0	69287	87124
GrLivArea	1	45.97896	3.26522	14.08	<.0001	0.58506	0.34229	0.34229	1.00000	39.55885	52.39907

With out Observations 131 and 139

SAS Code Display 13 (see appendix for code)



The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	381
Number of Observations Used	381

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.530999E11	2.530999E11	319.35	<.0001
Error	379	3.003789E11	792556375		
Corrected Total	380	5.534788E11			

Root MSE	28152	R-Square	0.4573
Dependent Mean	137882	Adj R-Sq	0.4559
Coeff Var	20.41768		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Semi-partial Corr Type I	Squared Semi-partial Corr Type II	Variance Inflation	95% Confidence Limits	
Intercept	1	54415	4888.34799	11.13	<.0001	0	.	.	0	44803	64027
GrLivArea	1	65.12810	3.64450	17.87	<.0001	0.67623	0.45729	0.45729	1.00000	57.96214	72.29407

Comparing Competing Models

The three model we will be comparing are

1. Model with full data set

SAS Code Display 14 (see appendix for code)

Root MSE	28552
Dependent Mean	138063
R-Square	0.4474
Adj R-Sq	0.4400
AIC	8249.72388
AICC	8250.02255
SBC	7888.41209
CV PRESS	3.350022E11

1. Model with Living area < 4500 sf

SAS Code Display 14 (see appendix for code)

Root MSE	27241
Dependent Mean	137882
R-Square	0.4945
Adj R-Sq	0.4905
AIC	8168.88300
AICC	8169.04300
SBC	7801.65420
CV PRESS	2.85477E11

1. Model with Living area < 4500 sf with variable interactions

SAS Code Display 15 (see appendix for code)

Root MSE	26825
Dependent Mean	137882
R-Square	0.5125
Adj R-Sq	0.5060
AIC	8159.14081
AICC	8159.44108
SBC	7799.79761
CV PRESS	2.82083E11

Looking at the above tables we can tell that the Model with Living Area < 4500 sf with variable interactions has the least Adjusted R-Square of the three models and so is the CV Press, AIC and SBC statistics. Hence we choose Model 3 as the most favorable model of the three.

Parameters Estimates

SAS Code Display 16 (see appendix for code)

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	74676.40154	B	5954.52674	12.54	<.0001	62967.95510	86384.84798
GrLivArea	54.31586	B	4.33457	12.53	<.0001	45.79276	62.83896
Neighborhood BrkSide	-54704.88774	B	13042.61747	-4.19	<.0001	-80350.71900	-29059.05648
Neighborhood Edwards	-43247.84694	B	11671.23793	-3.71	0.0002	-66197.12068	-20298.57320
Neighborhood NAmes	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho BrkSide	32.84667	B	10.16117	3.23	0.0013	12.86665	52.82669
GrLivArea*Neighborho Edwards	21.66057	B	8.79973	2.46	0.0143	4.35757	38.96358
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.	.	.

Interpretation

For every 100 sf increase in the Living Area in the NAmes neighborhood, the estimated / predicted SalePrice of the home increases by 5,400 dollars.

Estimate of the difference between the NAmes and BrkSide Intercepts, $\beta 2$ = -54705

BrkSide's Intercept is estimated to be 54705 dollars lower than NAmes intercept of 74,676 dollars.

Estimate of BrkSide Intercept = 19971

Adjustment to NAmes Slope to get BrkSide Slope = 33

Slope of BrkSide, = 87

Estimated Sale Price of home in BrkSideNeighborhood is (74676 –54705) + (54 +33) * GrLivArea

For every 100 sf increase in the Living Area in the BrkSide neighborhood, the estimated / predicted SalePrice of the home increases by 8,700 dollars.

Estimate of the difference between the NAmes and Edwards Intercepts, $\beta 3$= -43248

Edwards Intercept is estimated to be 43,248 dollars lower than NAmes intercept of 74,676 dollars.

Estimate of Edwards Intercept = 31428

Adjustment to NAmes Slope to get Edwards Slope = 22

Slope of Edwards, $\beta 5$= 76

Estimated Sale Price of home in BrkSideNeighborhood is (74676 –43248) + (54 +22) * GrLivArea

For every 100 sf increase in the Living Area in the Edwards neighborhood, the estimated / predicted SalePrice of the home increases by 7,600 dollars.

Confidence Intervals

The 95% confidence interval for predicted increase in the SalePrice of the home for NAmes Neighborhood is 4,600 and 6,300 dollars for every 100 sf increase in Living Area

The 95% confidence interval for predicted increase in the SalePrice of the home for BrkSide Neighborhood is 5,900 and 11,600 dollars for every 100 sf increase in Living Area

The 95% confidence interval for predicted increase in the SalePrice of the home for Edwards Neighborhood is 5,000 and 10,200 dollars for every 100 sf increase in Living Area

Conclusion

From our analysis we see that there is a linear relationship between the Living Area and Sale Price of the homes, for the three neighborhoods in the data set. We also see from the parameter estimate table that there is strong evidence at p-value <0.001, 0.0013 and 0.0143 that the linear relationship exists in all three neighborhoods. In conclusion looks like our predicted sales price per square footage with a confidence interval of (45.79sqft,62.83sqft) for square feet with respect to neighborhoods which are NAmes $62,967-$86,384, BrkSide is $29,059-$80,350, and Edwards$20,298-$66197. So, the sales price per 100 square feet is the most for Names neighborhood. The sales agents can use these to value the homes.

R Shiny: Price v. Living Area Chart

Shiny App - https://banu.shinyapps.io/SFDS_Final_Project/

Analysis Question 2

Restatement of Problem

The main objectives of the analysis are as follows:

Develop four different types of linear regression model (Stepwise, Forward, Backward, and Custom) to predict SalePrice based on GrLivArea, taking into account all the different neighborhoods.

Ensure that the assumptions of linear regression, such as linearity, independence, constant variance, and normality of residuals, are met.

Identify and address any suspicious observations, including outliers and influential observations.

Present a well-written conclusion that quantifies the relationship between living area and SalePrice for the three specific neighborhoods.

By conducting this analysis, Century 21 Ames aims to gain valuable insights to support their business decisions and better understand the housing market in the selected neighborhoods.

Model Selection

Type of Selection **(Stepwise)**

SAS Code Display 17(see appendix for code and output)

Type of Selection **(Forward)**

SAS Code Display 18 (see appendix for code and output)

Type of Selection **(Backward)**

SAS Code Display 19 (see appendix for code and output)

Type of Selection **(Custom)**

SAS Code Display 20 (see appendix for code and output)

Checking Assumptions

Independence – We will assume the observations in the data set provided are independent of each other.

Linearity – Looking at the fit diagnostics, we do believe a linear correlation exists between the variables.

Constant Variance – From the below plot we see some evidence of increased standard deviation.

Normality of Residuals – Looking at the qq-plot and histogram there is evidence that the data is normally distributed. The Residual plot also suggest that, with some outliers

Influential point analysis (Cook's D and Leverage)

From the plots we see evidence of high Cook's D for two observations and they also show up as having high leverage and as outliers

Comparing Competing Models (Adj R2, Internal CV Press, Kaggle Score)

Predictive Models	Adjusted R2	CV PRESS	Kaggle Score
Forward	0.8387	1255386000000	0.1766
Backward	0.9020	1194469000000	0.17903
Stepwise	0.7997	1760988000000	0.17903
CUSTOM	0.8291	1675250900000	0.16946

We can observe from the above that the custom model has the lowest Kaggle score, so it is the best model.

Conclusion:

In summary it looks like the variables that we selected for the custom model did a sufficient job in predicting the sales prices. We had a Kaggle score of 0.16946 which is respectable.

APPENDIX A: SAS and R Code

SAS Code Display 1

```
/* Step 1: Data Selection */
data houses;
  set trainkaggle2; /* Replace YourDataset with the name of your dataset containing the required variables */
  where Neighborhood in ('NAmes', 'Edwards', 'BrkSide');
run;

/* Step 2: Model Building */
proc glm data=houses plots = all;
  class Neighborhood;
  model SalePrice = GrLivArea Neighborhood / cli solution;
  output out =  results p = Predict;
run;

data results3;
set results;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 10000;
keep id SalePrice;
where id > 1460;
;

proc means data = results3;
var SalePrice;
run;
```

The GLM Procedure

Class Level Information		
Class	Levels	Values
Neighborhood	3	BrkSide Edwards NAmes

Number of Observations Read	745
Number of Observations Used	383

Sum of Residuals	-4.074536E-9
Sum of Squared Residuals	335644522201
Sum of Squared Residuals - Error SS	-0.000061035
PRESS Statistic	363653838311
First Order Autocorrelation	-0.059318972
Durbin-Watson D	2.1170411803

The GLM Procedure

Dependent Variable: SalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	220507606457	73502535486	83.00	<.0001
Error	379	335644522201	885605599.47		
Corrected Total	382	556152128658			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.396488	21.55482	29759.13	138062.5

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GrLivArea	1	190367576250	190367576250	214.96	<.0001
Neighborhood	2	30140030207	15070015103	17.02	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GrLivArea	1	187035079034	187035079034	211.19	<.0001
Neighborhood	2	30140030207	15070015103	17.02	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	85887.15908	B	4578.116845	18.76	<.0001
GrLivArea	45.76006		3.148801	14.53	<.0001
Neighborhood BrkSide	-16105.62078	B	4395.351815	-3.66	0.0003
Neighborhood Edwards	-18987.77587	B	3577.826684	-5.31	<.0001
Neighborhood NAmes	0.00000	B	.	.	.

The MEANS Procedure

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
362	137723.27	21302.23	89687.90	300046.91

SAS Code Display 2

```
/* Step 2: Model Building Forward Selection with CV press */
proc glmselect data=houses plots = all;
  class Neighborhood;
  model SalePrice = GrLivArea Neighborhood
  /selection= Forward (stop=CV) cvmethod=random(5) stats=adjrsq;
run;

data results1;
set results;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 10000;
keep id SalePrice;
where id > 1460;
;

proc means data = results1;
var SalePrice;
run;
```

Dimensions	
Number of Effects	71
Number of Parameters	341

The GLMSELECT Procedure

Forward Selection Summary

Step	Effect Entered	Number Effects In	Number Parms In	Adjusted R-Square	SBC	ASE	Test ASE	CV PRESS
0	Intercept	1	1	0.0000	24623.7146	5919805925	7388211567	6.47984E12
1	OverallQual	2	2	0.6070	23608.0995	2324621178	2583334414	2.55222E12
2	GrLivArea	3	3	0.7024	23309.6952	1758385978	1980610770	2.03425E12
3	Neighborhood	4	27	0.7715*	23164.4072*	1320591591	1565096116	1.62323E12*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	RoofMatl	1.67698E12	>	1.62323E12

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept GrLivArea OverallQual Neighborhood

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	26	5.03154E12	1.935208E11	142.92
Error	1067	1.444727E12	1354008623	
Corrected Total	1093	6.476268E12		

Root MSE	36797
Dependent Mean	185601
R-Square	0.7769
Adj R-Sq	0.7715
AIC	24125
AICC	24127
SBC	23164
ASE (Train)	1320591591
ASE (Test)	1565096116
CV PRESS	1.623234E12

SAS Code Display 3 (see appendix for code)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Neighborhood	3	BrkSide Edwards NAmes

Number of Observations Read	383
Number of Observations Used	383

The GLM Procedure

Dependent Variable: SalePrice

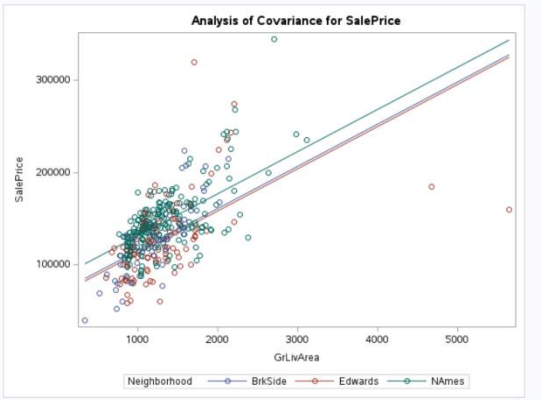
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	220507606457	73502535486	83.00	<.0001
Error	379	335644522201	885605599.47		
Corrected Total	382	556152128658			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.396488	21.55482	29759.13	138062.5

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GrLivArea	1	190367576250	190367576250	214.96	<.0001
Neighborhood	2	30140030207	15070015103	17.02	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GrLivArea	1	187035079034	187035079034	211.19	<.0001
Neighborhood	2	30140030207	15070015103	17.02	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	85887.15908	B	4578.116845	18.76	<.0001	76885.46901	94888.84915
GrLivArea	45.76006		3.148801	14.53	<.0001	39.56876	51.95137
Neighborhood BrkSide	-16105.62078	B	4395.351815	-3.66	0.0003	-24747.95040	-7463.29116
Neighborhood Edwards	-18987.77587	B	3577.826684	-5.31	<.0001	-26022.65245	-11952.89930
Neighborhood NAmes	0.00000	B	.	.	.	.	.



The MEANS Procedure

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1459	178206.04	70195.92	829.3645081	439665.44

SAS Code Display 4 (see appendix for code)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Neighborhood	3	BrkSide Edwards NAmes

Number of Observations Read	381
Number of Observations Used	381

The GLM Procedure

Dependent Variable: SalePrice

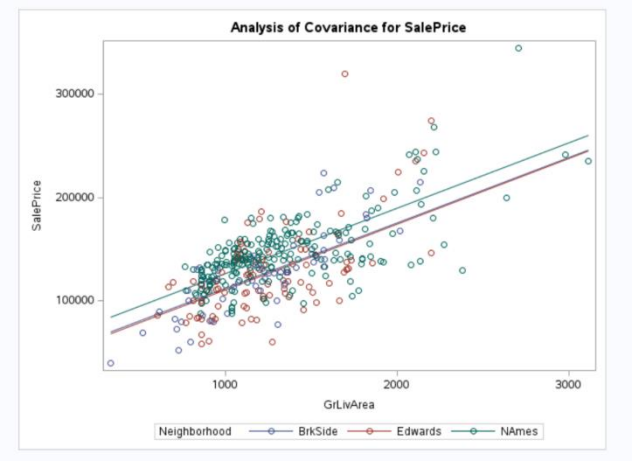
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	273720881458	91240293819	122.95	<.0001
Error	377	279757910212	742063422.31		
Corrected Total	380	553478791670			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.494546	19.75658	27240.84	137882.4

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GrLivArea	1	253099925462	253099925462	341.08	<.0001
Neighborhood	2	20620955996	10310477998	13.89	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GrLivArea	1	238636449532	238636449532	321.58	<.0001
Neighborhood	2	20620955996	10310477998	13.89	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	62577.22112	B	4985.940829	12.55	<.0001	52773.48340	72380.95885
GrLivArea	63.54969		3.543770	17.93	<.0001	56.58165	70.51772
Neighborhood BrkSide	-14197.82366	B	4029.477402	-3.52	0.0005	-22120.88993	-6274.75739
Neighborhood Edwards	-15464.83732	B	3301.412173	-4.68	<.0001	-21956.32612	-8973.34852
Neighborhood NAmes	0.00000	B	.	.	.	.	.



SAS Code Display 5 (see appendix for code)

The GLM Procedure

Class Level Information		
Class	Levels	Values
Neighborhood	3	BrkSide Edwards NAmes

Number of Observations Read	381
Number of Observations Used	381

The GLM Procedure

Dependent Variable: SalePrice

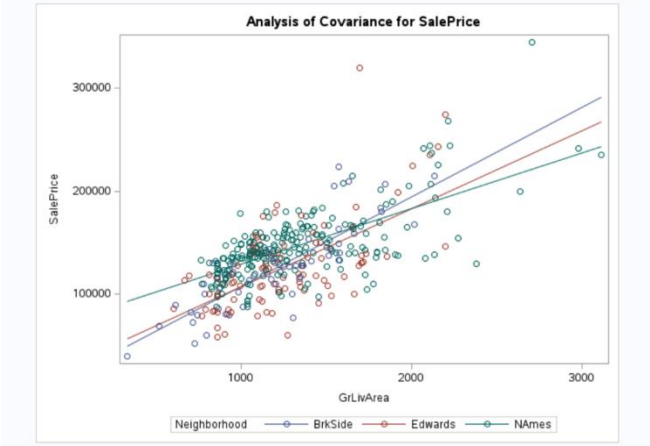
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	283631588727	56726317745	78.83	<.0001
Error	375	269847202943	719592541.18		
Corrected Total	380	553478791670			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.512452	19.45515	26825.22	137882.4

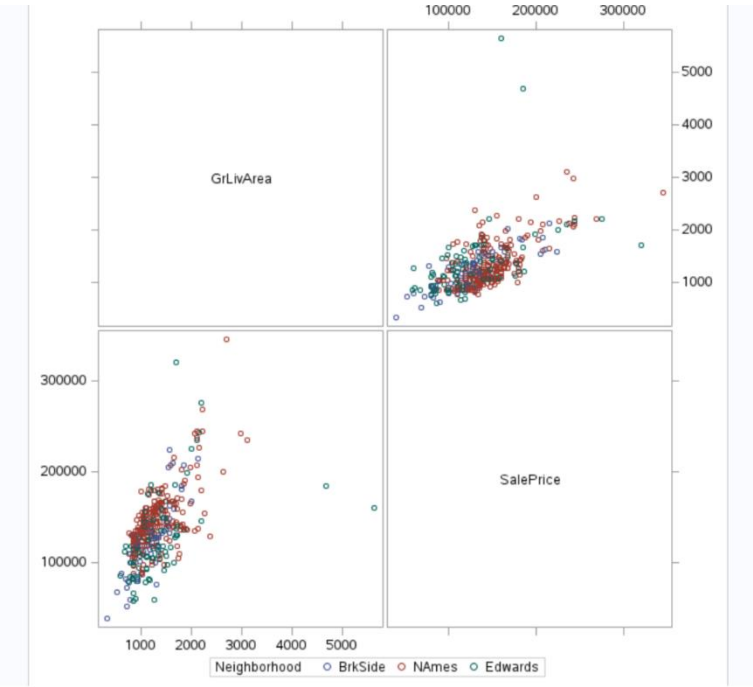
Source	DF	Type I SS	Mean Square	F Value	Pr > F
GrLivArea	1	253099925462	253099925462	351.73	<.0001
Neighborhood	2	20620955996	10310477998	14.33	<.0001
GrLivArea*Neighborho	2	9910707269.1	4955353634.6	6.89	0.0012

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GrLivArea	1	210178319563	210178319563	292.08	<.0001
Neighborhood	2	18323983921	9161991960.5	12.73	<.0001
GrLivArea*Neighborho	2	9910707269.1	4955353634.6	6.89	0.0012

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	74676.40154	B	5954.52674	12.54	<.0001	62967.95510	86384.84798
GrLivArea	54.31586	B	4.33457	12.53	<.0001	45.79276	62.83896
Neighborhood BrkSide	-54704.88774	B	13042.61747	-4.19	<.0001	-80350.71900	-29059.05648
Neighborhood Edwards	-43247.84694	B	11671.23793	-3.71	0.0002	-66197.12068	-20298.57320
Neighborhood NAmes	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho BrkSide	32.84667	B	10.16117	3.23	0.0013	12.86665	52.82669
GrLivArea*Neighborho Edwards	21.66057	B	8.79973	2.46	0.0143	4.35757	38.96358
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.	.	.



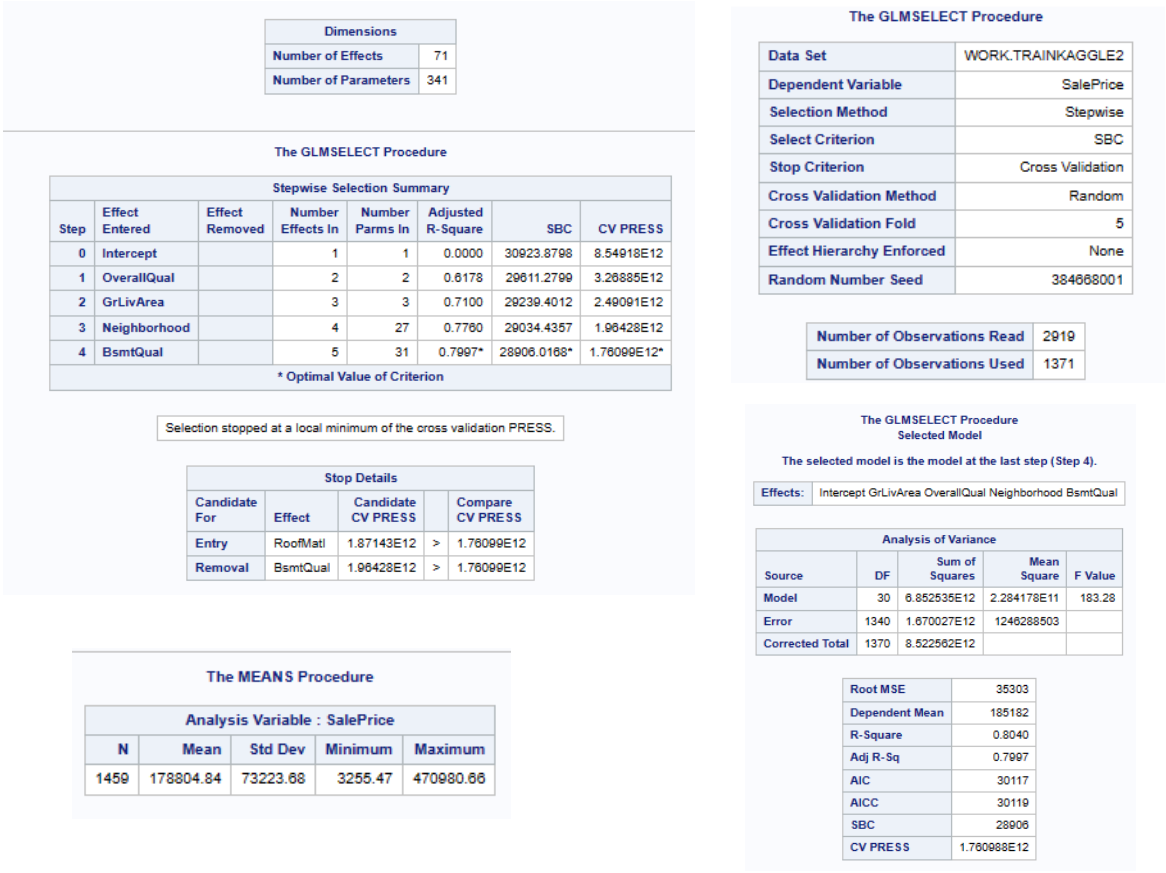
SAS Code Display 6 (see appendix for code)



SAS Code Display 17

```
Stepwise
proc glmselect data=trainkaggle2 seed=384668001 plots=all;
class  LotFrontage MSZoning Street 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 GarageType GarageFinish GarageQual GarageCond PavedDrive
SaleType SaleCondition;
model SalePrice = LotFrontage GrLivArea MSSubClass LotArea OverallQual OverallCond YearBuilt
YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2
BsmtUnfSF TotalBsmtSF LowQualFinSF BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF
EnclosedPorch ScreenPorch PoolArea MiscVal MoSold YrSold Neighborhood MSZoning Street LotShape
LandContour Utilities LotConfig LandSlope Condition1 BldgType HouseStyle RoofStyle RoofMatl Exterior1st
Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1
BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish
GarageQual GarageCond PavedDrive SaleType SaleCondition
/ selection= Stepwise (stop=CV) cvmethod=random(5) stats=adjrsq;
output out =  results p = Predict;
run;
/* Can't have negative predictions because of RMLSE */
/* Also must have only two columns with appropraite labels. */
data results3;
set results;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 10000;
keep id SalePrice;
where id > 1460;
;
proc means data = results3;
var SalePrice;
```

run;



Dimensions	
Number of Effects	71
Number of Parameters	341

The GLMSELECT Procedure

Stepwise Selection Summary							
Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept		1	1	0.0000	30923.8798	8.54918E12
1	OverallQual		2	2	0.6178	29611.2799	3.26885E12
2	GrLivArea		3	3	0.7100	29239.4012	2.49091E12
3	Neighborhood		4	27	0.7760	29034.4357	1.96428E12
4	BsmtQual		5	31	0.7997*	28906.0168*	1.76099E12*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details				
Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	RoofMatl	1.87143E12	>	1.76099E12
Removal	BsmtQual	1.96428E12	>	1.76099E12

The MEANS Procedure

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1459	178804.84	73223.68	3255.47	470980.66

The GLMSELECT Procedure

Data Set	WORK.TRAINKAGGLE2
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	384668001

Number of Observations Read	2919
Number of Observations Used	1371

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 4).

Effects:	Intercept GrLivArea OverallQual Neighborhood BsmtQual

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	30	6.852535E12	2.284178E11	183.28
Error	1340	1.670027E12	1246288503	
Corrected Total	1370	8.522562E12		

Root MSE	35303
Dependent Mean	185182
R-Square	0.8040
Adj R-Sq	0.7997
AIC	30117
AICC	30119
SBC	28906
CV PRESS	1.760988E12

SAS Code Display 18

```
/* Analysis part 2 */
/* Step 2: Model Building Forward Selection with CV press with all Neighborhoods */
proc glmselect data=trainkaggle2 seed=952011000 plots=all;
partition fraction(test=.2);
class  LotFrontage MSZoning Street 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 GarageType GarageFinish GarageQual GarageCond PavedDrive
SaleType SaleCondition;
model SalePrice =  LotFrontage GrLivArea MSSubClass LotArea OverallQual OverallCond YearBuilt
YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF LowQualFinSF BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt
GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea MiscVal MoSold
YrSold Neighborhood MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Condition1
BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation
BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition
/ selection= Forward (stop=CV) cvmethod=random(5) stats=adjrsq;
output out =  results p = Predict;
run;

data results4;
```

```
set results;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 10000;
keep id SalePrice;
where id > 1460;
;

proc means data = results4;
var SalePrice;
run;
```

The GLMSELECT Procedure

Data Set	WORK.TRAINKAGGLE2
Dependent Variable	SalePrice
Selection Method	Forward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	952011000

Number of Observations Read	2919
Number of Observations Used	1371
Number of Observations Used for Training	1091
Number of Observations Used for Testing	280

Dimensions	
Number of Effects	71
Number of Parameters	341

The GLMSELECT Procedure

Forward Selection Summary

Step	Effect Entered	Number Effects In	Number Parms In	Adjusted R-Square	SBC	ASE	Test ASE	CV PRESS
0	Intercept	1	1	0.0000	24536.9626	5816299981	7775184938	6.36201E12
1	OverallQual	2	2	0.6245	23474.2355	2181837801	3130831235	2.4033E12
2	GrLivArea	3	3	0.7088	23203.0582	1690793042	2267339496	1.93608E12
3	Neighborhood	4	27	0.7819	23030.9883	1238134903	1918786258	1.48643E12
4	BsmtQual	5	31	0.8065	22924.1968	1094257980	1761134882	1.34129E12
5	MSSubClass	6	32	0.8199	22851.8131	1017469673	1647615227	1.26838E12
6	RoofMatl	7	39	0.8387*	22773.6394*	905548631	1512185855	1.25539E12*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	BsmtFinSF1	1.28133E12	>	1.25539E12

The selected model is the model at the last step (Step 6).

Effects:	Intercept GrLivArea MSSubClass OverallQual Neighborhood RoofMatl BsmtQual

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	38	5.35763E12	1.409903E11	150.13
Error	1052	9.879536E11	939119350	
Corrected Total	1090	6.345583E12		

Root MSE	30645
Dependent Mean	184951
R-Square	0.8443
Adj R-Sq	0.8387
AIC	23672
AICC	23675
SBC	22774
ASE (Train)	905548631
ASE (Test)	1512185855
CV PRESS	1.255386E12

The MEANS Procedure

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1459	178310.77	72974.24	15172.67	495416.82

SAS Code Display 19

```
/* Step 2: Model Building Backward Selection with CV press with all Neighborhoods */
proc glmselect data=trainkaggle2 seed=631197001 plots=all;
partition fraction(test=.2);
class  LotFrontage MSZoning Street 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 GarageType GarageFinish GarageQual GarageCond PavedDrive
SaleType SaleCondition;
model SalePrice =  LotFrontage GrLivArea MSSubClass LotArea OverallQual OverallCond YearBuilt
YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF LowQualFinSF BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt
```

```
GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch PoolArea MiscVal MoSold
YrSold Neighborhood MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Condition1
BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation
BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition
/ selection= Backward (stop=CV) cvmethod=random(5) stats=adjrsq;
run;

data results5;
set results;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 10000;
keep id SalePrice;
where id > 1460;
;

proc means data = results5;
var SalePrice;
run;
```

The GLMSELECT Procedure

Data Set	WORK.TRAINKAGGLE2
Dependent Variable	SalePrice
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	631197001

Number of Observations Read	2919
Number of Observations Used	1371
Number of Observations Used for Training	1104
Number of Observations Used for Testing	267

Dimensions	
Number of Effects	71
Number of Parameters	341

The GLMSELECT Procedure

Backward Selection Summary

Step	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	ASE	Test ASE	CV PRESS
0		71	291	0.9063*	23871.5956	387779979	1047262619	1.50136E12
	TotalBsmtSF	70	291	0.9063*	23871.5956	387779979	1047262619	1.50136E12
1	LotFrontage	69	213	0.9015	23480.8800	446548986	1020694743	1.32318E12
2	Exterior2nd	68	199	0.9022	23392.3097	450425894	991438595	1.29343E12
3	HouseStyle	67	192	0.9028	23345.2516	451238048	994682906	1.28279E12
4	Exterior1st	66	181	0.9011	23301.0362	464870022	1004136481	1.25272E12
5	BsmtFinType2	65	175	0.9012	23264.6998	467277979	1016376682	1.24682E12
6	RoofStyle	64	170	0.9015	23232.4615	468462562	1005722191	1.23289E12
7	Heating	63	165	0.9017	23200.3086	469686502	1005984481	1.22851E12
8	Foundation	62	160	0.9019	23169.0843	471309849	1015774937	1.21198E12
9	GarageType	61	155	0.9021	23138.0856	473035503	1007301945	1.19835E12
10	SaleType	60	150	0.9020	23109.3471*	475740448	1007982475	1.19447E12*

* Optimal Value of Criterion

Note: Effects dropped at step 0 are redundant.

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Removal	Electrical	1.19721E12	>	1.19447E12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	149	5.673075E12	38074327221	69.16
Error	954	5.252175E11	550542406	
Corrected Total	1103	6.198292E12		

Root MSE	23464
Dependent Mean	182962
R-Square	0.9153
Adj R-Sq	0.9020
AIC	23464
AICC	23513
SBC	23109
ASE (Train)	475740448
ASE (Test)	1007982475
CV PRESS	1.194469E12

The MEANS Procedure

Analysis Variable : SalePrice

N	Mean	Std Dev	Minimum	Maximum
1459	178145.31	73046.64	7078.36	475727.99

SAS Code Display 20

```
/* Custom model add and remove variables at will */
proc glm data = trainkaggle2 plots = all;
class Neighborhood BsmtQual HouseStyle RoofMatl;
model SalePrice = Neighborhood BsmtQual HouseStyle OverallQual GrLivArea LotArea MSSubClass / cli
solution;
output out =  results p = Predict;
run;

/* Can't have negative predictions because of RMLSE */
/* Also must have only two columns with appropraite labels. */

data results7;
set results;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 10000;
keep id SalePrice;
where id > 1460;
;

proc means data = results7;
var SalePrice;
run;
```

Number of Observations Read	2919
Number of Observations Used	1460

The GLM Procedure

Dependent Variable: SalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	39	7.6767279E12	196839176783	182.55	<.0001
Error	1420	1.5311834E12	1078298197.2		
Corrected Total	1459	9.2079113E12			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.833710	18.15014	32837.45	180921.2

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Neighborhood	24	5.0236061E12	209316922573	194.12	<.0001
BsmtQual	4	922130654789	230532663697	213.79	<.0001
HouseStyle	7	152738385682	21819769383	20.24	<.0001
OverallQual	1	807795575865	807795575865	749.14	<.0001
GrLivArea	1	668729606672	668729606672	620.17	<.0001
LotArea	1	45095525718	45095525718	41.82	<.0001
MSSubClass	1	56632004070	56632004070	52.52	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Neighborhood	24	350322864446	14596786019	13.54	<.0001
BsmtQual	4	161549246286	40387311571	37.45	<.0001
HouseStyle	7	57577805644	8225400806.3	7.63	<.0001
OverallQual	1	216628052134	216628052134	200.90	<.0001
GrLivArea	1	526737771446	526737771446	488.49	<.0001
LotArea	1	41741941760	41741941760	38.71	<.0001
MSSubClass	1	56632004070	56632004070	52.52	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	39542.88588	B	13049.23882	3.03	0.0025
Neighborhood Blmngtn	-27436.49773	B	12940.97404	-2.12	0.0342
Neighborhood Blueste	-35403.82924	B	25466.76380	-1.39	0.1647
Neighborhood BrDale	-43768.99795	B	13547.70985	-3.23	0.0013
Neighborhood BrkSide	-45862.49331	B	11252.31984	-4.08	<.0001
Neighborhood ClearCr	-28949.81942	B	11876.90044	-2.44	0.0149
Neighborhood CollgCr	-30448.53185	B	10344.78962	-2.94	0.0033
Neighborhood Crawfor	-22233.41276	B	11080.22812	-2.01	0.0450
Neighborhood Edwards	-59675.03248	B	10682.57654	-5.59	<.0001
Neighborhood Gilbert	-36305.49795	B	10737.56961	-3.38	0.0007
Neighborhood IDOTRR	-61781.44546	B	11737.41232	-5.26	<.0001
Neighborhood MeadowV	-41930.07344	B	13293.72438	-3.15	0.0016
Neighborhood Mitchel	-49745.99324	B	11086.81497	-4.49	<.0001
Neighborhood NAmes	-48980.78670	B	10408.21732	-4.71	<.0001
Neighborhood NPkVill	-30778.65041	B	15009.40080	-2.05	0.0405
Neighborhood NWAmes	-46996.80504	B	10725.99827	-4.38	<.0001
Neighborhood NoRidge	24093.66138	B	11393.31831	2.11	0.0346
Neighborhood NridgHt	11423.95068	B	10850.13571	1.05	0.2926
Neighborhood OldTown	-60753.37753	B	10815.44907	-5.62	<.0001
Neighborhood SWISU	-62472.91786	B	12423.35852	-5.03	<.0001
Neighborhood Sawyer	-48435.57101	B	10897.84272	-4.44	<.0001
Neighborhood SawyerW	-37209.61361	B	10873.39333	-3.42	0.0006
Neighborhood Somerst	-16053.01347	B	10650.10010	-1.51	0.1320
Neighborhood StoneBr	25722.77939	B	12041.29798	2.14	0.0328
Neighborhood Timber	-32299.75797	B	11317.37312	-2.85	0.0044
Neighborhood Veenker	0.00000	B	.	.	.
BsmtQual Ex	51352.65855	B	4806.19231	10.68	<.0001
BsmtQual Fa	-4938.93565	B	5965.57426	-0.83	0.4079
BsmtQual Gd	4299.73410	B	2942.96341	1.46	0.1442
BsmtQual NA	-13506.61129	B	5781.16358	-2.34	0.0196
BsmtQual TA	0.00000	B	.	.	.
HouseStyle 1.5Fin	-23383.82383	B	5329.40485	-4.39	<.0001
HouseStyle 1.5Unf	-16844.44291	B	10115.30090	-1.67	0.0961
HouseStyle 1Story	-8240.89673	B	4636.93155	-1.78	0.0757
HouseStyle 2.5Fin	-36592.56463	B	13550.34419	-2.70	0.0070
HouseStyle 2.5Unf	-35208.97067	B	11266.01758	-3.13	0.0018
HouseStyle 2Story	-20595.90158	B	4722.53981	-4.36	<.0001
HouseStyle SFoyer	12084.47331	B	7065.97203	1.71	0.0874
HouseStyle SLvl	0.00000	B	.	.	.
OverallQual	15899.04271		1121.71640	14.17	<.0001
GrLivArea	62.09395		2.80945	22.10	<.0001
LotArea	0.61308		0.09854	6.22	<.0001
MSSubClass	-207.49520		28.63166	-7.25	<.0001

Sum of Residuals	-9.626092E-8
Sum of Squared Residuals	1.5311834E12
Sum of Squared Residuals - Error SS	-0.001464844
PRESS Statistic	1.6752509E12
First Order Autocorrelation	0.0211610827
Durbin-Watson D	1.957481468

Observations	1460
Parameters	40
Error DF	1420
MSE	1.08E9
R-Square	0.8337
Adj R-Square	0.8291

SHINEY APP CODE

```
library(class)
library(shiny)
library(caret)
library(ggplot2)
library(dplyr)
library(magrittr)
library(tidyverse)
library(readr)
library(plotly)

ui <- fluidPage(
  # App title ----
  titlePanel("House Prices - Advanced Regression Techniques!"),

  # Sidebar layout with input and output definitions ----
  sidebarLayout(

    # Sidebar panel for inputs ----
```

```
    sidebarPanel(

      # Input: Select a file ----
  fileInput("file1", "Choose the Training CSV file", multiple = FALSE, accept = c(".csv", ".txt")),
#  fileInput("file2", "Choose the Test CSV file", multiple = FALSE, accept = c(".csv", ".txt")),


  selectizeInput(
    inputId = "ExpVar",
    label = "Select an Explanatory Variable",
    choices = c(),
    multiple = FALSE,
    selected = "GrLivArea",
    options = list(maxItems = 1)
  ),

  selectizeInput(inputId = "ResVar", label = "Select a Response Variable", choices = c(),
              multiple = FALSE, selected = "SalePrice",options = list(maxItems = 1)),

  checkboxInput("SepNeighborhood", "Look at Neighborhoods separately?")
    ),

  # Main panel for displaying outputs ----
    tabsetPanel(
      tabPanel("Data Review",
               # content for the first tab goes here
               tableOutput("headTrain")
               #tableOutput("headTest"),
               #tableOutput("Selection"),
     ),
      tabPanel("Plots",
               # content for the second tab goes here
               fluidRow(
                 column(6, plotOutput(outputId = "Plot11")),
                 column(6, plotOutput(outputId = "Plot12"))
               ),
               fluidRow(
               column(6,plotOutput(outputId = "Plot21")),
               column(6,plotOutput(outputId = "Plot22"))

              )
            )
    )

),
)


server <- function(input, output, session) {

  HPA <- reactive({
    req(input$file1)

    ext <- tools::file_ext(input$file1$name)
    switch(ext,
           csv = vroom::vroom(input$file1$datapath, delim = ","),
```

```
        txt = vroom::vroom(input$file1$datapath, delim = "\t"),
        validate("Invalid file; Please upload a .csv or .txt file")
    )
    inFile <- input$file1
    df <- read_csv(inFile$datapath, col_types = cols())

    updateSelectInput(session, "ExpVar", choices = colnames(df))
    updateSelectInput(session, "ResVar", choices = colnames(df))

    # Check if each column has only integer values and convert to integer if true
    #df <- purrr::map_df(df, ~ if (all(is.na(.x) | is.numeric(.x)) && all(floor(.x) == .x)) {as.integer(.x)} else {.x})
    return(df)

  })

  output$headTrain <- renderTable({
    req(HPA())
    head(HPA(), 5)
    #as.data.frame(as.list(head(HPA(), 5)))
    })


  output$Plot11 <- renderPlot({
      {HPA() %>%  filter(Neighborhood == 'NAmes'| Neighborhood == 'Edwards' | Neighborhood == 'BrkSide')
%>% ggplot(aes(x = GrLivArea, y = SalePrice)) + geom_point(aes(color = Neighborhood)) +
geom_smooth(method = "lm") + theme(legend.position = "right")  + ggtitle("Home Price Analysis by
Neighborhood: Sale Price v. Gross Living Area")}

  })

  output$Plot12 <- renderPlot({
    if(input$SepNeighborhood == TRUE)
    {HPA() %>%  filter(Neighborhood == 'NAmes') %>% ggplot(aes(x = GrLivArea, y = SalePrice)) +
geom_point(aes(color = Neighborhood)) + geom_smooth(method = "lm")  + theme(legend.position = "right")  +
ggtitle("Home Price Analysis for NAmes Neighborhood: Sale Price v. Gross Living Area")}

  })

  output$Plot21 <- renderPlot({
    if(input$SepNeighborhood == TRUE)
    {HPA() %>%  filter(Neighborhood == 'Edwards') %>% ggplot(aes(x = GrLivArea, y = SalePrice)) +
geom_point(aes(color = Neighborhood)) + geom_smooth(method = "lm")  + theme(legend.position = "right")  +
ggtitle("Home Price Analysis for Edwards Neighborhood: Sale Price v. Gross Living Area")}


  })

  output$Plot22 <- renderPlot({
    if(input$SepNeighborhood == TRUE)
    {HPA() %>%  filter(Neighborhood == 'BrkSide')  %>% ggplot(aes(x = GrLivArea, y = SalePrice)) +
geom_point(aes(color = Neighborhood)) + geom_smooth(method = "lm") + theme(legend.position = "right")  +
ggtitle("Home Price Analysis for BrookkSide Neighborhood: Sale Price v. Gross Living Area")}


  })
```

```
  }

shinyApp(ui, server)
```

SAS Code

```
/*Scatter Plot*/
proc sgplot data = HPA;
scatter x= GrLivArea y = SalePrice / group=Neighborhood;
run;

/Matrix Plot - Display 6*/
proc sgscatter data=HPA;
matrix GrLivArea SalePrice / group=Neighborhood;
run;

/*Correlation Coefficient - Display 8*/
proc corr data = HPA;
var SalePrice GrLivArea;
run;

/*Fit a Model for Initial Analysis - Display 7, 9, 10, 12*/
PROC REG DATA=HPA plots=all;
  MODEL SalePrice = GrLivArea / STB;
RUN;

/*Get Outliers and Leverage Observtions on a plot - Display 11*/
proc reg data=HPA  plots(only label) =(CooksD RStudentByLeverage);
   model SalePrice = GrLivArea; /* can also use INFLUENCE option */
run;

/*Get DFFITS, COOKSD and H-Leverage - Display 11*/
PROC REG DATA=HPA;
  MODEL SalePrice = GrLivArea / STB;
  OUTPUT OUT=OutputDataset DFFITS=DFFITS COVRATIO=COOKD H=H;
RUN;

proc print data = OutputDataset;
run;

/*Get Variance Inflation, Partial Residuals*/
PROC REG DATA=HPA;
  MODEL SalePrice = GrLivArea / STB clb vif scorr1 scorr2;
run;

PROC REG DATA=HPA;
where GrLivArea < 4500;
  MODEL SalePrice = GrLivArea / STB clb vif scorr1 scorr2;
run;

/*Correlation improvement by limiting analysis to < 4500 sf*/
proc corr data = HPA;
where GrLivArea < 4500;
var SalePrice GrLivArea;
run;

/*Build a model with Conf Intervals - Display 3*/
proc glm data = HPA plots = all;
class Neighborhood;
model SalePrice = GrLivArea Neighborhood /solution clparm;
run;
```

```
/*Build a model with Conf Intervals by limiting analysis to < 4500 sf - Display 4*/
proc glm data = HPA plots = all;
where GrLivArea < 4500;
class Neighborhood;
model SalePrice = GrLivArea Neighborhood /solution clparm;
run;

/*Build a model with Conf Intervals by limiting analysis to < 4500 sf and adding interactions - Display 5, 13, 17*/
proc glm data = HPA plots = all;
where GrLivArea < 4500;
class Neighborhood;
model SalePrice = GrLivArea | Neighborhood /solution clparm;
run;

/*Fit a model and get diagnostic statistics - Display 14*/
proc glmselect data=HPA plots = all;
  class Neighborhood;
  model SalePrice = GrLivArea|Neighborhood/selection= Forward (stop=CV) cvmethod=random(5) stats=adjrsq;
run;

/*Fit a model and get diagnostic statistics by limiting analysis to < 4500 sf - Display 15*/
proc glmselect data=HPA plots = all;
where GrLivArea < 4500;
  class Neighborhood;
  model SalePrice = GrLivArea Neighborhood/selection= Forward (stop=CV) cvmethod=random(5) stats=adjrsq;
run;

/*Fit a model and get diagnostic statistics by limiting analysis to < 4500 sf and adding interactions - Display 16*/
proc glmselect data=HPA plots = all;
where GrLivArea < 4500;
  class Neighborhood;
  model SalePrice = GrLivArea|Neighborhood/selection= Forward (stop=CV) cvmethod=random(5) stats=adjrsq;
run;
```

APPENDIX B: Data Dictionary

Data fields
Here's a brief version of what you'll find in the data description file.
SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
MSSubClass: The building class
MSZoning: The general zoning classification
LotFrontage: Linear feet of street connected to property
LotArea: Lot size in square feet
Street: Type of road access
Alley: Type of alley access
LotShape: General shape of property
LandContour: Flatness of the property
Utilities: Type of utilities available
LotConfig: Lot configuration
LandSlope: Slope of property
Neighborhood: Physical locations within Ames city limits
Condition1: Proximity to main road or railroad
Condition2: Proximity to main road or railroad (if a second is present)
BldgType: Type of dwelling
HouseStyle: Style of dwelling

OverallQual: Overall material and finish quality
OverallCond: Overall condition rating
YearBuilt: Original construction date
YearRemodAdd: Remodel date
RoofStyle: Type of roof
RoofMatl: Roof material
Exterior1st: Exterior covering on house
Exterior2nd: Exterior covering on house (if more than one material)
MasVnrType: Masonry veneer type
MasVnrArea: Masonry veneer area in square feet
ExterQual: Exterior material quality
ExterCond: Present condition of the material on the exterior
Foundation: Type of foundation
BsmtQual: Height of the basement
BsmtCond: General condition of the basement
BsmtExposure: Walkout or garden level basement walls
BsmtFinType1: Quality of basement finished area
BsmtFinSF1: Type 1 finished square feet
BsmtFinType2: Quality of second finished area (if present)
BsmtFinSF2: Type 2 finished square feet
BsmtUnfSF: Unfinished square feet of basement area
TotalBsmtSF: Total square feet of basement area
Heating: Type of heating
HeatingQC: Heating quality and condition
CentralAir: Central air conditioning
Electrical: Electrical system
1stFlrSF: First Floor square feet
2ndFlrSF: Second floor square feet
LowQualFinSF: Low quality finished square feet (all floors)
GrLivArea: Above grade (ground) living area square feet
BsmtFullBath: Basement full bathrooms
BsmtHalfBath: Basement half bathrooms
FullBath: Full bathrooms above grade
HalfBath: Half baths above grade
Bedroom: Number of bedrooms above basement level
Kitchen: Number of kitchens
KitchenQual: Kitchen quality
TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
Functional: Home functionality rating
Fireplaces: Number of fireplaces
FireplaceQu: Fireplace quality
GarageType: Garage location
GarageYrBlt: Year garage was built
GarageFinish: Interior finish of the garage
GarageCars: Size of garage in car capacity
GarageArea: Size of garage in square feet
GarageQual: Garage quality
GarageCond: Garage condition
PavedDrive: Paved driveway
WoodDeckSF: Wood deck area in square feet
OpenPorchSF: Open porch area in square feet
EnclosedPorch: Enclosed porch area in square feet
3SsnPorch: Three season porch area in square feet
ScreenPorch: Screen porch area in square feet
PoolArea: Pool area in square feet
PoolQC: Pool quality

Fence: Fence quality
MiscFeature: Miscellaneous feature not covered in other categories
MiscVal: $Value of miscellaneous feature
MoSold: Month Sold
YrSold: Year Sold
SaleType: Type of sale
SaleCondition: Condition of sale

APPENDIX C: **Data Description File**

MSSubClass: Identifies the type of dwelling involved in the sale.

20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A	Agriculture
C	Commercial
FV	Floating Village Residential
I	Industrial
RH	Residential High Density
RL	Residential Low Density
RP	Residential Low Density Park
RM	Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl	Gravel
Pave	Paved

Alley: Type of alley access to property

Grvl	Gravel
Pave	Paved
NA	No alley access

LotShape: General shape of property

Reg Regular
IR1 Slightly irregular
IR2 Moderately Irregular
IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level
Bnk Banked - Quick and significant rise from street grade to building
HLS Hillside - Significant slope from side to side
LowDepression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)
NoSewr Electricity, Gas, and Water (Septic Tank)
NoSeWa Electricity and Gas Only
ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot
Corner Corner lot
CulDSac Cul-de-sac
FR2 Frontage on 2 sides of property
FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope
Mod Moderate Slope
Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights
Blueste Bluestem
BrDale Briardale
BrkSide Brookside
ClearCr Clear Creek
CollgCr College Creek
Crawfor Crawford
Edwards Edwards
Gilbert Gilbert
IDOTRR Iowa DOT and Rail Road
MeadowV Meadow Village
Mitchel Mitchell
Names North Ames
NoRidge Northridge
NPkVill Northpark Villa
NridgHt Northridge Heights
NWAmes Northwest Ames
OldTown Old Town
SWISU South & West of Iowa State University
Sawyer Sawyer

SawyerW	Sawyer West
Somerst	Somerset
StoneBr	Stone Brook
Timber	Timberland
Veenker	Veenker

Condition1: Proximity to various conditions

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street
Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street
Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam	Single-family Detached
2FmCon	Two-family Conversion; originally built as one-family dwelling
Duplx	Duplex
TwnhsE	Townhouse End Unit
TwnhsI	Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story	One story
1.5Fin	One and one-half story: 2nd level finished
1.5Unf	One and one-half story: 2nd level unfinished
2Story	Two story
2.5Fin	Two and one-half story: 2nd level finished
2.5Unf	Two and one-half story: 2nd level unfinished
SFoyer	Split Foyer
SLvl	Split Level

OverallQual: Rates the overall material and finish of the house

10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average

5 Average
4 Below Average
3 Fair
2 Poor
1 Very Poor

OverallCond: Rates the overall condition of the house

10 Very Excellent
9 Excellent
8 Very Good
7 Good
6 Above Average
5 Average
4 Below Average
3 Fair
2 Poor
1 Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat	Flat
Gable	Gable
Gambrel	Gabrel (Barn)
Hip	Hip
Mansard	Mansard
Shed	Shed

RoofMatl: Roof material

ClyTile	Clay or Tile
CompShg	Standard (Composite) Shingle
Membran	Membrane
Metal	Metal
Roll	Roll
Tar&Grv	Gravel & Tar
WdShake	Wood Shakes
WdShngl	Wood Shingles

Exterior1st: Exterior covering on house

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood

PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
None	None
Stone	Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

Foundation: Type of foundation

BrkTil	Brick & Tile

CBlock Cinder Block
PConc Poured Contrete
Slab Slab
Stone Stone
Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)
Gd Good (90-99 inches)
TA Typical (80-89 inches)
Fa Fair (70-79 inches)
Po Poor (<70 inches
NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent
Gd Good
TA Typical - slight dampness allowed
Fa Fair - dampness or some cracking or settling
Po Poor - Severe cracking, settling, or wetness
NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure
Av Average Exposure (split levels or foyers typically score average or above)
Mn Mimimum Exposure
No No Exposure
NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters
ALQ Average Living Quarters
BLQ Below Average Living Quarters
Rec Average Rec Room
LwQ Low Quality
Unf Unfinshed
NA No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ Good Living Quarters
ALQ Average Living Quarters
BLQ Below Average Living Quarters
Rec Average Rec Room
LwQ Low Quality
Unf Unfinshed
NA No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor	Floor Furnace
GasA	Gas forced warm air furnace
GasW	Gas hot water or steam heat
Grav	Gravity furnace
OthW	Hot water or steam heat other than gas
Wall	Wall furnace

HeatingQC: Heating quality and condition

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

CentralAir: Central air conditioning

N	No
Y	Yes

Electrical: Electrical system

SBrkr	Standard Circuit Breakers & Romex
FuseA	Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF	60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP	60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix	Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality
Min1 Minor Deductions 1
Min2 Minor Deductions 2
Mod Moderate Deductions
Maj1 Major Deductions 1
Maj2 Major Deductions 2
Sev Severely Damaged
Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace
Gd Good - Masonry Fireplace in main level
TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa Fair - Prefabricated Fireplace in basement
Po Poor - Ben Franklin Stove
NA No Fireplace

GarageType: Garage location

2Types More than one type of garage
Attchd Attached to home
Basment Basement Garage
BuiltIn Built-In (Garage part of house - typically has room above garage)
CarPort Car Port
Detchd Detached from home
NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished
RFn Rough Finished
Unf Unfinished
NA No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent

Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

GarageCond: Garage condition

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

PavedDrive: Paved driveway

Y Paved
P Partial Pavement
N Dirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
NA No Pool

Fence: Fence quality

GdPrv Good Privacy
MnPrv Minimum Privacy
GdWo Good Wood
MnWw Minimum Wood/Wire
NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

ElevElevator
Gar2 2nd Garage (if not described in garage section)
OthrOther
Shed Shed (over 100 SF)
TenC Tennis Court

NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD	Warranty Deed - Conventional
CWD	Warranty Deed - Cash
VWD	Warranty Deed - VA Loan
New	Home just constructed and sold
COD	Court Officer Deed/Estate
Con	Contract 15% Down payment regular terms
ConLw	Contract Low Down payment and low interest
ConLI	Contract Low Interest
ConLD	Contract Low Down
Oth	Other

SaleCondition: Condition of sale

Normal	Normal Sale
Abnorml	Abnormal Sale - trade, foreclosure, short sale
AdjLand	Adjoining Land Purchase
Alloca	Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family	Sale between family members
Partial	Home was not completed when last assessed (associated with New Homes)