Data Analysis With SAS: A Home Value Modeling Application

Date: March 13, 2020

Author: Armand Tossou

Contents

1. Purpose 2
2. Data (Ames Residential Home Sales) 2
3. Setup 2
4. Data exploration: Graphical analysis of associations 3
5. Analysis of Variance (ANOVA) and Regression 40
6. Post hoc pairewise comparisons to find out which heating quality has the higher sale price 45
7. Correlation analysis 48
8. Simple linear regression analysis 56
9. Two-way ANOVA: Without Interaction 60
10. Two-way ANOVA: With Interaction 66
11. Multiple Linear Regression Analysis 75
12. Multiple Regression Analysis: With Variable Selection 82
13. 15. Using AIC, BIC Information Criteria To Select Variables 126
14. Model Residual Diagnostics/post fitting for inference 150
15. Identifying outliers and influential observations 158
16. Dealing with collinearity 172
17. Building a predictive model 180
18. Categorical data analysis 187
19. Performing a Chi-square Test of Association 193
20. Simple logistic regression model 197
21. Simple logistic regression model 201
22. Multiple logistic regression model: with interactions 205

Reference: 215

1. Purpose

We use the Statistical Analysis Software (SAS/SAT) to model home sale price. Codes and outputs are included in the project.

2. Data (Ames Residential Home Sales)

All residential home sales in Ames, Iowa between 2006 and 2010. The data set contains many explanatory variables on the quality and quantity of physical attributes of residential homes in Iowa sold between 2006 and 2010. Most of the variables describe information a typical home buyer would like to know about a property (square footage, number of bedrooms and bathrooms, size of lot, etc.). A detailed discussion of variables can be found in the original paper referenced below.

Data Source:

De Cock D. 2011. Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education; 19(3).

Loading data:

- Click this link to download the data in .csv format: CSV Download
- To access the data directly in SAS, type:

```
filename amesh url 'http://www.openintro.org/books/statdata/ames_sas.csv'
```

- To access the data in R, type:

```
download.file("http://www.openintro.org/books/statdata/ames.RData",
              destfile = "ames.RData")
load("ames.RData")
```

3. Setup

To set up our data analysis folder path:

```
%let path=/home/&sysuserid/EST142;


%let homefolder=~/EST142/data;
libname STAT1 "&homefolder";

%include "&homefolder/_1stat_data.sas";

options fmtsearch=(stat1.myfmts);

proc format library=stat1.myfmts;
run;
```

/*options nomprint nosymbolgen nonotes nosource dlcreatedir;*/

```
options mprint symbolgen notes source;
```

/* create macro variables to hold the names of the interval and */

/* categorical variables used in the project and practice programs */

/*%let statements define macro variables containing lists of dataset variables*/

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area
Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;

%let categorical=House_Style2 Overall_Qual2 Overall_Cond2
Fireplaces
         Season_Sold Garage_Type_2 Foundation_2 Heating_QC
         Masonry_Veneer Lot_Shape_2 Central_Air;
```

4. Data exploration: Graphical analysis of associations

/* display all variables in the dataset */

```
proc contents data=stat1._all_ nods;
run;
```

Directory	
Libref	STAT1
Engine	V9
Physical Name	/home/u58887350/EST142/data
Filename	/home/u58887350/EST142/data
Inode Number	15046429239
Access Permission	rwxr-xr-x
Owner Name	u58887350
File Size	4KB
File Size (bytes)	4096

#	Name	Member Type	File Size	Last Modified
1	AMESALTUSE	DATA	256KB	08/20/2021 20:16:20
2	AMESHOUSING	DATA	2MB	08/20/2021 20:16:19
3	AMESHOUSING2	DATA	512KB	08/20/2021 20:16:19
4	AMESHOUSING3	DATA	256KB	08/20/2021 20:16:20
5	AMESHOUSING4	DATA	256KB	08/20/2021 20:16:20
6	BODYFAT	DATA	256KB	08/20/2021 20:16:20
7	BODYFAT2	DATA	256KB	08/20/2021 20:16:20
8	CONCRETE	DATA	256KB	08/20/2021 20:16:20
9	DRUG	DATA	256KB	08/20/2021 20:16:20
10	EXACT	DATA	256KB	08/20/2021 20:16:20
11	GARLIC	DATA	256KB	08/20/2021 20:16:20
12	GERMAN	DATA	256KB	08/20/2021 20:16:20
13	HOSP	DATA	256KB	08/20/2021 20:16:20
14	MARKET	DATA	256KB	08/20/2021 20:16:20
15	MYFMTS	CATALOG	76KB	08/20/2021 20:16:20
16	NORMTEMP	DATA	256KB	08/20/2021 20:16:20
17	SAFETY	DATA	256KB	08/20/2021 20:16:20
18	SAT	DATA	256KB	08/20/2021 20:16:20
19	SAT2013	DATA	256KB	08/20/2021 20:16:20
20	SPENDING2011	DATA	256KB	08/20/2021 20:16:20
21	VEN	DATA	256KB	08/20/2021 20:16:20

/*Exploration of all variables that are available for analysis */

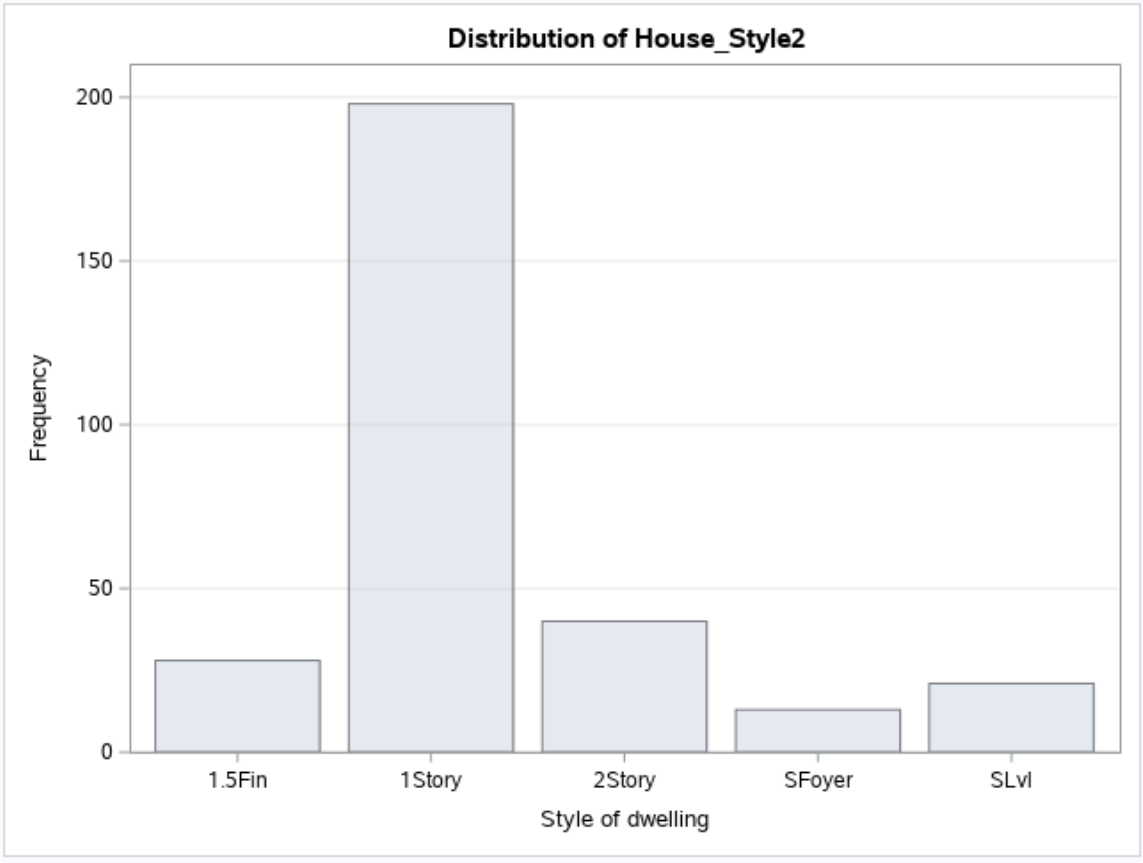
/*PROC FREQ is used with categorical variables*/

```
ods graphics;

proc freq data=STAT1.ameshousing3;
    tables &categorical / plots=freqplot ;
```

```
    format House_Style $House_Style.
           Overall_Qual Overall.
           Overall_Cond Overall.
           Heating_QC $Heating_QC.
           Central_Air $NoYes.
           Masonry_Veneer $NoYes.
           ;
    title "Categorical Variable Frequency Analysis";
run;
```

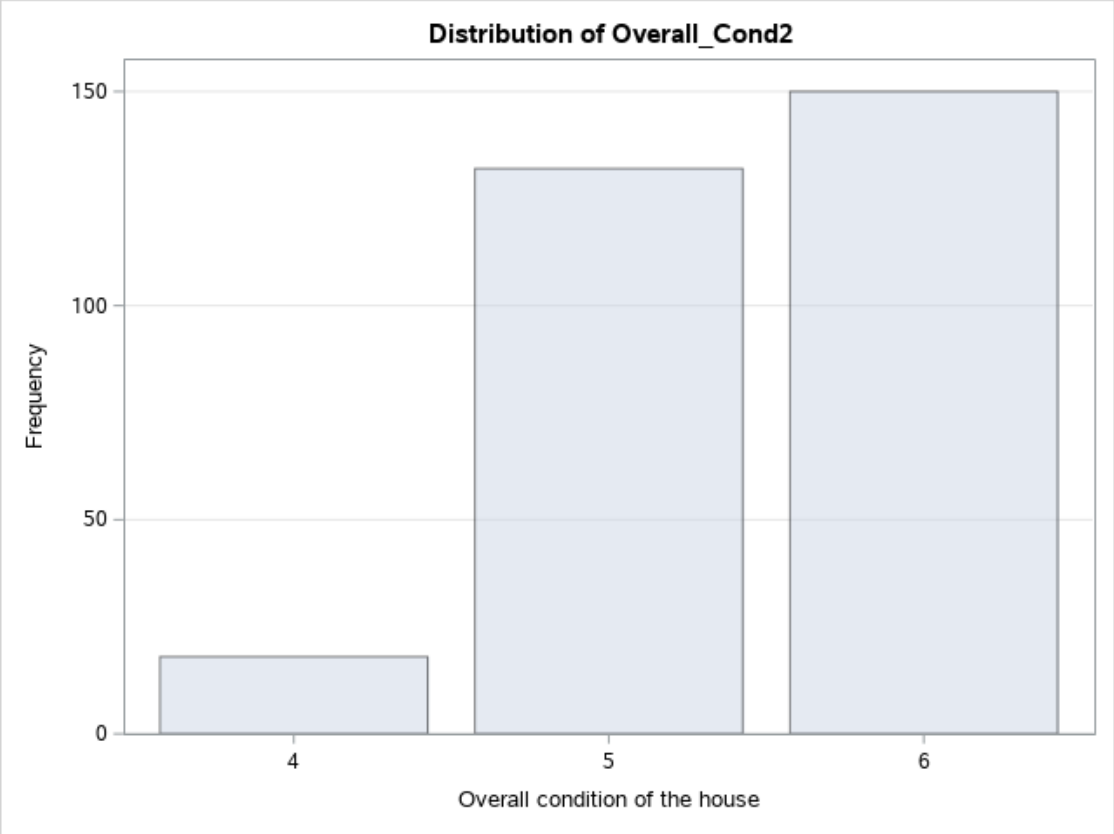
Categorical Variable Frequency Analysis

Style of dwelling				
House_Style2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.5Fin	28	9.33	28	9.33
1Story	198	66.00	226	75.33
2Story	40	13.33	266	88.67
SFoyer	13	4.33	279	93.00
SLvl	21	7.00	300	100.00

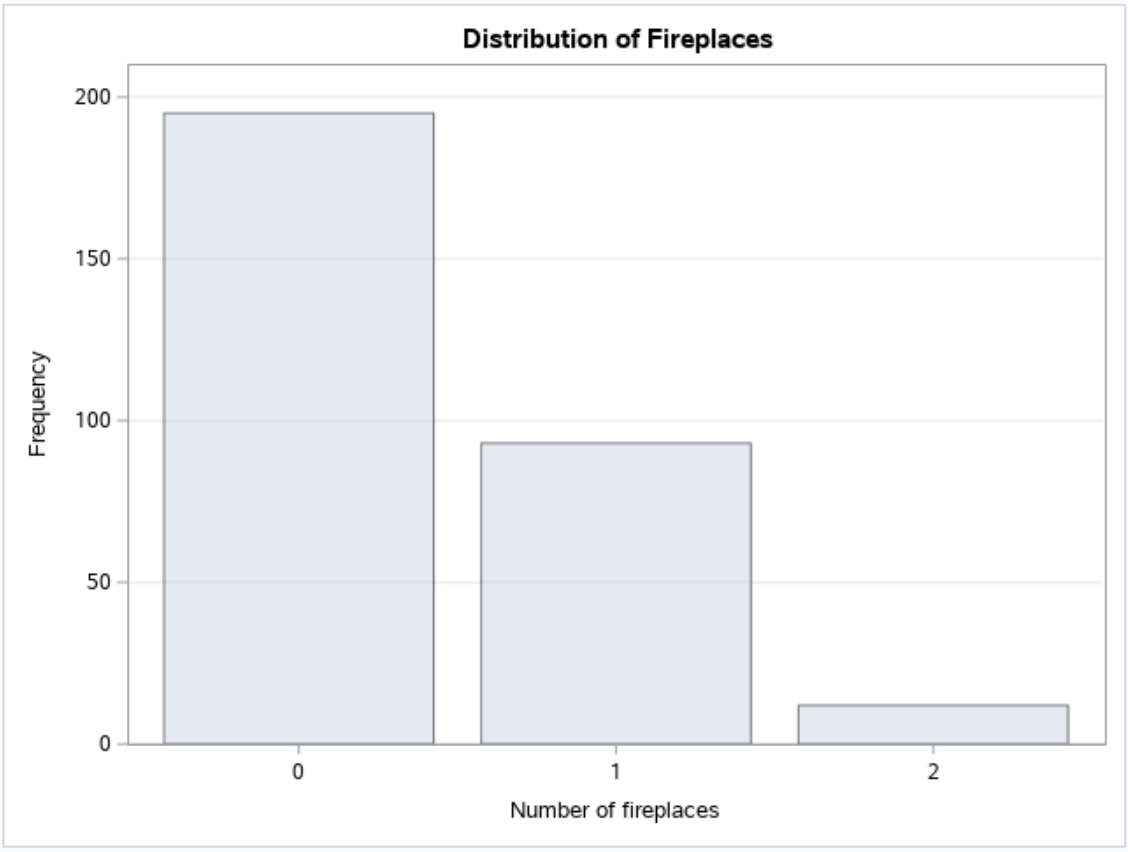


Overall material and finish of the house				
Overall_Qual2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	45	15.00	45	15.00
5	113	37.67	158	52.67
6	142	47.33	300	100.00

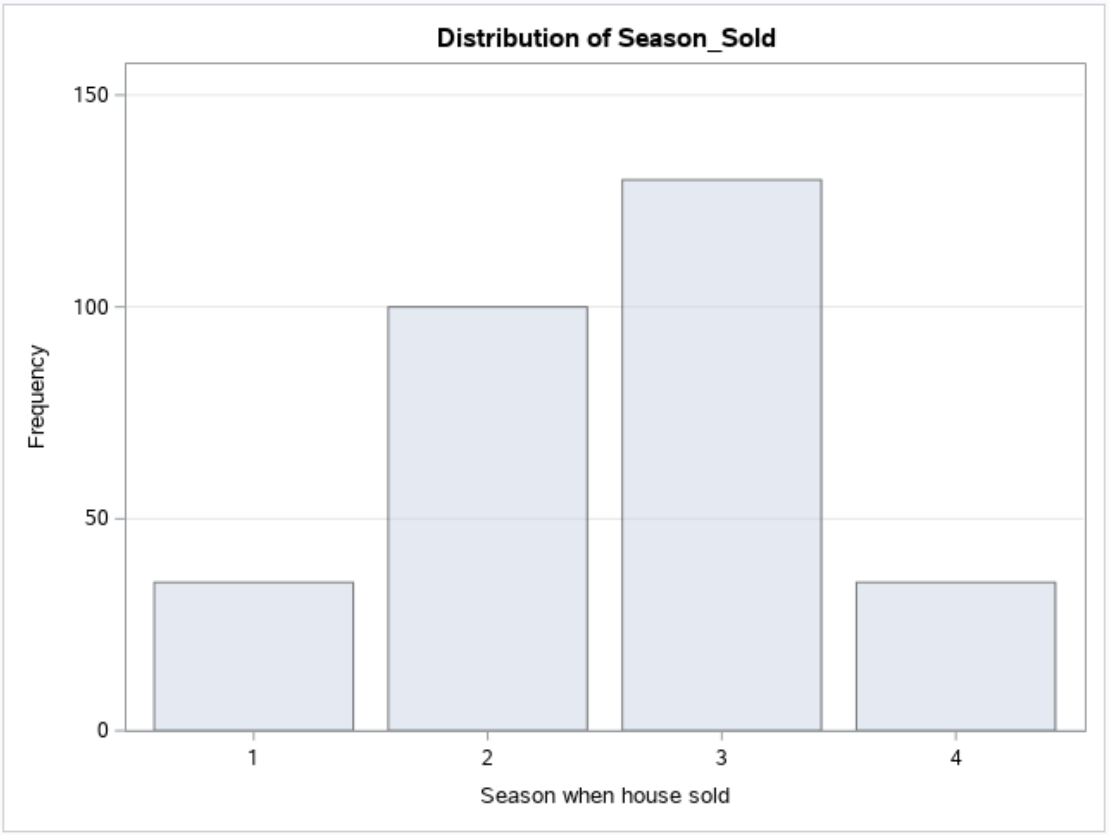
Overall condition of the house				
Overall_Cond2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	18	6.00	18	6.00
5	132	44.00	150	50.00
6	150	50.00	300	100.00



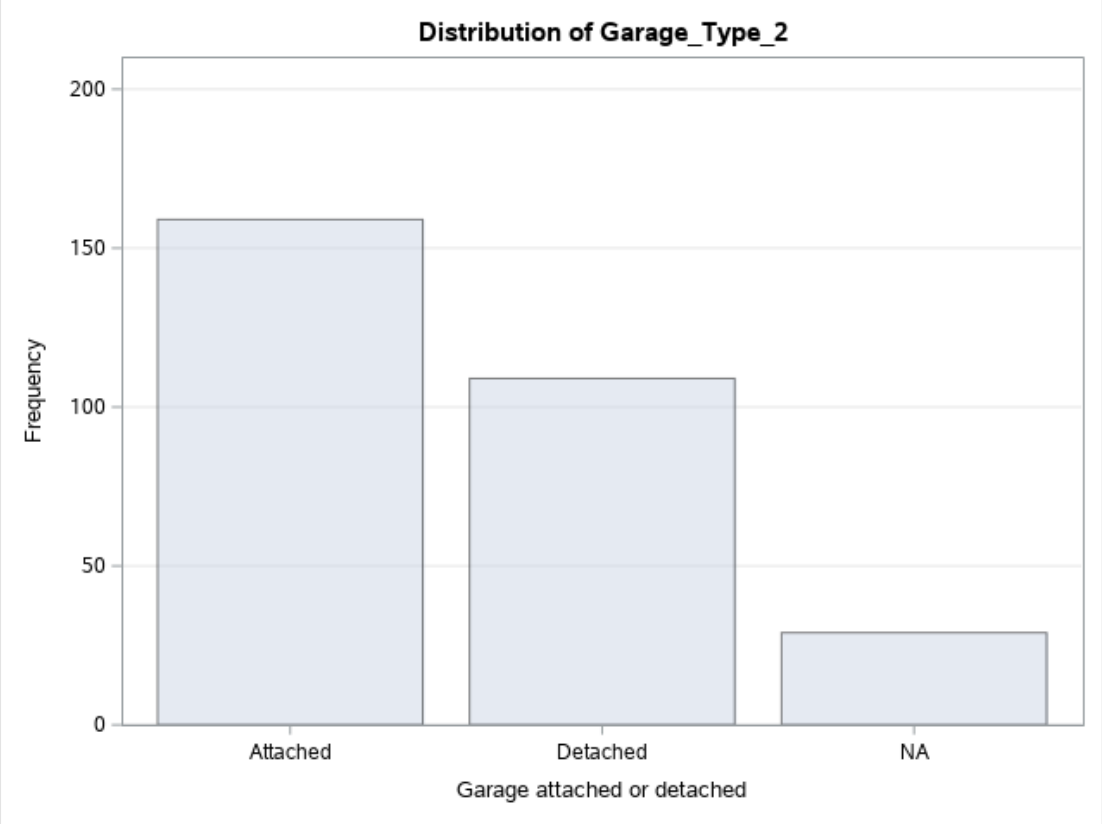
Number of fireplaces				
Fireplaces	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	195	65.00	195	65.00
1	93	31.00	288	96.00
2	12	4.00	300	100.00



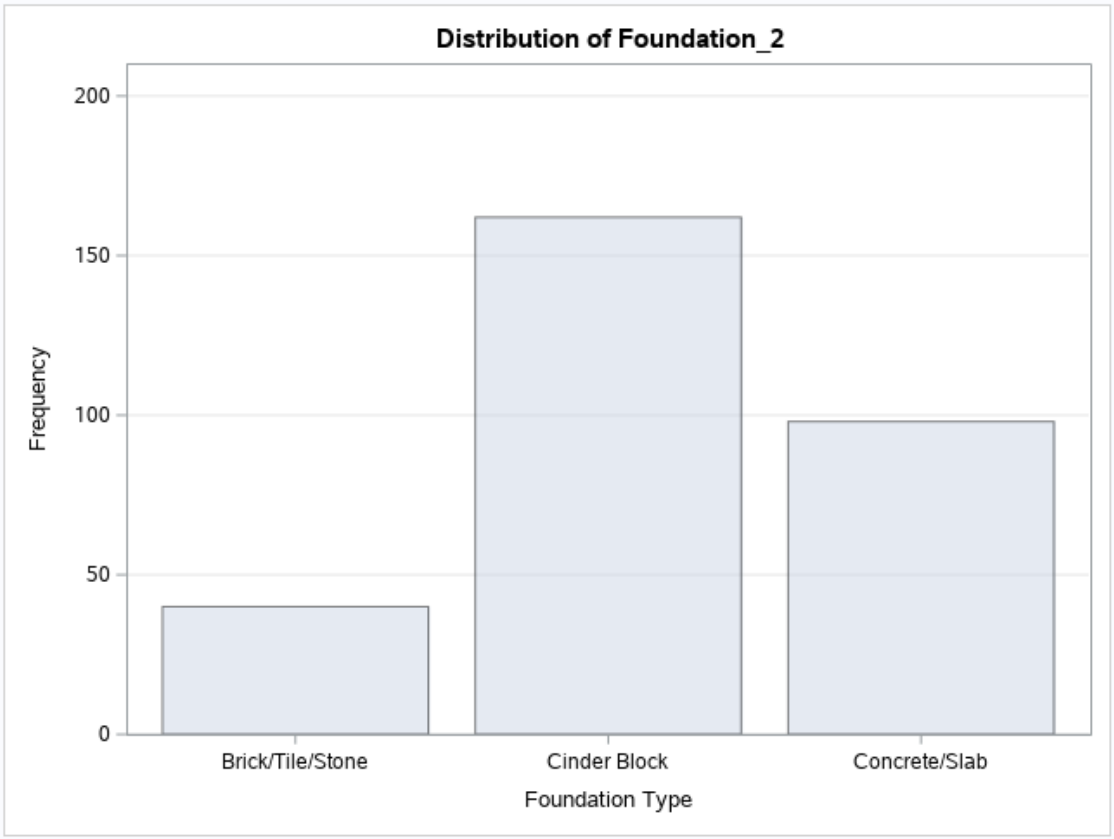
Season when house sold				
Season_Sold	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	35	11.67	35	11.67
2	100	33.33	135	45.00
3	130	43.33	265	88.33
4	35	11.67	300	100.00



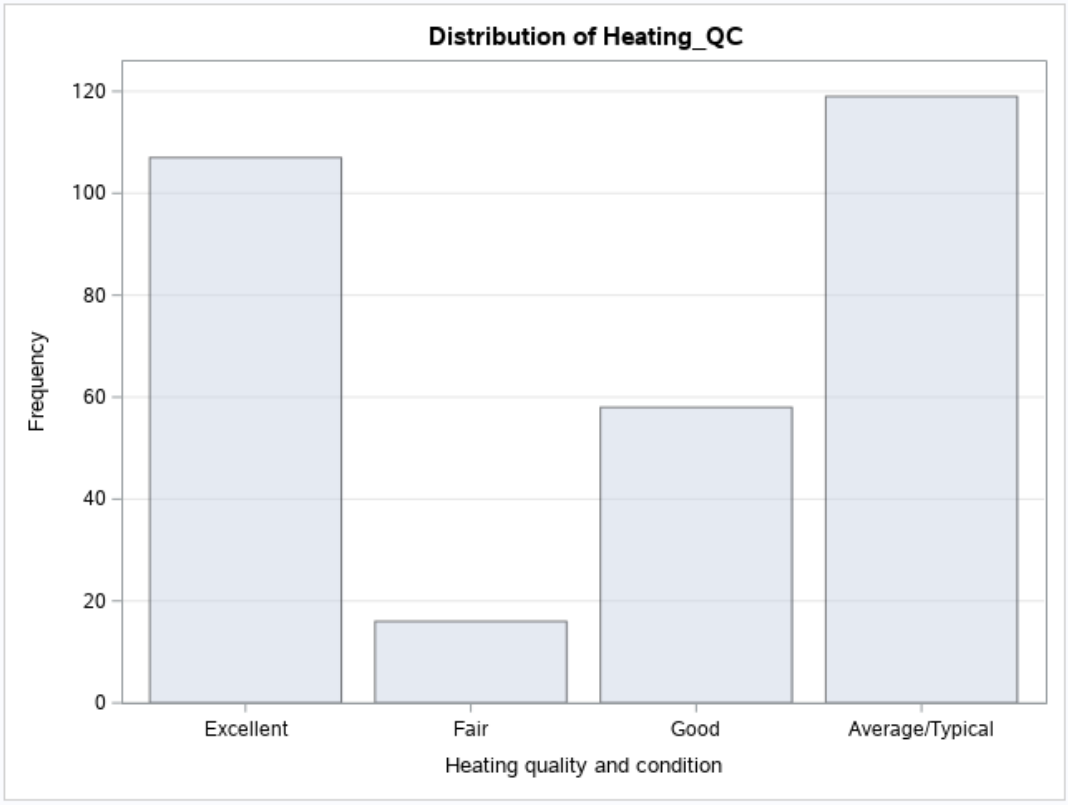
Garage attached or detached				
Garage_Type_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Attached	159	53.54	159	53.54
Detached	109	36.70	268	90.24
NA	29	9.76	297	100.00
Frequency Missing = 3				



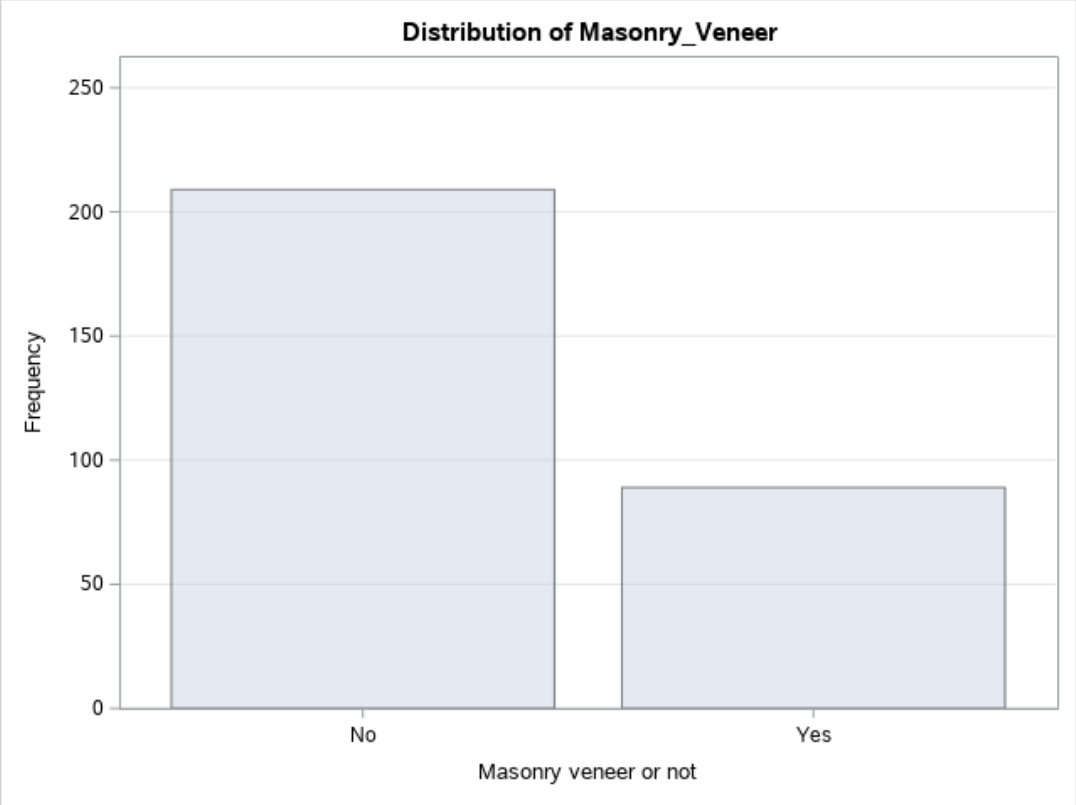
Foundation Type				
Foundation_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Brick/Tile/Stone	40	13.33	40	13.33
Cinder Block	162	54.00	202	67.33
Concrete/Slab	98	32.67	300	100.00



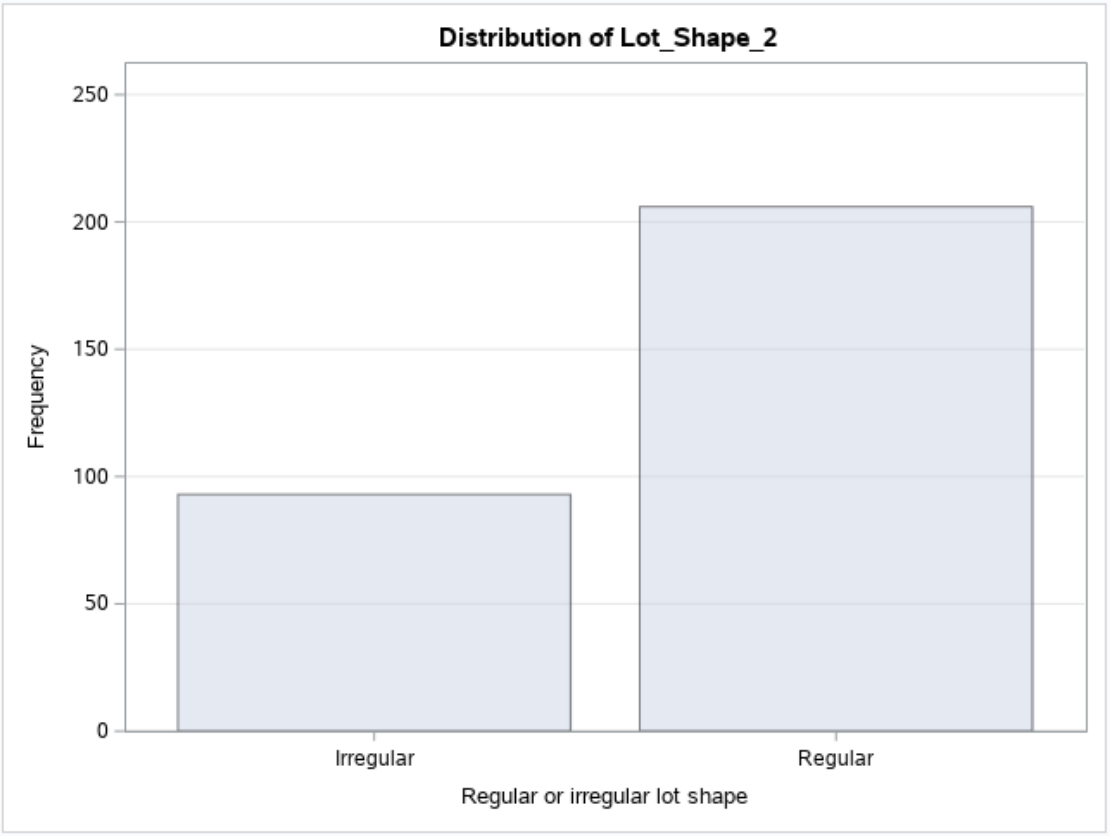
Heating quality and condition				
Heating_QC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Excellent	107	35.67	107	35.67
Fair	16	5.33	123	41.00
Good	58	19.33	181	60.33
Average/Typical	119	39.67	300	100.00



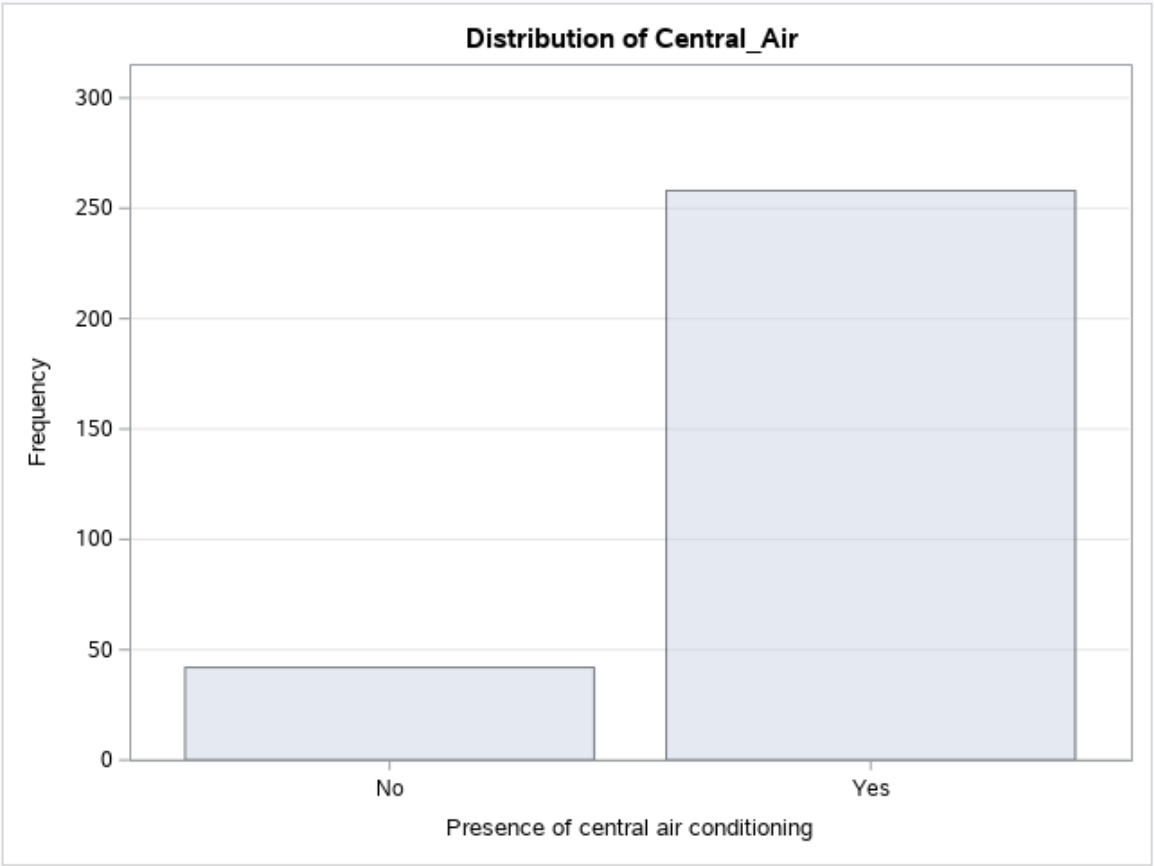
Masonry veneer or not				
Masonry_Veneer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	209	70.13	209	70.13
Yes	89	29.87	298	100.00
Frequency Missing = 2				



Regular or irregular lot shape				
Lot_Shape_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Irregular	93	31.10	93	31.10
Regular	206	68.90	299	100.00
Frequency Missing = 1				



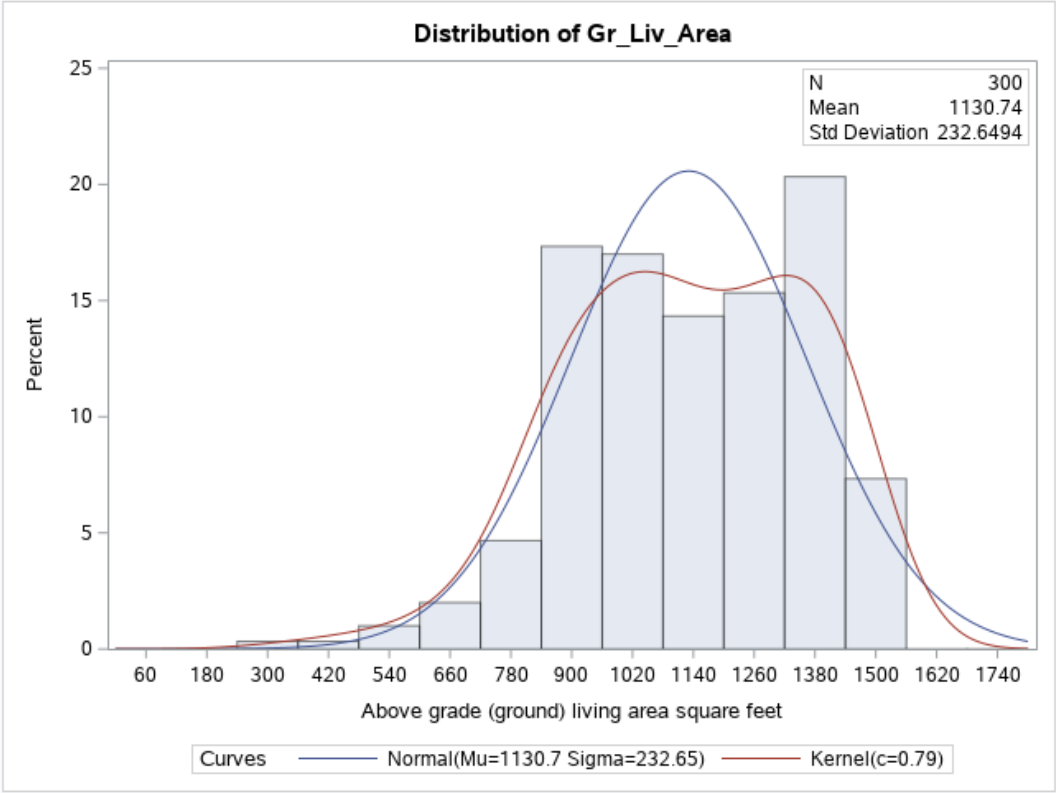
Presence of central air conditioning				
Central_Air	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	42	14.00	42	14.00
Yes	258	86.00	300	100.00

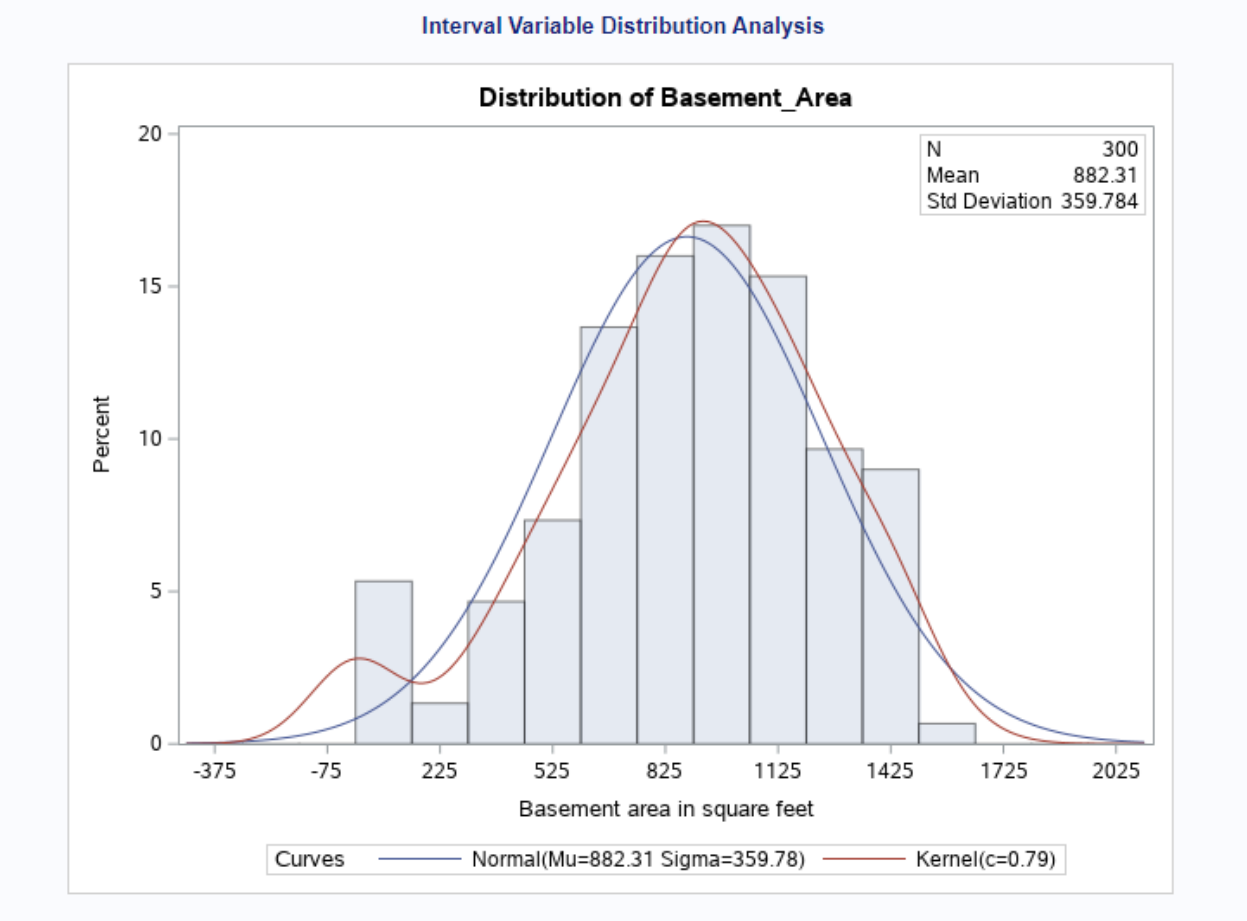


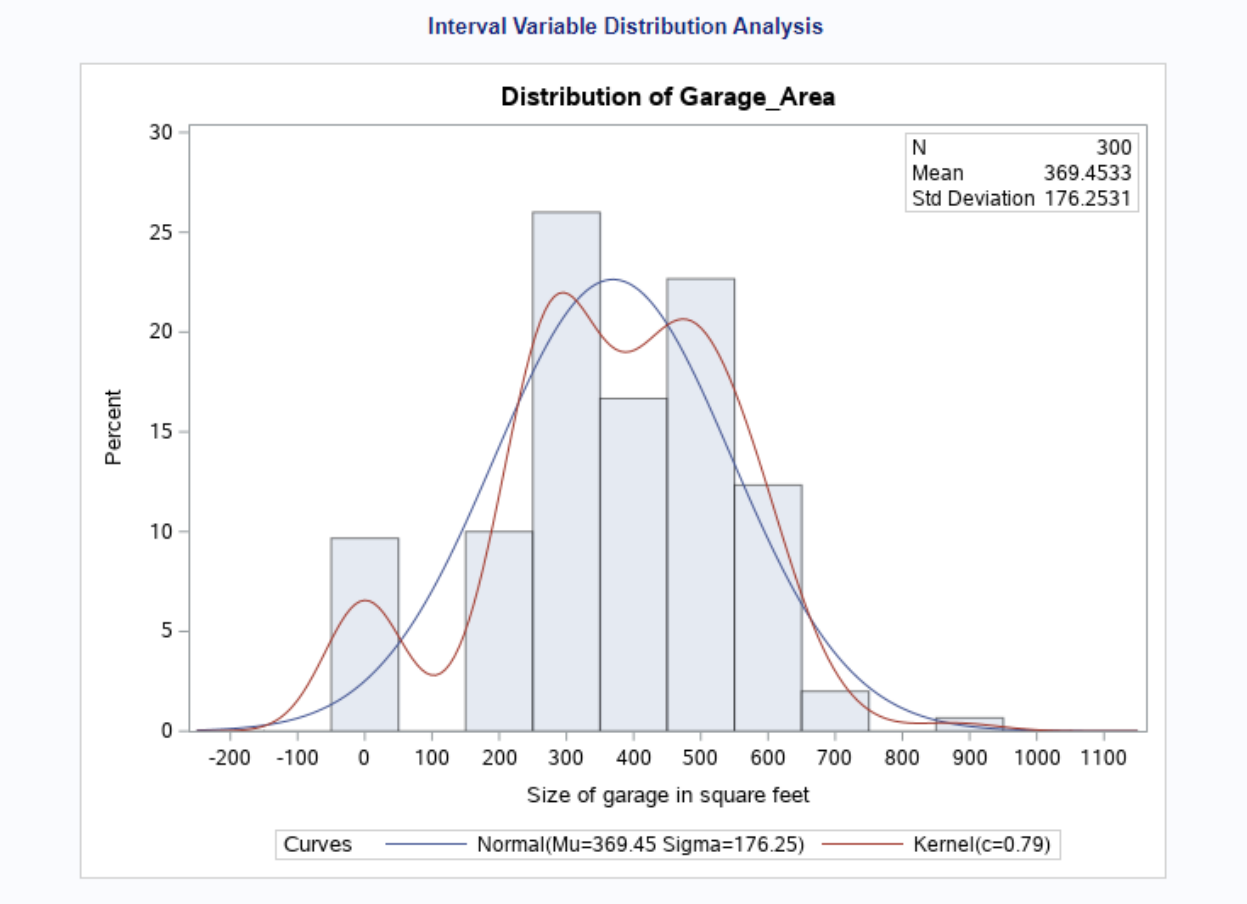
/*PROC UNIVARIATE provides summary statistics and plots for */
/*interval variables. The ODS statement specifies that only */
/*the histogram be displayed. The INSET statement requests */
/*summary statistics without having to print out tables.*/

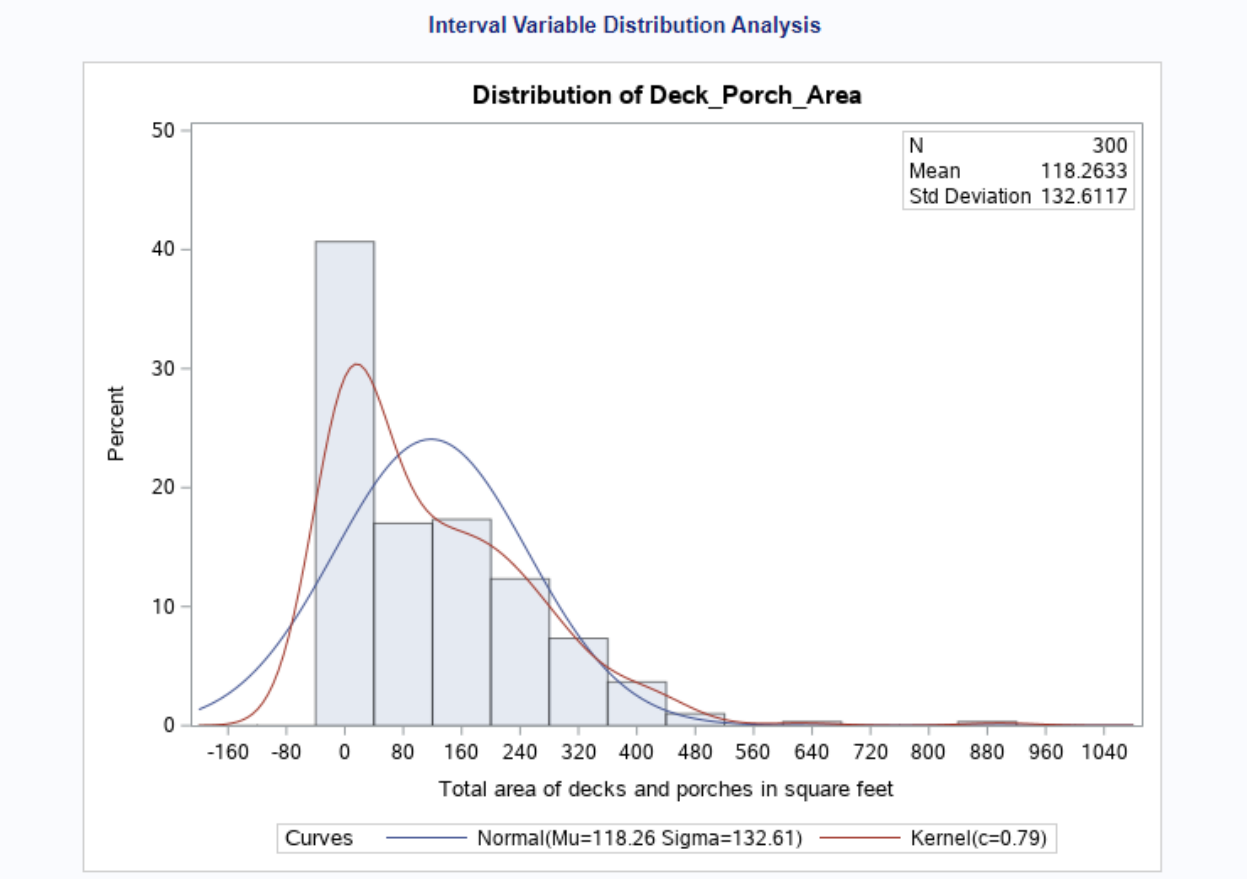
```
ods select histogram;
proc univariate data=STAT1.ameshousing3 noprint;
    var &interval;
    histogram &interval / normal kernel;
    inset n mean std / position=ne;
    title "Interval Variable Distribution Analysis";
run;

title;
```

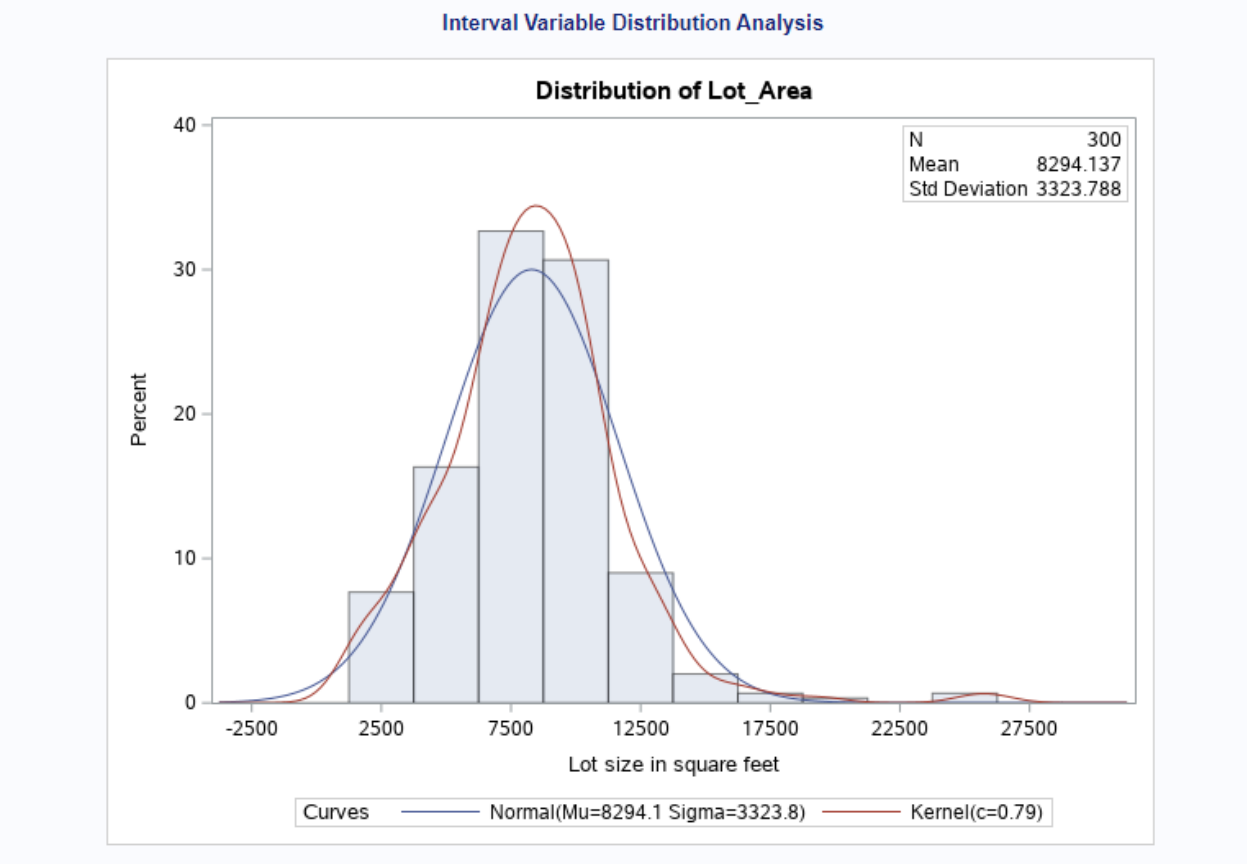
Interval Variable Distribution Analysis

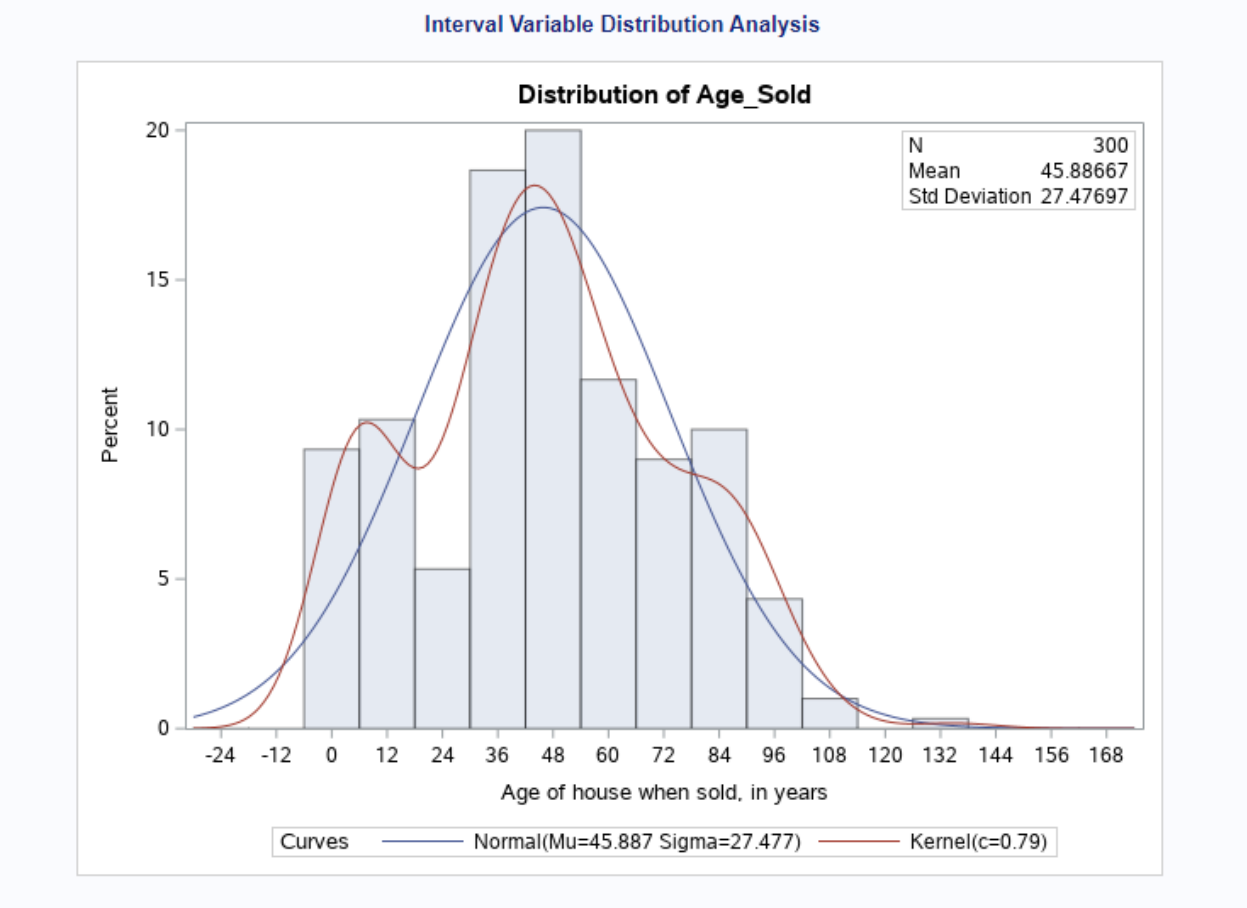


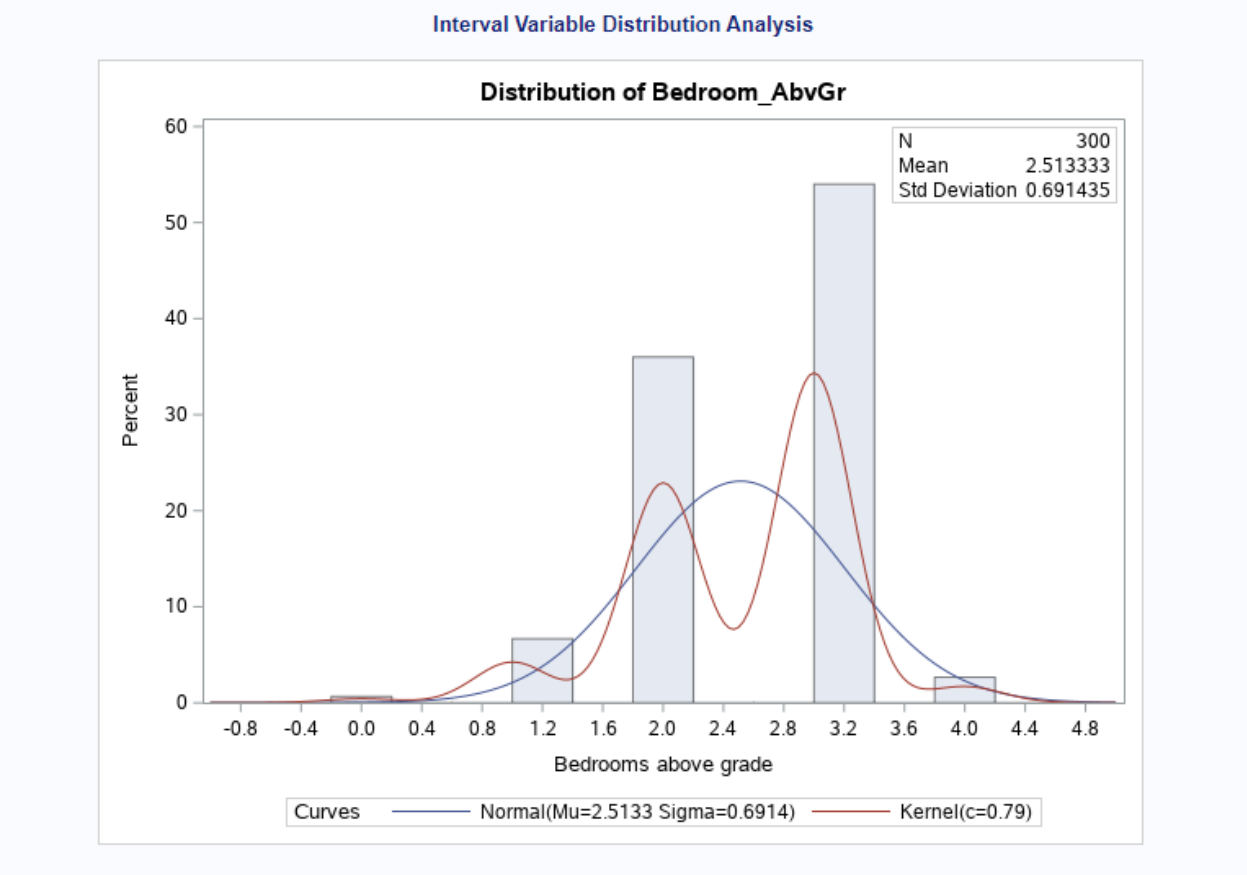

Interval Variable Distribution Analysis
Distribution of Basement_Area
N 300
Mean 882.31
Std Deviation 359.784
Percent
20
15
10
5
0
-375
-75
225
525
825
1125
1425
1725
2025
Basement area in square feet
Curves
Normal(Mu=882.31 Sigma=359.78)
Kernel(c=0.79)

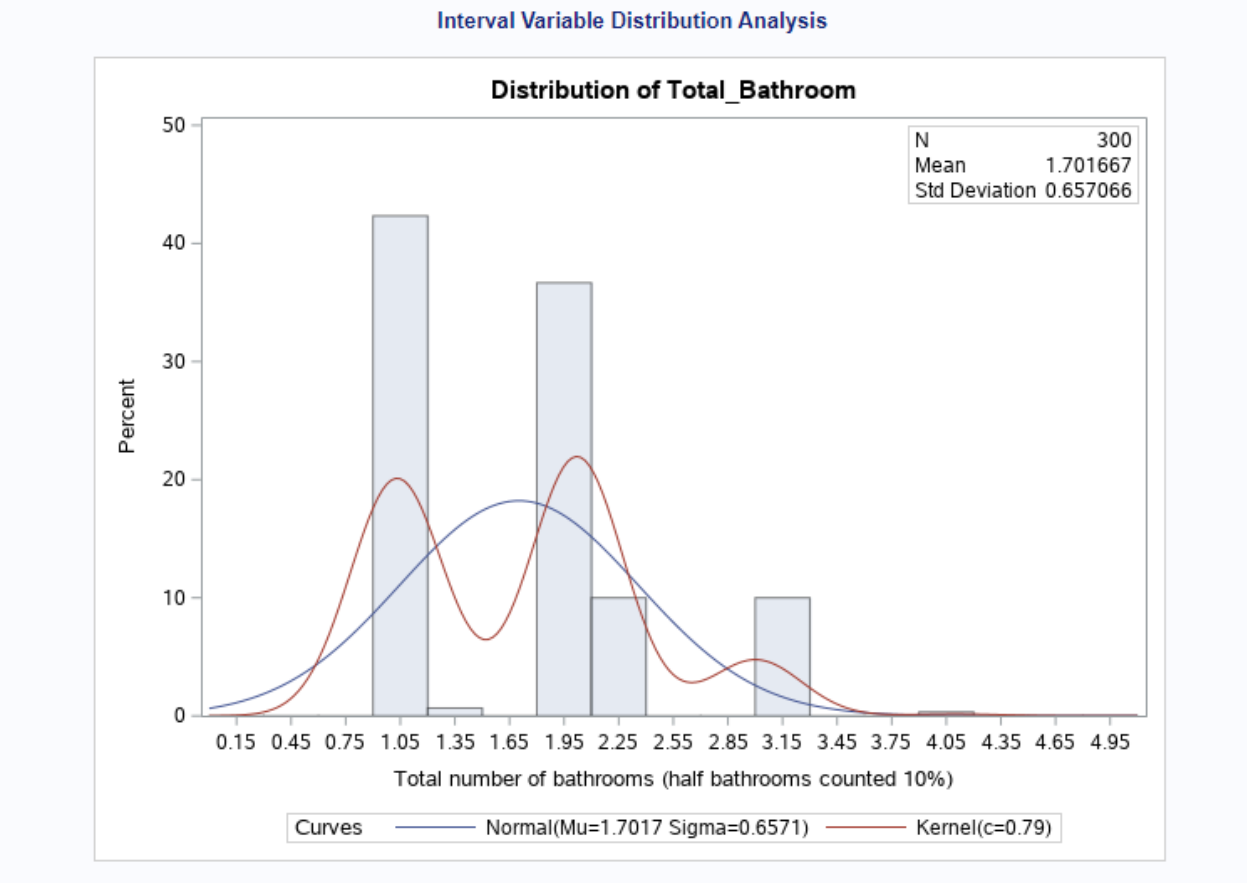

Interval Variable Distribution Analysis
Distribution of Garage_Area
N 300
Mean 369.4533
Std Deviation 176.2531
Percent
30
25
20
15
10
5
0
-200 -100 0 100 200 300 400 500 600 700 800 900 1000 1100
Size of garage in square feet
Curves Normal(Mu=369.45 Sigma=176.25) Kernel(c=0.79)


Interval Variable Distribution Analysis
Distribution of Deck_Porch_Area
N 300
Mean 118.2633
Std Deviation 132.6117
Percent
50
40
30
20
10
0
-160 -80 0 80 160 240 320 400 480 560 640 720 800 880 960 1040
Total area of decks and porches in square feet
Curves
Normal(Mu=118.26 Sigma=132.61)
Kernel(c=0.79)


Interval Variable Distribution Analysis
Distribution of Lot_Area
N 300
Mean 8294.137
Std Deviation 3323.788
Percent
40
30
20
10
0
-2500
2500
7500
12500
17500
22500
27500
Lot size in square feet
Curves
Normal(Mu=8294.1 Sigma=3323.8)
Kernel(c=0.79)


Interval Variable Distribution Analysis
Distribution of Age_Sold
N 300
Mean 45.88667
Std Deviation 27.47697
Percent
20
15
10
5
0
-24 -12 0 12 24 36 48 60 72 84 96 108 120 132 144 156 168
Age of house when sold, in years
Curves
Normal(Mu=45.887 Sigma=27.477)
Kernel(c=0.79)


Interval Variable Distribution Analysis
Distribution of Bedroom_AbvGr
N 300
Mean 2.513333
Std Deviation 0.691435
Percent
60
50
40
30
20
10
0
-0.8 -0.4 0.0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8
Bedrooms above grade
Curves
Normal(Mu=2.5133 Sigma=0.6914)
Kernel(c=0.79)




```
/* One-Sample t-test testing whether mean SalePrice=$135,000 */
ods graphics;

proc ttest data=STAT1.ameshousing3
           plots(shownull)=interval
           H0=135000;
    var SalePrice;
    title "One-Sample t-test testing whether mean
SalePrice=$135,000";
run;

title;
```

Conclusion: we fail to reject the null hypothesis that the mean SalePrice=$135,000, at the 5% level of statistical significance.

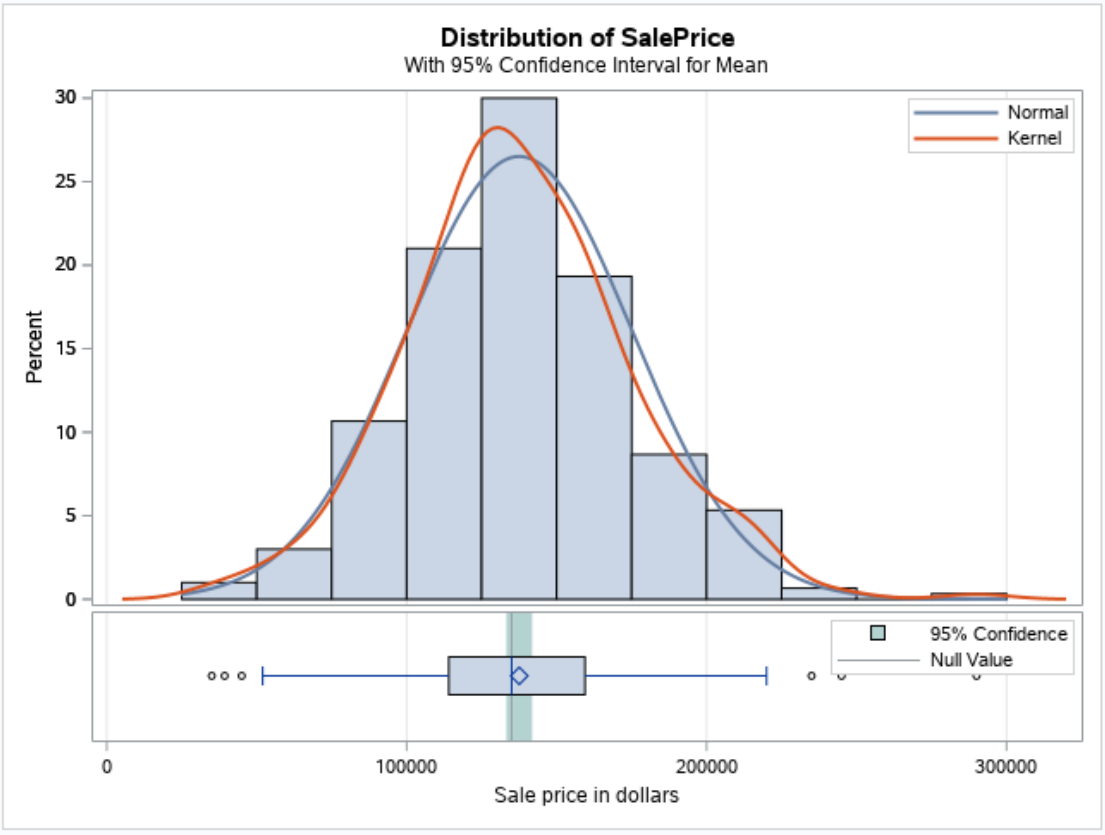
One-Sample t-test testing whether mean SalePrice=$135,000

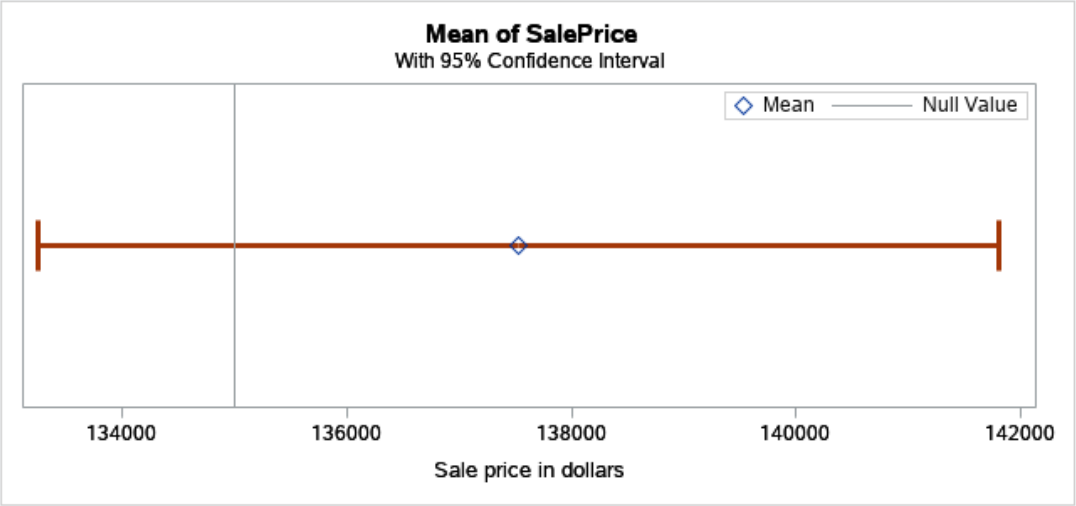
Variable: SalePrice (Sale price in dollars)

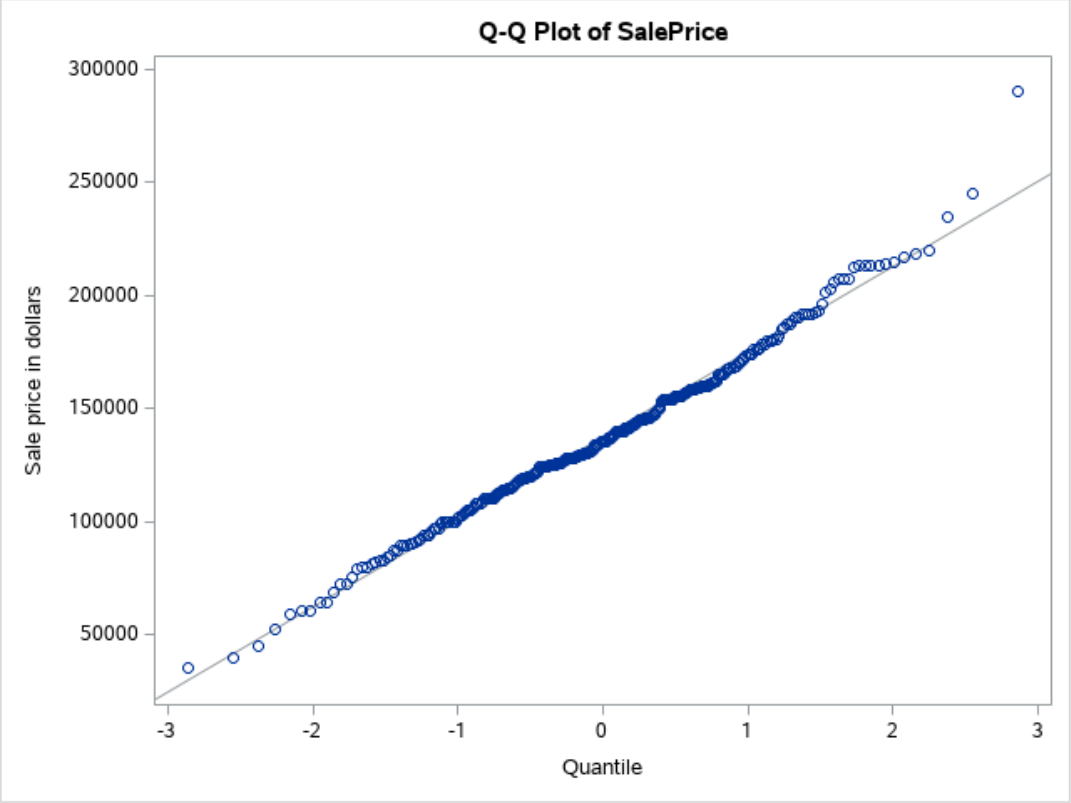
N	Mean	Std Dev	Std Err	Minimum	Maximum
300	137525	37622.6	2172.1	35000.0	290000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
137525	133250	141799	37622.6	34833.7	40900.7

DF	t Value	Pr > |t|
299	1.16	0.2460







/* Two-Sample t-test Comparing Masonry Veneer, No vs. Yes */

```
ods graphics;

proc ttest data=STAT1.ameshousing3 plots(shownull)=interval;
    class Masonry_Veneer;
    var SalePrice;
```

```
    format Masonry_Veneer $NoYes.;
    title "Two-Sample t-test Comparing Masonry Veneer, No vs.
Yes";
run;

title;
```

Conclusion:

- using the pooled F test: we fail to reject the null hypothesis that the homes with vs without masonry veneer have equal variances, at the 5% level of statistical significance.
- using the pooled t-test: we reject the null hypothesis that the homes with vs without masonry veneer have equal mean sales price, at the 5% level of statistical significance.

Two-Sample t-test Comparing Masonry Veneer, No vs. Yes

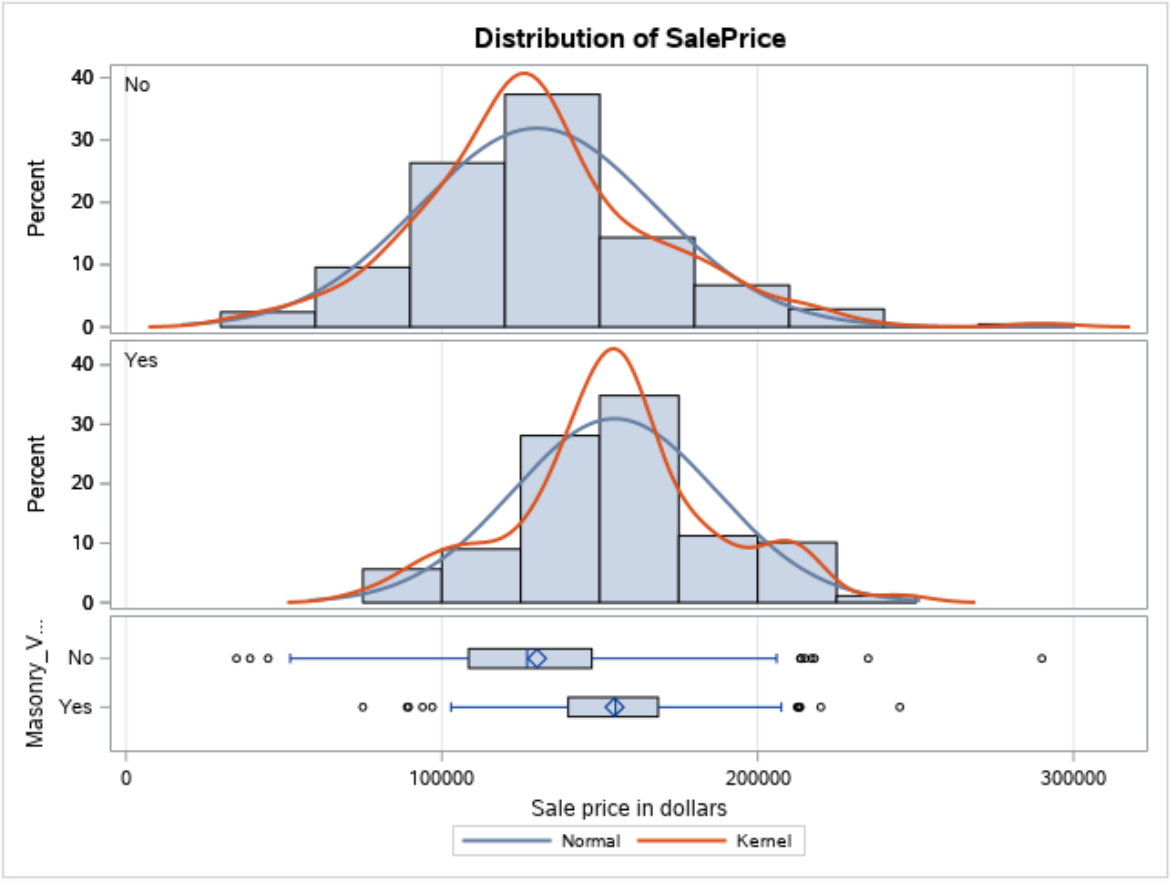
Variable: SalePrice (Sale price in dollars)

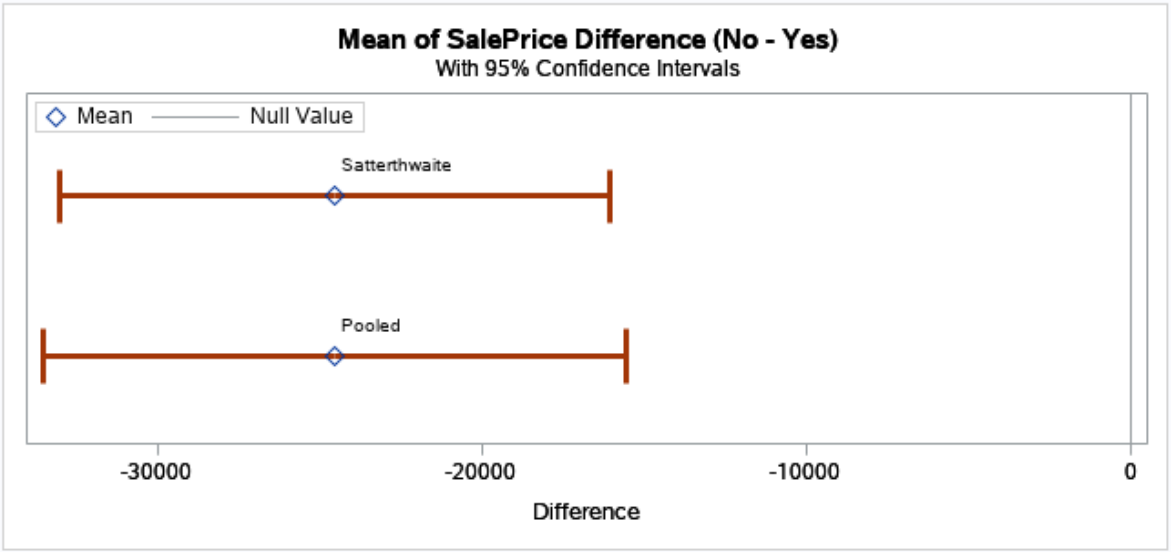
Masonry_Veneer	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
No		209	130172	37531.7	2596.1	35000.0	290000
Yes		89	154705	32239.8	3417.4	75000.0	245000
Diff (1-2)	Pooled		-24533.0	36039.6	4561.6		
Diff (1-2)	Satterthwaite		-24533.0		4291.7		

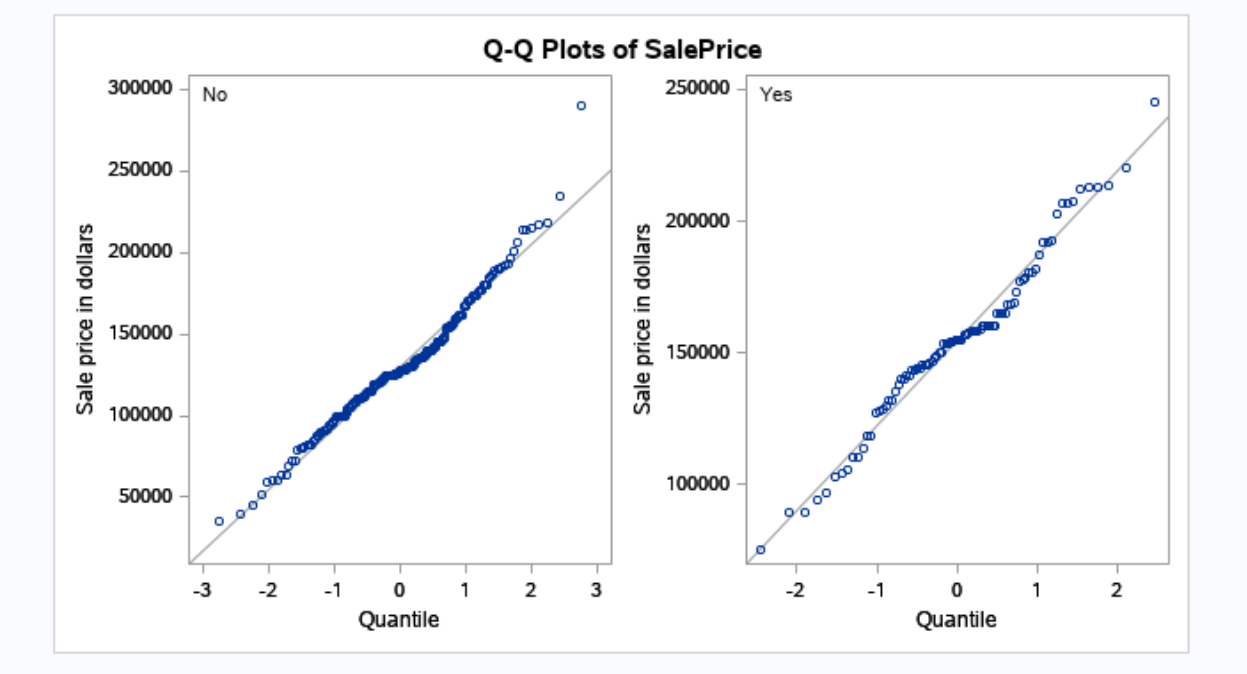
Masonry_Veneer	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
No		130172	125054	135290	37531.7	34245.4	41521.0
Yes		154705	147914	161496	32239.8	28099.7	37821.9
Diff (1-2)	Pooled	-24533.0	-33510.3	-15555.6	36039.6	33355.6	39197.1
Diff (1-2)	Satterthwaite	-24533.0	-32997.9	-16068.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	296	-5.38	<.0001
Satterthwaite	Unequal	191.85	-5.72	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	208	88	1.36	0.1039

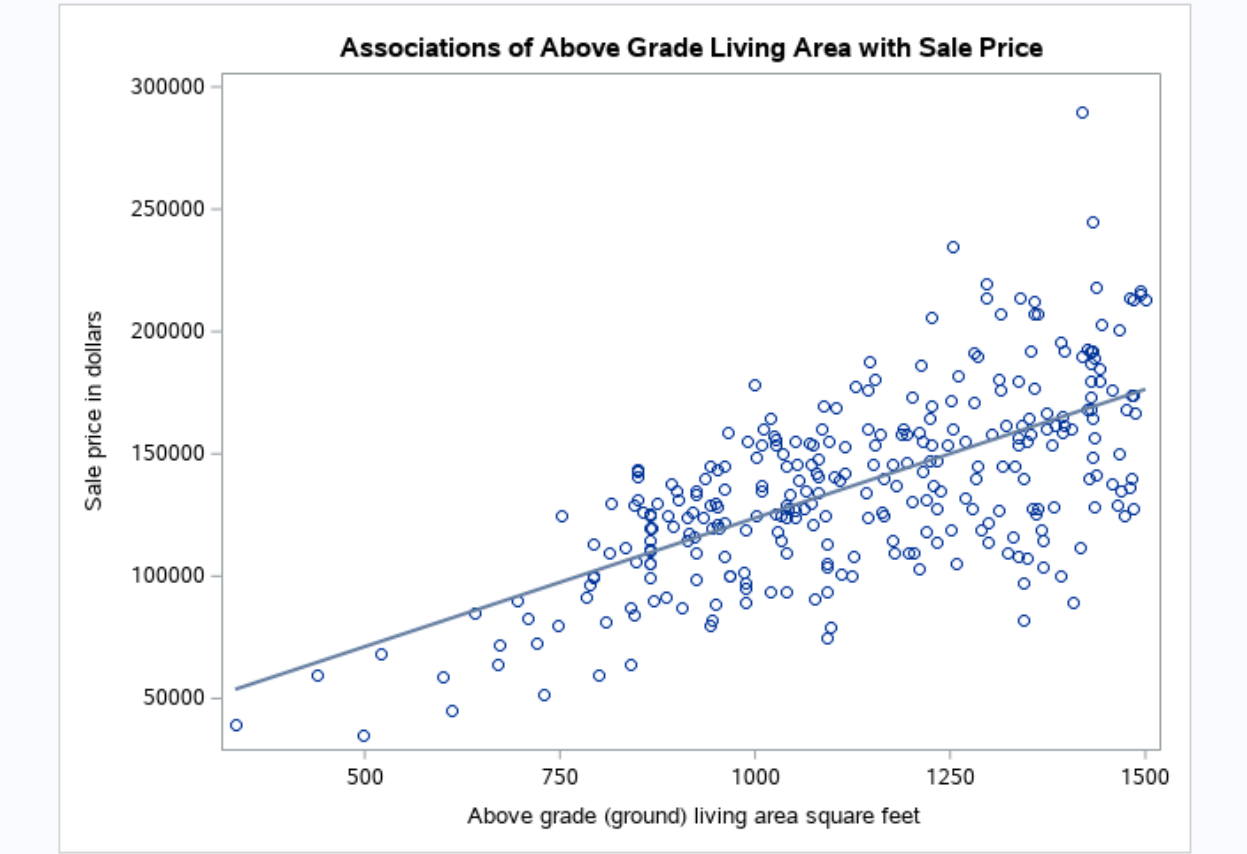

Distribution of SalePrice
No
Yes
Percent
Masonry_V...
No
Yes
0
100000
200000
300000
Sale price in dollars
Normal
Kernel


Mean of SalePrice Difference (No - Yes)
With 95% Confidence Intervals
Mean
Null Value
Satterthwaite
Pooled
-30000
-20000
-10000
0
Difference



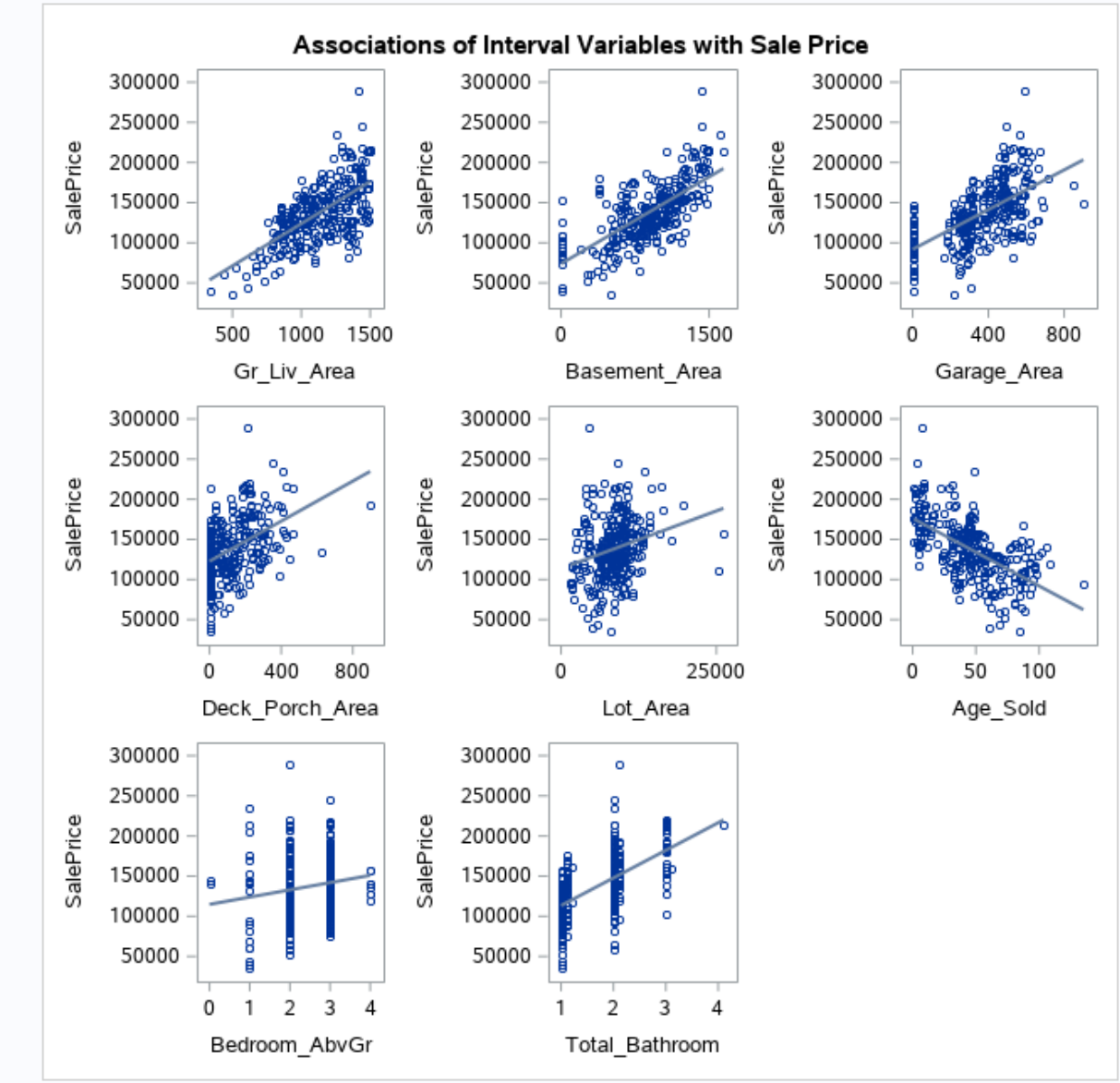
/*PROC SGSCATTER is used to explore relationships among continuous variables*/
/* here we check the relationship between the Above Grade Living Area and Sale Price */

```
proc sgscatter data=STAT1.ameshousing3;
    plot SalePrice*Gr_Liv_Area / reg;
    title "Associations of Above Grade Living Area with Sale
Price";
run;
```



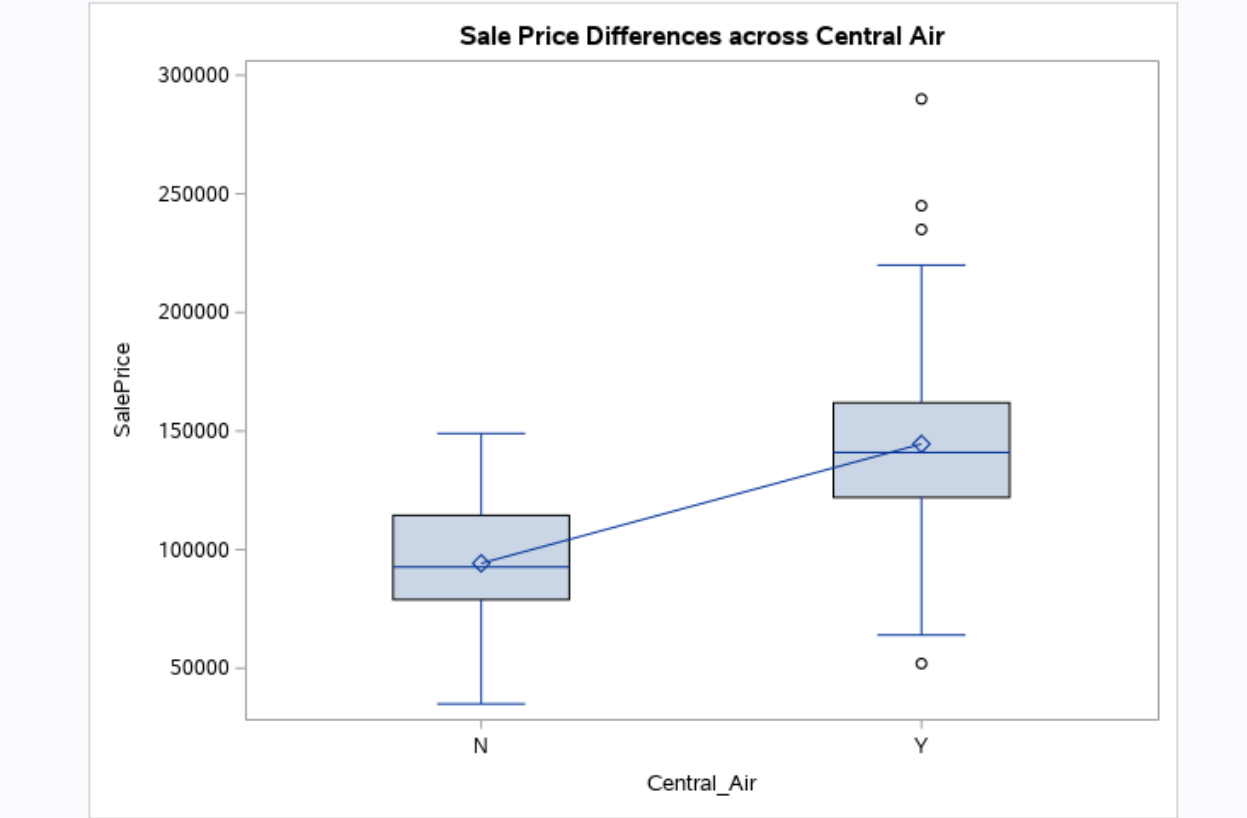
/* Let's explore relationships between the interval variables in our dataset with sale price */

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area
Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;

/*using scatter plots*/
options nolabel;
proc sgscatter data=STAT1.ameshousing3;
    plot SalePrice*(&interval) / reg;
    title "Associations of Interval Variables with Sale Price";
run;
```



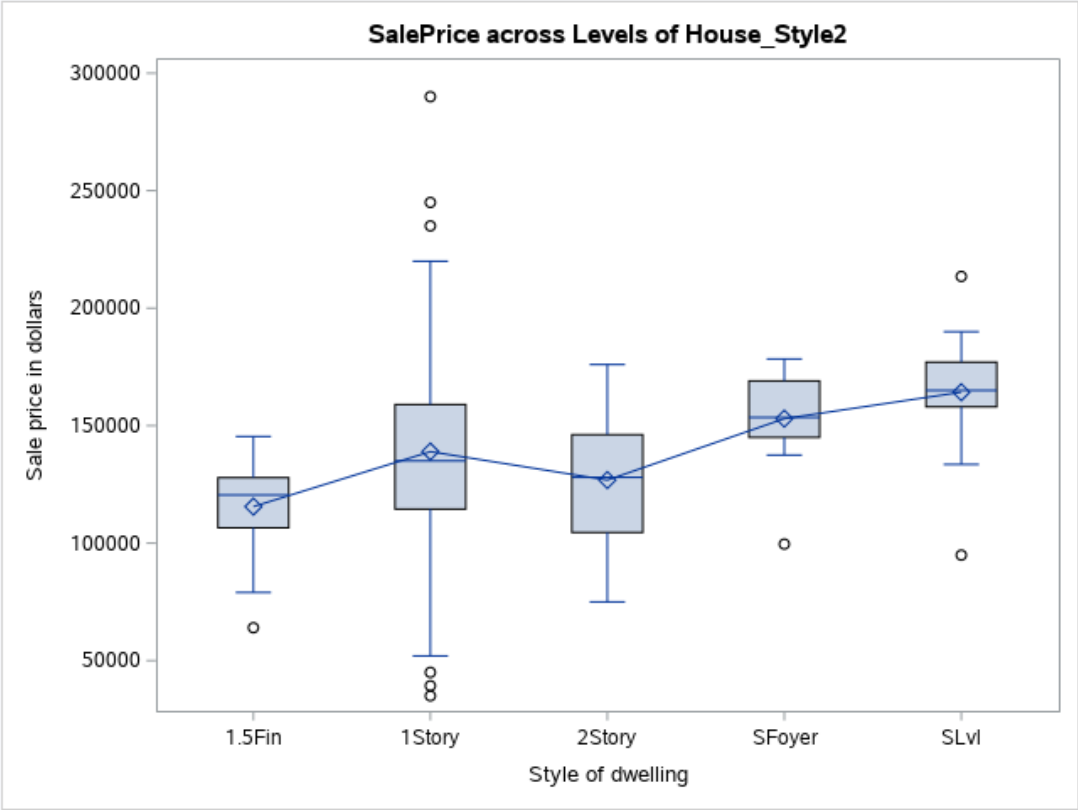
/* Let's create box plots of sale price by central air systems*/

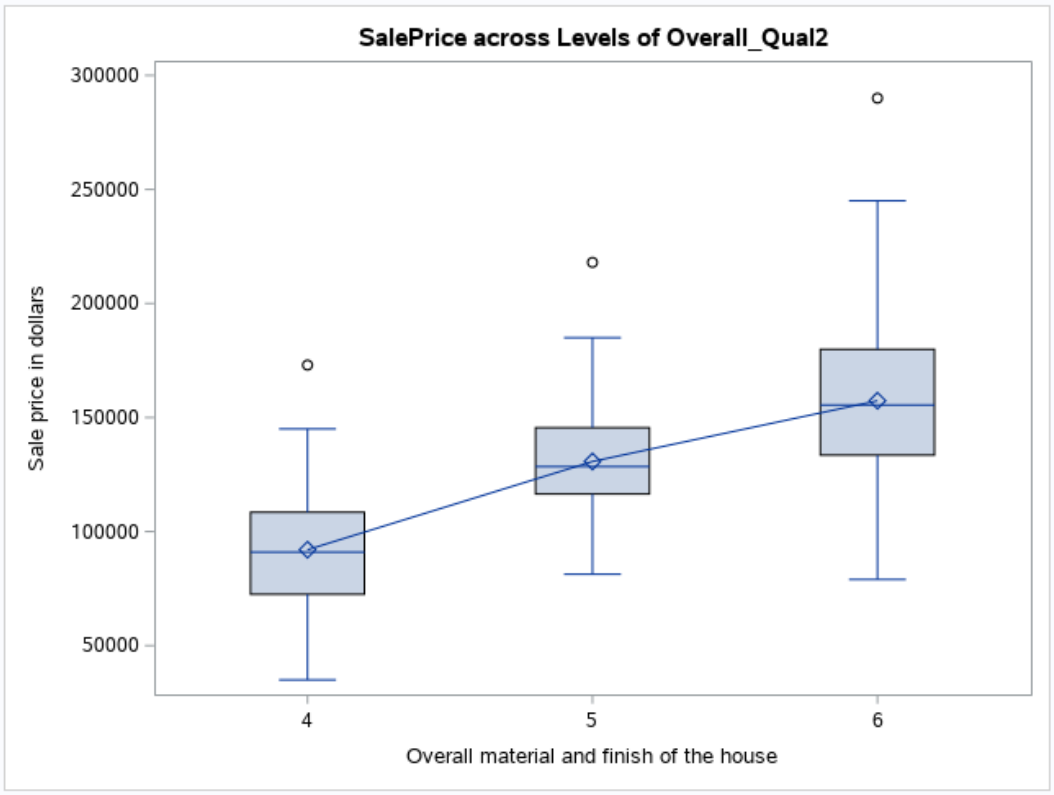
```
proc sgplot data=STAT1.ameshousing3;
    vbox SalePrice / category=Central_Air
                     connect=mean;
    title "Sale Price Differences across Central Air";
run;
```

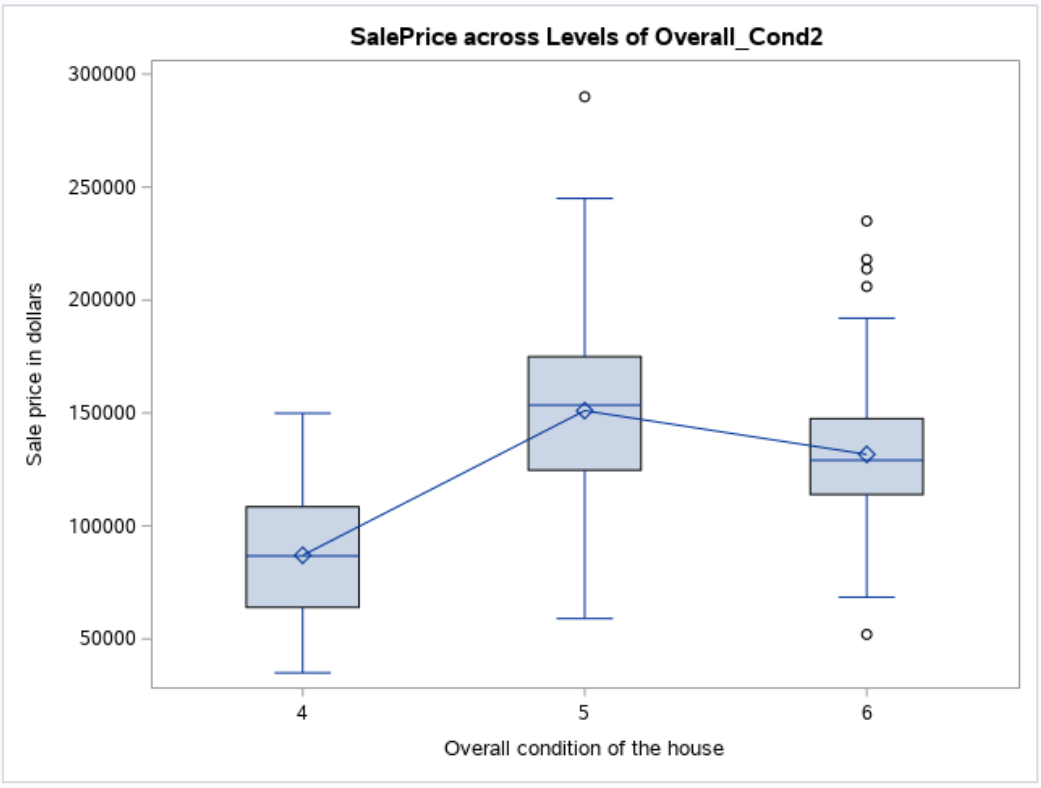


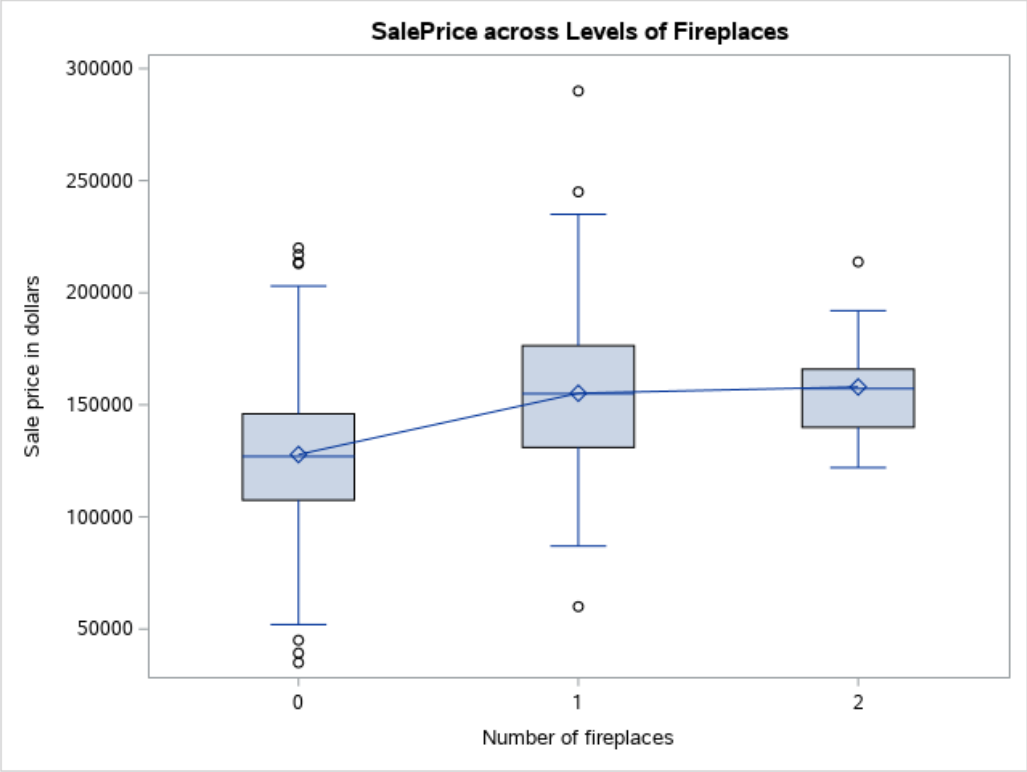

```
/* save categorical variables list in a macro */
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2
Fireplaces
         Season_Sold Garage_Type_2 Foundation_2 Heating_QC
         Masonry_Veneer Lot_Shape_2 Central_Air;

/*PROC SGPLOT is used here with the VBAR statement to produce vertical bar charts*/
/*PROC SGPLOT can only produce one plot at a time and so the macro is written to*/
/*produce one plot for each member in the list of the &categorical macro variable.*/
/*
   Macro Usage:
   %box(DSN = <data set name>,
      Response = <response variable name>,
      CharVar = <bar chart grouping variable list>)
*/
%macro box(dsn      = ,
           response = ,
           Charvar  = );

%let i = 1 ;

%do %while(%scan(&charvar,&i,%str( )) ^= %str()) ;
```

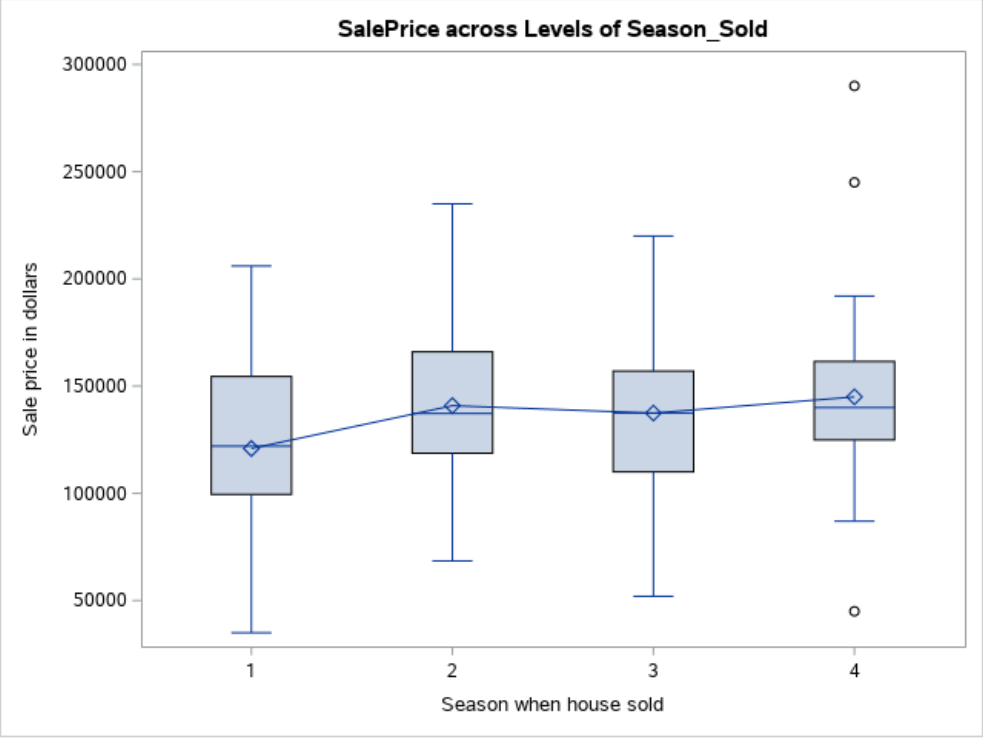
```
    %let var = %scan(&charvar,&i,%str( ));

    proc sgplot data=&dsn;
        vbox &response / category=&var
                         grouporder=ascending
                         connect=mean;
        title "&response across Levels of &var";
    run;

    %let i = %eval(&i + 1 ) ;

%end ;

%mend box;

%box(dsn      = STAT1.ameshousing3,
     response = SalePrice,
     charvar  = &categorical);

title;

options label;
```

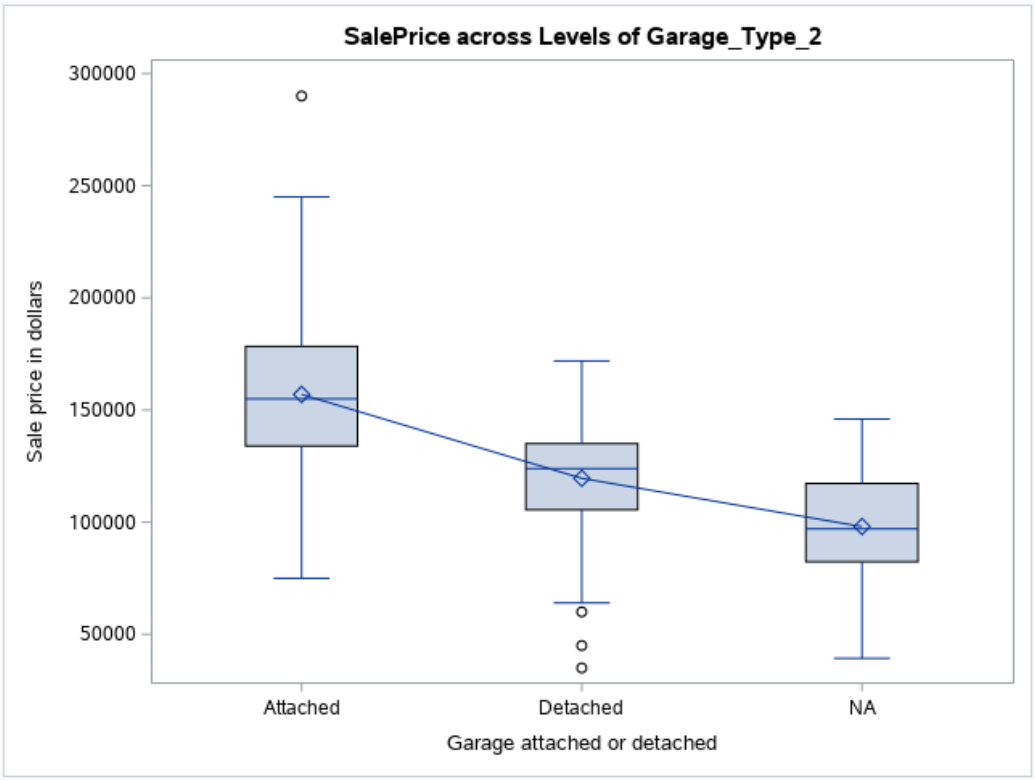

SalePrice across Levels of House_Style2
300000
250000
200000
150000
100000
50000
Sale price in dollars
1.5Fin
1Story
2Story
SFoyer
SLvl
Style of dwelling

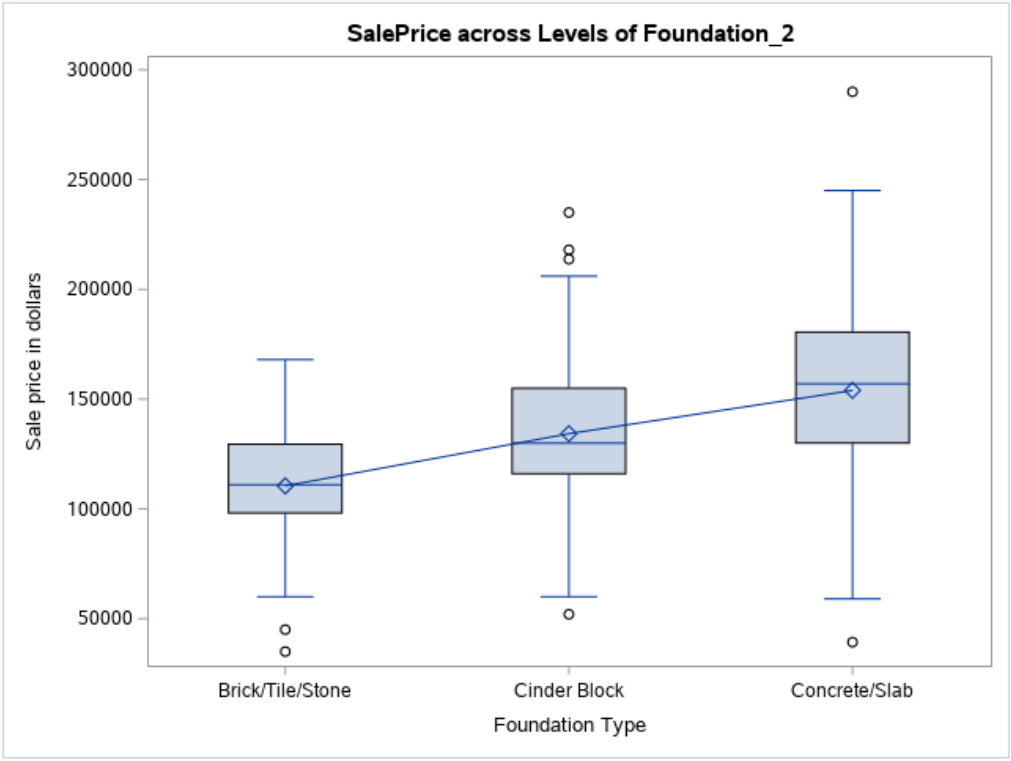

SalePrice across Levels of Overall_Qual2
300000
250000
200000
150000
100000
50000
Sale price in dollars
4
5
6
Overall material and finish of the house

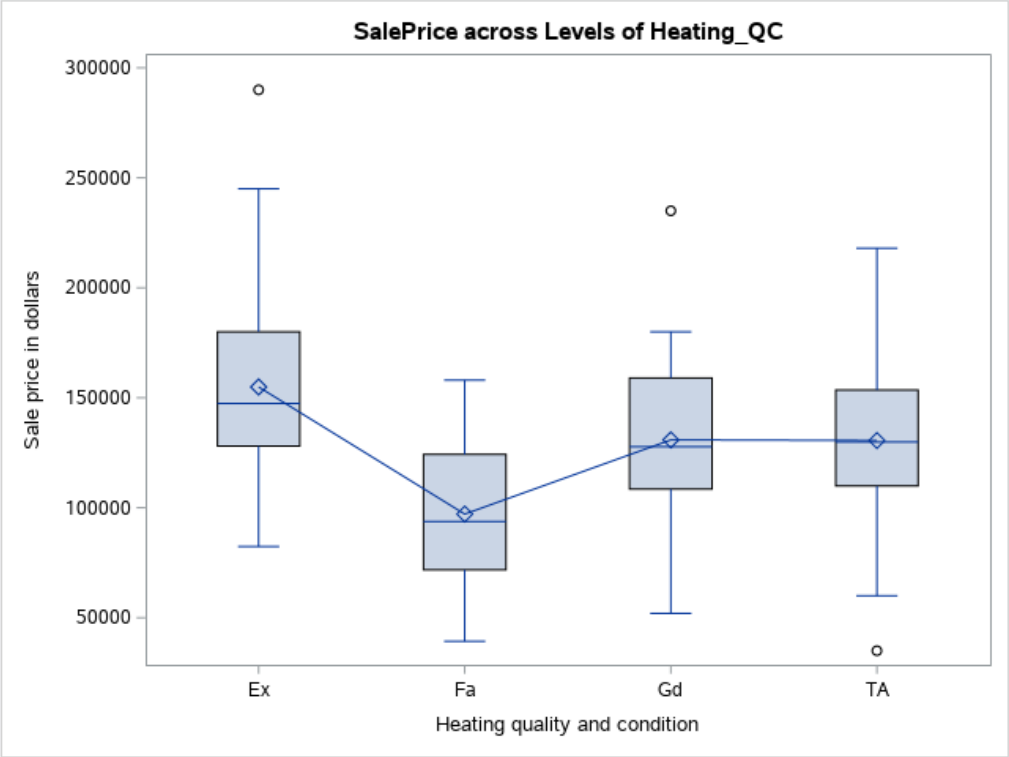

SalePrice across Levels of Overall_Cond2
300000
250000
200000
150000
100000
50000
Sale price in dollars
4
5
6
Overall condition of the house

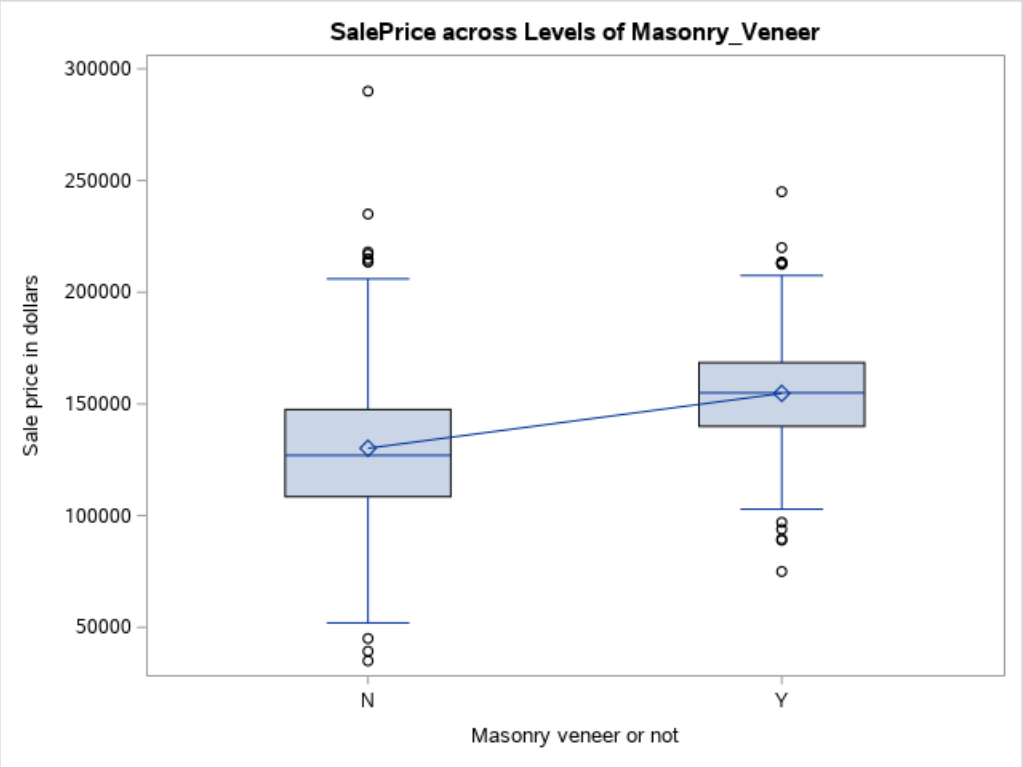

SalePrice across Levels of Fireplaces
300000
250000
200000
150000
100000
50000
Sale price in dollars
0
1
2
Number of fireplaces

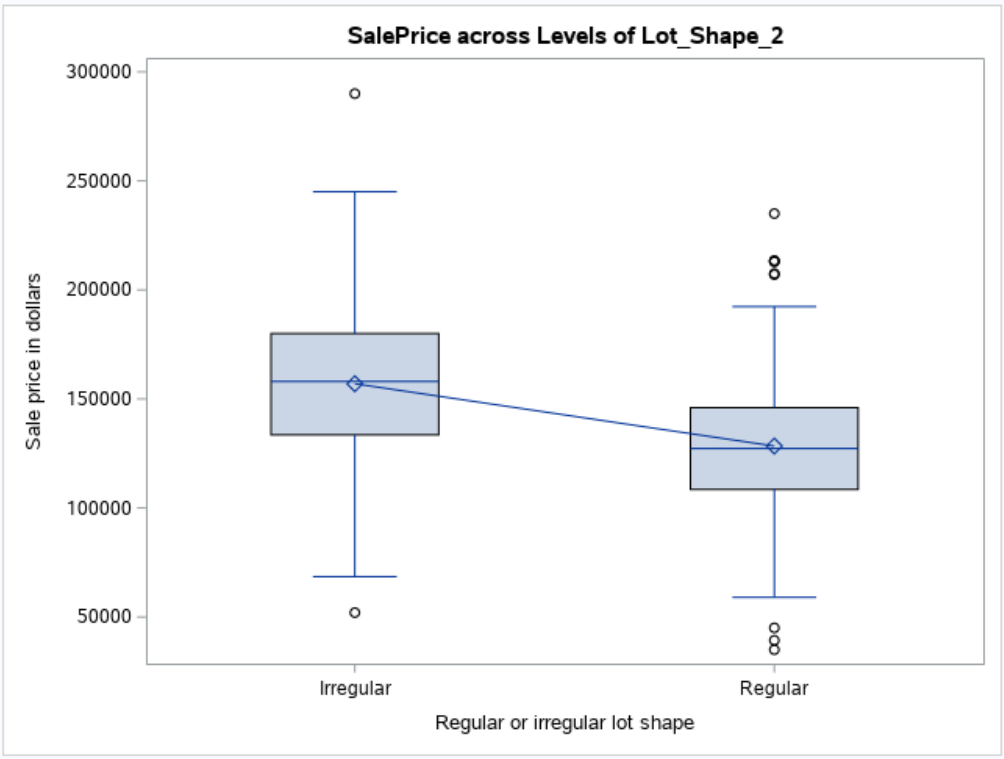

SalePrice across Levels of Season_Sold
300000
250000
200000
150000
100000
50000
Sale price in dollars
1
2
3
4
Season when house sold

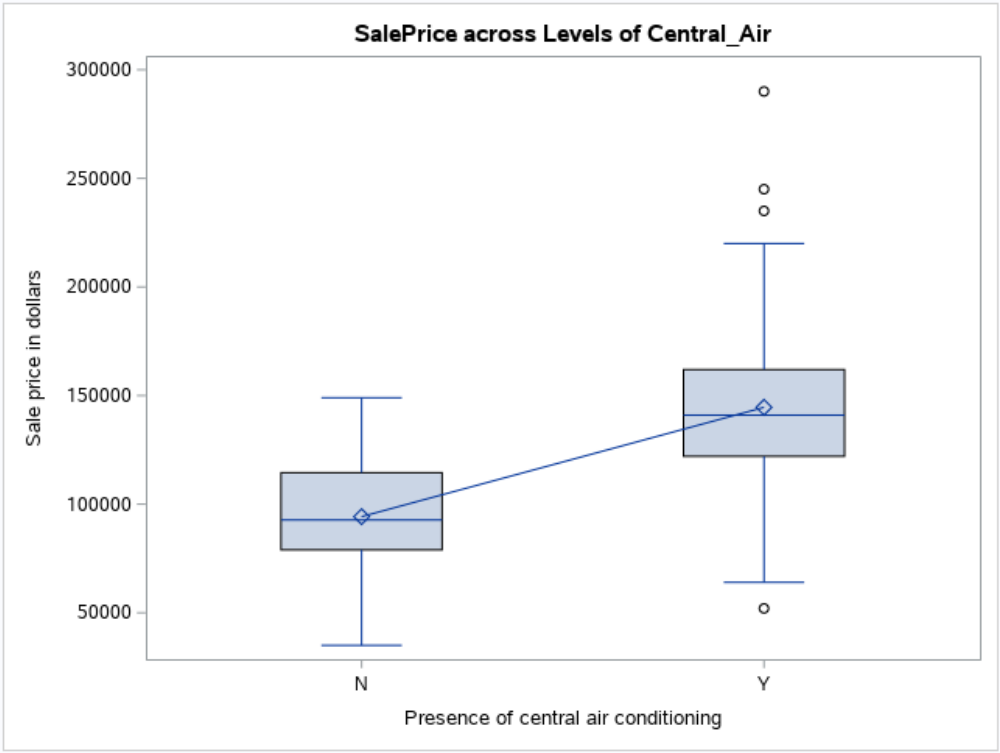

SalePrice across Levels of Garage_Type_2
300000
250000
200000
150000
100000
50000
Sale price in dollars
Attached
Detached
NA
Garage attached or detached


SalePrice across Levels of Foundation_2
300000
250000
200000
150000
100000
50000
Sale price in dollars
Brick/Tile/Stone
Cinder Block
Concrete/Slab
Foundation Type

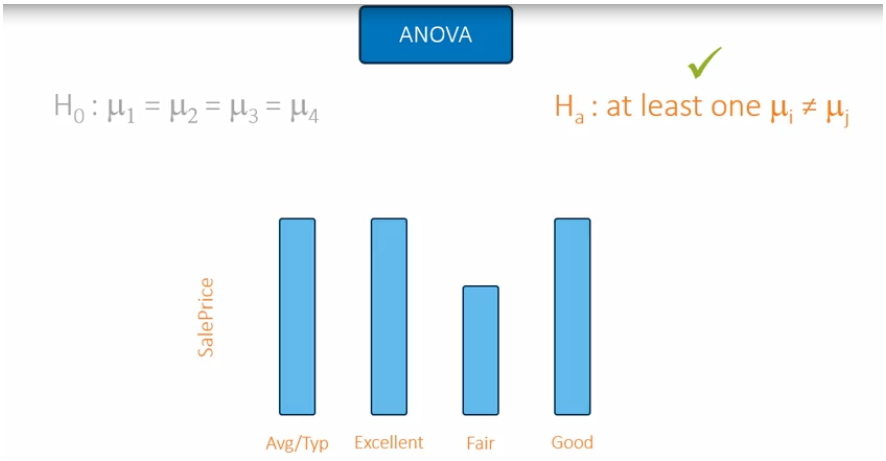

SalePrice across Levels of Heating_QC
300000
250000
200000
150000
100000
50000
Sale price in dollars
Ex
Fa
Gd
TA
Heating quality and condition


SalePrice across Levels of Masonry_Veneer
300000
250000
200000
150000
100000
50000
Sale price in dollars
N
Y
Masonry veneer or not


SalePrice across Levels of Lot_Shape_2
300000
250000
200000
150000
100000
50000
Sale price in dollars
Irregular
Regular
Regular or irregular lot shape


SalePrice across Levels of Central_Air
300000
250000
200000
150000
100000
50000
Sale price in dollars
N
Y
Presence of central air conditioning

5. Analysis of Variance (ANOVA) and Regression



Total Variation	Total Sum of Squares (SS_T)	$\sum\sum\left(Y_{ij} - \bar{\bar{Y}}\right)^2$
Between Group Variation	Model Sum of Squares (SS_M)	$\sum n_i \left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$
Within Group Variation	Error Sum of Squares (SS_E)	$\sum\sum\left(Y_{ij} - \bar{Y}_i\right)^2$

/* Let's use PROC GLM to perform an analysis of variance, testing whether house sale price differs as a function of the quality of their heating systems */

/* We use Levene's test of homogeneity to test for Constance of variance */

/* We can check normality by examining residuals: histograms, q-q plots, residual vs predicted values */

```
ods graphics;


proc glm data=STAT1.ameshousing3 plots=diagnostics;
    class Heating_QC;
    model SalePrice=Heating_QC;
     means Heating_QC / hovtest=levene;
```

```
    format Heating_QC $Heating_QC.;
    title "One-Way ANOVA with Heating Quality as Predictor";
run;
quit;
title;
```

CONCLUSION: Based on the ANOVA table below, we reject the null hypothesis of no difference between the means. At least one sale price is different for the 4 levels of heating quality.

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

Class Level Information		
Class	Levels	Values
Heating_QC	4	Average/Typical Excellent Fair Good

Number of Observations Read	300
Number of Observations Used	300

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

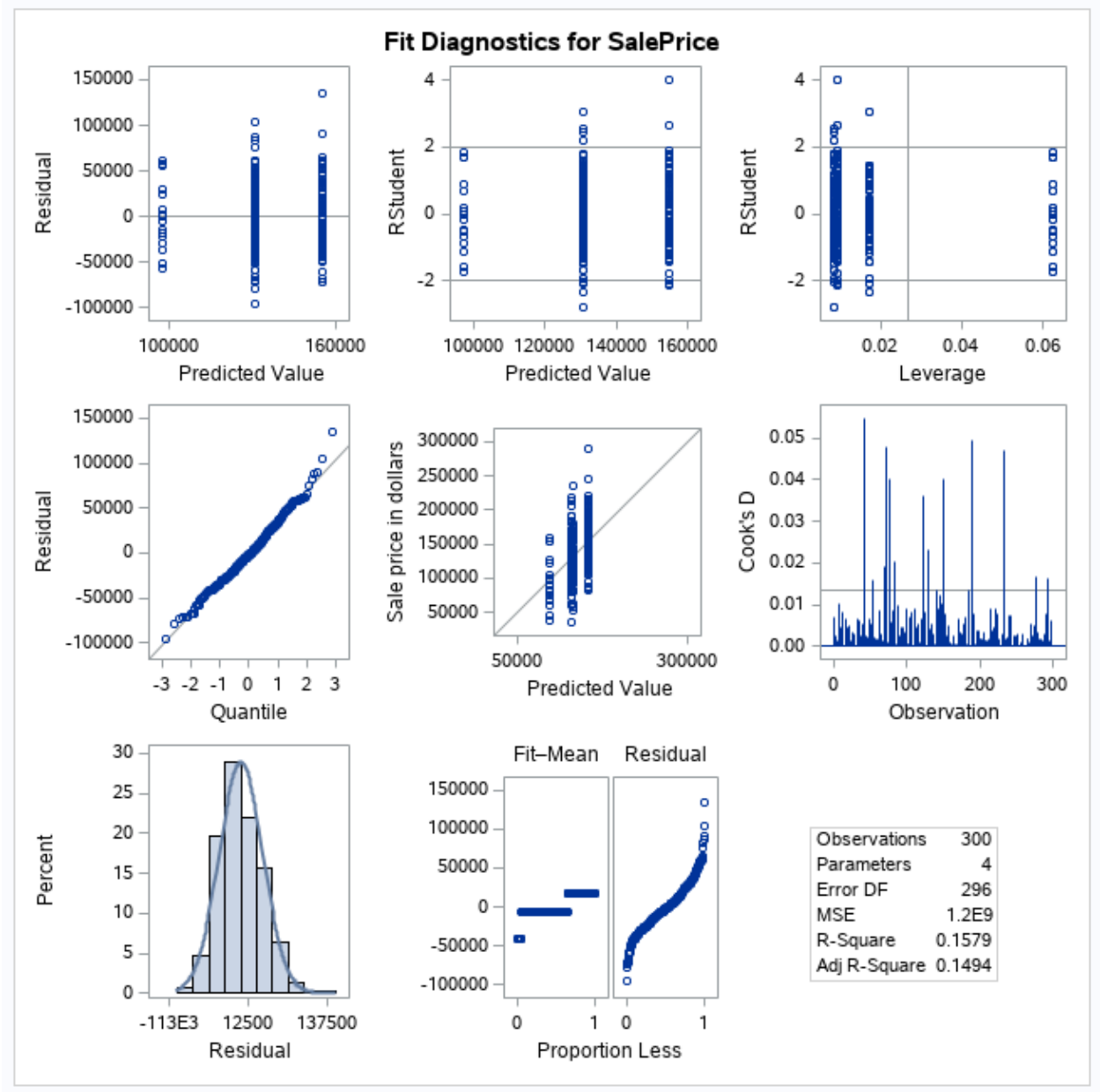
Dependent Variable: SalePrice Sale price in dollars

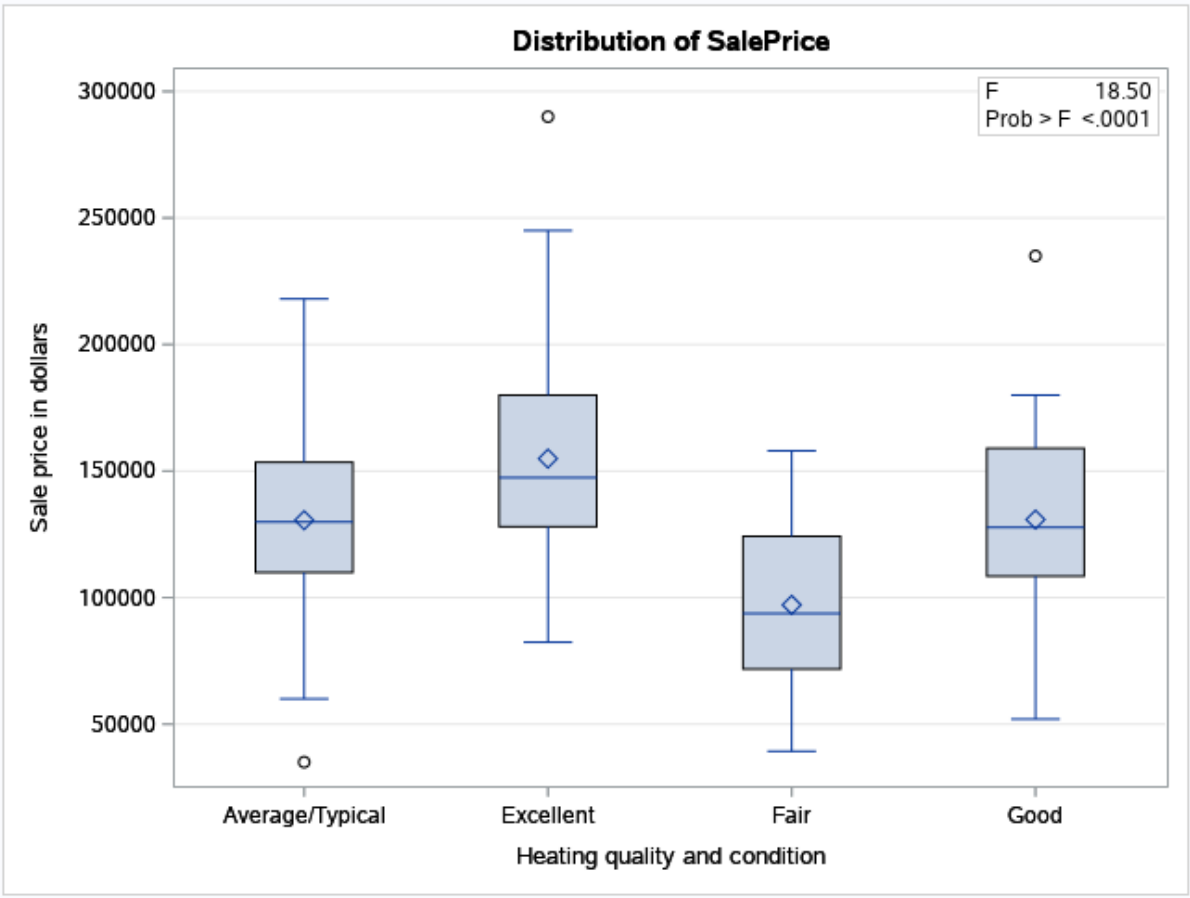
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	66835556221	22278518740	18.50	<.0001
Error	296	356387963289	1204013389.5		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.157920	25.23100	34698.90	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.50	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.50	<.0001


Fit Diagnostics for SalePrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Sale price in dollars
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 300
Parameters 4
Error DF 296
MSE 1.2E9
R-Square 0.1579
Adj R-Square 0.1494



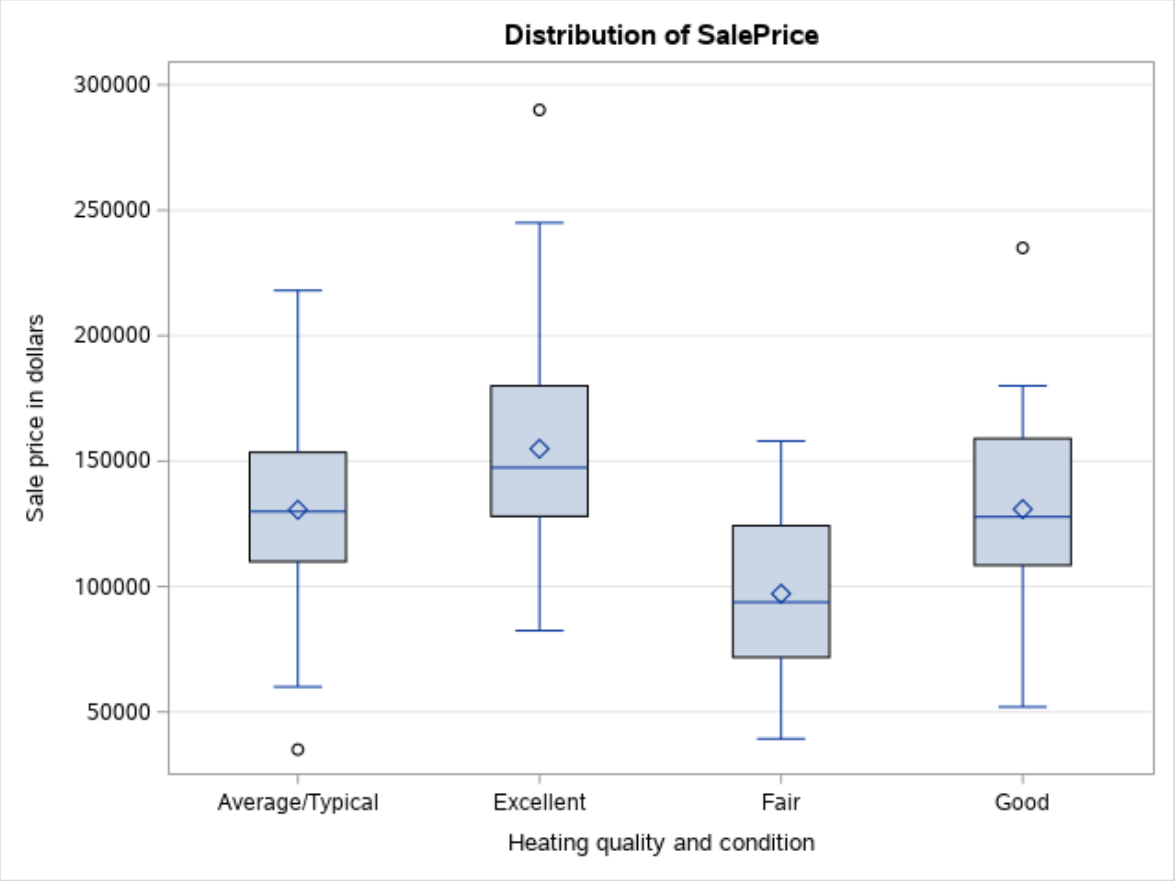
One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure

Levene's Test for Homogeneity of SalePrice Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Heating_QC	3	5.931E18	1.977E18	0.58	0.6305
Error	296	1.014E21	3.426E18		

One-Way ANOVA with Heating Quality as Predictor

The GLM Procedure



Level of Heating_QC	N	SalePrice	
		Mean	Std Dev
Average/Typical	119	130573.529	32177.4508
Excellent	107	154919.187	36822.8795
Fair	16	97118.750	37423.5437
Good	58	130844.086	34912.5027

6. Post hoc pairewise comparisons to find out which heating quality has the higher sale price

Post hoc methods to find which group has the higher mean for house sale price, based on heating quality:

- Tukey method: used to ajust Type I error while making multiple comparisons
    - Note: SAS applies the Tukey method by default, in case no option is selected to control for muliple comparisons

- Dunnett's method: used when making comparisons to a control group

Methods for pairewise comparisons:

- Diffogram
- Control plot

In the code below, we use the Tukey method to adjust for multiple comparisons, then the Dunnett method while making comparisons to a control group:

```
ods graphics;

ods select lsmeans diff diffplot controlplot;
proc glm data=STAT1.ameshousing3
         plots(only)=(diffplot(center) controlplot);
    class Heating_QC;
    model SalePrice=Heating_QC;
    lsmeans Heating_QC / pdiff=all
                         adjust=tukey;
    lsmeans Heating_QC / pdiff=control('Average/Typical')
                         adjust=dunnett;
    format Heating_QC $Heating_QC.;
    title "Post-Hoc Analysis of ANOVA - Heating Quality as
Predictor";
run;
quit;

title;
```

Post-Hoc Analysis of ANOVA - Heating Quality as Predictor

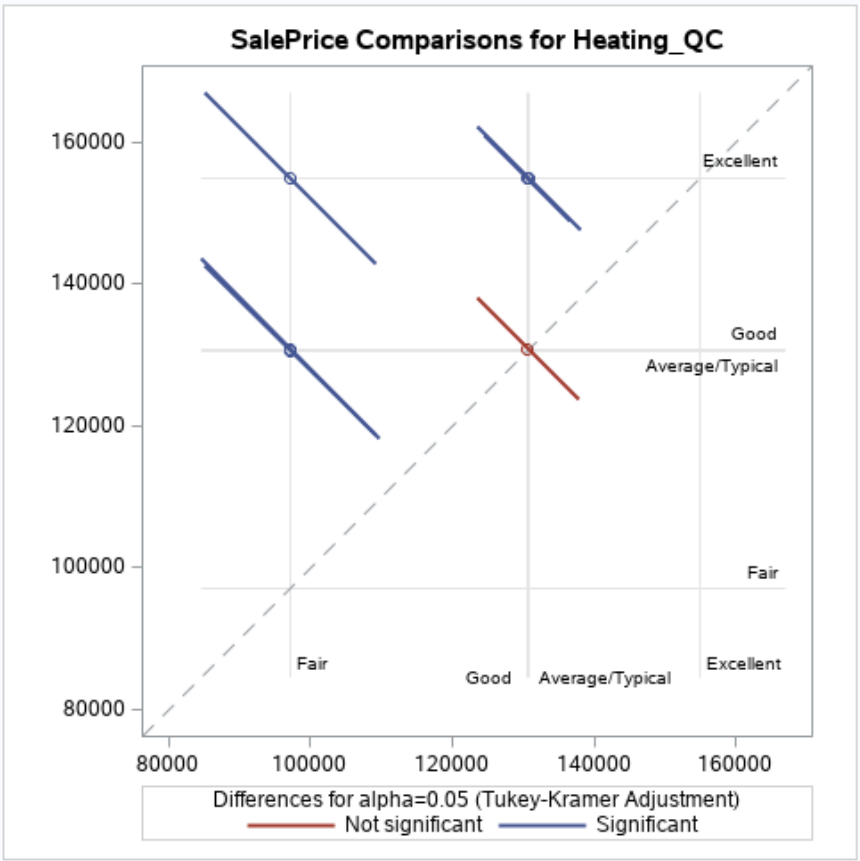
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Heating_QC	SalePrice LSMEAN	LSMEAN Number
Average/Typical	130573.529	1
Excellent	154919.187	2
Fair	97118.750	3
Good	130844.086	4

Least Squares Means for effect Heating_QC
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SalePrice

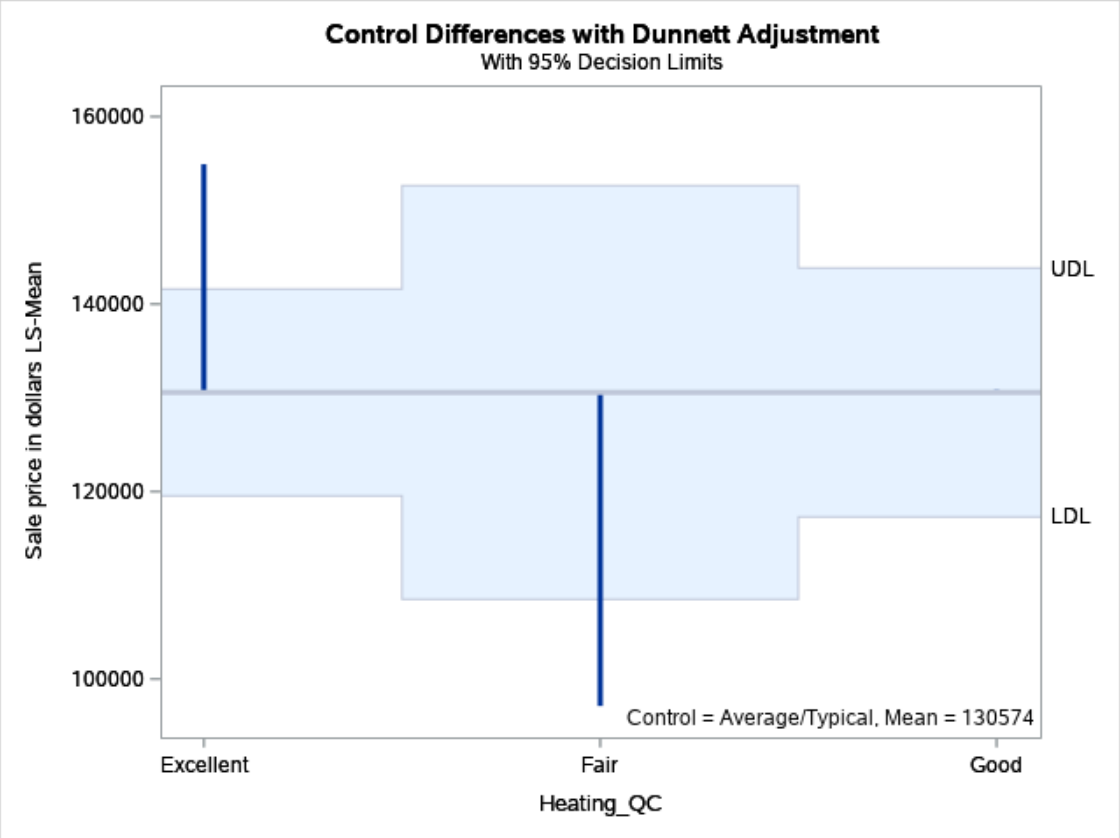
i/j	1	2	3	4
1		<.0001	0.0020	1.0000
2	<.0001		<.0001	0.0002
3	0.0020	<.0001		0.0037
4	1.0000	0.0002	0.0037	



Post-Hoc Analysis of ANOVA - Heating Quality as Predictor

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

Heating_QC	SalePrice LSMEAN	H0:LSMean=Control Pr > |t|
Average/Typical	130573.529	
Excellent	154919.187	<.0001
Fair	97118.750	0.0010
Good	130844.086	0.9999



7. Correlation analysis

/* Let's explore how continuous variables in the dataset might relate linearly to home sale price */

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;


ods graphics / reset=all imagemap;

proc corr data=STAT1.AmesHousing3 rank
```

```
          plots(only)=scatter(nvar=all ellipse=none);
   var &interval;
   with SalePrice;
   id PID;
   title "Correlations and Scatter Plots with SalePrice";
run;
title;
```

Correlations and Scatter Plots with SalePrice

The CORR Procedure

1 With Variables:	SalePrice
8 Variables:	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Simple Statistics

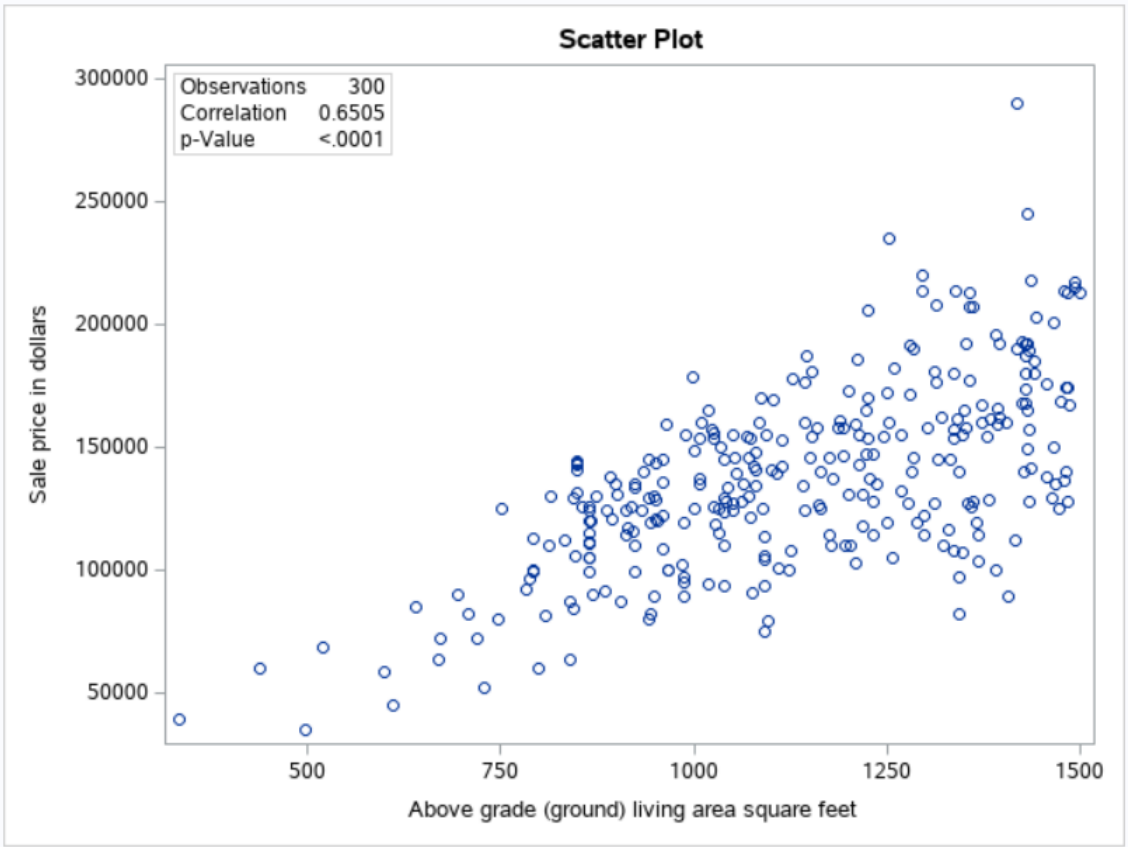
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
SalePrice	300	137525	37623	41257460	35000	290000	Sale price in dollars
Gr_Liv_Area	300	1131	232.64939	339222	334.00000	1500	Above grade (ground) living area square feet
Basement_Area	300	882.31000	359.78397	264693	0	1645	Basement area in square feet
Garage_Area	300	369.45333	176.25309	110836	0	902.00000	Size of garage in square feet
Deck_Porch_Area	300	118.26333	132.61169	35479	0	897.00000	Total area of decks and porches in square feet
Lot_Area	300	8294	3324	2488241	1495	26142	Lot size in square feet
Age_Sold	300	45.88667	27.47697	13766	1.00000	135.00000	Age of house when sold, in years
Bedroom_AbvGr	300	2.51333	0.69144	754.00000	0	4.00000	Bedrooms above grade
Total_Bathroom	300	1.70167	0.65707	510.50000	1.00000	4.10000	Total number of bathrooms (half bathrooms counted 10%)

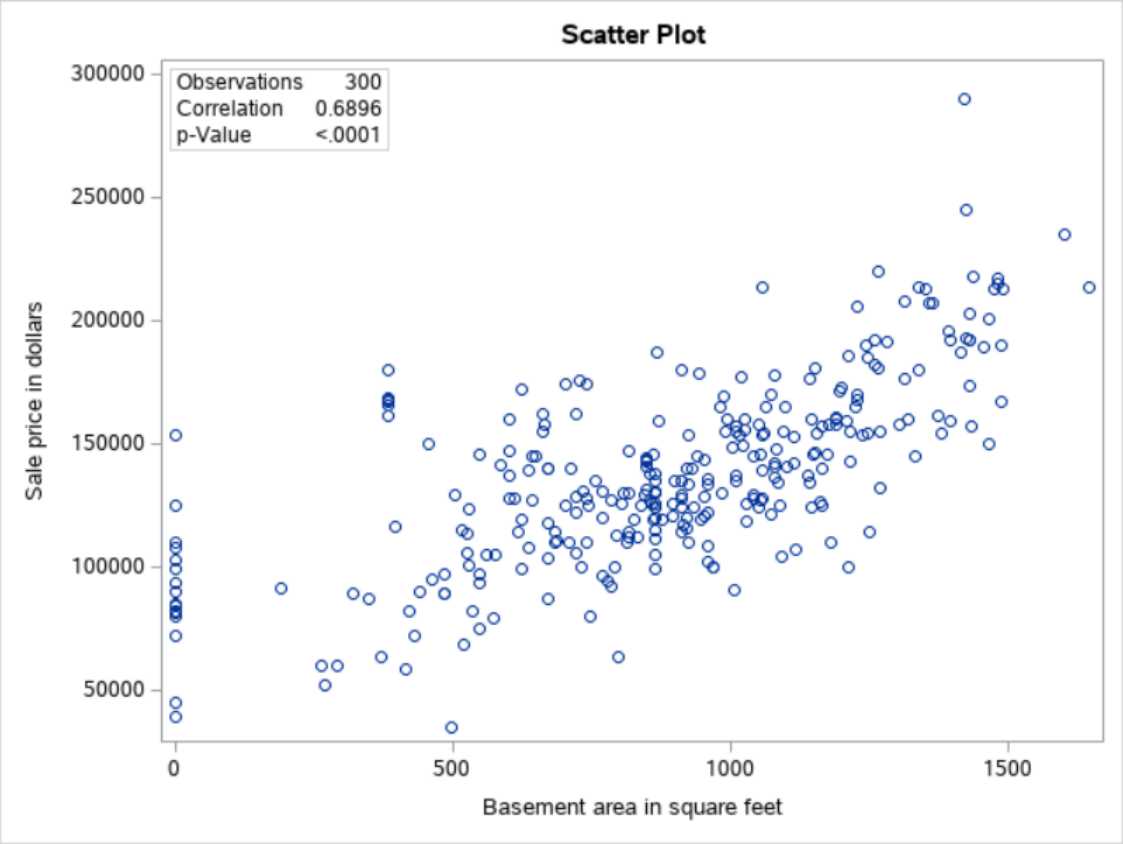
Pearson Correlation Coefficients, N = 300
Prob > |r| under H0: Rho=0

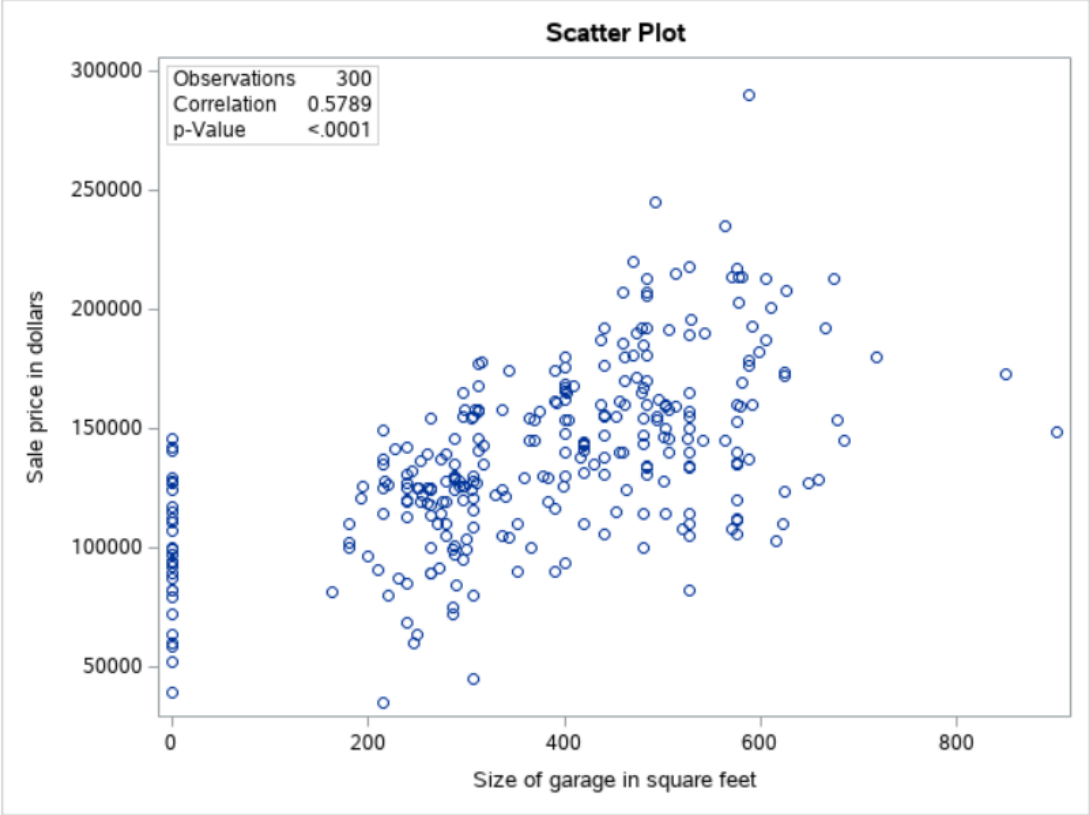
	Basement_Area	Gr_Liv_Area	Age_Sold	Total_Bathroom	Garage_Area	Deck_Porch_Area	Lot_Area	Bedroom_AbvGr
SalePrice Sale price in dollars	0.68956 <.0001	0.65046 <.0001	-0.61542 <.0001	0.60043 <.0001	0.57892 <.0001	0.43989 <.0001	0.25335 <.0001	0.16594 0.0040

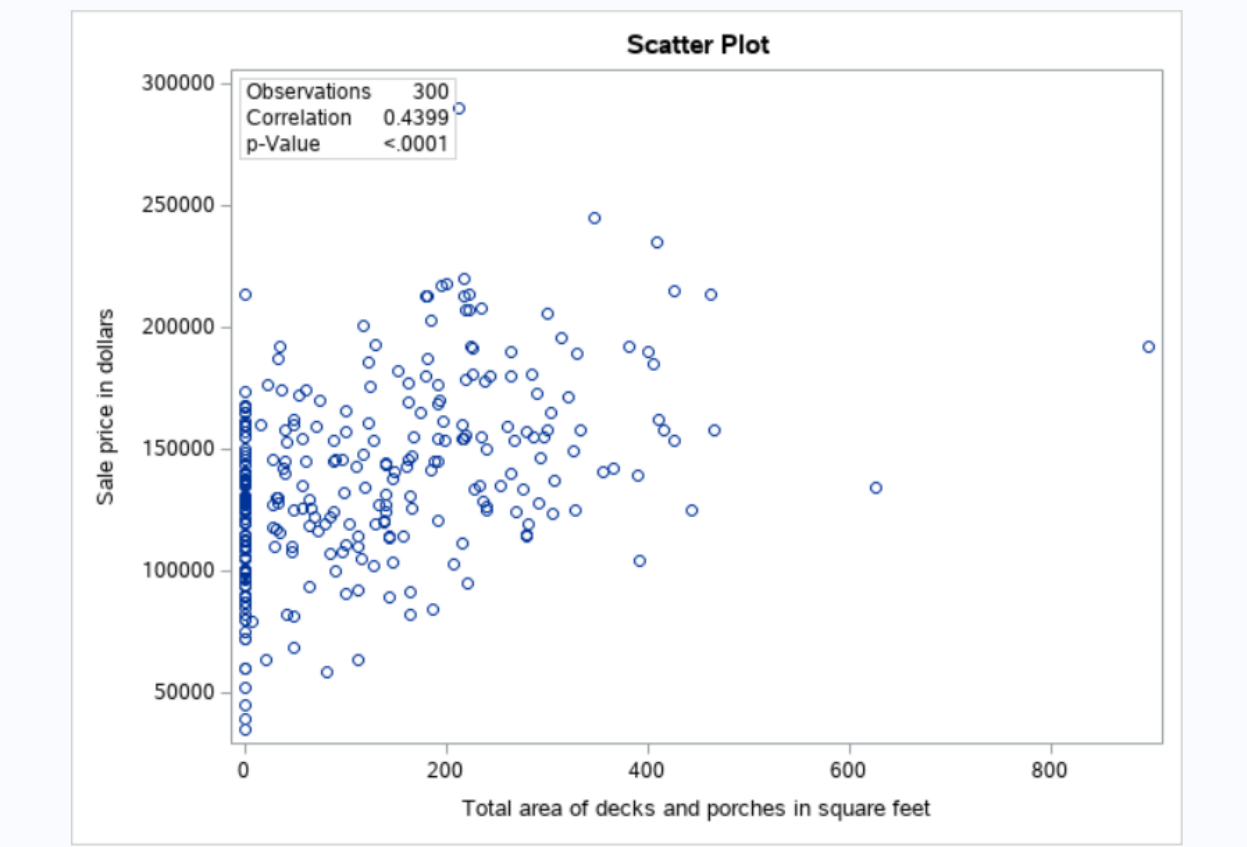
Correlations and Scatter Plots with SalePrice

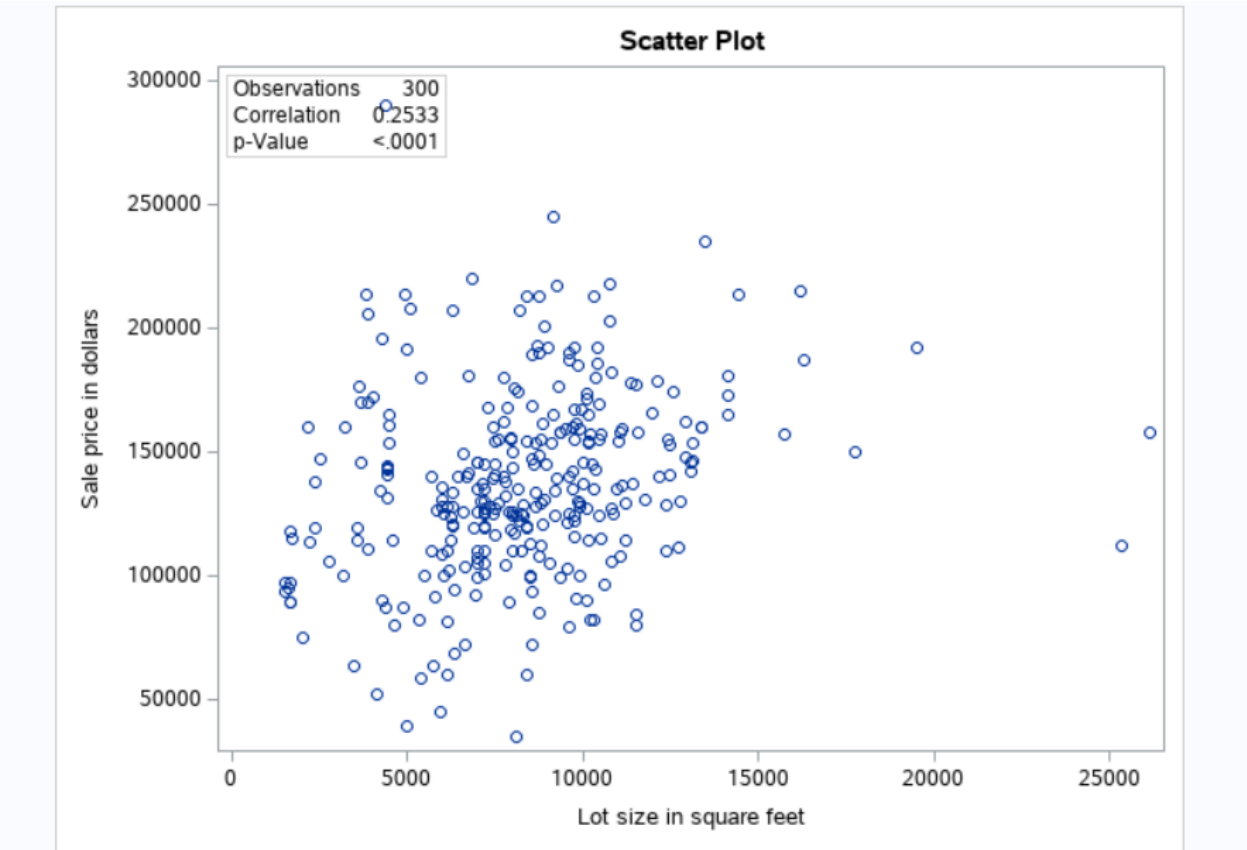
The CORR Procedure



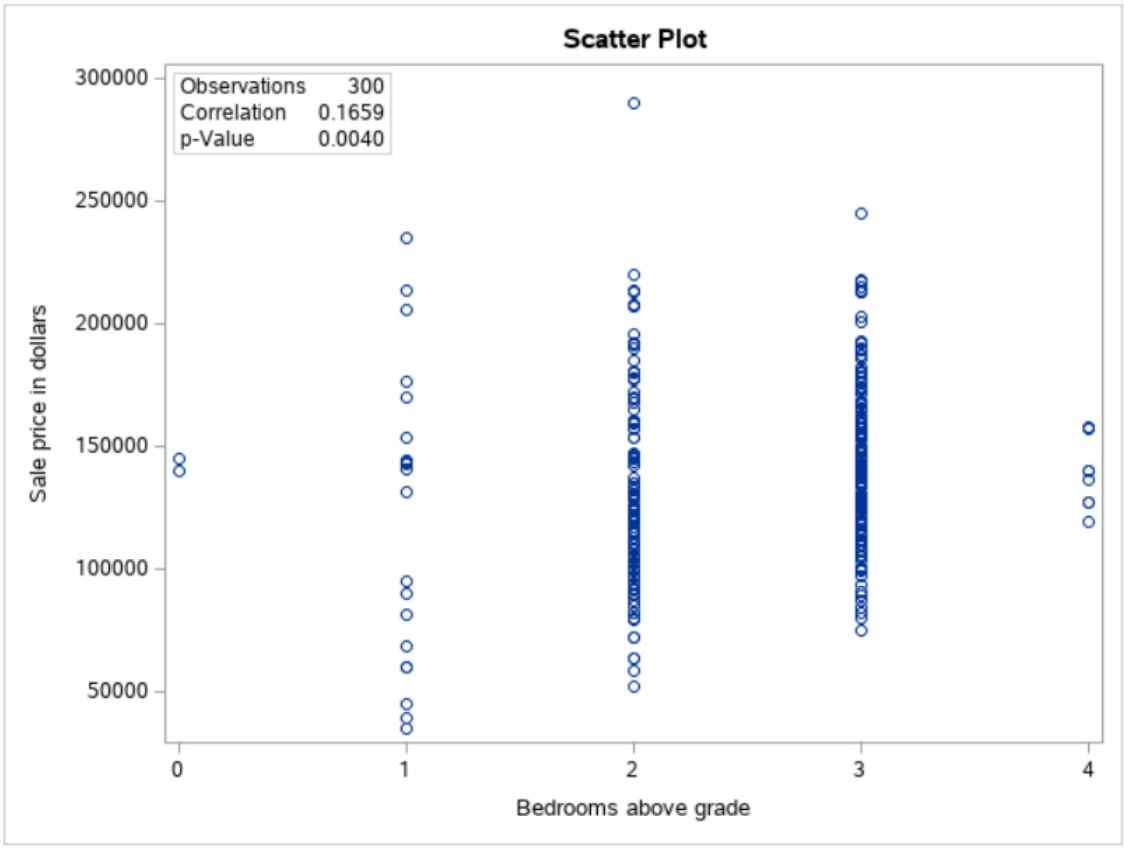

Scatter Plot
Observations 300
Correlation 0.6896
p-Value <.0001
Sale price in dollars
300000
250000
200000
150000
100000
50000
0
500
1000
1500
Basement area in square feet

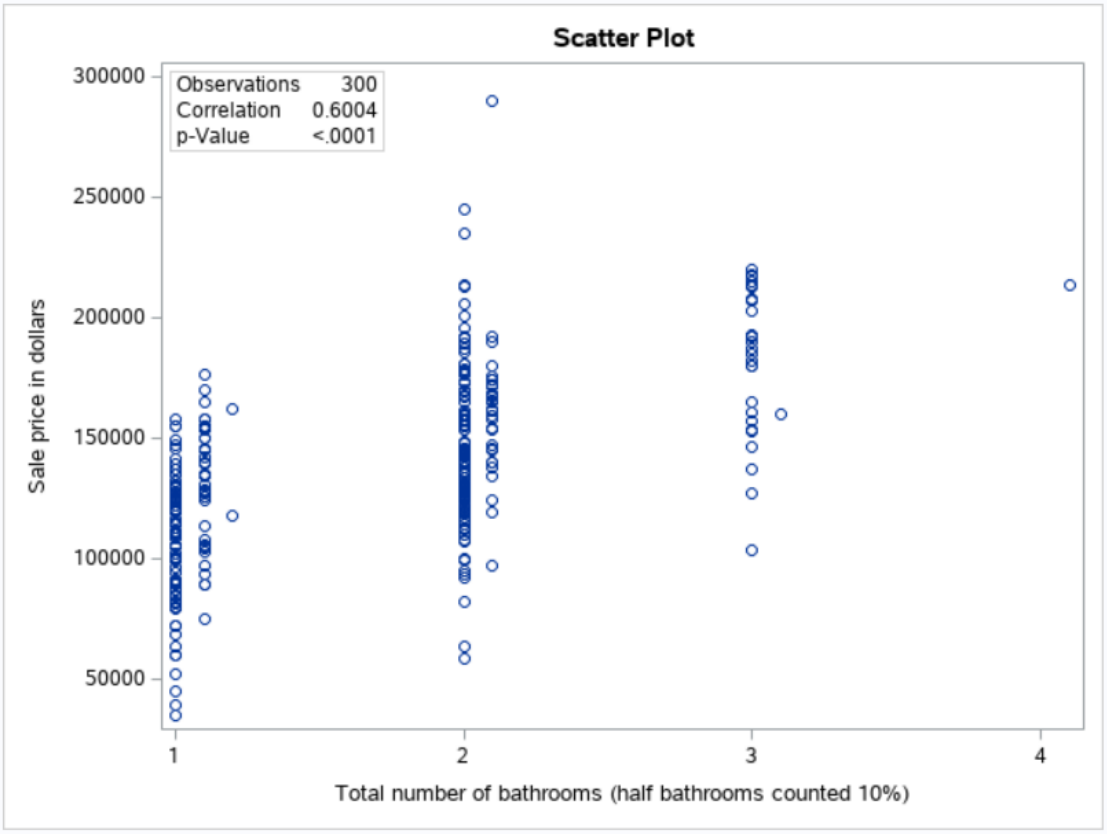

Scatter Plot
Observations 300
Correlation 0.5789
p-Value <.0001
Sale price in dollars
300000
250000
200000
150000
100000
50000
0
200
400
600
800
Size of garage in square feet


Scatter Plot
Observations 300
Correlation 0.4399
p-Value <.0001
Sale price in dollars
300000
250000
200000
150000
100000
50000
0
200
400
600
800
Total area of decks and porches in square feet


Scatter Plot
Observations 300
Correlation 0.2533
p-Value <.0001
Sale price in dollars
300000
250000
200000
150000
100000
50000
0
5000
10000
15000
20000
25000
Lot size in square feet


Scatter Plot
300000
250000
200000
150000
100000
50000
Sale price in dollars
Observations 300
Correlation -0.615
p-Value <.0001
0
25
50
75
100
125
Age of house when sold, in years


Scatter Plot
Observations 300
Correlation 0.1659
p-Value 0.0040
Sale price in dollars
300000
250000
200000
150000
100000
50000
0
1
2
3
4
Bedrooms above grade


Scatter Plot
Observations 300
Correlation 0.6004
p-Value <.0001
Sale price in dollars
300000
250000
200000
150000
100000
50000
1
2
3
4
Total number of bathrooms (half bathrooms counted 10%)

/* Let's check the correlation matrix for all predictors */

```
ods graphics off;
proc corr data=STAT1.AmesHousing3
          nosimple
          best=3;
   var &interval;
   title "Correlations and Scatter Plot Matrix of Predictors";
run;
title;
```

Correlations and Scatter Plot Matrix of Predictors

The CORR Procedure

8 Variables:	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Pearson Correlation Coefficients, N = 300
Prob > |r| under H0: Rho=0

Gr_Liv_Area Above grade (ground) living area square feet	Gr_Liv_Area 1.00000	Bedroom_AbvGr 0.48431 <.0001	Basement_Area 0.43985 <.0001
Basement_Area Basement area in square feet	Basement_Area 1.00000	Total_Bathroom 0.48500 <.0001	Gr_Liv_Area 0.43985 <.0001
Garage_Area Size of garage in square feet	Garage_Area 1.00000	Age_Sold -0.41346 <.0001	Total_Bathroom 0.36876 <.0001
Deck_Porch_Area Total area of decks and porches in square feet	Deck_Porch_Area 1.00000	Basement_Area 0.33689 <.0001	Gr_Liv_Area 0.28058 <.0001
Lot_Area Lot size in square feet	Lot_Area 1.00000	Bedroom_AbvGr 0.29801 <.0001	Basement_Area 0.27198 <.0001
Age_Sold Age of house when sold, in years	Age_Sold 1.00000	Total_Bathroom -0.52889 <.0001	Garage_Area -0.41346 <.0001
Bedroom_AbvGr Bedrooms above grade	Bedroom_AbvGr 1.00000	Gr_Liv_Area 0.48431 <.0001	Lot_Area 0.29801 <.0001
Total_Bathroom Total number of bathrooms (half bathrooms counted 10%)	Total_Bathroom 1.00000	Age_Sold -0.52889 <.0001	Basement_Area 0.48500 <.0001

8. Simple linear regression analysis

/* Let build a model to explore how lot area affects sale price */

```
ods graphics;

proc reg data=STAT1.ameshousing3;
    model SalePrice=Lot_Area;
    title "Simple Regression with Lot Area as Regressor";
run;
quit;

title;
```

Simple Regression with Lot Area as Regressor

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	300
Number of Observations Used	300

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27164711173	27164711173	20.44	<.0001
Error	298	3.960588E11	1329056404		
Corrected Total	299	4.232235E11			

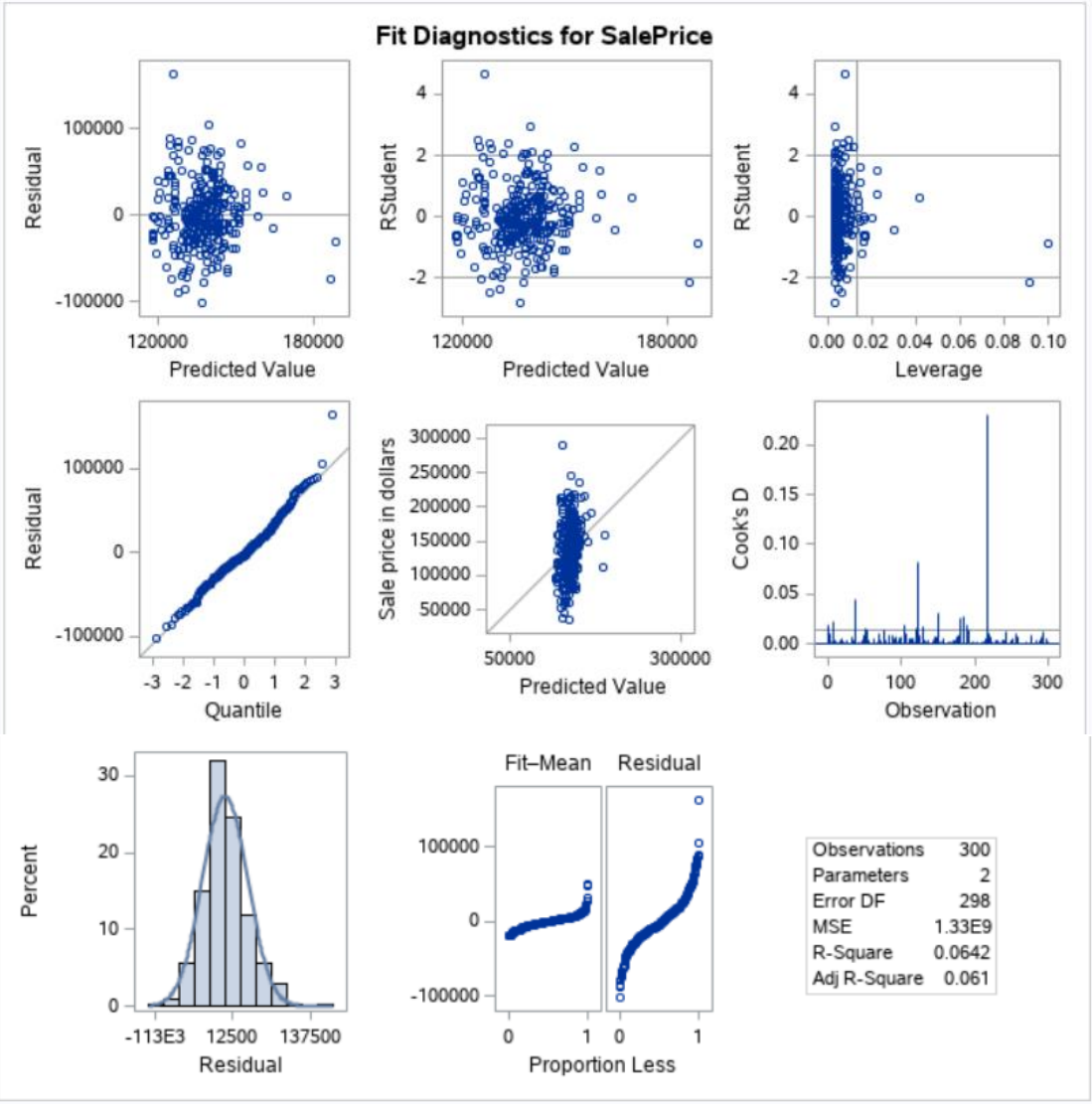
Root MSE	36456	R-Square	0.0642
Dependent Mean	137525	Adj R-Sq	0.0610
Coeff Var	26.50882		

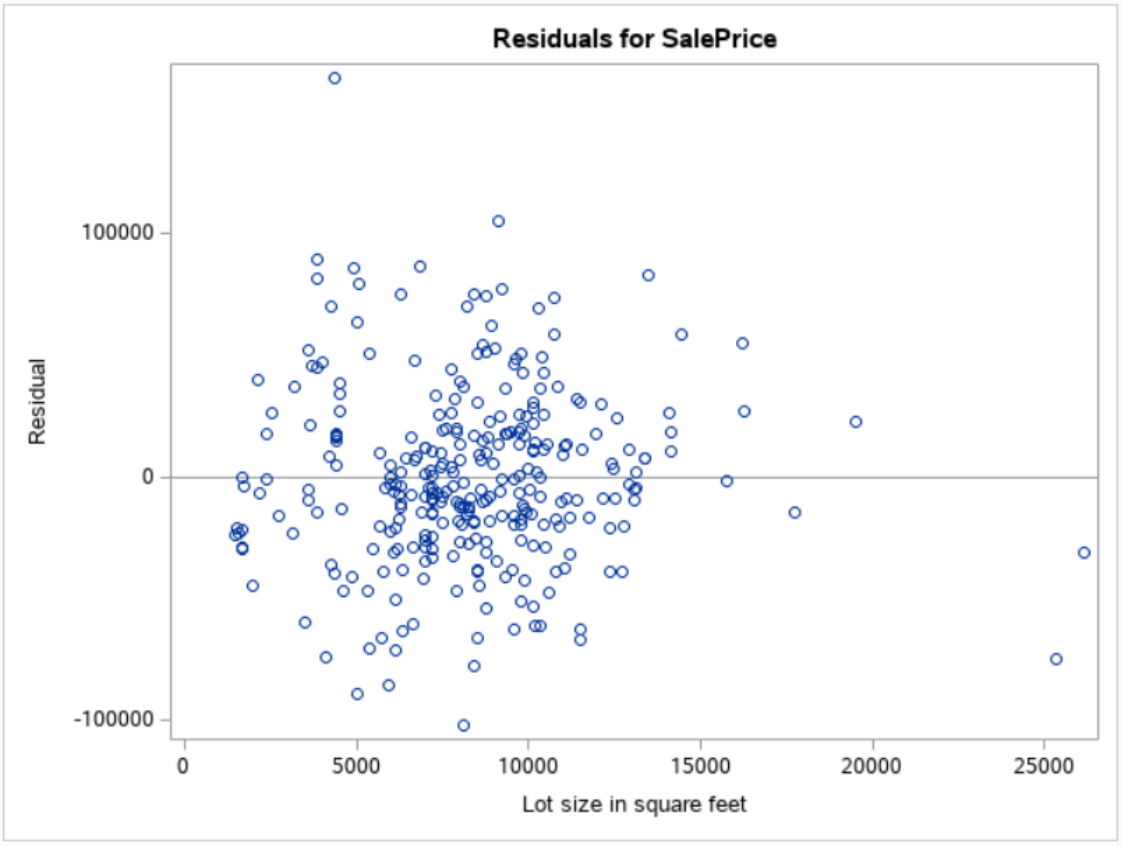
Parameter Estimates

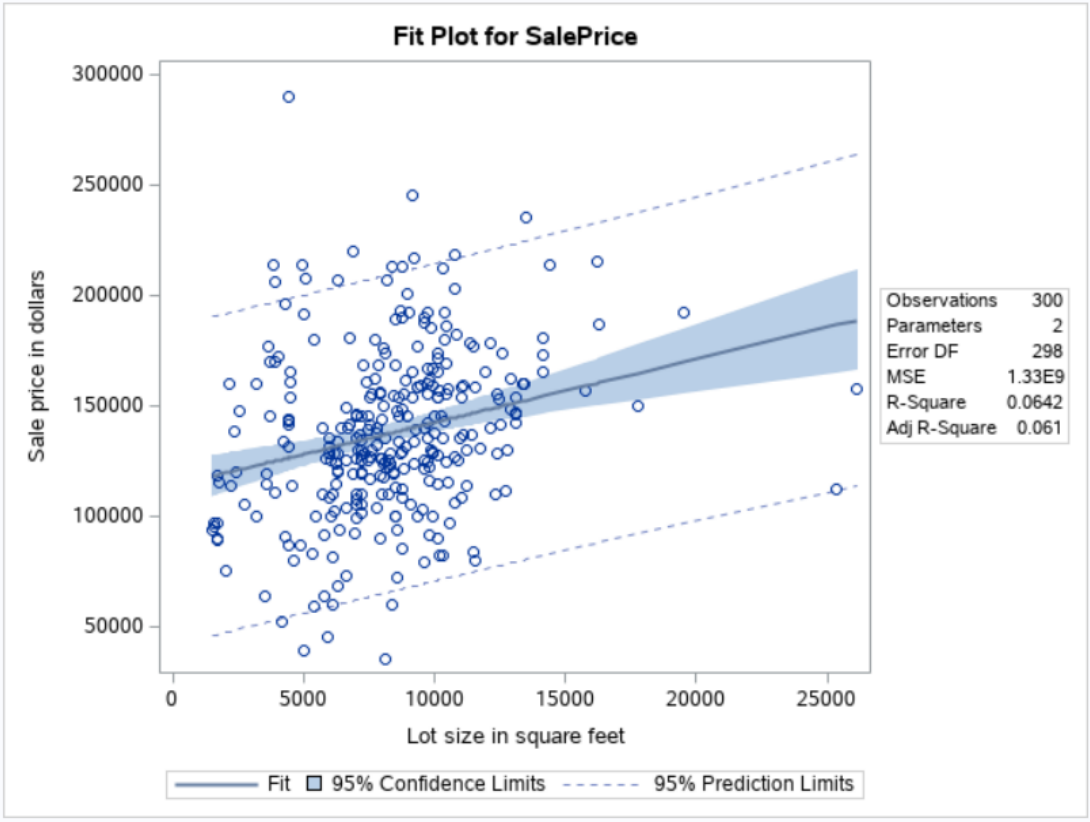
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	113740	5666.48352	20.07	<.0001
Lot_Area	Lot size in square feet	1	2.86770	0.63431	4.52	<.0001

Simple Regression with Lot Area as Regressor

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars




Residuals for SalePrice
Residual
100000
0
-100000
0
5000
10000
15000
20000
25000
Lot size in square feet


Fit Plot for SalePrice
Sale price in dollars
300000
250000
200000
150000
100000
50000
0
5000
10000
15000
20000
25000
Lot size in square feet
Observations 300
Parameters 2
Error DF 298
MSE 1.33E9
R-Square 0.0642
Adj R-Square 0.061
Fit
95% Confidence Limits
95% Prediction Limits

9. Two-way ANOVA: Without Interaction

N-way ANOVA analysis allows for exploring the effect of multiple (n) categorical predictors on a continuous response variable. Let's explore the effect of heating quality and season sold on sale price.

```
/* Start by checking summary statistics */
ods graphics off;
proc means data=STAT1.ameshousing3
           mean var std nway;
    class Season_Sold Heating_QC;
    var SalePrice;
    format Season_Sold Season.;
    title 'Selected Descriptive Statistics';
run;
```

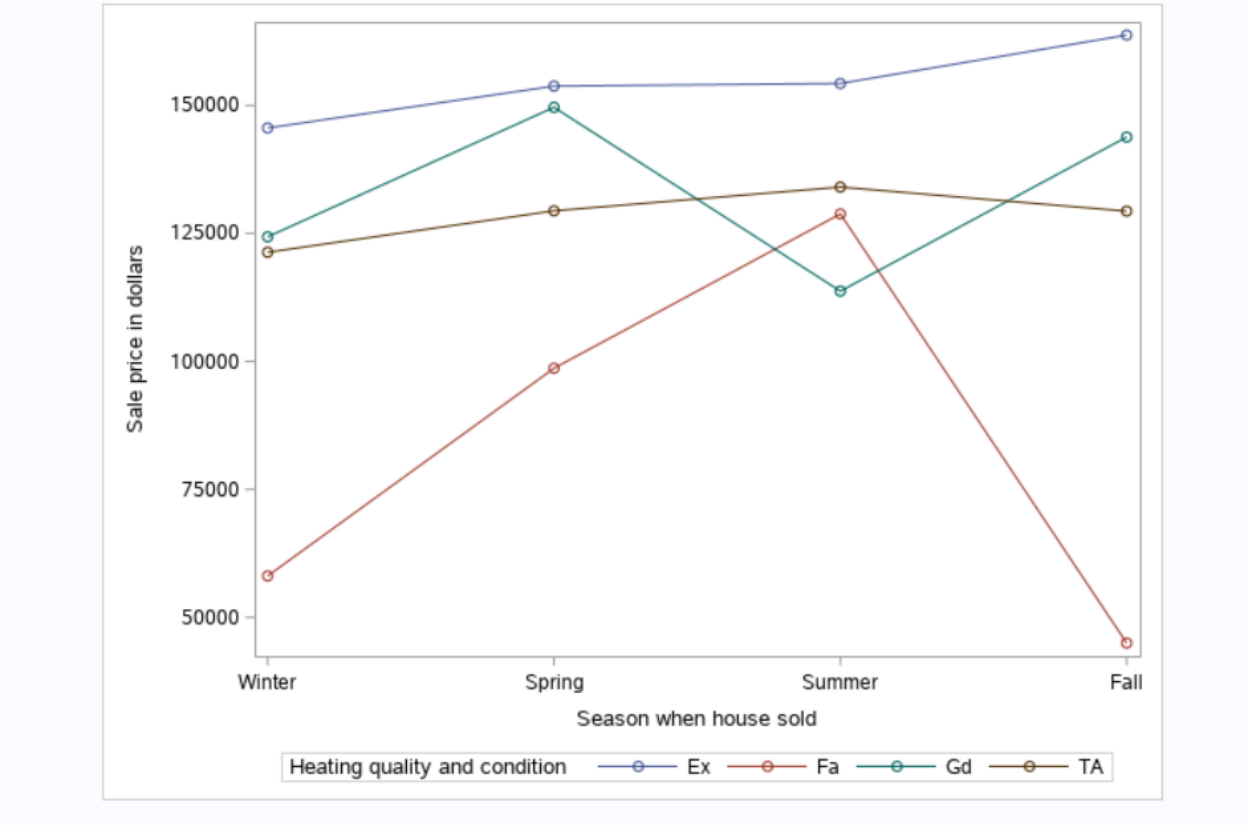
Selected Descriptive Statistics

The MEANS Procedure

Analysis Variable : SalePrice Sale price in dollars					
Season when house sold	Heating quality and condition	N Obs	Mean	Variance	Std Dev
Winter	Ex	6	145583.33	1579141667	39738.42
	Fa	3	58100.00	321330000	17925.68
	Gd	10	124330.00	935189000	30580.86
	TA	16	121312.50	1679295833	40979.21
Spring	Ex	41	153765.24	1129742652	33611.64
	Fa	7	98657.14	452506190	21272.19
	Gd	18	149619.83	1082782633	32905.66
	TA	34	129404.41	767370965	27701.46
Summer	Ex	45	154279.42	1244833504	35282.20
	Fa	5	128800.00	1332825000	36507.88
	Gd	22	113727.27	1155184935	33988.01
	TA	58	134046.55	1138642444	33743.78
Fall	Ex	15	163726.93	2436449681	49360.41
	Fa	1	45000.00	.	.
	Gd	8	143812.50	547495536	23398.62
	TA	11	129345.45	462560727	21507.23

/* Compare mean home sale prices by season and heating quality */

```
proc sgplot data=STAT1.ameshousing3;
    vline Season_Sold / group=Heating_QC
                        stat=mean
                        response=SalePrice
                        markers;
    format Season_Sold season.;
run;
```



/* Let's run the ANOVA analysis */

```
ods graphics on;

proc glm data=STAT1.ameshousing3 order=internal;
    class Season_Sold Heating_QC;
    model SalePrice = Heating_QC Season_Sold;
    lsmeans Season_Sold / diff adjust=tukey;
    format Season_Sold season.;
    title "Model with Heating Quality and Season as Predictors";
run;
quit;

title;
```

Model with Heating Quality and Season as Predictors

The GLM Procedure

Class Level Information		
Class	Levels	Values
Season_Sold	4	Winter Spring Summer Fall
Heating_QC	4	Ex Fa Gd TA

Number of Observations Read	300
Number of Observations Used	300

Model with Heating Quality and Season as Predictors

The GLM Procedure

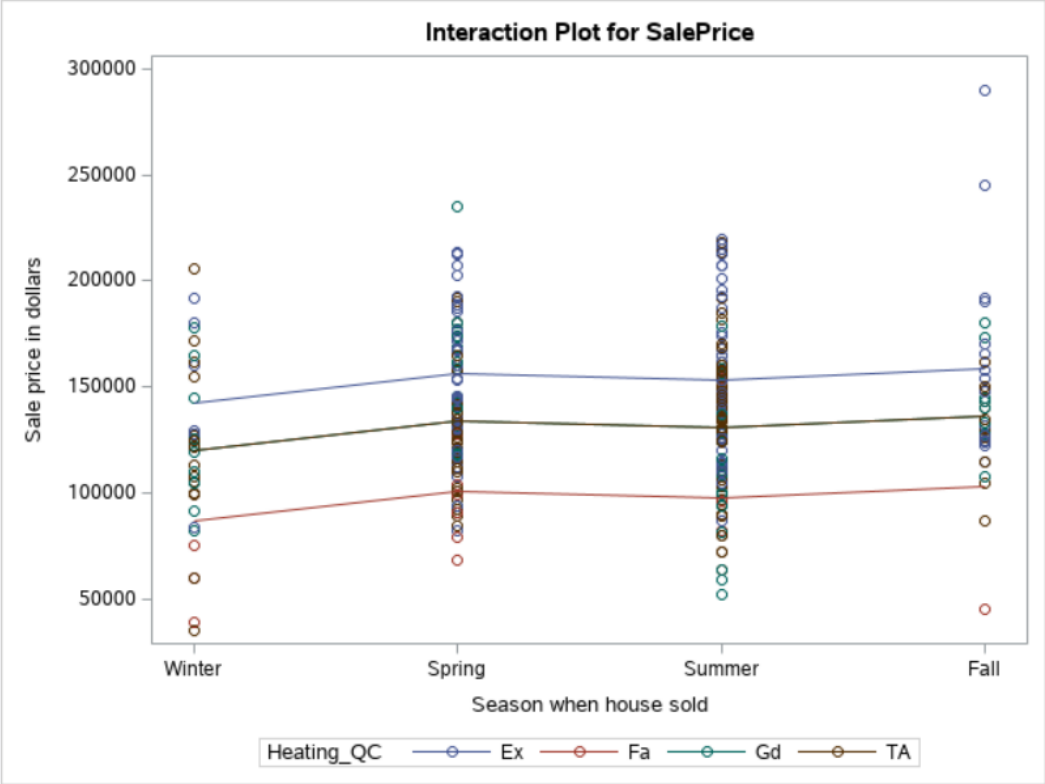
Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	72774816066	12129136011	10.14	<.0001
Error	293	350448703445	1196070660.2		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.171954	25.14764	34584.25	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	18.63	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	60050783038	20016927679	16.74	<.0001
Season_Sold	3	5939259845	1979753282	1.66	0.1768



Model with Heating Quality and Season as Predictors

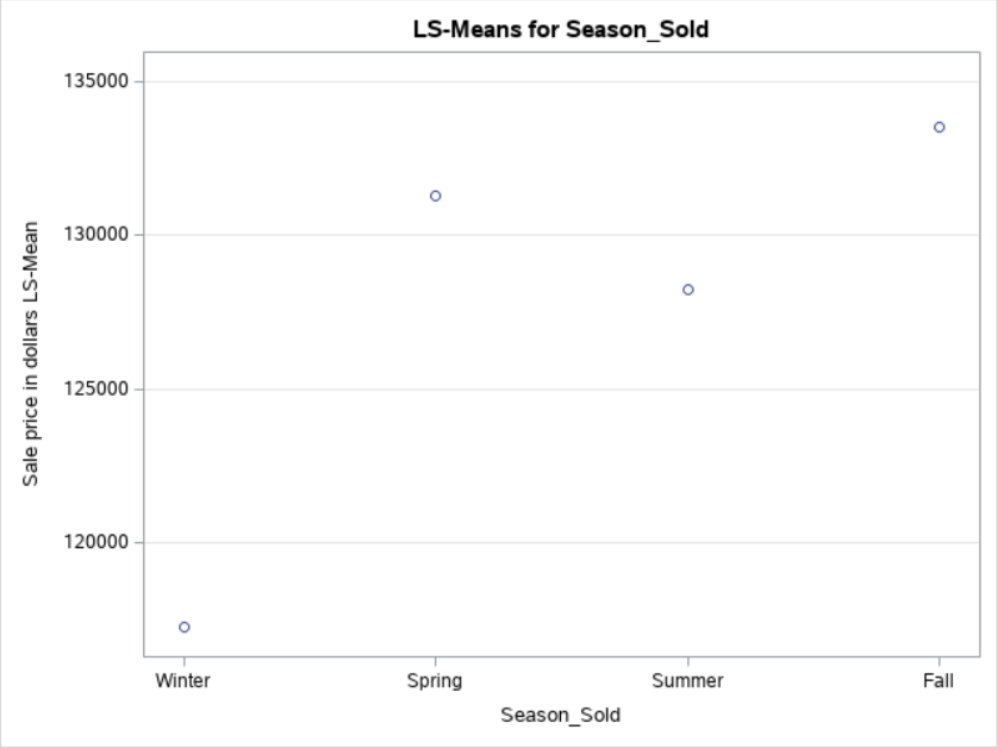
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

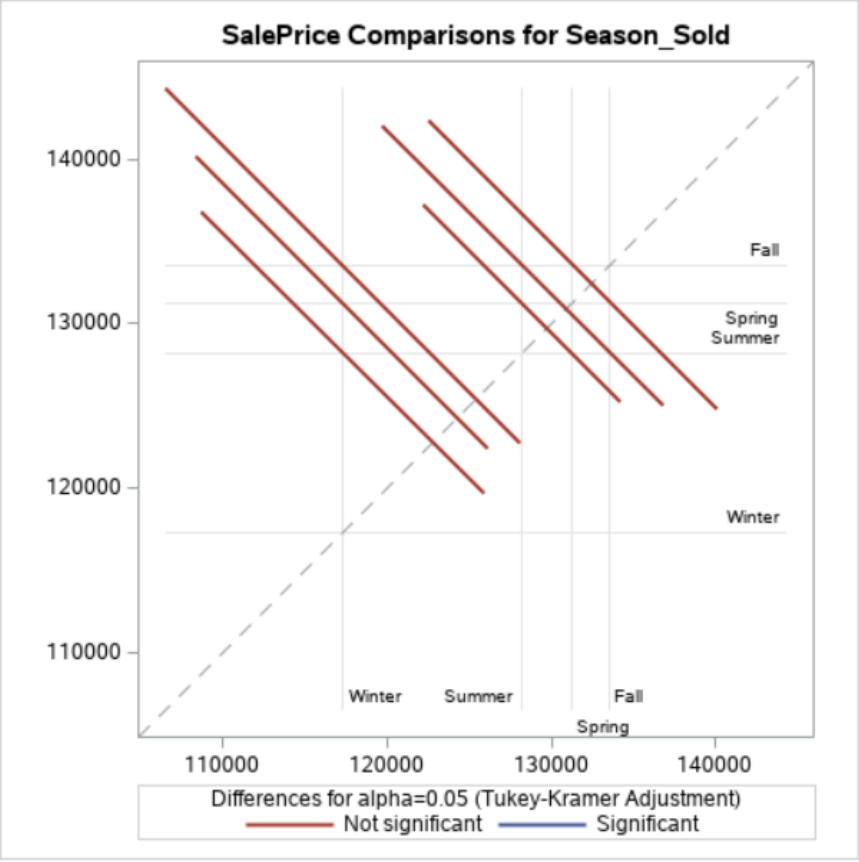
Season_Sold	SalePrice LSMEAN	LSMEAN Number
Winter	117255.605	1
Spring	131263.281	2
Summer	128216.231	3
Fall	133543.394	4

Least Squares Means for effect Season_Sold
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SalePrice

i/j	1	2	3	4
1		0.1760	0.3529	0.2089
2	0.1760		0.9124	0.9870
3	0.3529	0.9124		0.8517
4	0.2089	0.9870	0.8517	


LS-Means for Season_Sold
135000
130000
125000
120000
Sale price in dollars LS-Mean
Winter
Spring
Summer
Fall
Season_Sold


SalePrice Comparisons for Season_Sold
140000
130000
120000
110000
Fall
Spring
Summer
Winter
Winter
Summer
Fall
Spring
110000
120000
130000
140000
Differences for alpha=0.05 (Tukey-Kramer Adjustment)
Not significant
Significant

10. Two-way ANOVA: With Interaction

/* Let's perform the ANOVA again by including interaction terms between predictors */

```
ods graphics on;


proc glm data=STAT1.ameshousing3
         order=internal
         plots(only)=intplot;
    class Season_Sold Heating_QC;
    model SalePrice = Heating_QC Season_Sold Heating_QC*Season_Sold;
    lsmeans Heating_QC*Season_Sold / diff slice=Heating_QC;
    format Season_Sold Season.;
    store out=interact;
    title "Model with Heating Quality and Season as Interacting
Predictors";
run;
quit;
```

Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure

Class Level Information		
Class	Levels	Values
Season_Sold	4	Winter Spring Summer Fall
Heating_QC	4	Ex Fa Gd TA

Number of Observations Read	300
Number of Observations Used	300

Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure

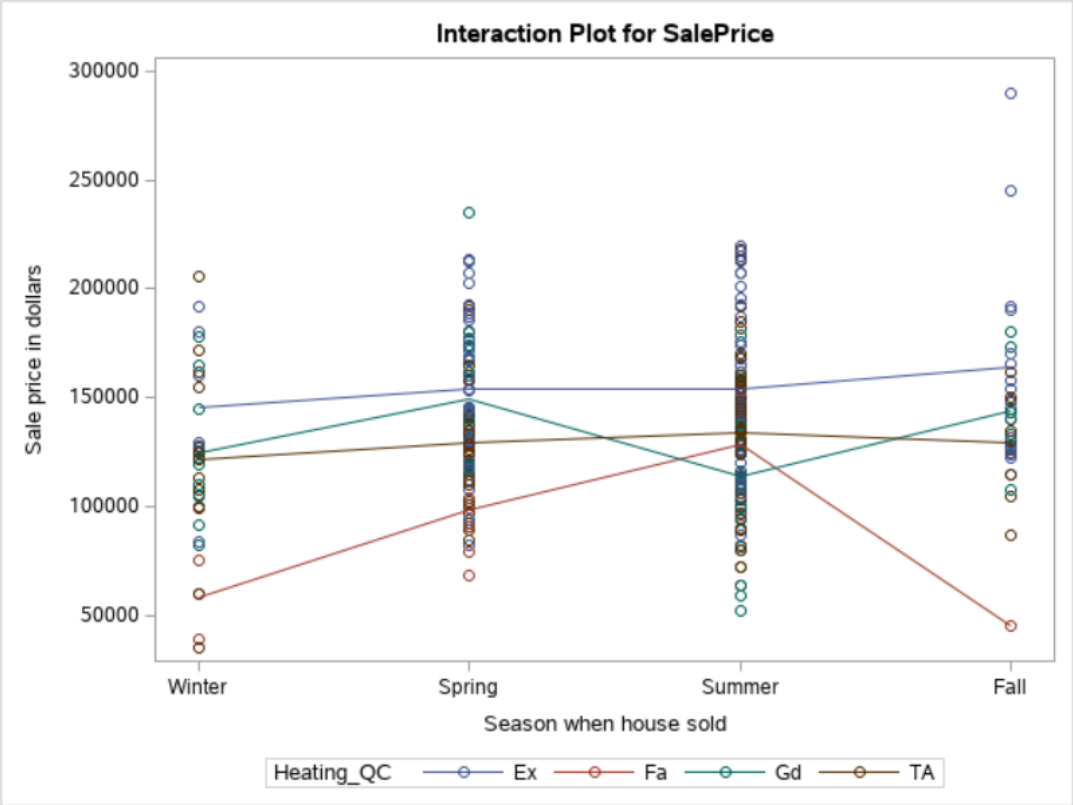
Dependent Variable: SalePrice Sale price in dollars

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	97609874155	6507324943.7	5.68	<.0001
Error	284	325613645356	1146526920.3		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.230634	24.62130	33860.40	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Heating_QC	3	66835556221	22278518740	19.43	<.0001
Season_Sold	3	5939259845	1979753282	1.73	0.1617
Season_So*Heating_QC	9	24835058089	2759450899	2.41	0.0121

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Heating_QC	3	51116493768	17038831256	14.86	<.0001
Season_Sold	3	9318181844	3106060615	2.71	0.0455
Season_So*Heating_QC	9	24835058089	2759450899	2.41	0.0121



Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure
Least Squares Means

Season_Sold	Heating_QC	SalePrice LSMEAN	LSMEAN Number
Winter	Ex	145583.333	1
Winter	Fa	58100.000	2
Winter	Gd	124330.000	3
Winter	TA	121312.500	4
Spring	Ex	153765.244	5
Spring	Fa	98657.143	6
Spring	Gd	149619.833	7
Spring	TA	129404.412	8
Summer	Ex	154279.422	9
Summer	Fa	128800.000	10
Summer	Gd	113727.273	11
Summer	TA	134046.552	12
Fall	Ex	163726.933	13
Fall	Fa	45000.000	14
Fall	Gd	143812.500	15
Fall	TA	129345.455	16

Least Squares Means for effect Season_So*Heating_QC
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: SalePrice

i/j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		0.0003	0.2252	0.1354	0.5808	0.0133	0.8005	0.2815	0.5550	0.4137	0.0420	0.4276	0.2682	0.0063	0.9229	0.3455
2	0.0003		0.0032	0.0033	<.0001	0.0837	<.0001	0.0005	<.0001	0.0046	0.0080	0.0002	<.0001	0.7378	0.0002	0.0014
3	0.2252	0.0032		0.8252	0.0143	0.1250	0.0593	0.6773	0.0119	0.8097	0.4123	0.4027	0.0047	0.0263	0.2261	0.7349
4	0.1354	0.0033	0.8252		0.0013	0.1409	0.0156	0.4312	0.0009	0.6664	0.4959	0.1840	0.0006	0.0296	0.1260	0.5452
5	0.5808	<.0001	0.0143	0.0013		<.0001	0.6654	0.0021	0.9440	0.1207	<.0001	0.0046	0.3304	0.0017	0.4476	0.0345
6	0.0133	0.0837	0.1250	0.1409	<.0001		0.0008	0.0295	<.0001	0.1295	0.3059	0.0095	<.0001	0.1394	0.0105	0.0619
7	0.8005	<.0001	0.0593	0.0156	0.6654	0.0008		0.0415	0.6221	0.2249	0.0010	0.0894	0.2344	0.0029	0.6868	0.1188
8	0.2815	0.0005	0.6773	0.4312	0.0021	0.0295	0.0415		0.0014	0.9703	0.0917	0.5261	0.0012	0.0146	0.2798	0.9960
9	0.5550	<.0001	0.0119	0.0009	0.9440	<.0001	0.6221	0.0014		0.1115	<.0001	0.0029	0.3502	0.0016	0.4211	0.0294
10	0.4137	0.0046	0.8097	0.6664	0.1207	0.1295	0.2249	0.9703	0.1115		0.3697	0.7398	0.0467	0.0246	0.4374	0.9762
11	0.0420	0.0080	0.4123	0.4959	<.0001	0.3059	0.0010	0.0917	<.0001	0.3697		0.0172	<.0001	0.0481	0.0322	0.2127
12	0.4276	0.0002	0.4027	0.1840	0.0046	0.0095	0.0894	0.5261	0.0029	0.7398	0.0172		0.0027	0.0096	0.4451	0.6732
13	0.2682	<.0001	0.0047	0.0006	0.3304	<.0001	0.2344	0.0012	0.3502	0.0467	<.0001	0.0027		0.0008	0.1802	0.0110
14	0.0063	0.7378	0.0263	0.0296	0.0017	0.1394	0.0029	0.0146	0.0016	0.0246	0.0481	0.0096	0.0008		0.0063	0.0177
15	0.9229	0.0002	0.2261	0.1260	0.4476	0.0105	0.6868	0.2798	0.4211	0.4374	0.0322	0.4451	0.1802	0.0063		0.3586
16	0.3455	0.0014	0.7349	0.5452	0.0345	0.0619	0.1188	0.9960	0.0294	0.9762	0.2127	0.6732	0.0110	0.0177	0.3586	

Model with Heating Quality and Season as Interacting Predictors

The GLM Procedure
Least Squares Means

Season_So*Heating_QC Effect Sliced by Heating_QC for SalePrice

Heating_QC	DF	Sum of Squares	Mean Square	F Value	Pr > F
Ex	3	1759608339	586536113	0.51	0.6746
Fa	3	12318827232	4106275744	3.58	0.0143
Gd	3	14560964166	4853654722	4.23	0.0060
TA	3	2134918196	711639399	0.62	0.6021

Note: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

/* We use PROC PLM to access the item stored in the previous step, and adjust for multiple testing in the ANOVA analysis, using the Tukey method */

```
proc plm restore=interact plots=all;
    slice Heating_QC*Season_Sold / sliceby=Heating_QC adjust=tukey;
    effectplot interaction(sliceby=Heating_QC) / clm;
run;


title;
```

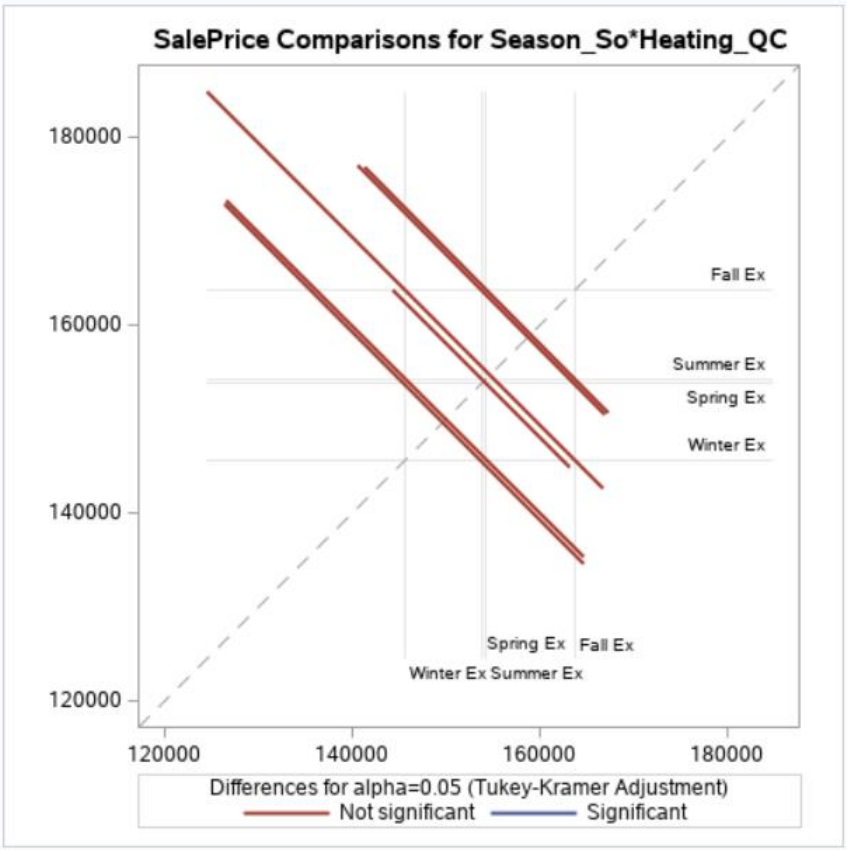
The PLM Procedure

Store Information	
Item Store	WORK.INTERACT
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLM
Date Created	21AUG21:04:53:27
Response Variable	SalePrice
Class Variables	Season_Sold Heating_QC
Model Effects	Intercept Heating_QC Season_Sold Season_So*Heating_QC

Class Level Information		
Class	Levels	Values
Season_Sold	4	Winter Spring Summer Fall
Heating_QC	4	Ex Fa Gd TA

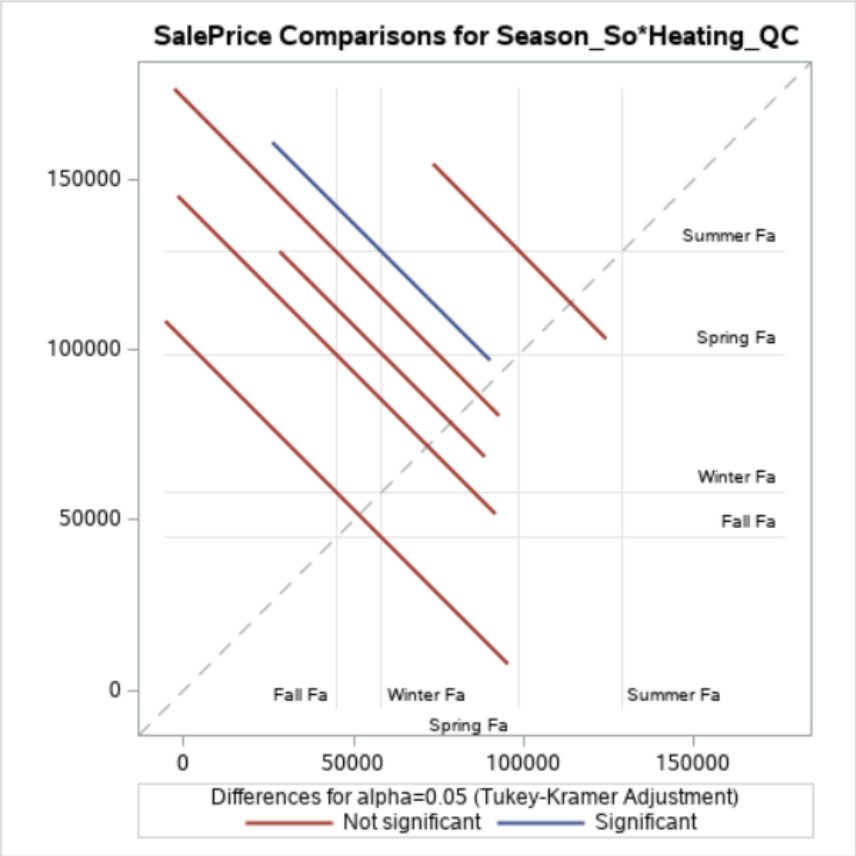
F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Ex	3	284	0.51	0.6746

Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer								
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC Ex	Winter	Spring	-8181.91	14800	284	-0.55	0.5808	0.9457
Heating_QC Ex	Winter	Summer	-8696.09	14716	284	-0.59	0.5550	0.9348
Heating_QC Ex	Winter	Fall	-18144	16356	284	-1.11	0.2682	0.6841
Heating_QC Ex	Spring	Summer	-514.18	7310.43	284	-0.07	0.9440	0.9999
Heating_QC Ex	Spring	Fall	-9961.69	10218	284	-0.97	0.3304	0.7638
Heating_QC Ex	Summer	Fall	-9447.51	10095	284	-0.94	0.3502	0.7856



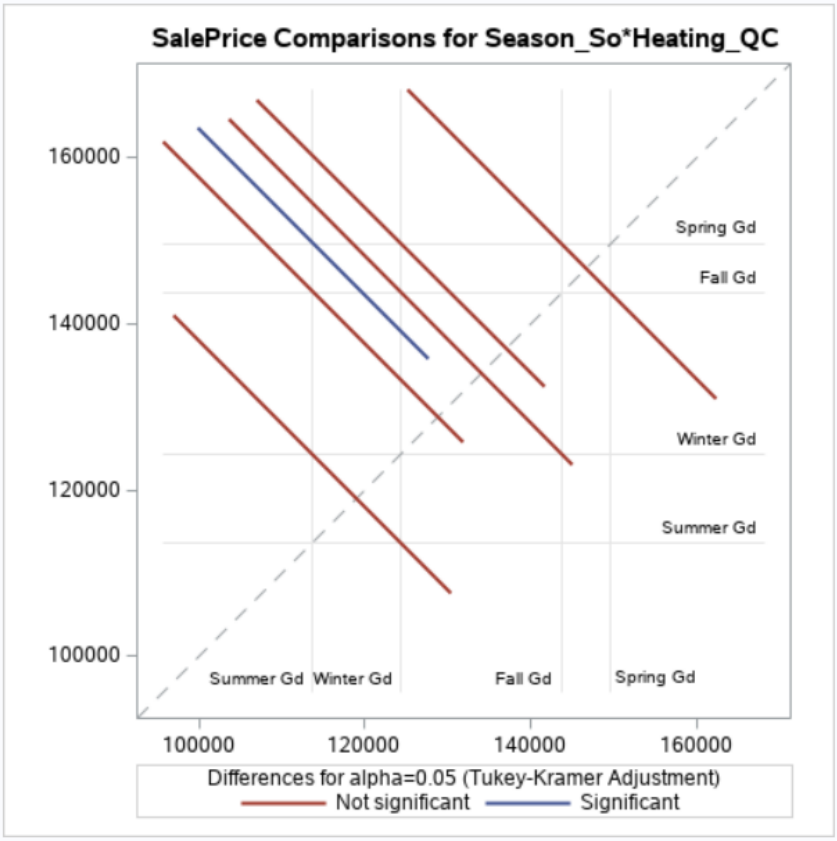
F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Fa	3	284	3.58	0.0143

Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer								
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC Fa	Winter	Spring	-40557	23366	284	-1.74	0.0837	0.3071
Heating_QC Fa	Winter	Summer	-70700	24728	284	-2.86	0.0046	0.0235
Heating_QC Fa	Winter	Fall	13100	39099	284	0.34	0.7378	0.9870
Heating_QC Fa	Spring	Summer	-30143	19827	284	-1.52	0.1295	0.4267
Heating_QC Fa	Spring	Fall	53657	36198	284	1.48	0.1394	0.4495
Heating_QC Fa	Summer	Fall	83800	37092	284	2.26	0.0246	0.1102



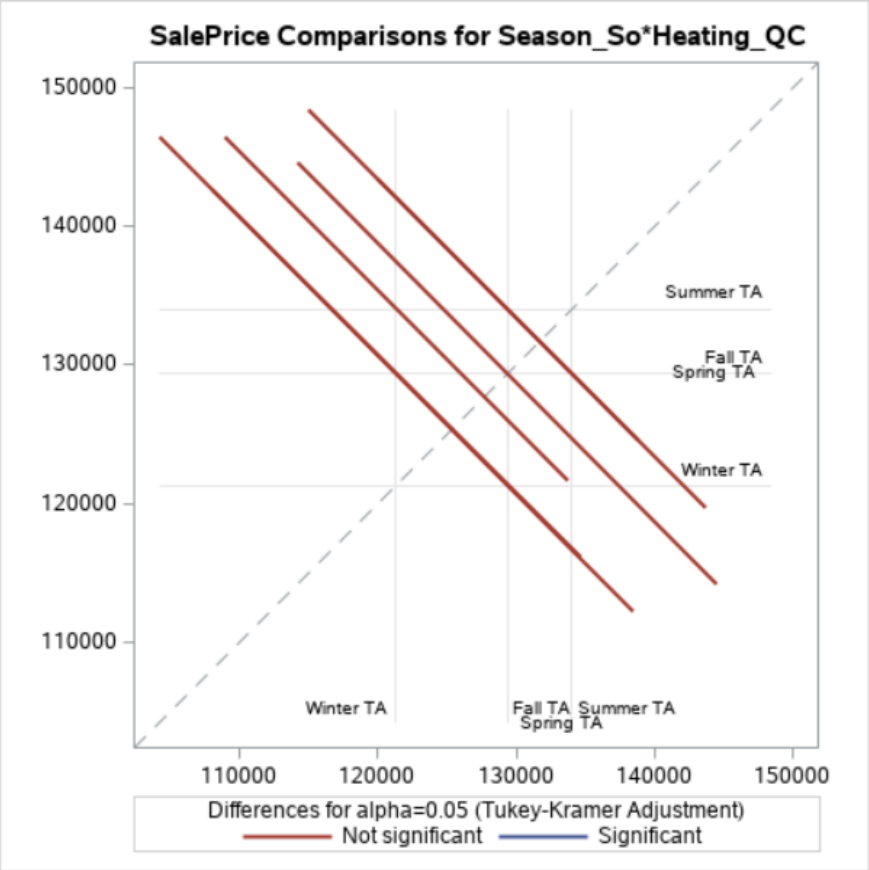
F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC Gd	3	284	4.23	0.0060

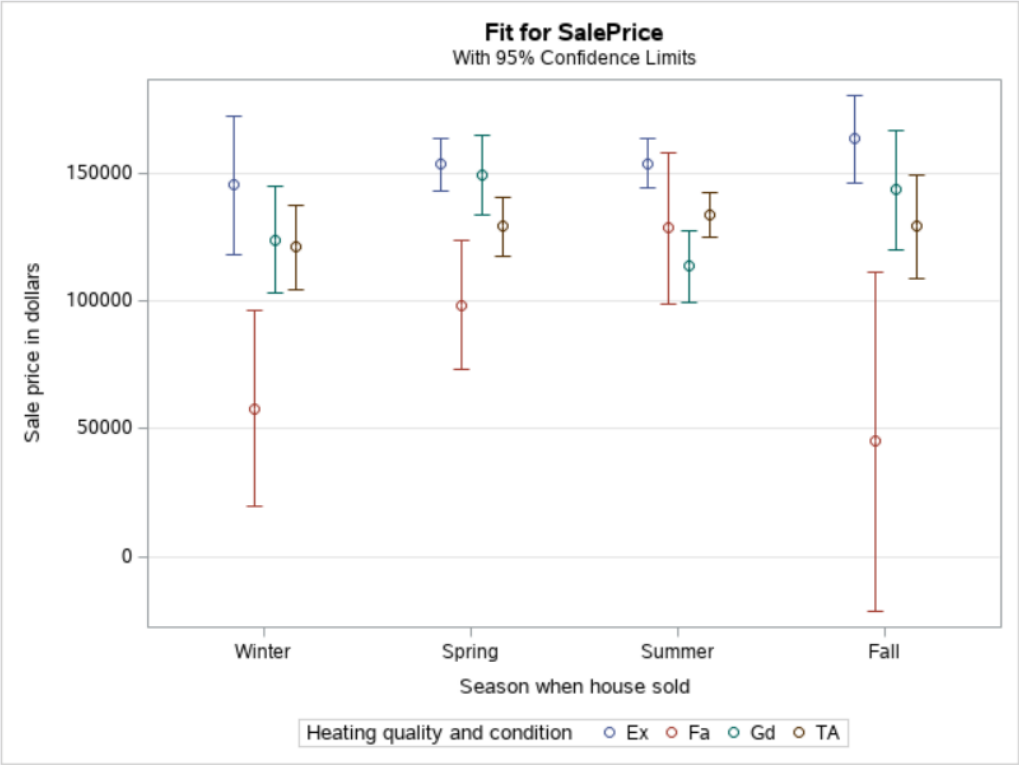
Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer								
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC Gd	Winter	Spring	-25290	13355	284	-1.89	0.0593	0.2330
Heating_QC Gd	Winter	Summer	10603	12914	284	0.82	0.4123	0.8445
Heating_QC Gd	Winter	Fall	-19483	16061	284	-1.21	0.2261	0.6191
Heating_QC Gd	Spring	Summer	35893	10762	284	3.34	0.0010	0.0053
Heating_QC Gd	Spring	Fall	5807.33	14388	284	0.40	0.6868	0.9777
Heating_QC Gd	Summer	Fall	-30085	13980	284	-2.15	0.0322	0.1394



F Test for Season_So*Heating_QC Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
Heating_QC TA	3	284	0.62	0.6021

Simple Differences of Season_So*Heating_QC Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer								
Slice	Season when house sold	Season when house sold	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P
Heating_QC TA	Winter	Spring	-8091.91	10265	284	-0.79	0.4312	0.8598
Heating_QC TA	Winter	Summer	-12734	9561.68	284	-1.33	0.1840	0.5434
Heating_QC TA	Winter	Fall	-8032.95	13262	284	-0.61	0.5452	0.9302
Heating_QC TA	Spring	Summer	-4642.14	7313.62	284	-0.63	0.5261	0.9207
Heating_QC TA	Spring	Fall	58.9572	11745	284	0.01	0.9960	1.0000
Heating_QC TA	Summer	Fall	4701.10	11135	284	0.42	0.6732	0.9747


SalePrice Comparisons for Season_So*Heating_QC
150000
140000
130000
120000
110000
Summer TA
Fall TA
Spring TA
Winter TA
Winter TA
Fall TA Summer TA
Spring TA
110000
120000
130000
140000
150000
Differences for alpha=0.05 (Tukey-Kramer Adjustment)
Not significant
Significant



11. Multiple Linear Regression Analysis

/* Model sale price as a function of lot area and basement area */

/* we start by using PROC REG */

```
ods graphics on;

proc reg data=STAT1.ameshousing3 ;
    model SalePrice=Basement_Area Lot_Area;
    title "Model with Basement Area and Lot Area";
run;
quit;
```

Model with Basement Area and Lot Area

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	300
Number of Observations Used	300

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.032206E11	1.016103E11	137.17	<.0001
Error	297	2.200029E11	740750509		
Corrected Total	299	4.232235E11			

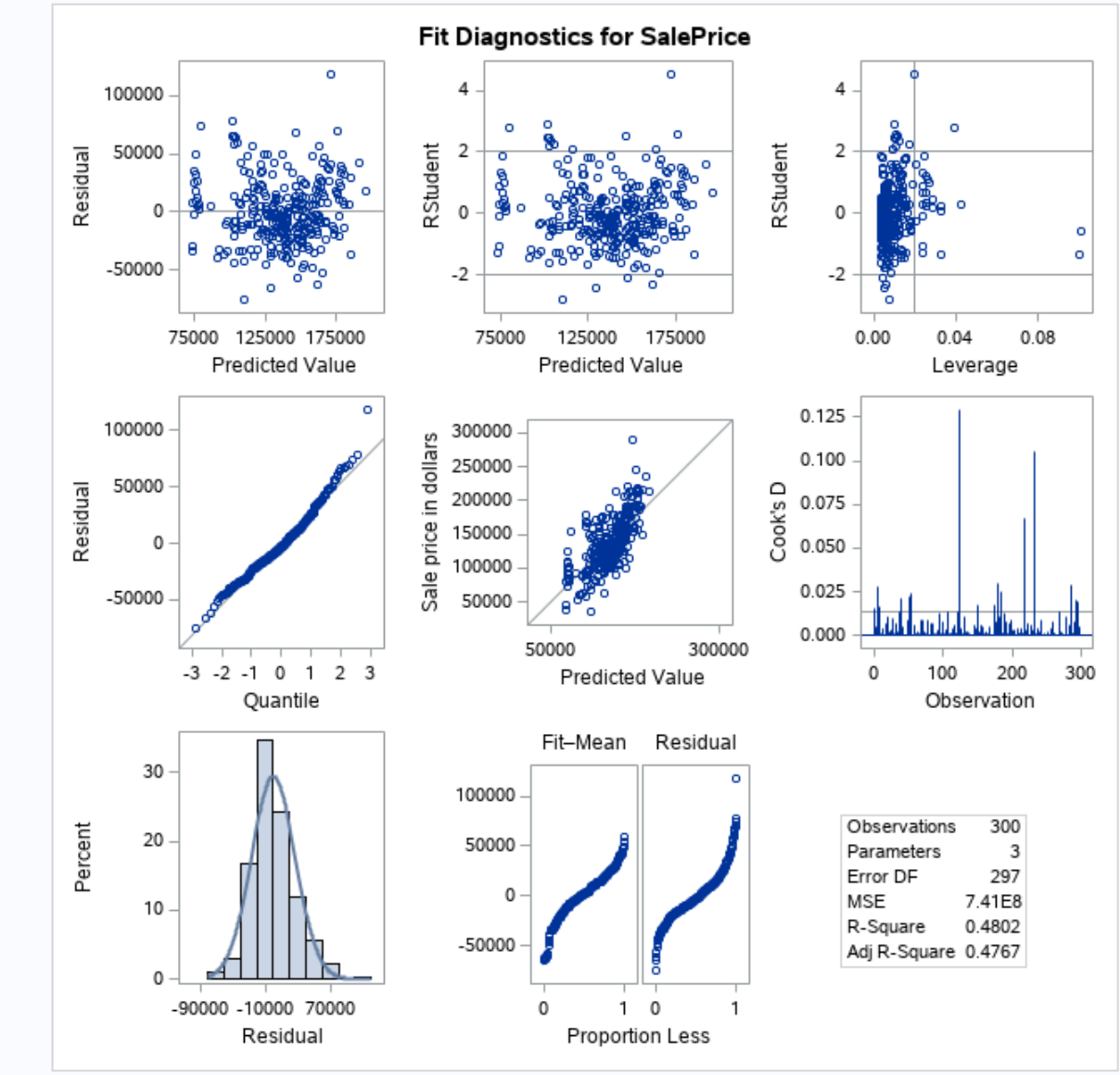
Root MSE	27217	R-Square	0.4802
Dependent Mean	137525	Adj R-Sq	0.4767
Coeff Var	19.79041		

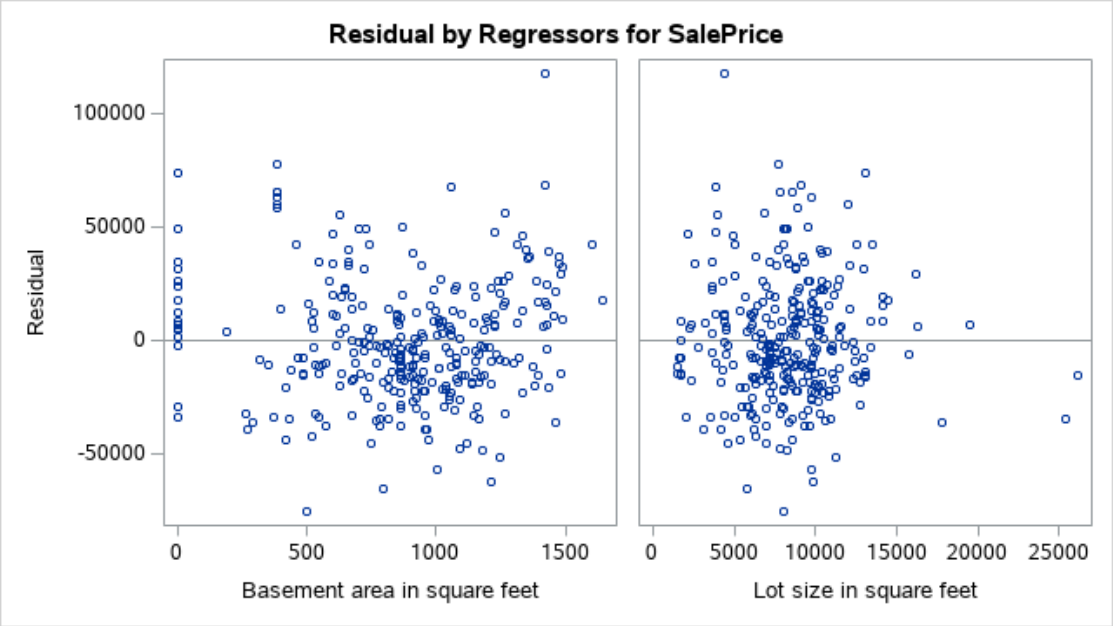
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	69016	5129.52179	13.45	<.0001
Basement_Area	Basement area in square feet	1	70.08680	4.54618	15.42	<.0001
Lot_Area	Lot size in square feet	1	0.80430	0.49210	1.63	0.1032

Model with Basement Area and Lot Area

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars


Fit Diagnostics for SalePrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Sale price in dollars
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 300
Parameters 3
Error DF 297
MSE 7.41E8
R-Square 0.4802
Adj R-Square 0.4767



/* next, we use PROC GLM to access additional features that are not available with PROC REG. We request a contour plot, and store the model output */

```
proc glm data=STAT1.ameshousing3
         plots(only)=(contourfit);
    model SalePrice=Basement_Area Lot_Area;
    store out=multiple;
    title "Model with Basement Area and Gross Living Area";
run;
quit;
```

Model with Basement Area and Gross Living Area

The GLM Procedure

Number of Observations Read	300
Number of Observations Used	300

Model with Basement Area and Gross Living Area

The GLM Procedure

Dependent Variable: SalePrice Sale price in dollars

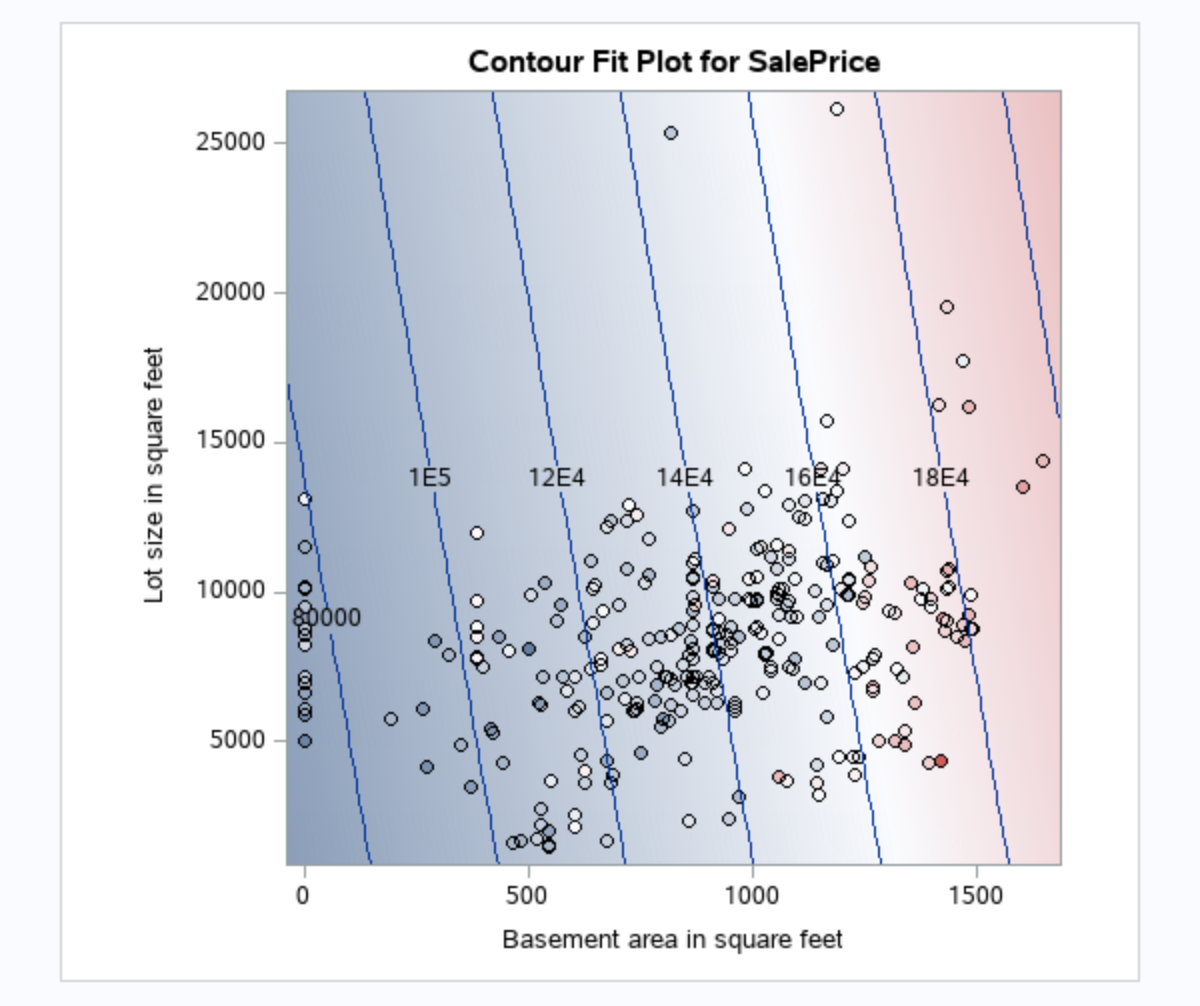
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	203220618262	101610309131	137.17	<.0001
Error	297	220002901249	740750509.26		
Corrected Total	299	423223519511			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.480173	19.79041	27216.73	137524.9

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Basement_Area	1	201241844480	201241844480	271.67	<.0001
Lot_Area	1	1978773781.7	1978773781.7	2.67	0.1032

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Basement_Area	1	176055907089	176055907089	237.67	<.0001
Lot_Area	1	1978773781.7	1978773781.7	2.67	0.1032

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	69015.61360	5129.521790	13.45	<.0001
Basement_Area	70.08680	4.546183	15.42	<.0001
Lot_Area	0.80430	0.492102	1.63	0.1032

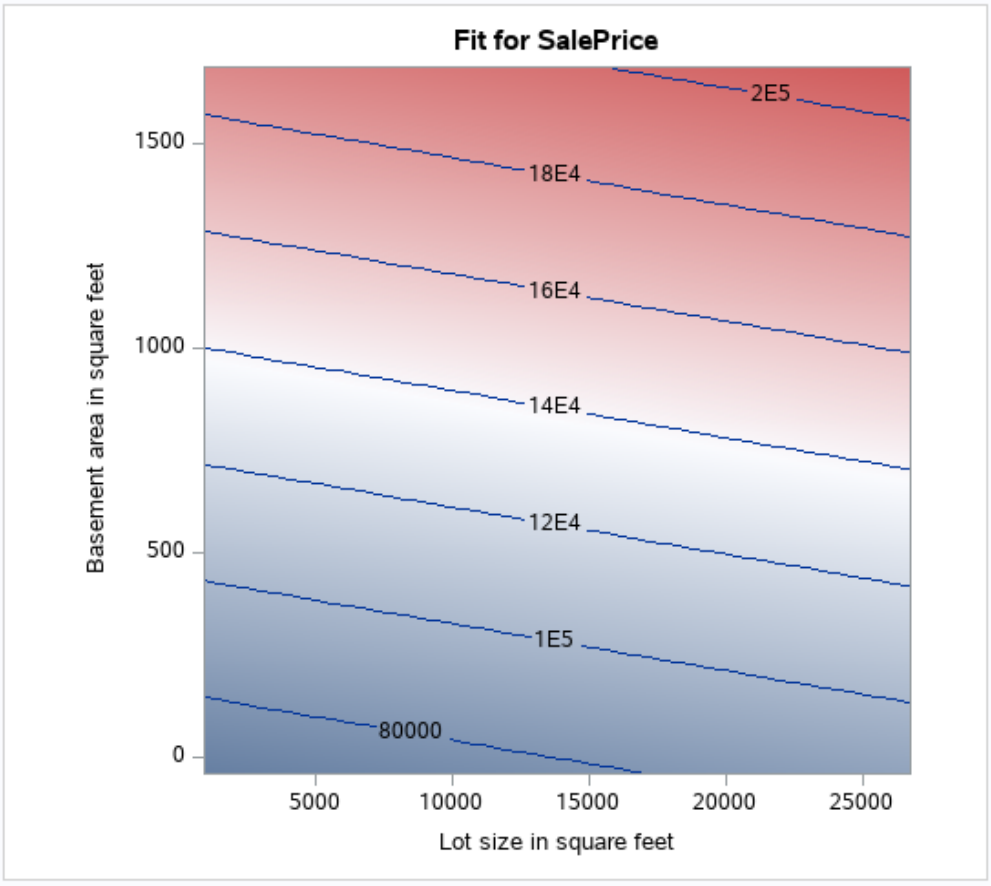


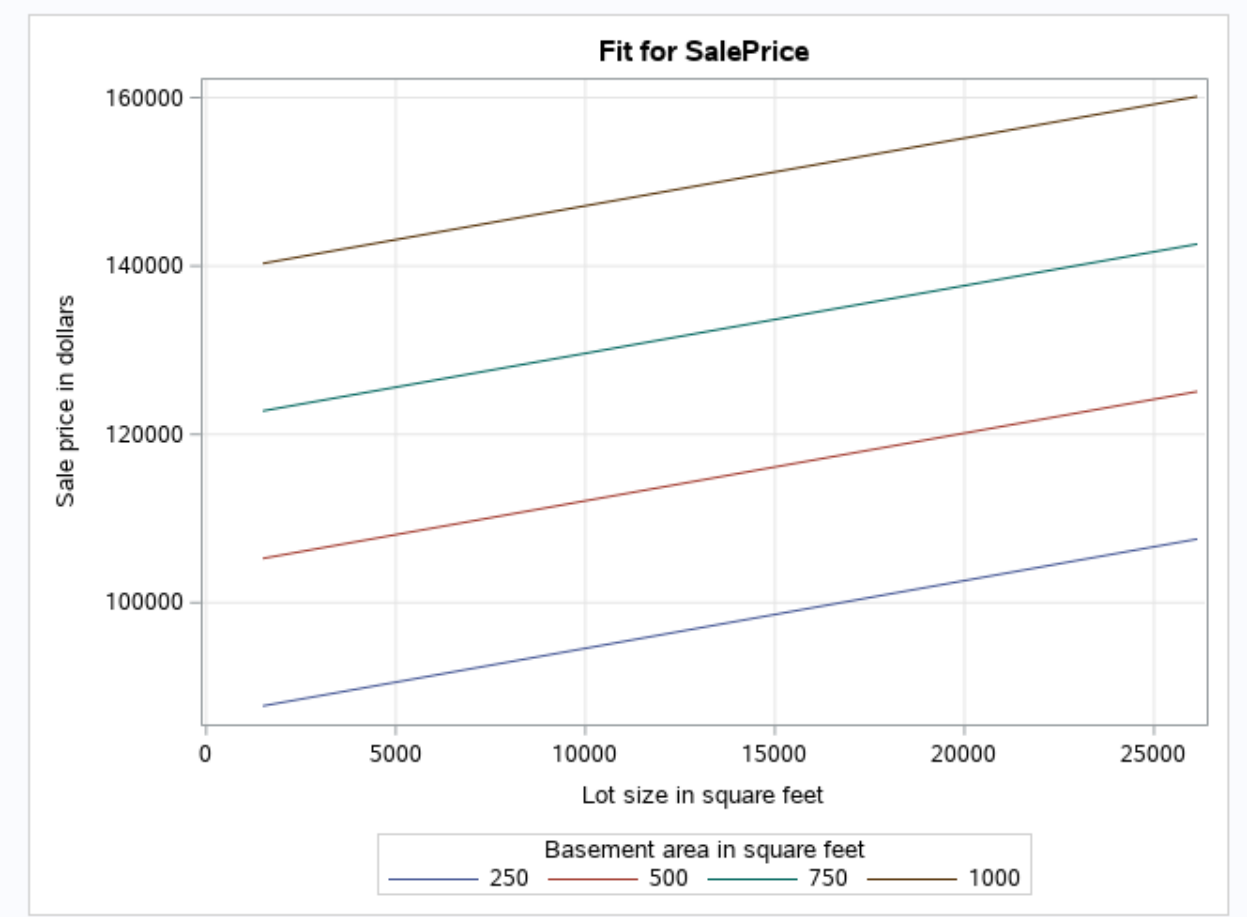
/* we use PROC PLM to further process the item created by PROC GLM */

```
proc plm restore=multiple plots=all;
    effectplot contour (y=Basement_Area x=Lot_Area);
    effectplot slicefit(x=Lot_Area sliceby=Basement_Area=250 to
1000 by 250);
run;

title;
```

The PLM Procedure

Store Information	
Item Store	WORK.MULTIPLE
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLM
Date Created	21AUG21:05:55:25
Response Variable	SalePrice
Model Effects	Intercept Basement_Area Lot_Area





12. Multiple Regression Analysis: With Variable Selection

/* Trying stepwise selection of variables */

```
ods graphics on;
proc glmselect data=STAT1.ameshousing3 plots=all;
      STEPWISE: model SalePrice = &interval / selection=stepwise
details=steps select=SL slstay=0.05 slentry=0.05;
      title "Stepwise Model Selection for SalePrice - SL 0.05";
run;
```

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	300
Number of Observations Used	300

Dimensions	
Number of Effects	9
Number of Parameters	9

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 0

Effect Entered: Intercept

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	299	4.232235E11	1415463276	
Corrected Total	299	4.232235E11		

Root MSE	37623
Dependent Mean	137525
R-Square	0.0000
Adj R-Sq	0.0000
AIC	6624.21515
AICC	6624.25555
SBC	6325.91893

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	137525	2172.144314	63.31

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 1

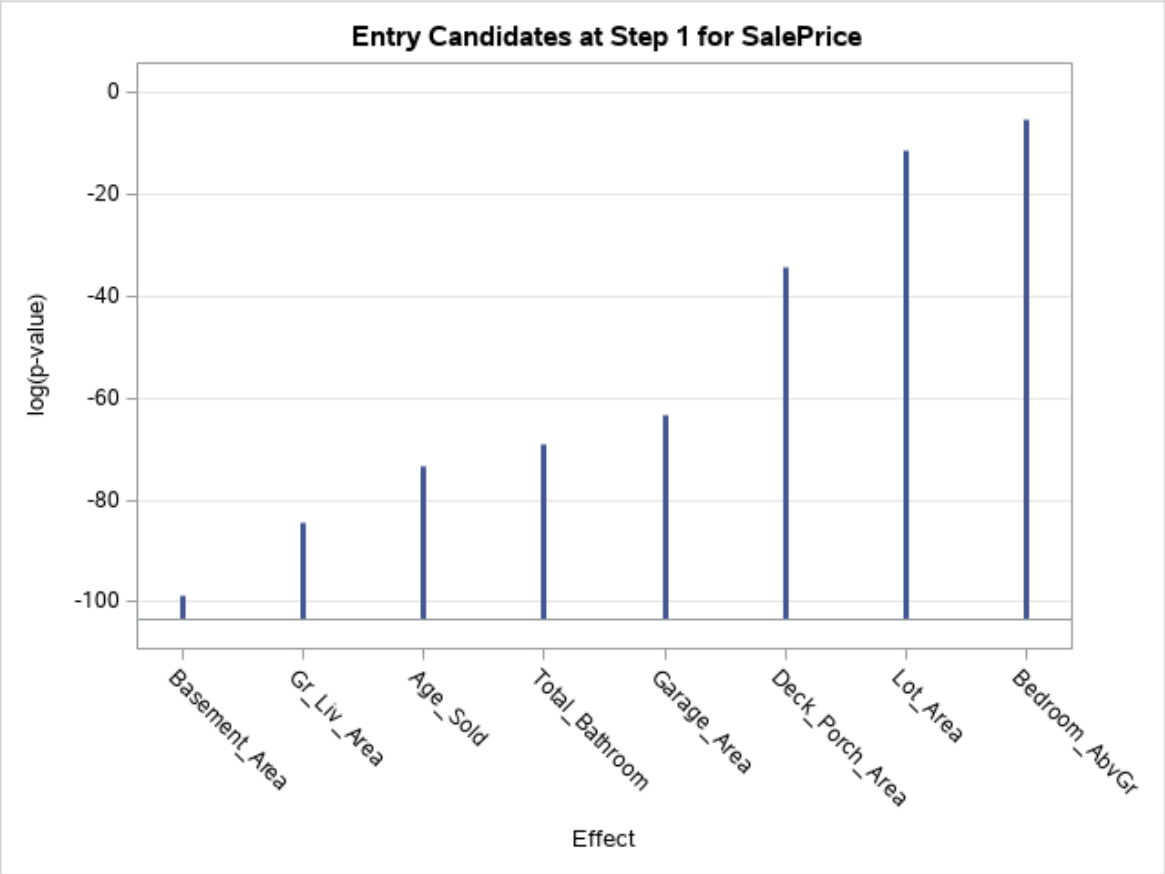
Effect Entered: Basement_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	2.012418E11	2.012418E11	270.16
Error	298	2.219817E11	744904950	
Corrected Total	299	4.232235E11		

Root MSE	27293
Dependent Mean	137525
R-Square	0.4755
Adj R-Sq	0.4737
AIC	6432.62346
AICC	6432.70454
SBC	6138.03102

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	73904	4179.193780	17.68
Basement_Area	1	72.107717	4.387055	16.44

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Basement_Area	-98.8577	<.0001
2	Gr_Liv_Area	-84.6132	<.0001
3	Age_Sold	-73.5219	<.0001
4	Total_Bathroom	-69.1880	<.0001
5	Garage_Area	-63.3558	<.0001
6	Deck_Porch_Area	-34.3105	<.0001
7	Lot_Area	-11.6303	<.0001
8	Bedroom_AbvGr	-5.5339	0.0040



Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 2

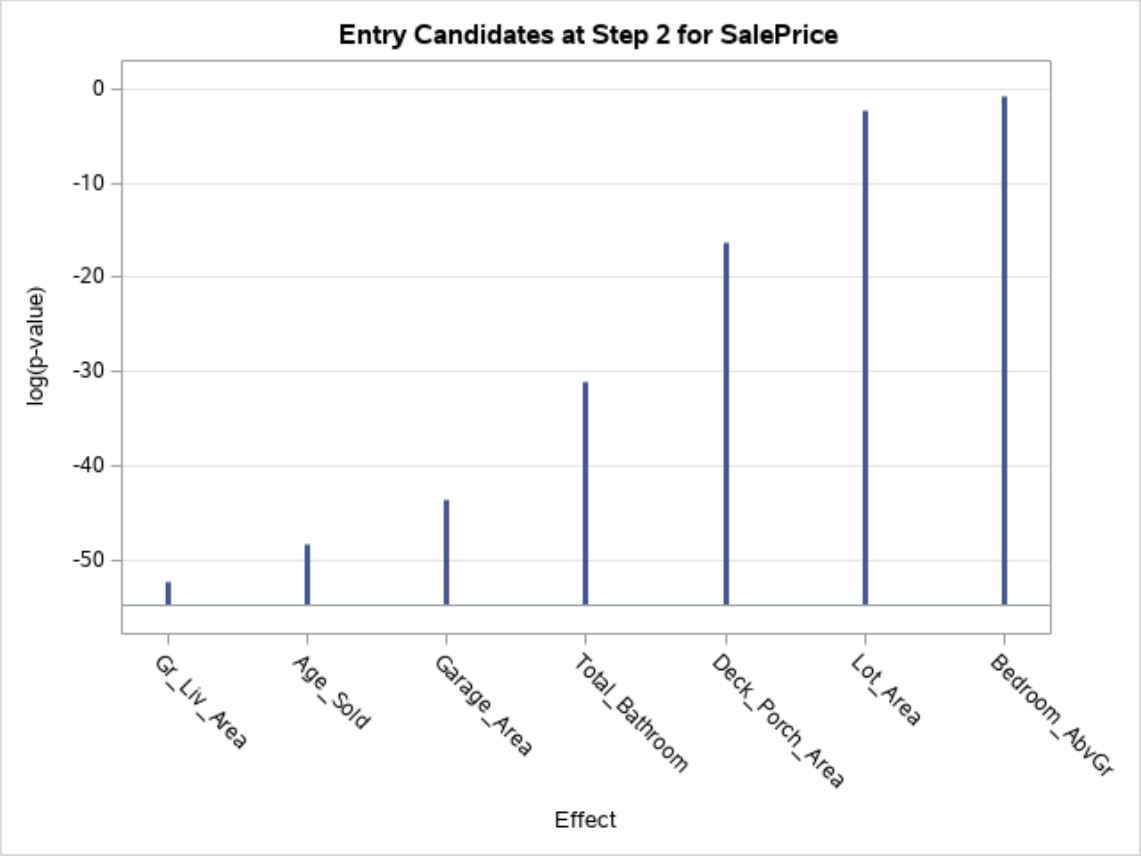
Effect Entered: Gr_Liv_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	2.64483E11	1.322415E11	247.42
Error	297	1.587405E11	534479711	
Corrected Total	299	4.232235E11		

Root MSE	23119
Dependent Mean	137525
R-Square	0.6249
Adj R-Sq	0.6224
AIC	6334.02620
AICC	6334.16179
SBC	6043.13755

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	12664	6650.339855	1.90
Gr_Liv_Area	1	69.606974	6.399091	10.88
Basement_Area	1	52.309702	4.137885	12.64

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Gr_Liv_Area	-52.2496	<.0001
2	Age_Sold	-48.2636	<.0001
3	Garage_Area	-43.6174	<.0001
4	Total_Bathroom	-31.0375	<.0001
5	Deck_Porch_Area	-16.3568	<.0001
6	Lot_Area	-2.2708	0.1032
7	Bedroom_AbvGr	-0.7570	0.4691


Entry Candidates at Step 2 for SalePrice
0
-10
-20
-30
-40
-50
log(p-value)
Gr_Liv_Area
Age_Sold
Garage_Area
Total_Bathroom
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Effect

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 3

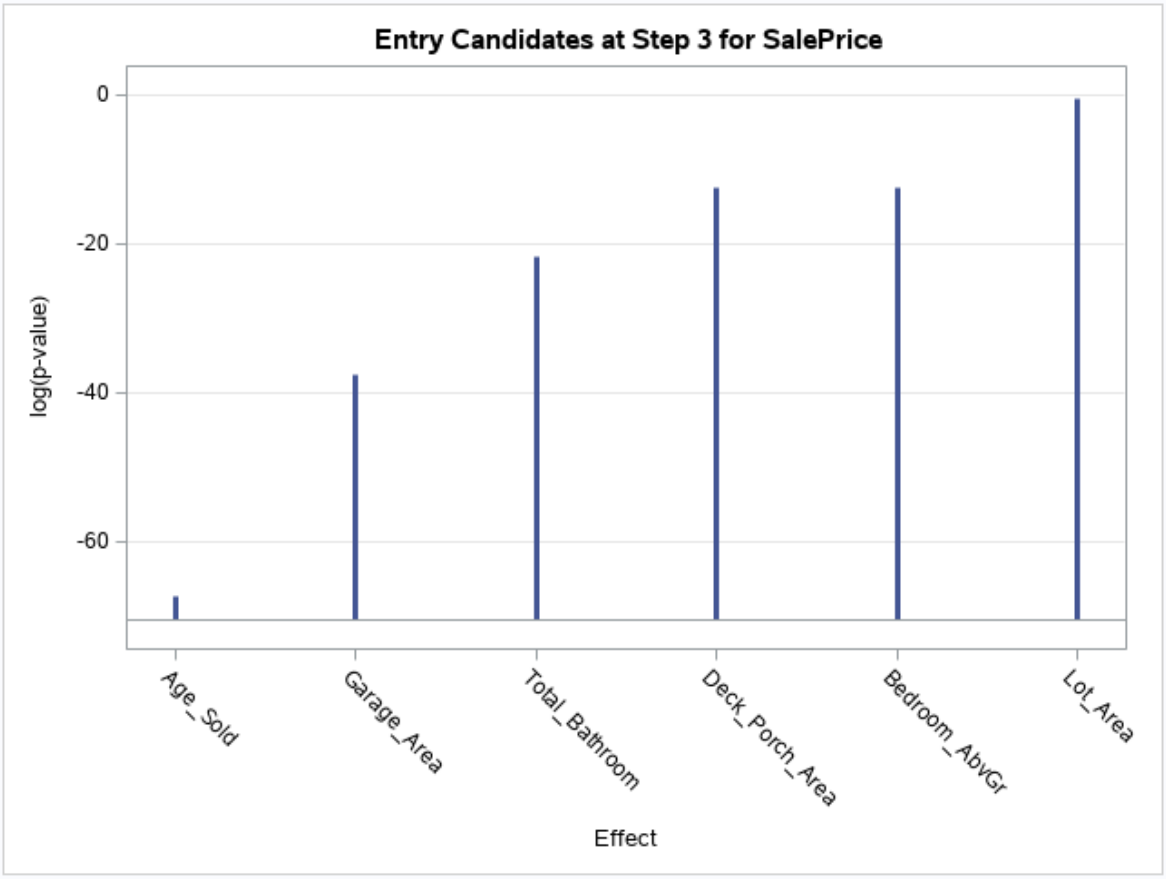
Effect Entered: Age_Sold

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	3.207148E11	1.069049E11	308.69
Error	296	1.025087E11	346313132	
Corrected Total	299	4.232235E11		

Root MSE	18609
Dependent Mean	137525
R-Square	0.7578
Adj R-Sq	0.7553
AIC	6204.82927
AICC	6205.03335
SBC	5917.64440

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	53400	6235.076995	8.56
Gr_Liv_Area	1	68.106646	5.152294	13.22
Basement_Area	1	36.329120	3.559067	10.21
Age_Sold	1	-543.493346	42.651840	-12.74

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Age_Sold	-67.2828	<.0001
2	Garage_Area	-37.5122	<.0001
3	Total_Bathroom	-21.6266	<.0001
4	Deck_Porch_Area	-12.5097	<.0001
5	Bedroom_AbvGr	-12.4446	<.0001



Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 4

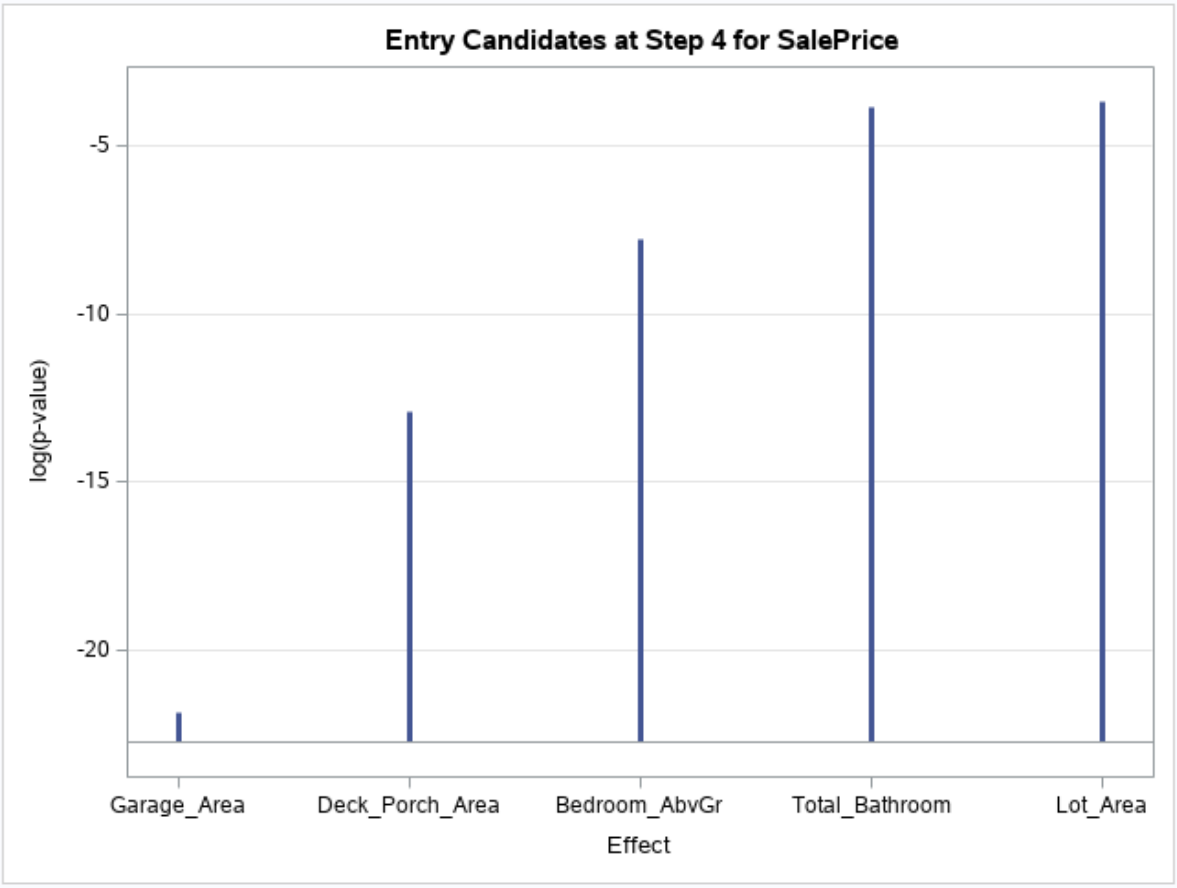
Effect Entered: Garage_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	4	3.33571E11	83392754480	274.40
Error	295	89652501590	303906785	
Corrected Total	299	4.232235E11		

Root MSE	17433
Dependent Mean	137525
R-Square	0.7882
Adj R-Sq	0.7853
AIC	6166.62734
AICC	6166.91403
SBC	5883.14625

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	43815	6023.907004	7.27
Gr_Liv_Area	1	61.238136	4.940722	12.39
Basement_Area	1	33.430181	3.363709	9.94
Garage_Area	1	42.984492	6.608851	6.50
Age_Sold	1	-455.704354	42.173481	-10.81

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Garage_Area	-21.8203	<.0001
2	Deck_Porch_Area	-12.9294	<.0001
3	Bedroom_AbvGr	-7.8057	0.0004
4	Total_Bathroom	-3.8856	0.0205
5	Lot_Area	-3.6980	0.0248



Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 5

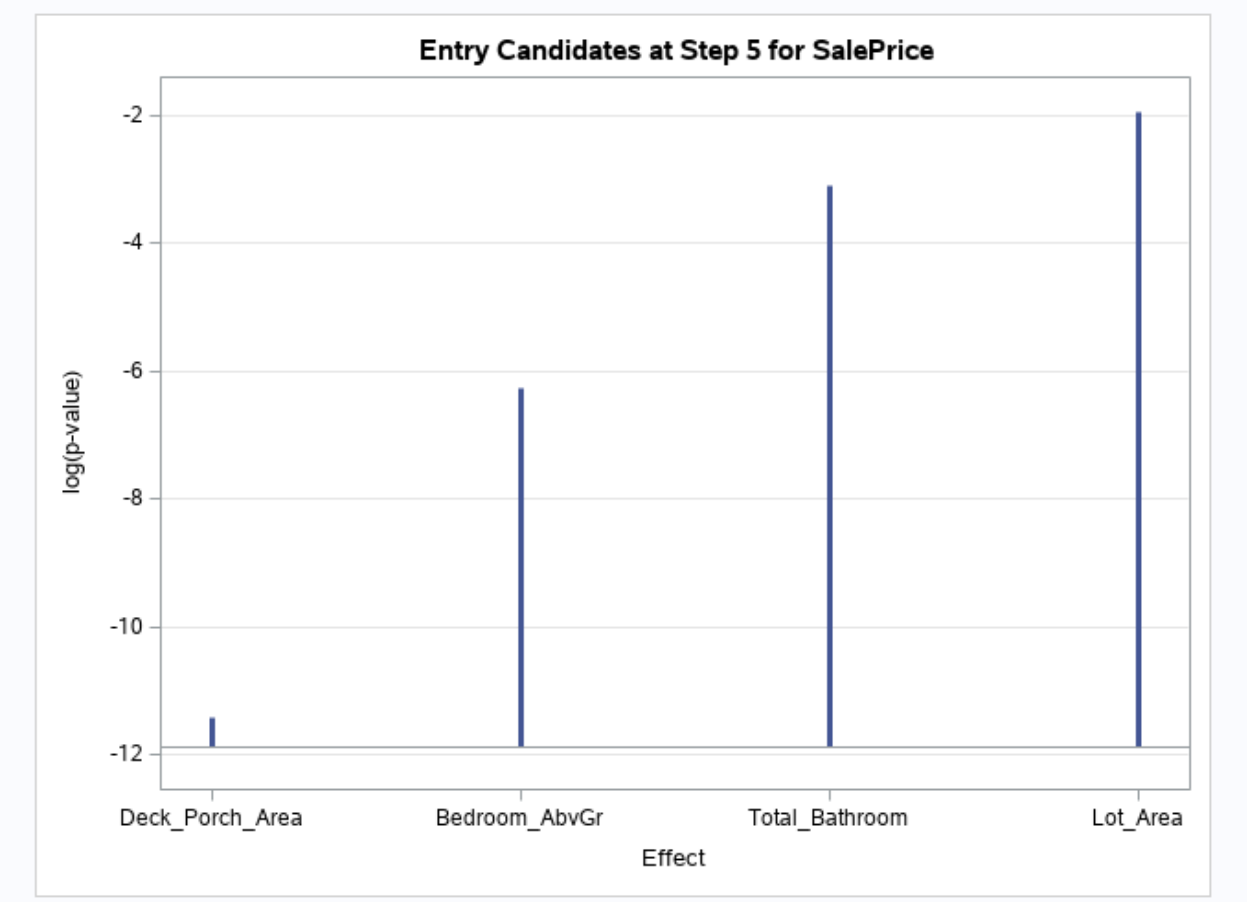
Effect Entered: Deck_Porch_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	5	3.392788E11	67855752389	237.65
Error	294	83944757568	285526386	
Corrected Total	299	4.232235E11		

Root MSE	16898
Dependent Mean	137525
R-Square	0.8017
Adj R-Sq	0.7983
AIC	6148.89269
AICC	6149.27625
SBC	5869.11538

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	46009	5859.485517	7.85
Gr_Liv_Area	1	58.386514	4.831268	12.09
Basement_Area	1	30.554240	3.323249	9.19
Garage_Area	1	40.158112	6.436997	6.24
Deck_Porch_Area	1	35.720258	7.989240	4.47
Age_Sold	1	-447.254040	40.921927	-10.93

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Deck_Porch_Area	-11.4060	<.0001
2	Bedroom_AbvGr	-6.2737	0.0019
3	Total_Bathroom	-3.1045	0.0448
4	Lot_Area	-1.9476	0.1426


Entry Candidates at Step 5 for SalePrice
-2
-4
-6
-8
-10
-12
log(p-value)
Deck_Porch_Area
Bedroom_AbvGr
Total_Bathroom
Lot_Area
Effect

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 6

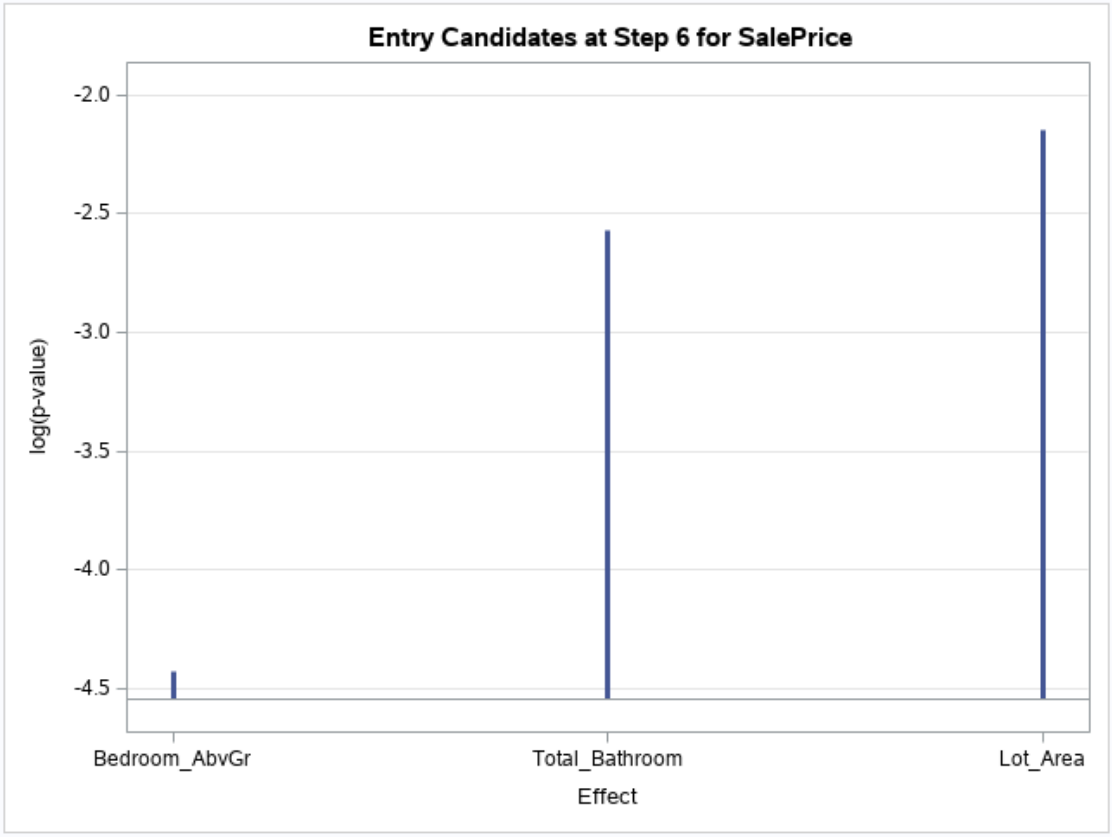
Effect Entered: Bedroom_AbvGr

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	6	3.410749E11	56845818595	202.75
Error	293	82148607939	280370676	
Corrected Total	299	4.232235E11		

Root MSE	16744
Dependent Mean	137525
R-Square	0.8059
Adj R-Sq	0.8019
AIC	6144.40398
AICC	6144.89882
SBC	5868.33046

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	48620	5897.324643	8.24
Gr_Liv_Area	1	65.097413	5.472624	11.90
Basement_Area	1	31.279351	3.305546	9.46
Garage_Area	1	38.728785	6.403565	6.05
Deck_Porch_Area	1	32.487956	8.019119	4.05
Age_Sold	1	-434.199118	40.877494	-10.62
Bedroom_AbvGr	1	-4189.095026	1655.065743	-2.53

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Bedroom_AbvGr	-4.4317	0.0119
2	Total_Bathroom	-2.5664	0.0768
3	Lot_Area	-2.1476	0.1168



Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 7

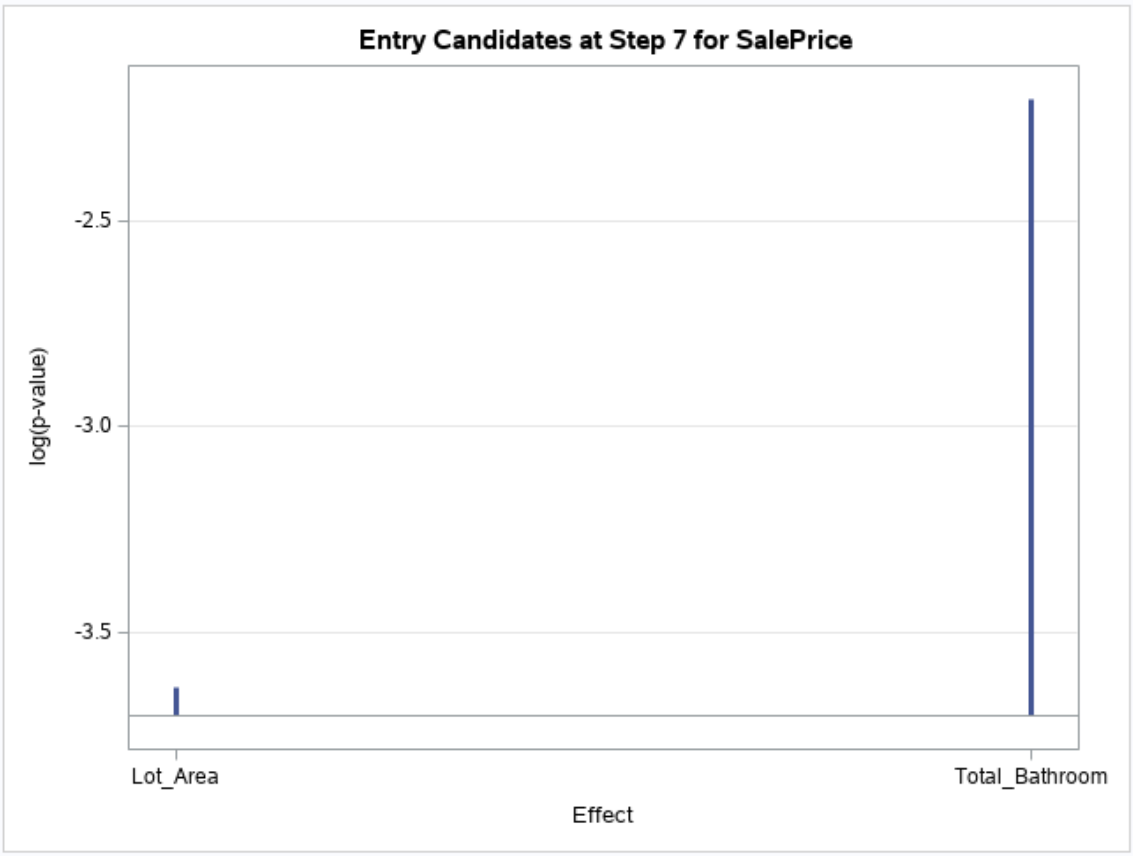
Effect Entered: Lot_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Lot_Area	-3.6309	0.0265
2	Total_Bathroom	-2.2036	0.1104



Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure

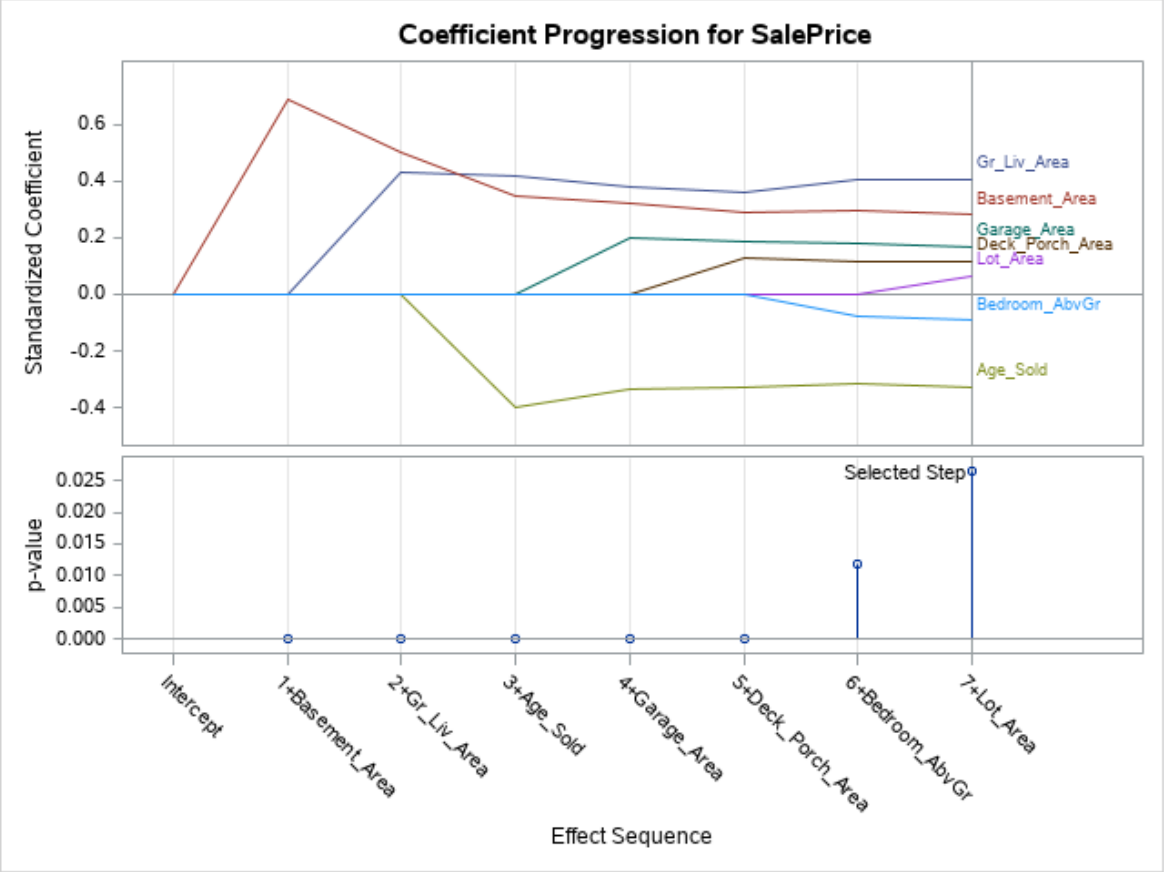
Stepwise Selection Summary

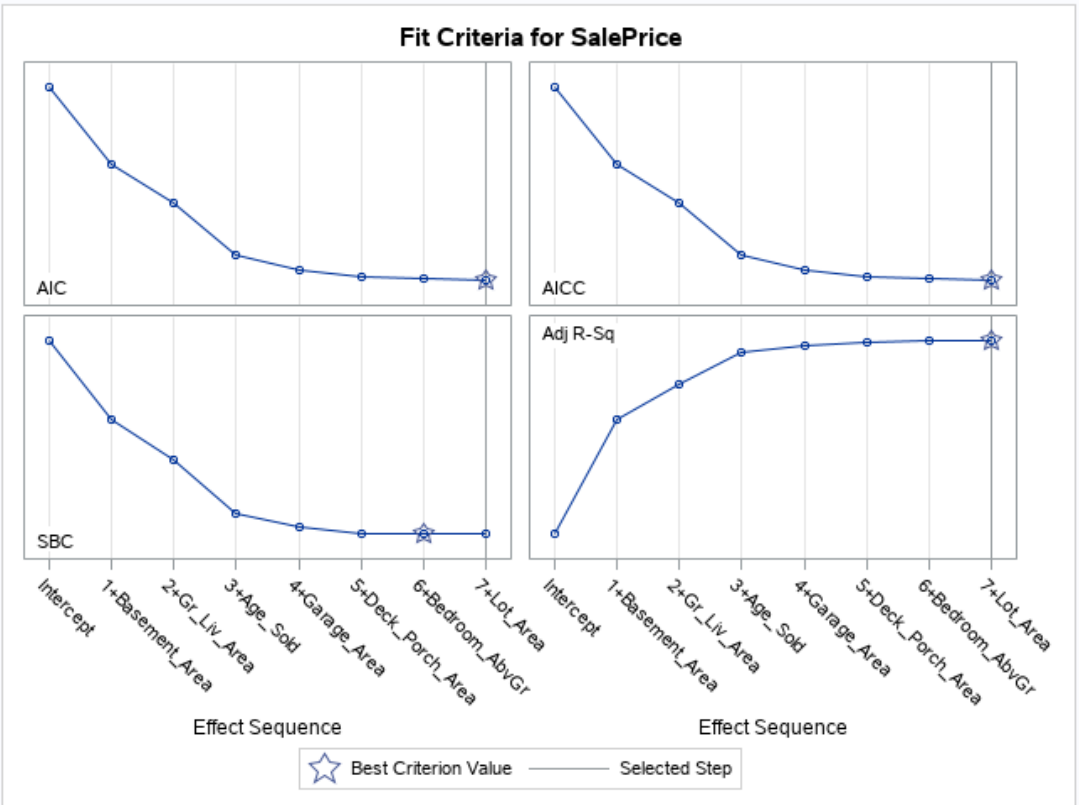
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	Basement_Area		2	270.16	<.0001
2	Gr_Liv_Area		3	118.32	<.0001
3	Age_Sold		4	162.37	<.0001
4	Garage_Area		5	42.30	<.0001
5	Deck_Porch_Area		6	19.99	<.0001
6	Bedroom_AbvGr		7	6.41	0.0119
7	Lot_Area		8	4.97	0.0265

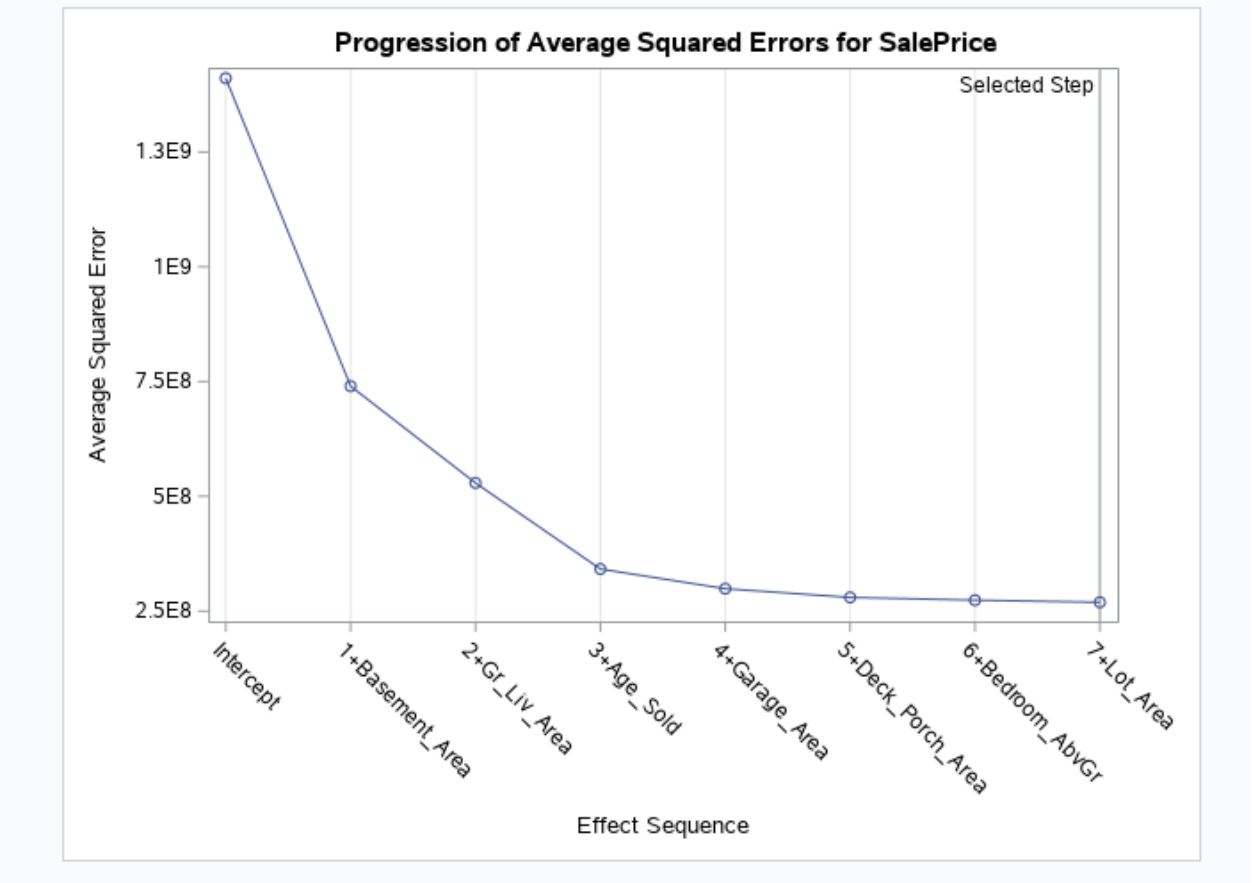
Selection stopped because the candidate for entry has SLE > 0.05 and the candidate for removal has SLS < 0.05.

Stop Details

Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	Total_Bathroom	0.1167	>	0.0500	(SLE)
Removal	Lot_Area	0.0265	<	0.0500	(SLS)


Coefficient Progression for SalePrice
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
-0.4
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Age_Sold
p-value
0.025
0.020
0.015
0.010
0.005
0.000
Selected Step
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence


Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for SalePrice
Selected Step
Average Squared Error
1.3E9
1E9
7.5E8
5E8
2.5E8
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 7).

Effects:	Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

/*Optional Code that will execute forward and backward selection
Each with slentry and slstay = 0.05. */

```
proc glmselect data=STAT1.ameshousing3 plots=all;
      FORWARD: model SalePrice = &interval / selection=forward
details=steps select=SL slentry=0.05;
      title "Forward Model Selection for SalePrice - SL 0.05";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
      BACKWARD: model SalePrice = &interval / selection=backward
details=steps select=SL slstay=0.05;
      title "Backward Model Selection for SalePrice - SL 0.05";
```

```
run;
```

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Forward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	300
Number of Observations Used	300

Dimensions	
Number of Effects	9
Number of Parameters	9

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 0

Effect Entered: Intercept

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	299	4.232235E11	1415463276	
Corrected Total	299	4.232235E11		

Root MSE	37623
Dependent Mean	137525
R-Square	0.0000
Adj R-Sq	0.0000
AIC	6624.21515
AICC	6624.25555
SBC	6325.91893

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	137525	2172.144314	63.31

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 1

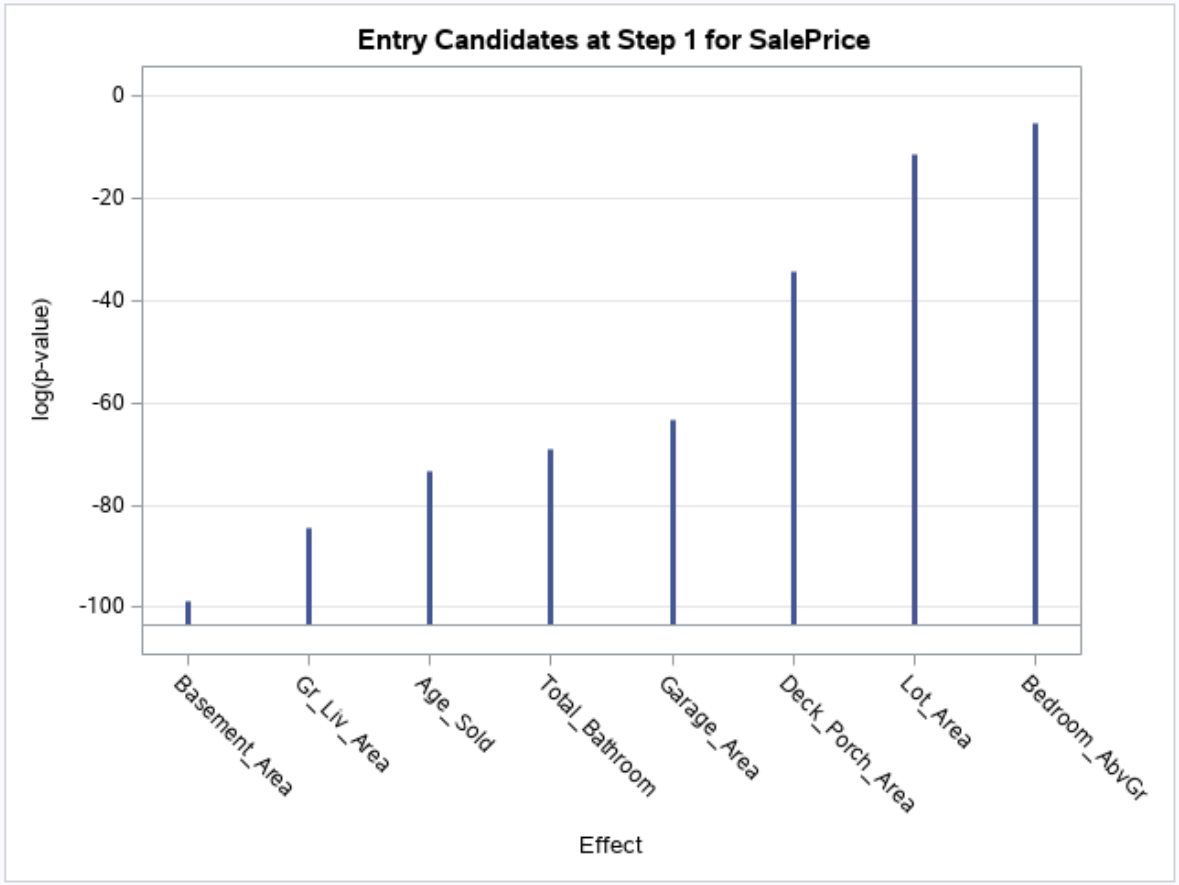
Effect Entered: Basement_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	2.012418E11	2.012418E11	270.16
Error	298	2.219817E11	744904950	
Corrected Total	299	4.232235E11		

Root MSE	27293
Dependent Mean	137525
R-Square	0.4755
Adj R-Sq	0.4737
AIC	6432.62346
AICC	6432.70454
SBC	6138.03102

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	73904	4179.193780	17.68
Basement_Area	1	72.107717	4.387055	16.44

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Basement_Area	-98.8577	<.0001
2	Gr_Liv_Area	-84.6132	<.0001
3	Age_Sold	-73.5219	<.0001
4	Total_Bathroom	-69.1880	<.0001
5	Garage_Area	-63.3558	<.0001
6	Deck_Porch_Area	-34.3105	<.0001
7	Lot_Area	-11.6303	<.0001
8	Bedroom_AbvGr	-5.5339	0.0040



Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 2

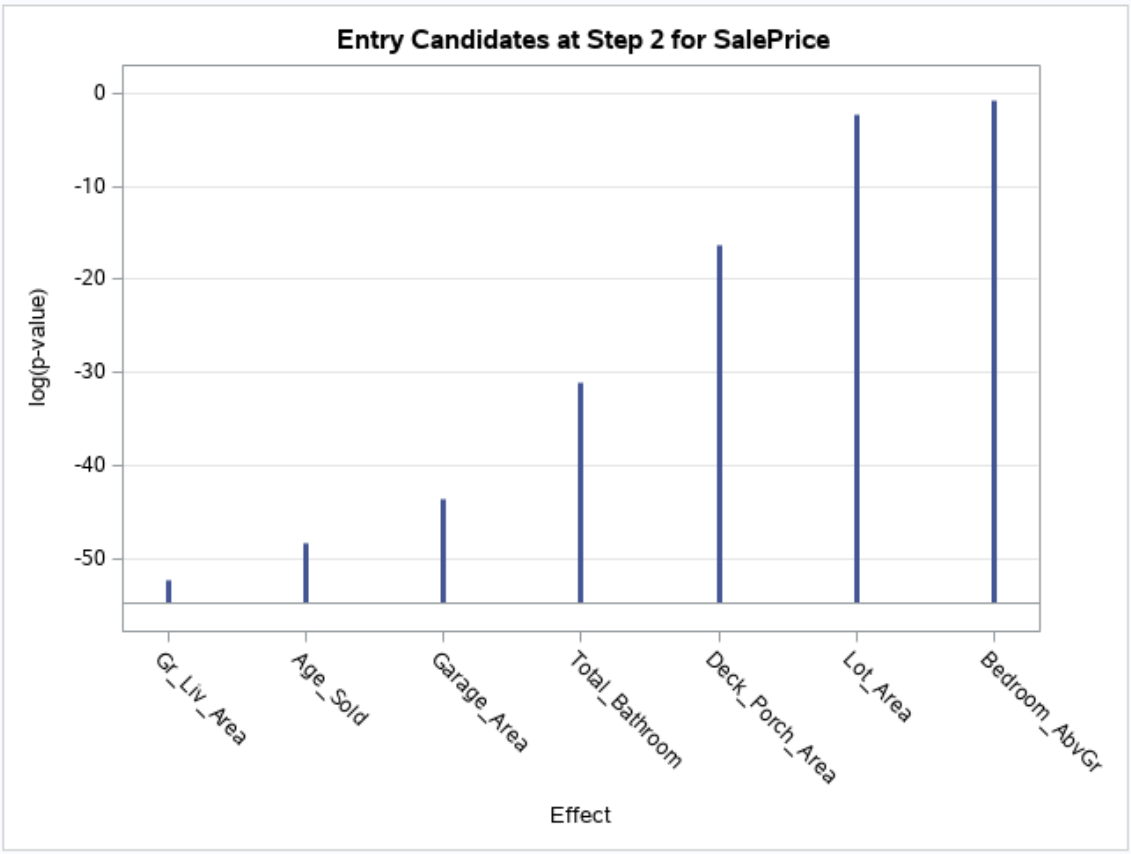
Effect Entered: Gr_Liv_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	2.64483E11	1.322415E11	247.42
Error	297	1.587405E11	534479711	
Corrected Total	299	4.232235E11		

Root MSE	23119
Dependent Mean	137525
R-Square	0.6249
Adj R-Sq	0.6224
AIC	6334.02620
AICC	6334.16179
SBC	6043.13755

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	12664	6650.339855	1.90
Gr_Liv_Area	1	69.606974	6.399091	10.88
Basement_Area	1	52.309702	4.137885	12.64

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Gr_Liv_Area	-52.2496	<.0001
2	Age_Sold	-48.2636	<.0001
3	Garage_Area	-43.6174	<.0001
4	Total_Bathroom	-31.0375	<.0001
5	Deck_Porch_Area	-16.3568	<.0001
6	Lot_Area	-2.2708	0.1032
7	Bedroom_AbvGr	-0.7570	0.4691


Entry Candidates at Step 2 for SalePrice
0
-10
-20
-30
-40
-50
log(p-value)
Gr_Liv_Area
Age_Sold
Garage_Area
Total_Bathroom
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Effect

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 3

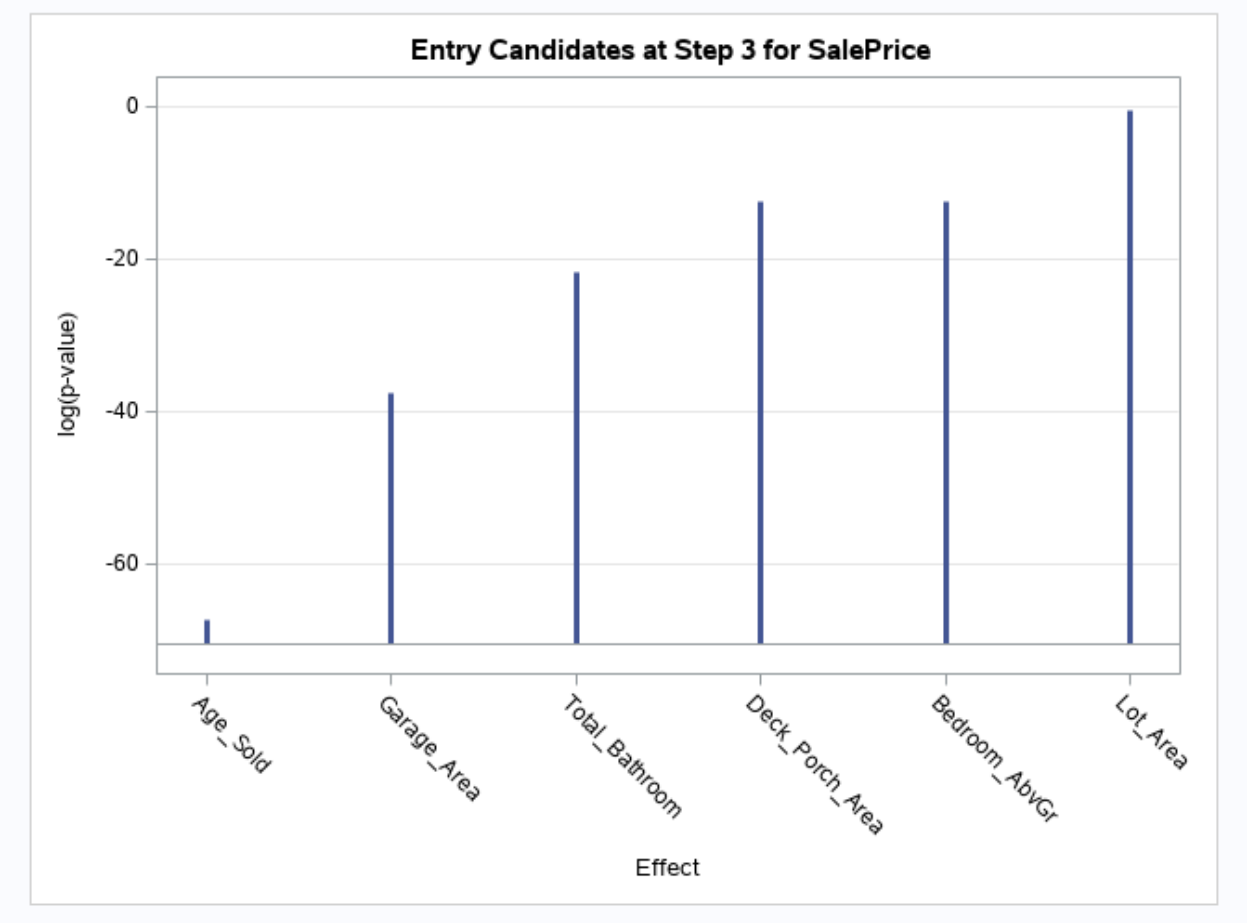
Effect Entered: Age_Sold

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	3.207148E11	1.069049E11	308.69
Error	296	1.025087E11	346313132	
Corrected Total	299	4.232235E11		

Root MSE	18609
Dependent Mean	137525
R-Square	0.7578
Adj R-Sq	0.7553
AIC	6204.82927
AICC	6205.03335
SBC	5917.64440

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	53400	6235.076995	8.56
Gr_Liv_Area	1	68.106646	5.152294	13.22
Basement_Area	1	36.329120	3.559067	10.21
Age_Sold	1	-543.493346	42.651840	-12.74

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Age_Sold	-67.2828	<.0001
2	Garage_Area	-37.5122	<.0001
3	Total_Bathroom	-21.6266	<.0001
4	Deck_Porch_Area	-12.5097	<.0001
5	Bedroom_AbvGr	-12.4446	<.0001
6	Lot_Area	-0.4524	0.6361


Entry Candidates at Step 3 for SalePrice
0
-20
-40
-60
log(p-value)
Age_Sold
Garage_Area
Total_Bathroom
Deck_Porch_Area
Bedroom_AbvGr
Lot_Area
Effect

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 4

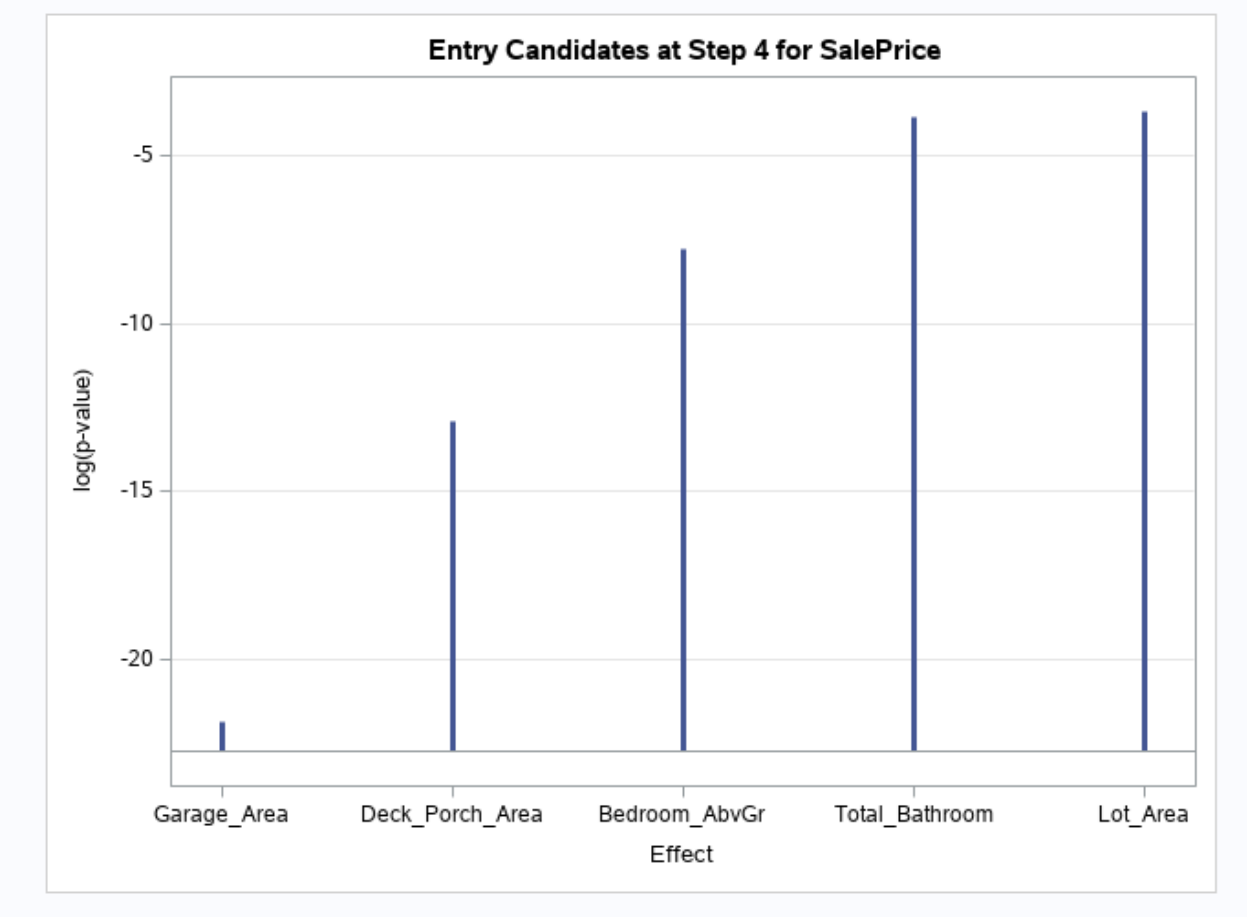
Effect Entered: Garage_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	4	3.33571E11	83392754480	274.40
Error	295	89652501590	303906785	
Corrected Total	299	4.232235E11		

Root MSE	17433
Dependent Mean	137525
R-Square	0.7882
Adj R-Sq	0.7853
AIC	6166.62734
AICC	6166.91403
SBC	5883.14625

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	43815	6023.907004	7.27
Gr_Liv_Area	1	61.238136	4.940722	12.39
Basement_Area	1	33.430181	3.363709	9.94
Garage_Area	1	42.984492	6.608851	6.50
Age_Sold	1	-455.704354	42.173481	-10.81

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Garage_Area	-21.8203	<.0001
2	Deck_Porch_Area	-12.9294	<.0001
3	Bedroom_AbvGr	-7.8057	0.0004
4	Total_Bathroom	-3.8856	0.0205
5	Lot_Area	-3.6980	0.0248


Entry Candidates at Step 4 for SalePrice
-5
-10
-15
-20
log(p-value)
Garage_Area
Deck_Porch_Area
Bedroom_AbvGr
Total_Bathroom
Lot_Area
Effect

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 5

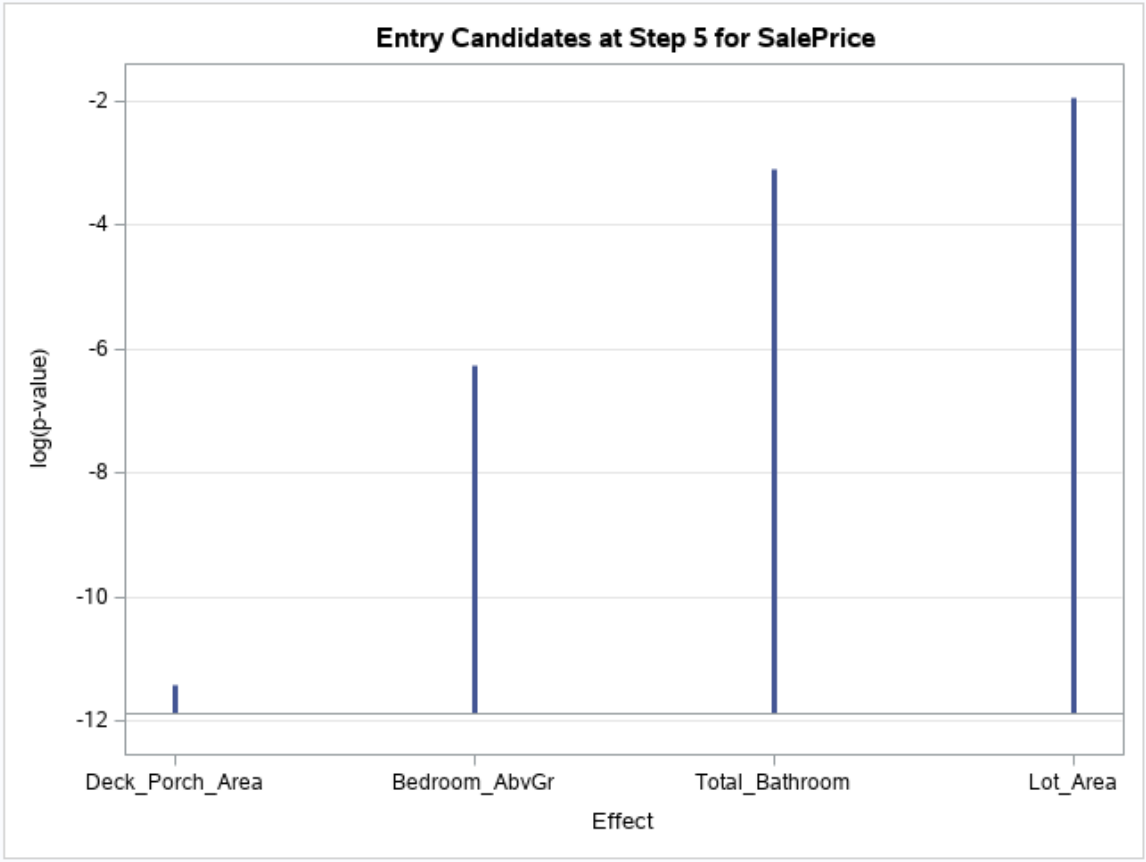
Effect Entered: Deck_Porch_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	5	3.392788E11	67855752389	237.65
Error	294	83944757568	285526386	
Corrected Total	299	4.232235E11		

Root MSE	16898
Dependent Mean	137525
R-Square	0.8017
Adj R-Sq	0.7983
AIC	6148.89269
AICC	6149.27625
SBC	5869.11538

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	46009	5859.485517	7.85
Gr_Liv_Area	1	58.386514	4.831268	12.09
Basement_Area	1	30.554240	3.323249	9.19
Garage_Area	1	40.158112	6.436997	6.24
Deck_Porch_Area	1	35.720258	7.989240	4.47
Age_Sold	1	-447.254040	40.921927	-10.93

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Deck_Porch_Area	-11.4060	<.0001
2	Bedroom_AbvGr	-6.2737	0.0019
3	Total_Bathroom	-3.1045	0.0448
4	Lot_Area	-1.9476	0.1426



Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 6

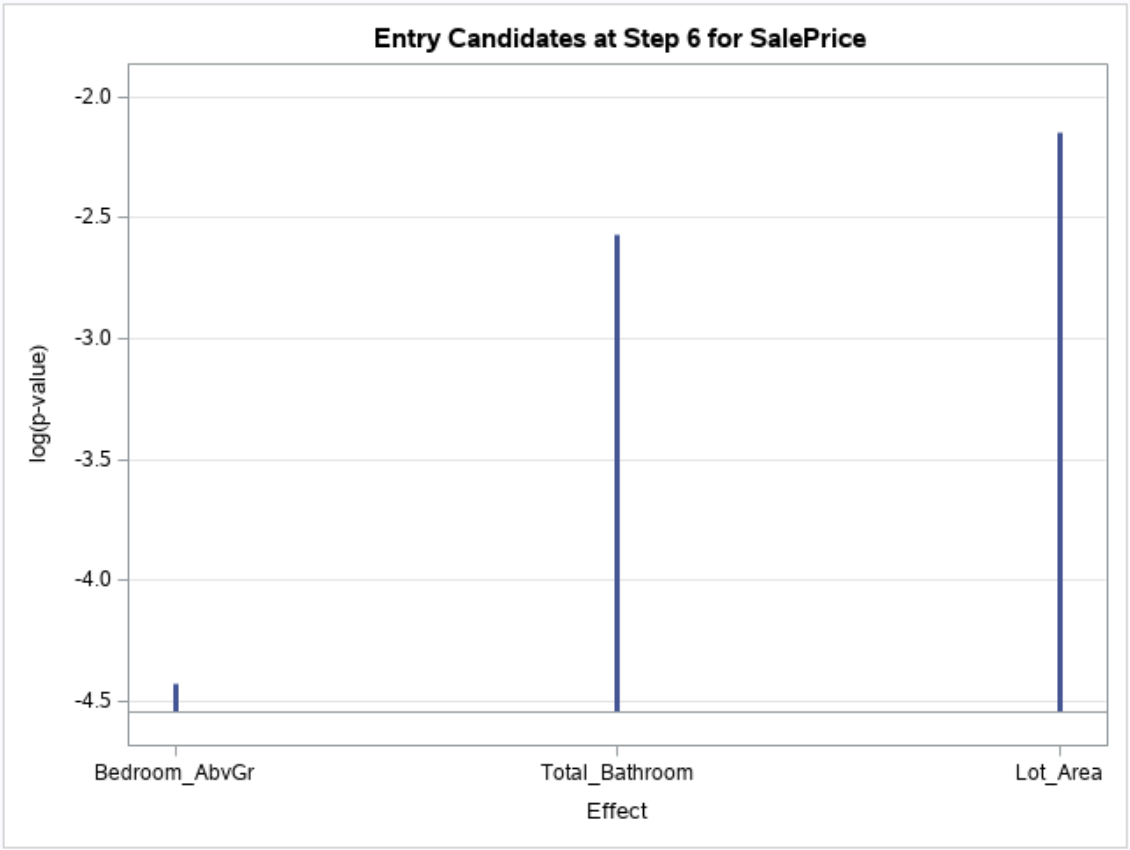
Effect Entered: Bedroom_AbvGr

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	6	3.410749E11	56845818595	202.75
Error	293	82148607939	280370676	
Corrected Total	299	4.232235E11		

Root MSE	16744
Dependent Mean	137525
R-Square	0.8059
Adj R-Sq	0.8019
AIC	6144.40398
AICC	6144.89882
SBC	5868.33046

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	48620	5897.324643	8.24
Gr_Liv_Area	1	65.097413	5.472624	11.90
Basement_Area	1	31.279351	3.305546	9.46
Garage_Area	1	38.728785	6.403565	6.05
Deck_Porch_Area	1	32.487956	8.019119	4.05
Age_Sold	1	-434.199118	40.877494	-10.62
Bedroom_AbvGr	1	-4189.095026	1655.065743	-2.53

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Bedroom_AbvGr	-4.4317	0.0119
2	Total_Bathroom	-2.5664	0.0768
3	Lot_Area	-2.1476	0.1168


Entry Candidates at Step 6 for SalePrice
-2.0
-2.5
-3.0
-3.5
-4.0
-4.5
log(p-value)
Bedroom_AbvGr
Total_Bathroom
Lot_Area
Effect

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 7

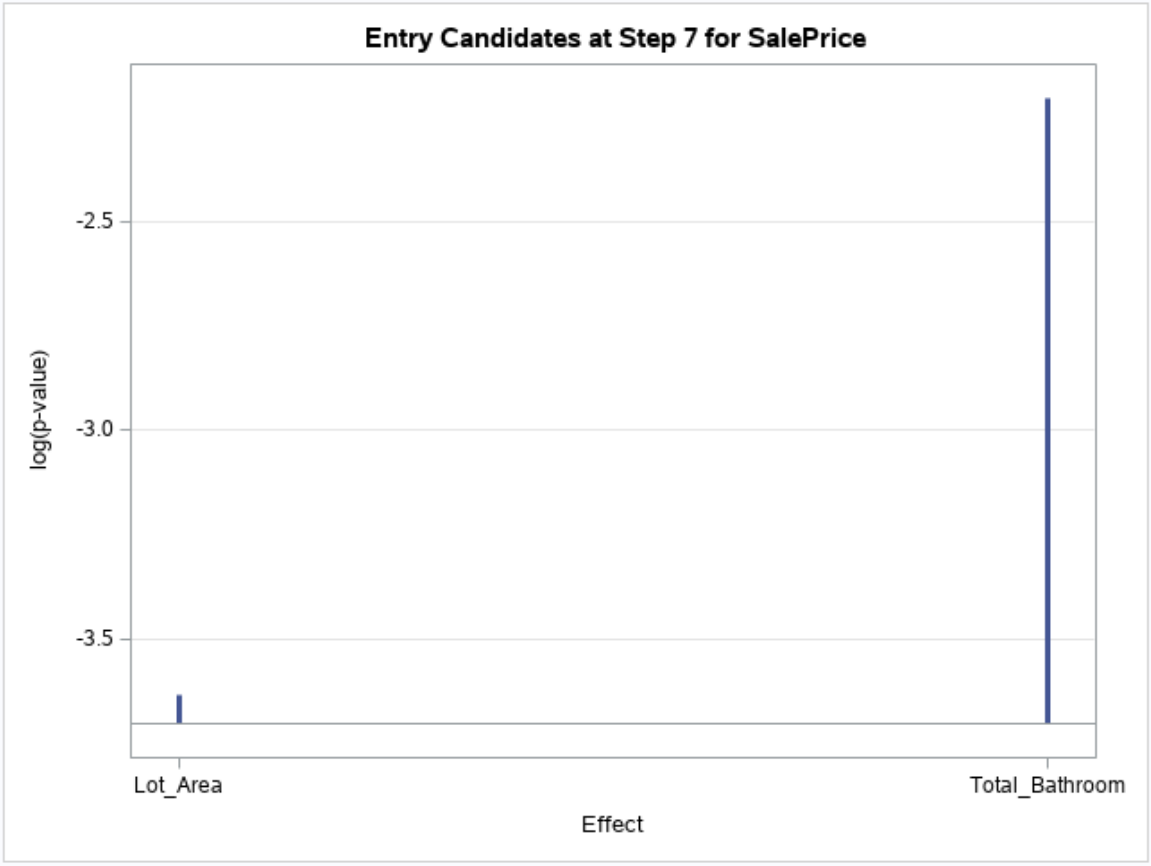
Effect Entered: Lot_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Lot_Area	-3.6309	0.0265
2	Total_Bathroom	-2.2036	0.1104



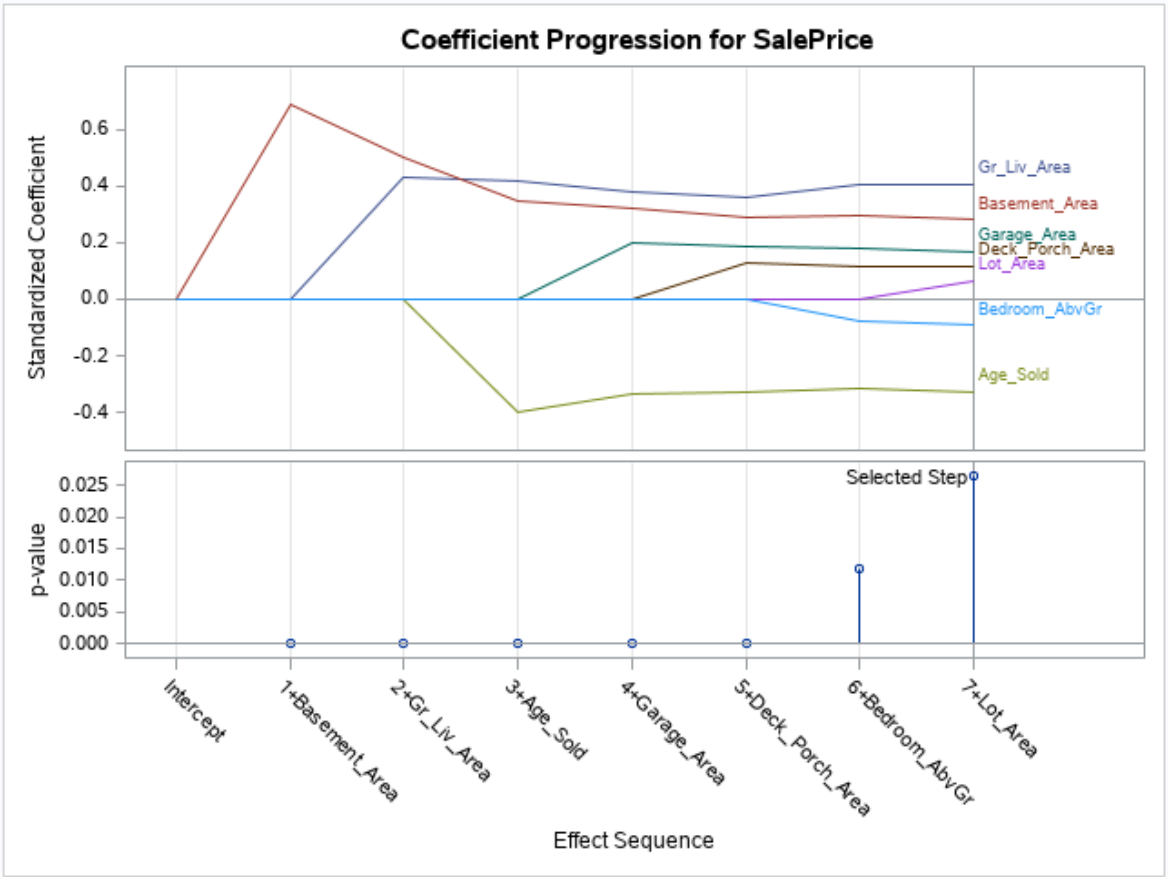
Forward Model Selection for SalePrice - SL 0.05

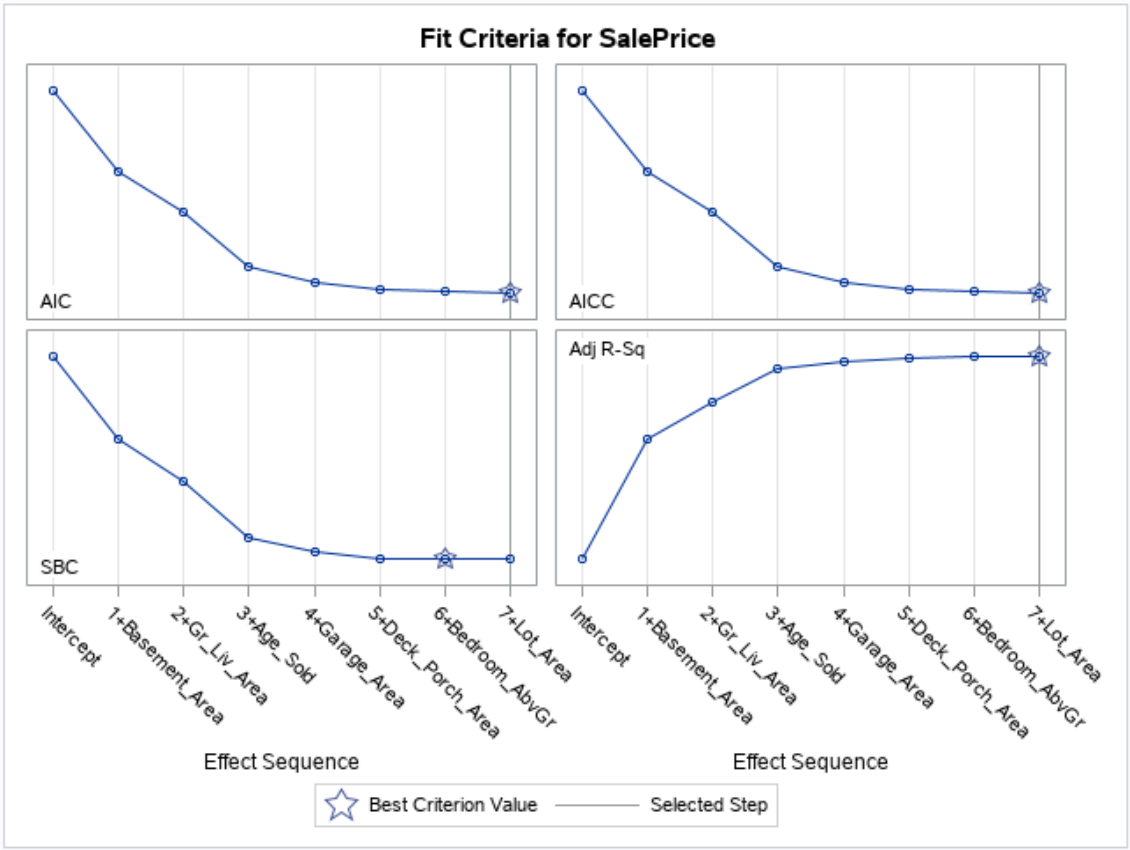
The GLMSELECT Procedure

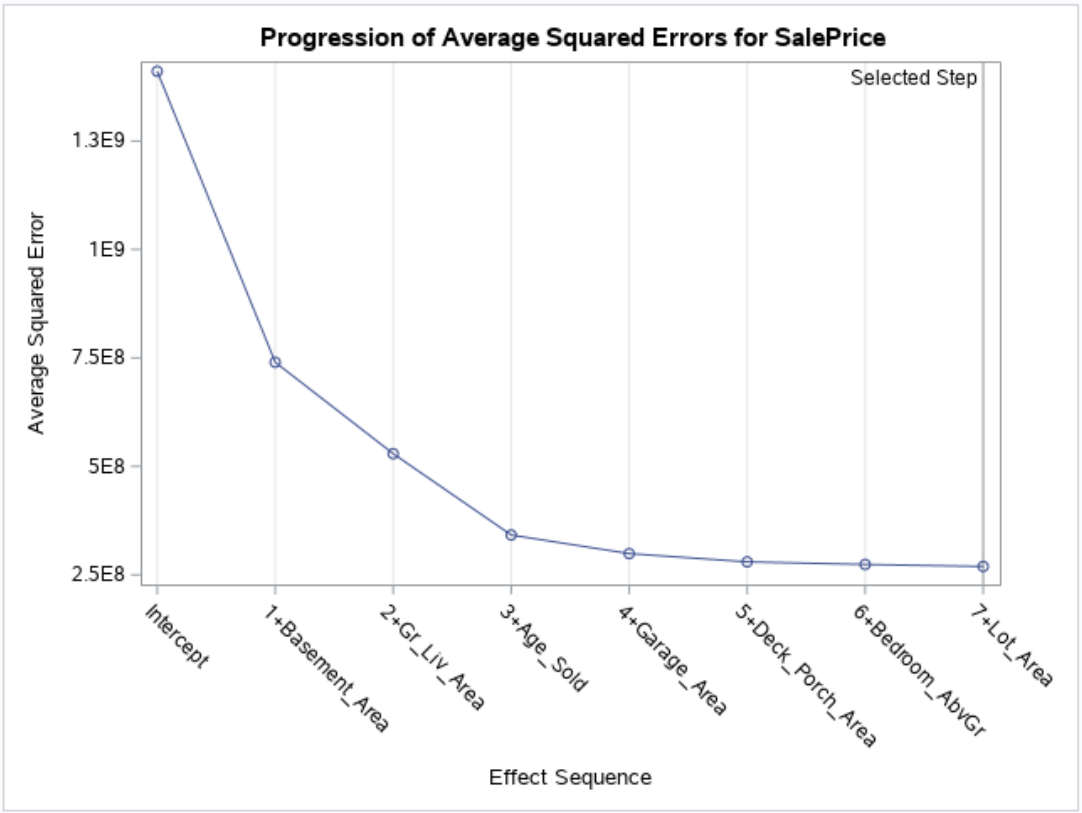
Forward Selection Summary				
Step	Effect Entered	Number Effects In	F Value	Pr > F
0	Intercept	1	0.00	1.0000
1	Basement_Area	2	270.16	<.0001
2	Gr_Liv_Area	3	118.32	<.0001
3	Age_Sold	4	162.37	<.0001
4	Garage_Area	5	42.30	<.0001
5	Deck_Porch_Area	6	19.99	<.0001
6	Bedroom_AbvGr	7	6.41	0.0119
7	Lot_Area	8	4.97	0.0265

Selection stopped as the candidate for entry has SLE > 0.05.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	Total_Bathroom	0.1167	>	0.0500	(SLE)




Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for SalePrice
Selected Step
Average Squared Error
1.3E9
1E9
7.5E8
5E8
2.5E8
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 7).

Effects:	Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Backward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	300
Number of Observations Used	300

Dimensions	
Number of Effects	9
Number of Parameters	9

Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Backward Selection: Step 0

Full Least Squares Model

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	8	3.431321E11	42891512314	155.84
Error	291	80091420996	275228251	
Corrected Total	299	4.232235E11		

Root MSE	16590
Dependent Mean	137525
R-Square	0.8108
Adj R-Sq	0.8056
AIC	6140.79563
AICC	6141.55688
SBC	5872.12967

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	44347	6191.271944	7.16
Gr_Liv_Area	1	63.197764	5.585739	11.31
Basement_Area	1	28.692184	3.417034	8.40
Garage_Area	1	35.754191	6.445840	5.55
Deck_Porch_Area	1	31.370539	7.959436	3.94
Lot_Area	1	0.699495	0.316761	2.21
Age_Sold	1	-420.815037	44.219144	-9.52
Bedroom_AbvGr	1	-4834.848748	1688.858227	-2.86
Total_Bathroom	1	3022.124723	1920.839066	1.57

Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Backward Selection: Step 1

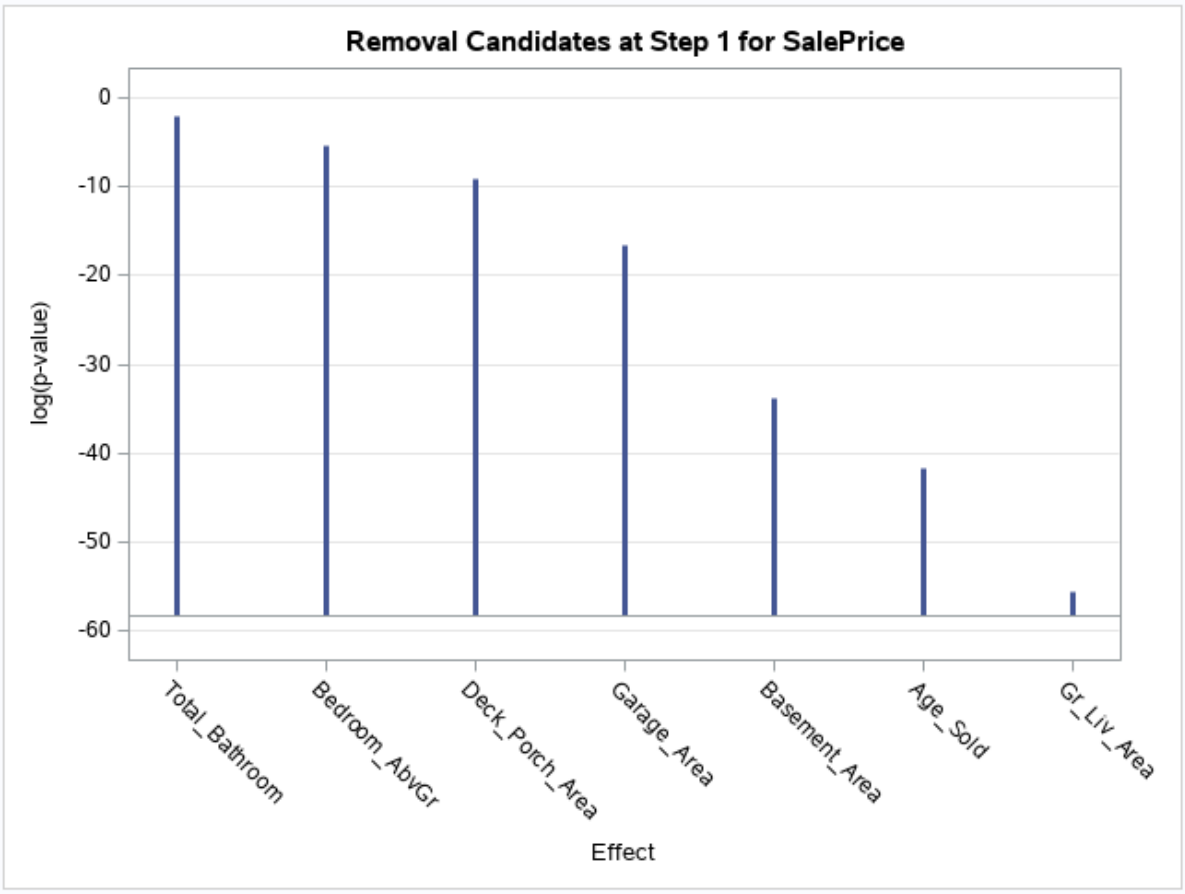
Effect Removed: Total_Bathroom

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

Removal Candidates			
Rank	Effect	Log pValue	Pr > F
1	Total_Bathroom	-2.1479	0.1167
2	Bedroom_AbvGr	-5.4027	0.0045
3	Deck_Porch_Area	-9.1949	0.0001
4	Garage_Area	-16.5434	<.0001
5	Basement_Area	-33.8268	<.0001
6	Age_Sold	-41.7794	<.0001
7	Gr_Liv_Area	-55.5230	<.0001



Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure

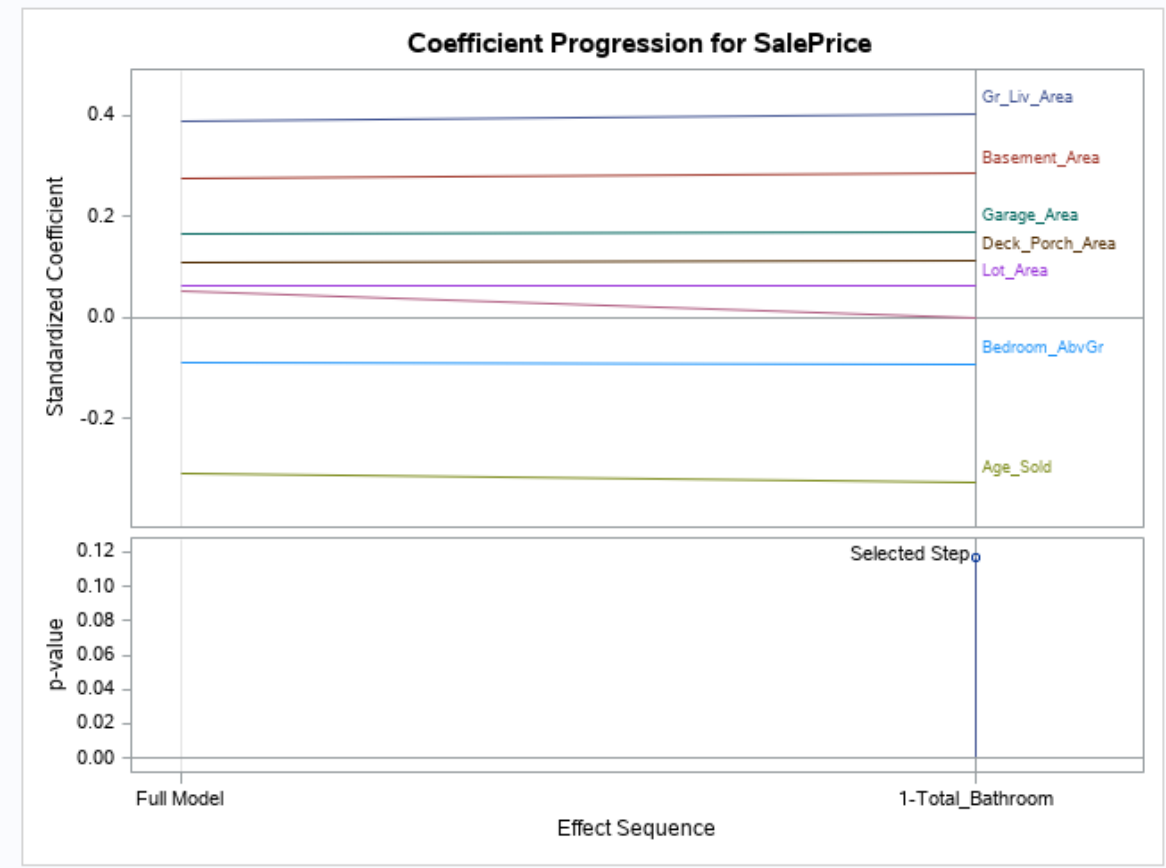
Backward Selection Summary

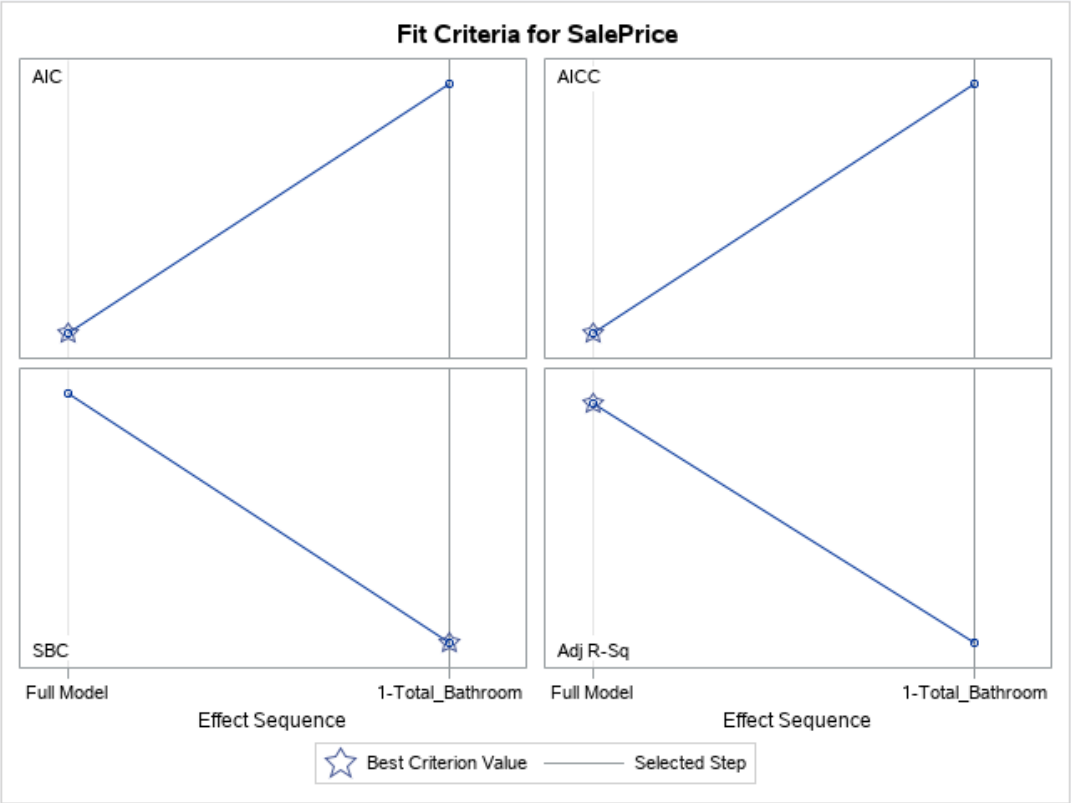
Step	Effect Removed	Number Effects In	F Value	Pr > F
0		9		
1	Total_Bathroom	8	2.48	0.1167

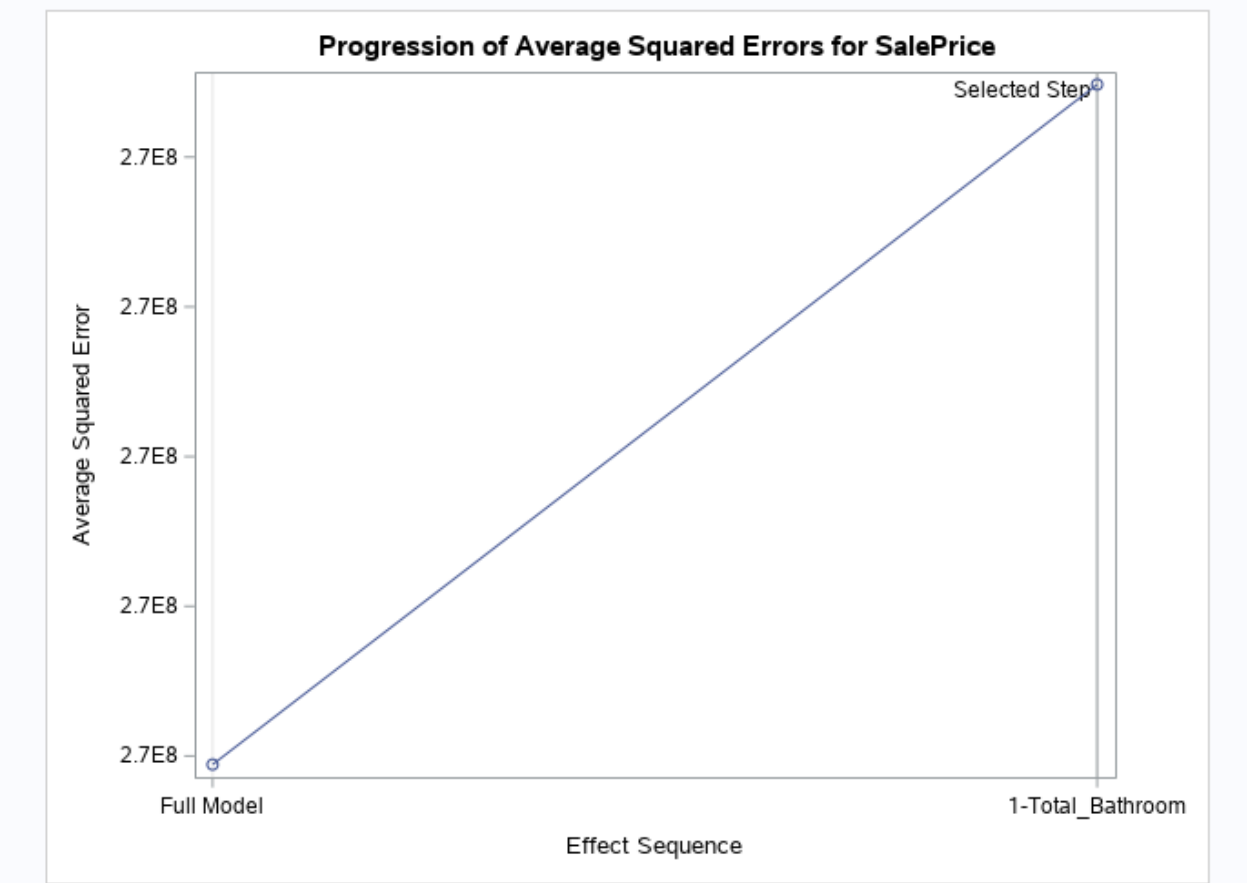
Selection stopped because the next candidate for removal has SLS < 0.05.

Stop Details

Candidate For	Effect	Candidate Significance		Compare Significance	
Removal	Lot_Area	0.0265	<	0.0500	(SLS)


Coefficient Progression for SalePrice
Standardized Coefficient
0.4
0.2
0.0
-0.2
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Age_Sold
p-value
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Selected Step
Full Model
1-Total_Bathroom
Effect Sequence


Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Full Model
1-Total_Bathroom
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for SalePrice
Selected Step
2.7E8
2.7E8
2.7E8
2.7E8
2.7E8
Average Squared Error
Full Model
1-Total_Bathroom
Effect Sequence

Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

13. 15. Using AIC, BIC Information Criteria To Select Variables

Several information criteria are available with the PROC GLMSELCT procedure:

PROC GLMSELECT

Akaike's information criterion (AIC)

corrected Akaike's information criterion (AICC)

Sawa Bayesian information criterion (BIC)

Schwarz Bayesian information criterion (SBC)

$$n\log\left(\frac{SSE}{n}\right)$$

information criterion	penalty component
AIC	$2p+n+2$
AICC	$\frac{n(n+p)}{n-p-2}$
BIC	$2(p+2)q-2q^2$
SBC	$p\log(n)$

```
/* save all interval variables in a macro */
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;


/* Let's use the AIC criterion for variable selection */
ods graphics on;
proc glmselect data=STAT1.ameshousing3 plots=all;
     STEPWISEAIC: model SalePrice = &interval / selection=stepwise
details=steps select=AIC;
     title "Stepwise Model Selection for SalePrice - AIC";
run;
```

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	AIC
Stop Criterion	AIC
Effect Hierarchy Enforced	None

Number of Observations Read	300
Number of Observations Used	300

Dimensions	
Number of Effects	9
Number of Parameters	9

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 0

Effect Entered: Intercept

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	299	4.232235E11	1415463276	
Corrected Total	299	4.232235E11		

Root MSE	37623
Dependent Mean	137525
R-Square	0.0000
Adj R-Sq	0.0000
AIC	6624.21515
AICC	6624.25555
SBC	6325.91893

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	137525	2172.144314	63.31

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 1

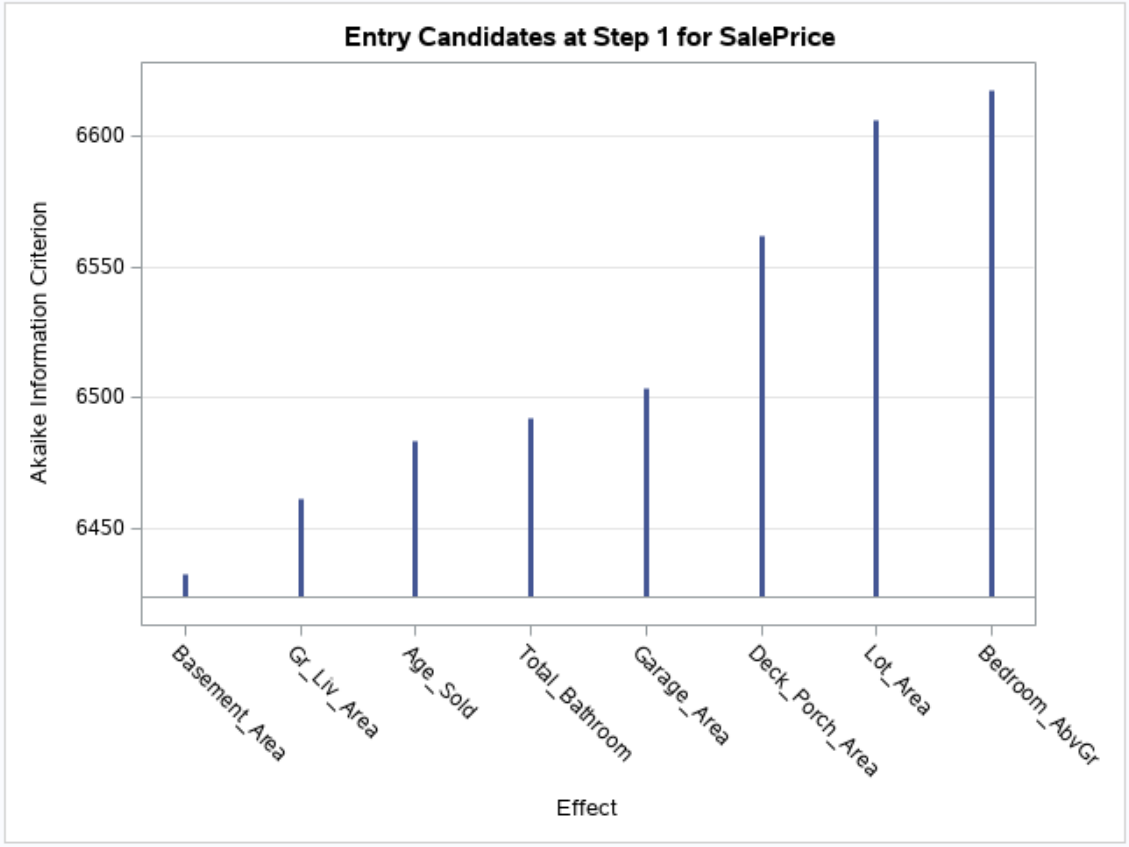
Effect Entered: Basement_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	2.012418E11	2.012418E11	270.16
Error	298	2.219817E11	744904950	
Corrected Total	299	4.232235E11		

Root MSE	27293
Dependent Mean	137525
R-Square	0.4755
Adj R-Sq	0.4737
AIC	6432.62346
AICC	6432.70454
SBC	6138.03102

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	73904	4179.193780	17.68
Basement_Area	1	72.107717	4.387055	16.44

Entry Candidates		
Rank	Effect	AIC
1	Basement_Area	6432.6235
2	Gr_Liv_Area	6461.1877
3	Age_Sold	6483.4097
4	Total_Bathroom	6492.0868
5	Garage_Area	6503.7574
6	Deck_Porch_Area	6561.6989
7	Lot_Area	6606.3138
8	Bedroom_AbvGr	6617.8389



Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 2

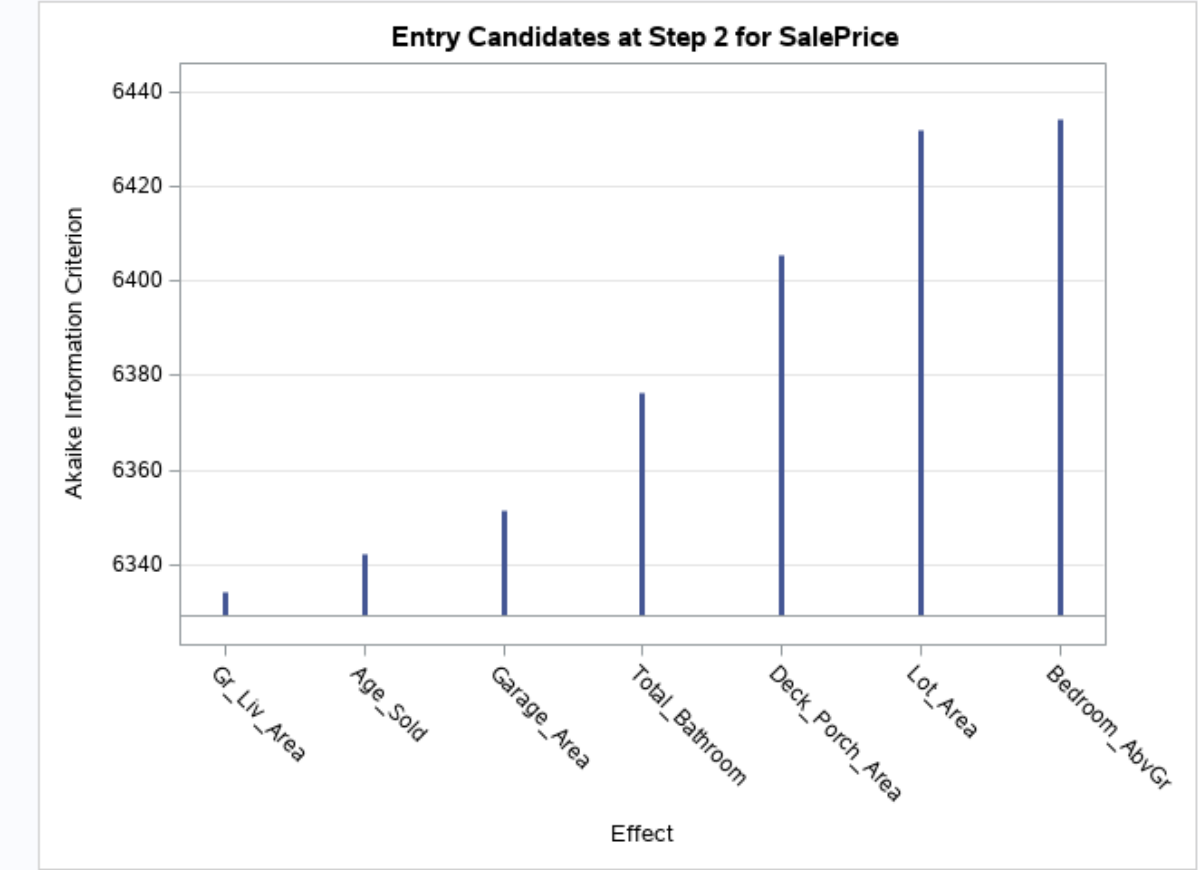
Effect Entered: Gr_Liv_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	2.64483E11	1.322415E11	247.42
Error	297	1.587405E11	534479711	
Corrected Total	299	4.232235E11		

Root MSE	23119
Dependent Mean	137525
R-Square	0.6249
Adj R-Sq	0.6224
AIC	6334.02620
AICC	6334.16179
SBC	6043.13755

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	12664	6650.339855	1.90
Gr_Liv_Area	1	69.606974	6.399091	10.88
Basement_Area	1	52.309702	4.137885	12.64

Entry Candidates		
Rank	Effect	AIC
1	Gr_Liv_Area	6334.0262
2	Age_Sold	6342.0095
3	Garage_Area	6351.3061
4	Total_Bathroom	6376.4084
5	Deck_Porch_Area	6405.4472
6	Lot_Area	6431.9372
7	Bedroom_AbvGr	6434.0931


Entry Candidates at Step 2 for SalePrice
Akaike Information Criterion
6440
6420
6400
6380
6360
6340
Gr_Liv_Area
Age_Sold
Garage_Area
Total_Bathroom
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Effect

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 3

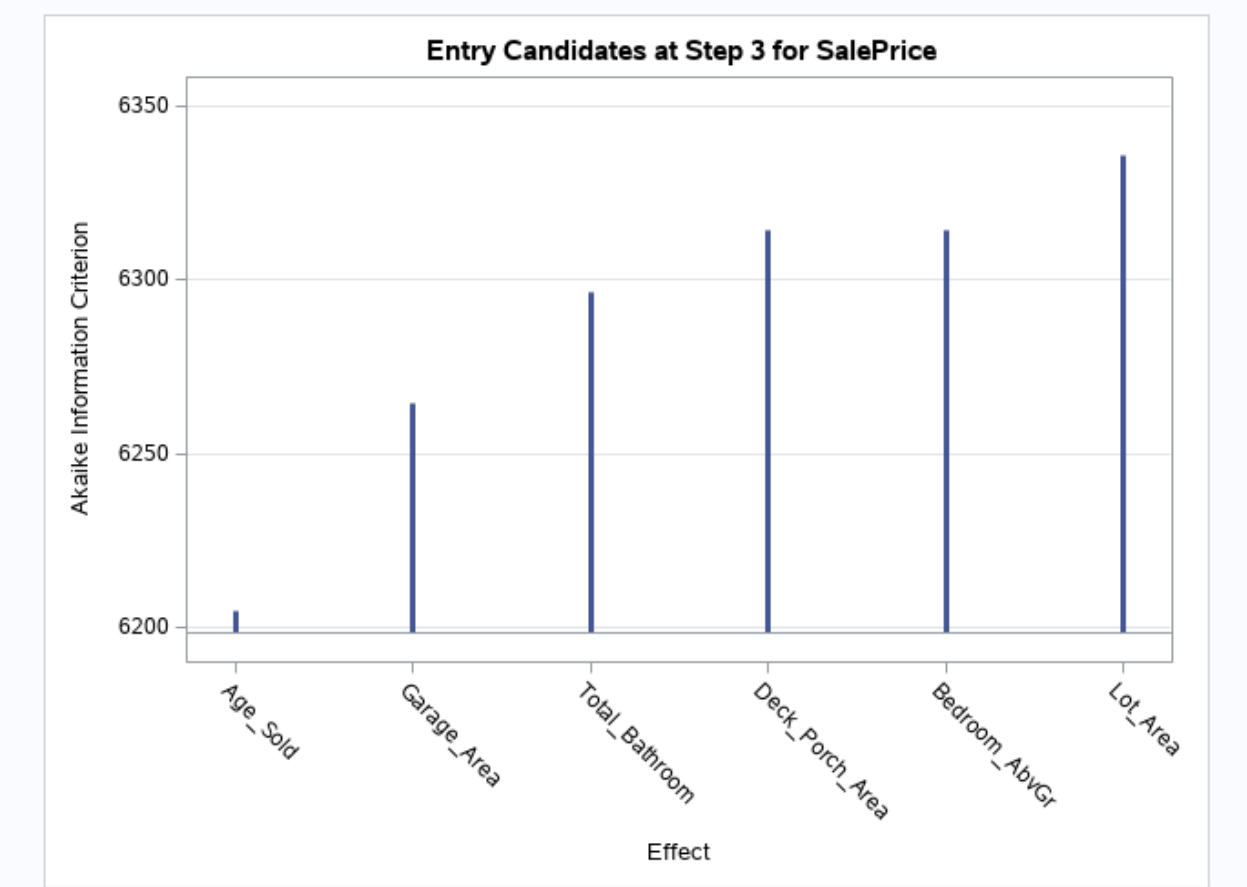
Effect Entered: Age_Sold

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	3.207148E11	1.069049E11	308.69
Error	296	1.025087E11	346313132	
Corrected Total	299	4.232235E11		

Root MSE	18609
Dependent Mean	137525
R-Square	0.7578
Adj R-Sq	0.7553
AIC	6204.82927
AICC	6205.03335
SBC	5917.64440

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	53400	6235.076995	8.56
Gr_Liv_Area	1	68.106646	5.152294	13.22
Basement_Area	1	36.329120	3.559067	10.21
Age_Sold	1	-543.493346	42.651840	-12.74

Entry Candidates		
Rank	Effect	AIC
1	Age_Sold	6204.8293
2	Garage_Area	6264.6656
3	Total_Bathroom	6296.3441
4	Deck_Porch_Area	6314.2811
5	Bedroom_AbvGr	6314.4078
6	Lot_Area	6335.7989


Entry Candidates at Step 3 for SalePrice
Akaike Information Criterion
6350
6300
6250
6200
Age_Sold
Garage_Area
Total_Bathroom
Deck_Porch_Area
Bedroom_AbvGr
Lot_Area
Effect

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 4

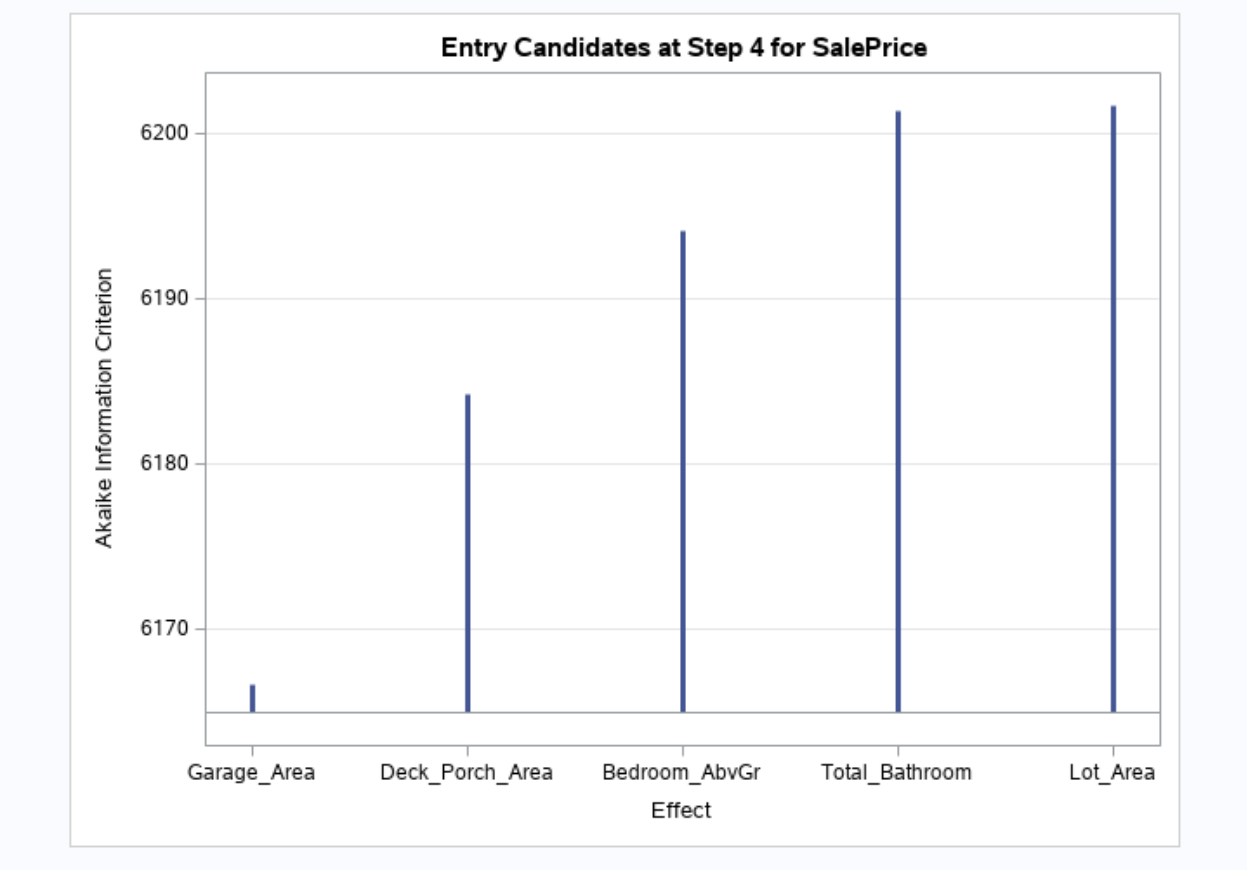
Effect Entered: Garage_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	4	3.33571E11	83392754480	274.40
Error	295	89652501590	303906785	
Corrected Total	299	4.232235E11		

Root MSE	17433
Dependent Mean	137525
R-Square	0.7882
Adj R-Sq	0.7853
AIC	6166.62734
AICC	6166.91403
SBC	5883.14625

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	43815	6023.907004	7.27
Gr_Liv_Area	1	61.238136	4.940722	12.39
Basement_Area	1	33.430181	3.363709	9.94
Garage_Area	1	42.984492	6.608851	6.50
Age_Sold	1	-455.704354	42.173481	-10.81

Entry Candidates		
Rank	Effect	AIC
1	Garage_Area	6166.6273
2	Deck_Porch_Area	6184.1900
3	Bedroom_AbvGr	6194.0980
4	Total_Bathroom	6201.3633
5	Lot_Area	6201.6954


Entry Candidates at Step 4 for SalePrice
Akaike Information Criterion
6200
6190
6180
6170
Garage_Area
Deck_Porch_Area
Bedroom_AbvGr
Total_Bathroom
Lot_Area
Effect

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 5

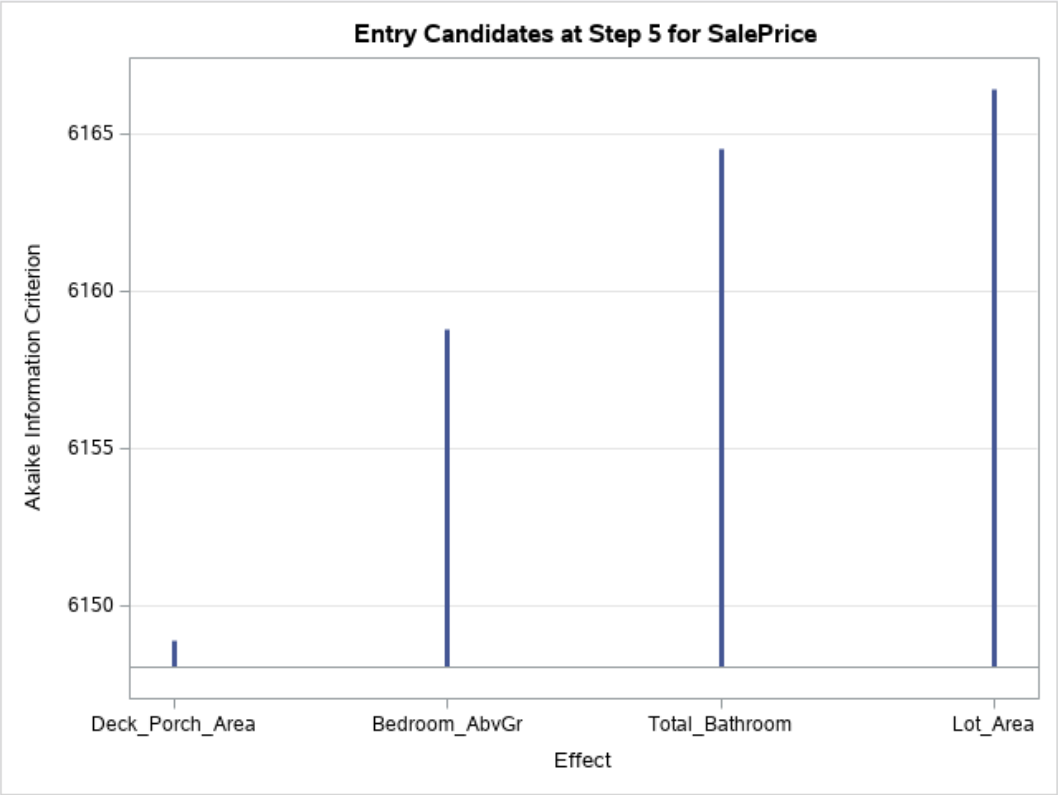
Effect Entered: Deck_Porch_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	5	3.392788E11	67855752389	237.65
Error	294	83944757568	285526386	
Corrected Total	299	4.232235E11		

Root MSE	16898
Dependent Mean	137525
R-Square	0.8017
Adj R-Sq	0.7983
AIC	6148.89269
AICC	6149.27625
SBC	5869.11538

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	46009	5859.485517	7.85
Gr_Liv_Area	1	58.386514	4.831268	12.09
Basement_Area	1	30.554240	3.323249	9.19
Garage_Area	1	40.158112	6.436997	6.24
Deck_Porch_Area	1	35.720258	7.989240	4.47
Age_Sold	1	-447.254040	40.921927	-10.93

Entry Candidates		
Rank	Effect	AIC
1	Deck_Porch_Area	6148.8927
2	Bedroom_AbvGr	6158.7554
3	Total_Bathroom	6164.5138
4	Lot_Area	6166.4302


Entry Candidates at Step 5 for SalePrice
Akaike Information Criterion
6165
6160
6155
6150
Deck_Porch_Area
Bedroom_AbvGr
Total_Bathroom
Lot_Area
Effect

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 6

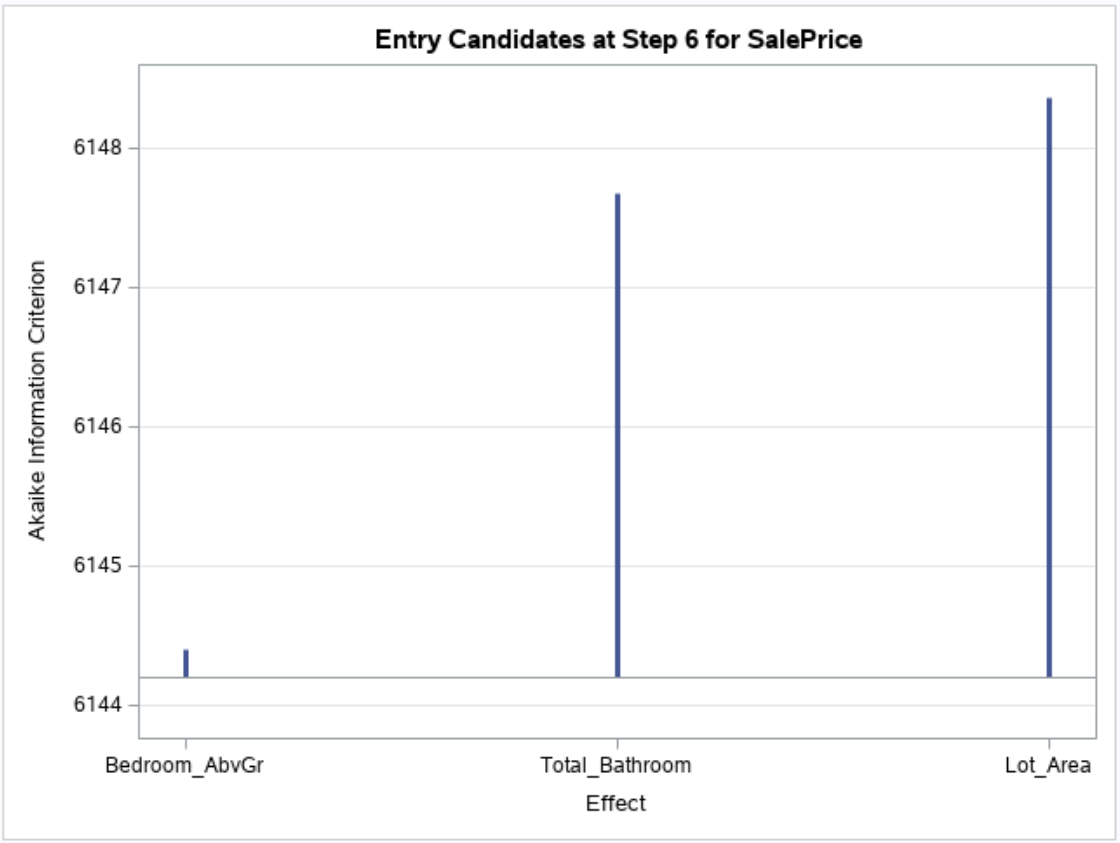
Effect Entered: Bedroom_AbvGr

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	6	3.410749E11	56845818595	202.75
Error	293	82148607939	280370676	
Corrected Total	299	4.232235E11		

Root MSE	16744
Dependent Mean	137525
R-Square	0.8059
Adj R-Sq	0.8019
AIC	6144.40398
AICC	6144.89882
SBC	5868.33046

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	48620	5897.324643	8.24
Gr_Liv_Area	1	65.097413	5.472624	11.90
Basement_Area	1	31.279351	3.305546	9.46
Garage_Area	1	38.728785	6.403565	6.05
Deck_Porch_Area	1	32.487956	8.019119	4.05
Age_Sold	1	-434.199118	40.877494	-10.62
Bedroom_AbvGr	1	-4189.095026	1655.065743	-2.53

Entry Candidates		
Rank	Effect	AIC
1	Bedroom_AbvGr	6144.4040
2	Total_Bathroom	6147.6813
3	Lot_Area	6148.3694



Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 7

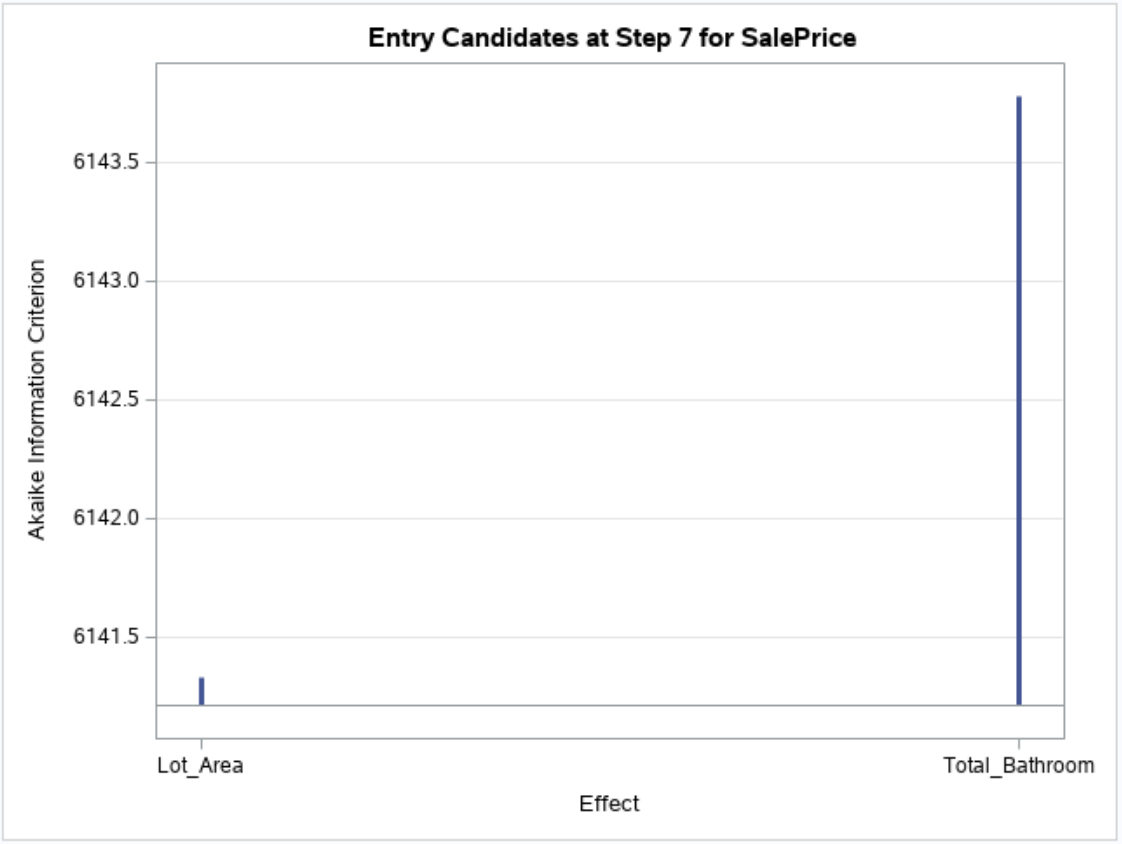
Effect Entered: Lot_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

Entry Candidates		
Rank	Effect	AIC
1	Lot_Area	6141.3368
2	Total_Bathroom	6143.7813



Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 8

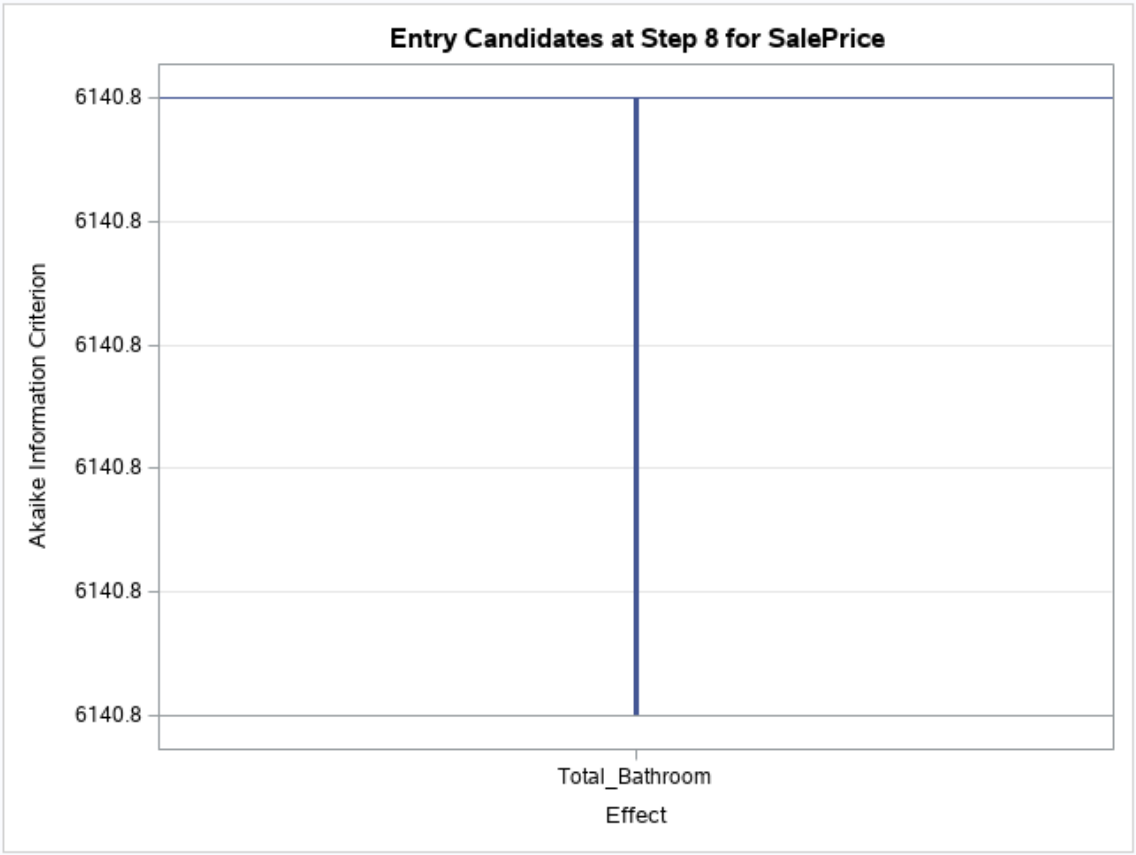
Effect Entered: Total_Bathroom

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	8	3.431321E11	42891512314	155.84
Error	291	80091420996	275228251	
Corrected Total	299	4.232235E11		

Root MSE	16590
Dependent Mean	137525
R-Square	0.8108
Adj R-Sq	0.8056
AIC	6140.79563
AICC	6141.55688
SBC	5872.12967

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	44347	6191.271944	7.16
Gr_Liv_Area	1	63.197764	5.585739	11.31
Basement_Area	1	28.692184	3.417034	8.40
Garage_Area	1	35.754191	6.445840	5.55
Deck_Porch_Area	1	31.370539	7.959436	3.94
Lot_Area	1	0.699495	0.316761	2.21
Age_Sold	1	-420.815037	44.219144	-9.52
Bedroom_AbvGr	1	-4834.848748	1688.858227	-2.86
Total_Bathroom	1	3022.124723	1920.839066	1.57

Entry Candidates		
Rank	Effect	AIC
1	Total_Bathroom	6140.7956

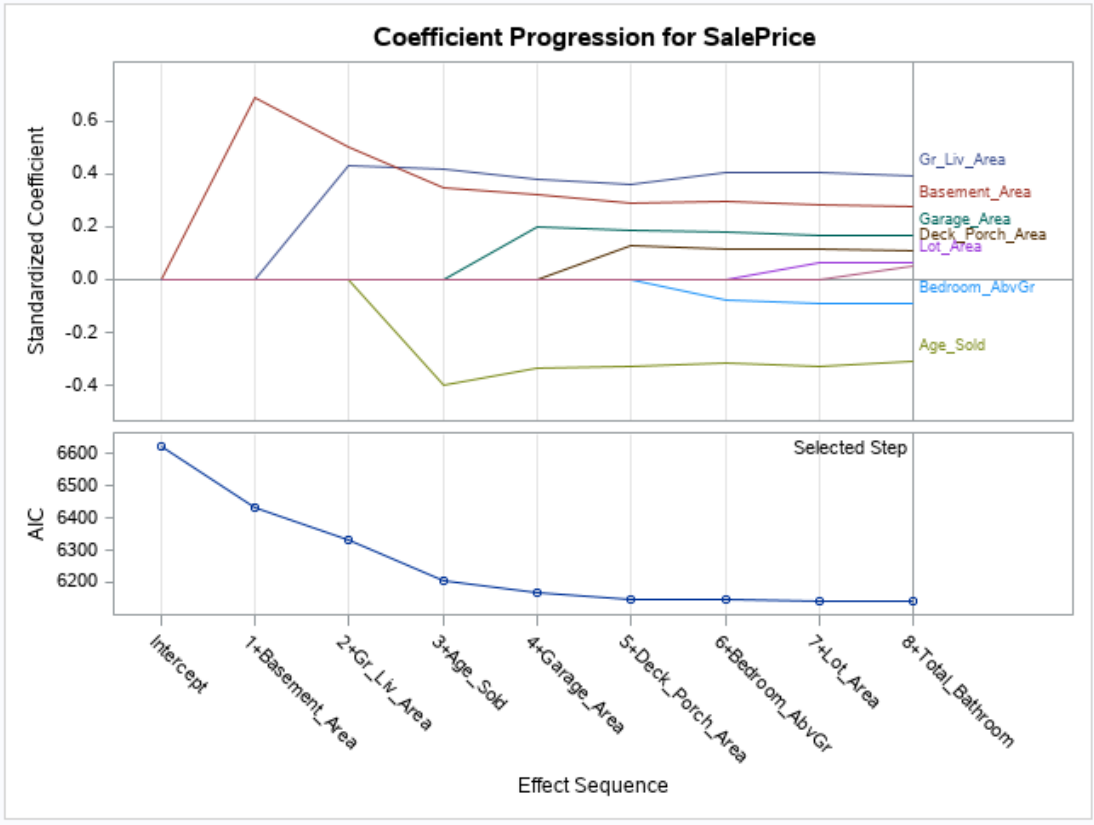


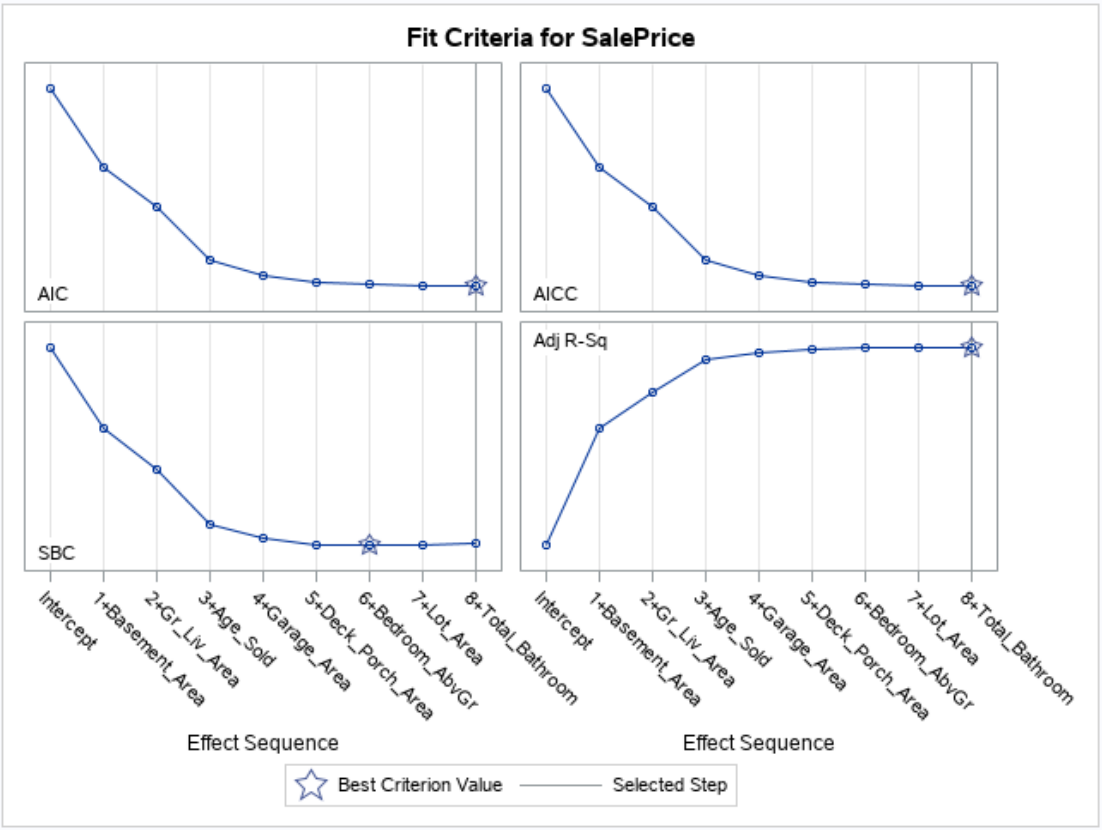
Stepwise Model Selection for SalePrice - AIC

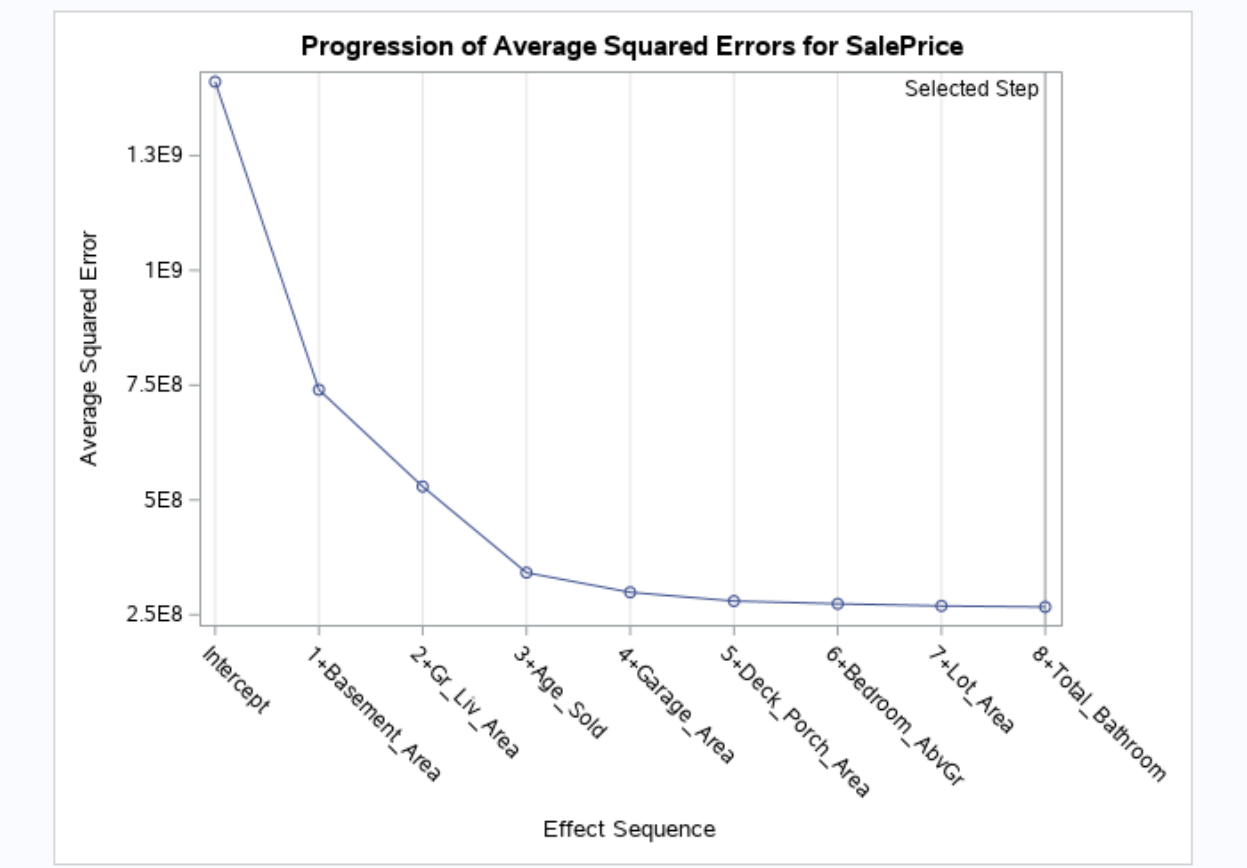
The GLMSELECT Procedure

Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	AIC
0	Intercept		1	6624.2151
1	Basement_Area		2	6432.6235
2	Gr_Liv_Area		3	6334.0262
3	Age_Sold		4	6204.8293
4	Garage_Area		5	6166.6273
5	Deck_Porch_Area		6	6148.8927
6	Bedroom_AbvGr		7	6144.4040
7	Lot_Area		8	6141.3368
8	Total_Bathroom		9	6140.7956*
* Optimal Value of Criterion				

Selection stopped because all effects are in the final model.


Coefficient Progression for SalePrice
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
-0.4
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Age_Sold
Selected Step
AIC
6600
6500
6400
6300
6200
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
8+Total_Bathroom
Effect Sequence


Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
8+Total_Bathroom
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for SalePrice
Selected Step
Average Squared Error
1.3E9
1E9
7.5E8
5E8
2.5E8
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
8+Total_Bathroom
Effect Sequence

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 8).

Effects:	Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	8	3.431321E11	42891512314	155.84
Error	291	80091420996	275228251	
Corrected Total	299	4.232235E11		

Root MSE	16590
Dependent Mean	137525
R-Square	0.8108
Adj R-Sq	0.8056
AIC	6140.79563
AICC	6141.55688
SBC	5872.12967

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	44347	6191.271944	7.16
Gr_Liv_Area	1	63.197764	5.585739	11.31
Basement_Area	1	28.692184	3.417034	8.40
Garage_Area	1	35.754191	6.445840	5.55
Deck_Porch_Area	1	31.370539	7.959436	3.94
Lot_Area	1	0.699495	0.316761	2.21
Age_Sold	1	-420.815037	44.219144	-9.52
Bedroom_AbvGr	1	-4834.848748	1688.858227	-2.86
Total_Bathroom	1	3022.124723	1920.839066	1.57

/* Let's use the BIC information criterion to select variables */

```
proc glmselect data=STAT1.ameshousing3 plots=all;
     STEPWISEBIC: model SalePrice = &interval / selection=stepwise
details=steps select=BIC;
     title "Stepwise Model Selection for SalePrice - BIC";
run;
```

/* Let's use the AICC information criterion to select variables */

```
proc glmselect data=STAT1.ameshousing3 plots=all;
```

```
      STEPWISEAICC: model SalePrice = &interval / selection=stepwise
details=steps select=AICC;

      title "Stepwise Model Selection for SalePrice - AICC";

run;
```

/* Let's use the SBC information criterion to select variables */

```
proc glmselect data=STAT1.ameshousing3 plots=all;

      STEPWISESBC: model SalePrice = &interval / selection=stepwise
details=steps select=SBC;

      title "Stepwise Model Selection for SalePrice - SBC";

run;
```

/* we can also explore all possible models and determine the "best" one, by using the 'ALLPOSS' specification in SAS */

```
ods graphics on;

proc reg data=STAT1.ameshousing3 plots(only)=(rsquare adjrsq
cp);
       ALLPOSS: model SalePrice = &interval / selection=rsquare
adjrsq cp;
       title "All Possible Model Selection for SalePrice";
run;
quit;
```

/* here, we keep only the 20 best models */

```
proc reg data=STAT1.ameshousing3 plots(only)=(cp);
       ALLPOSS: model SalePrice = &interval / selection=cp
rsquare adjrsq best=20;
       title "Best Models Using All Possible Selection for
SalePrice";
run;
quit;
```

14. Model Residual Diagnostics/post fitting for inference

```
/* save the list of interval variables in a macro */
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/* Let's create plots of disgnostic statistics for sale price */
ods graphics on;
proc reg data=STAT1.ameshousing3;
    CONTINUOUS: model SalePrice
                  = &interval;
    title 'SalePrice Model - Plots of Diagnostic Statistics';
run;
quit;
```

SalePrice Model - Plots of Diagnostic Statistics

The REG Procedure
Model: CONTINUOUS
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	300
Number of Observations Used	300

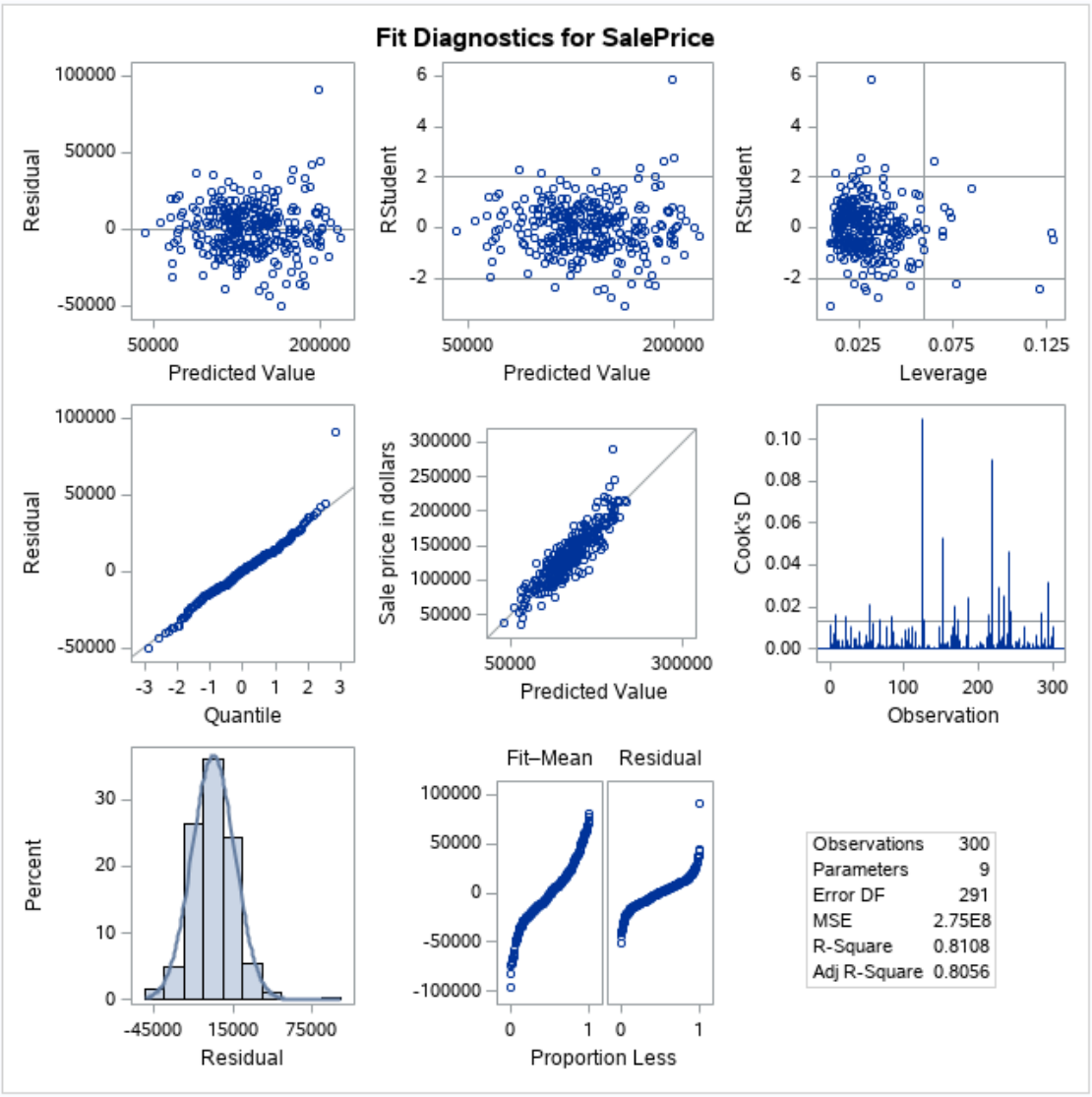
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	3.431321E11	42891512314	155.84	<.0001
Error	291	80091420996	275228251		
Corrected Total	299	4.232235E11			

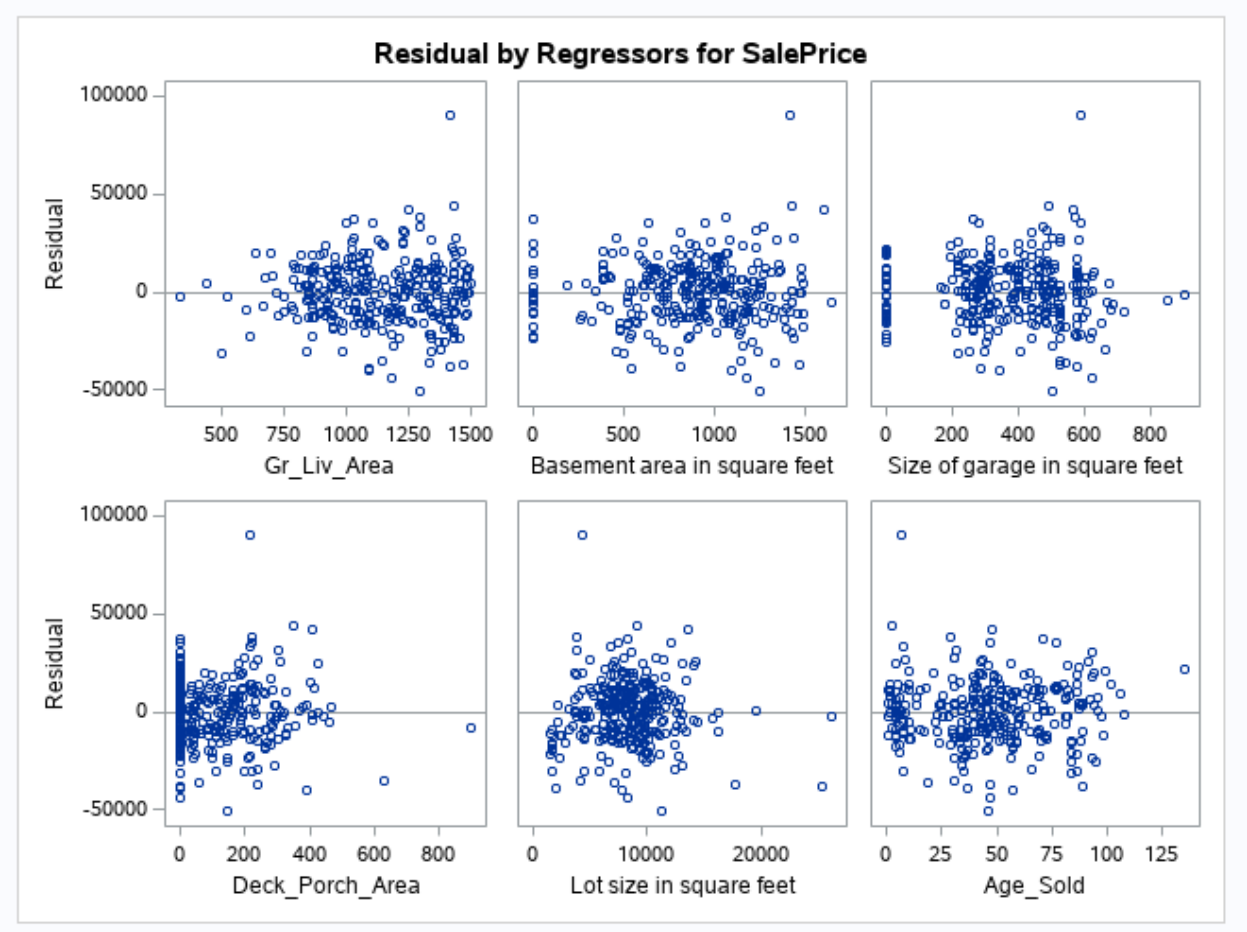
Root MSE	16590	R-Square	0.8108
Dependent Mean	137525	Adj R-Sq	0.8056
Coeff Var	12.06328		

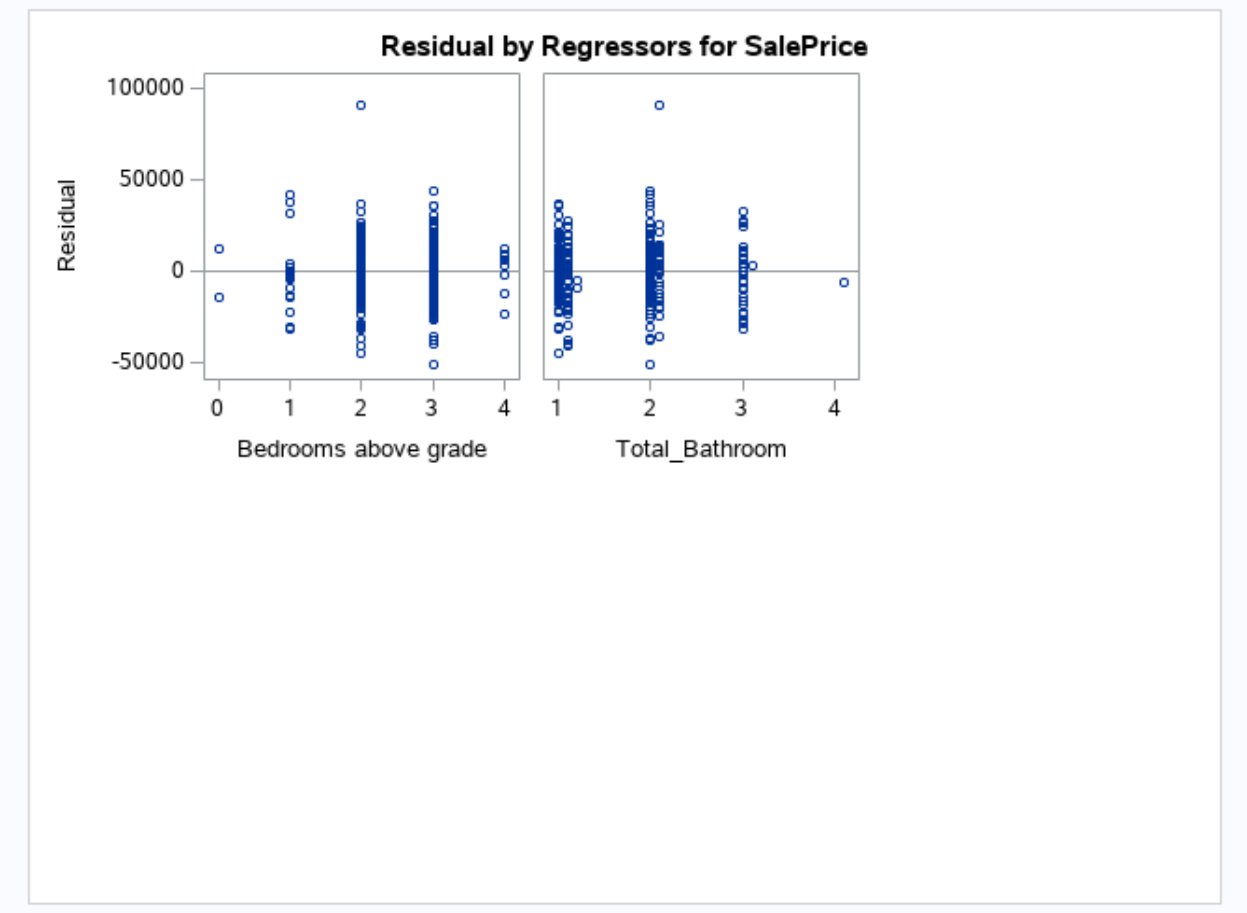
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	44347	6191.27194	7.16	<.0001
Gr_Liv_Area	Above grade (ground) living area square feet	1	63.19776	5.58574	11.31	<.0001
Basement_Area	Basement area in square feet	1	28.69218	3.41703	8.40	<.0001
Garage_Area	Size of garage in square feet	1	35.75419	6.44584	5.55	<.0001
Deck_Porch_Area	Total area of decks and porches in square feet	1	31.37054	7.95944	3.94	0.0001
Lot_Area	Lot size in square feet	1	0.69950	0.31676	2.21	0.0280
Age_Sold	Age of house when sold, in years	1	-420.81504	44.21914	-9.52	<.0001
Bedroom_AbvGr	Bedrooms above grade	1	-4834.84875	1688.85823	-2.86	0.0045
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3022.12472	1920.83907	1.57	0.1167

SalePrice Model - Plots of Diagnostic Statistics

The REG Procedure
Model: CONTINUOUS
Dependent Variable: SalePrice Sale price in dollars


Fit Diagnostics for SalePrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Sale price in dollars
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 300
Parameters 9
Error DF 291
MSE 2.75E8
R-Square 0.8108
Adj R-Square 0.8056


Residual by Regressors for SalePrice
Residual
100000
50000
0
-50000
500
750
1000
1250
1500
Gr_Liv_Area
0
500
1000
1500
Basement area in square feet
0
200
400
600
800
Size of garage in square feet
0
200
400
600
800
Deck_Porch_Area
0
10000
20000
Lot size in square feet
0
25
50
75
100
125
Age_Sold



/* this time, let's request selected plots only */

```
proc reg data=STAT1.ameshousing3
         plots(only)=(QQ RESIDUALBYPREDICTED RESIDUALS);
    CONTINUOUS: model SalePrice
                  = &interval;
    title 'SalePrice Model - Plots of Diagnostic Statistics';
run;
quit;
```

SalePrice Model - Plots of Diagnostic Statistics

The REG Procedure
Model: CONTINUOUS
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	300
Number of Observations Used	300

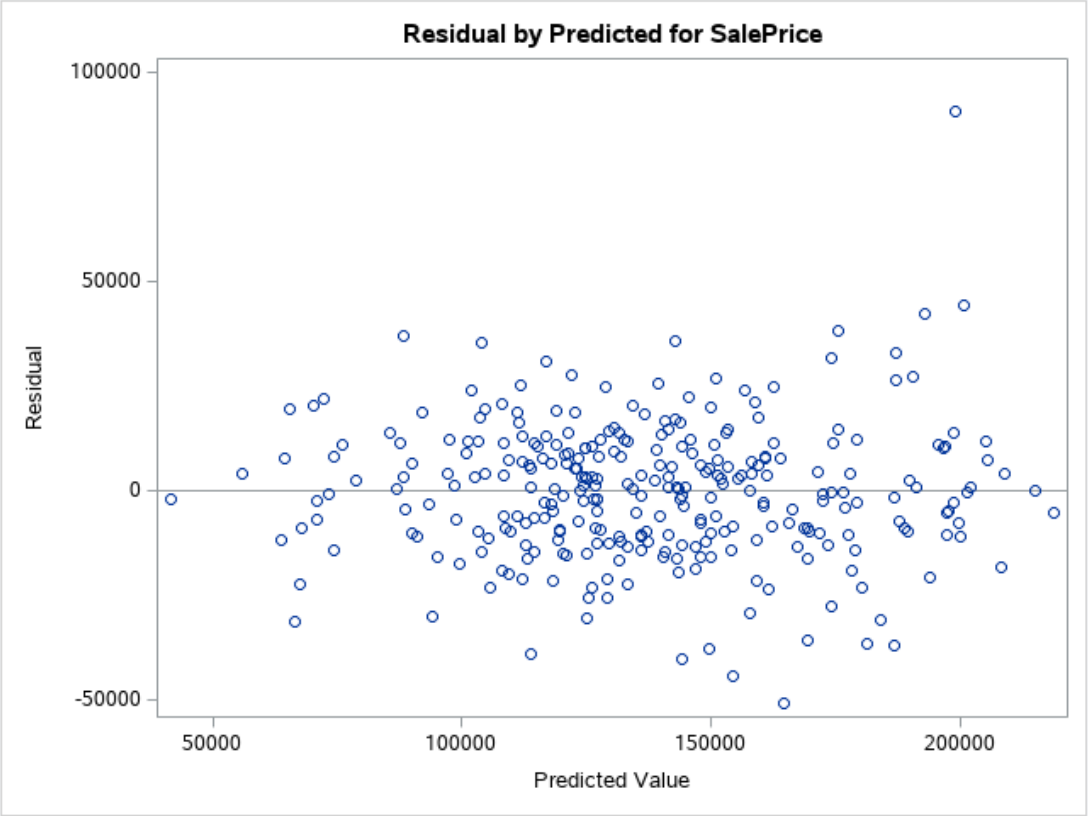
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	3.431321E11	42891512314	155.84	<.0001
Error	291	80091420996	275228251		
Corrected Total	299	4.232235E11			

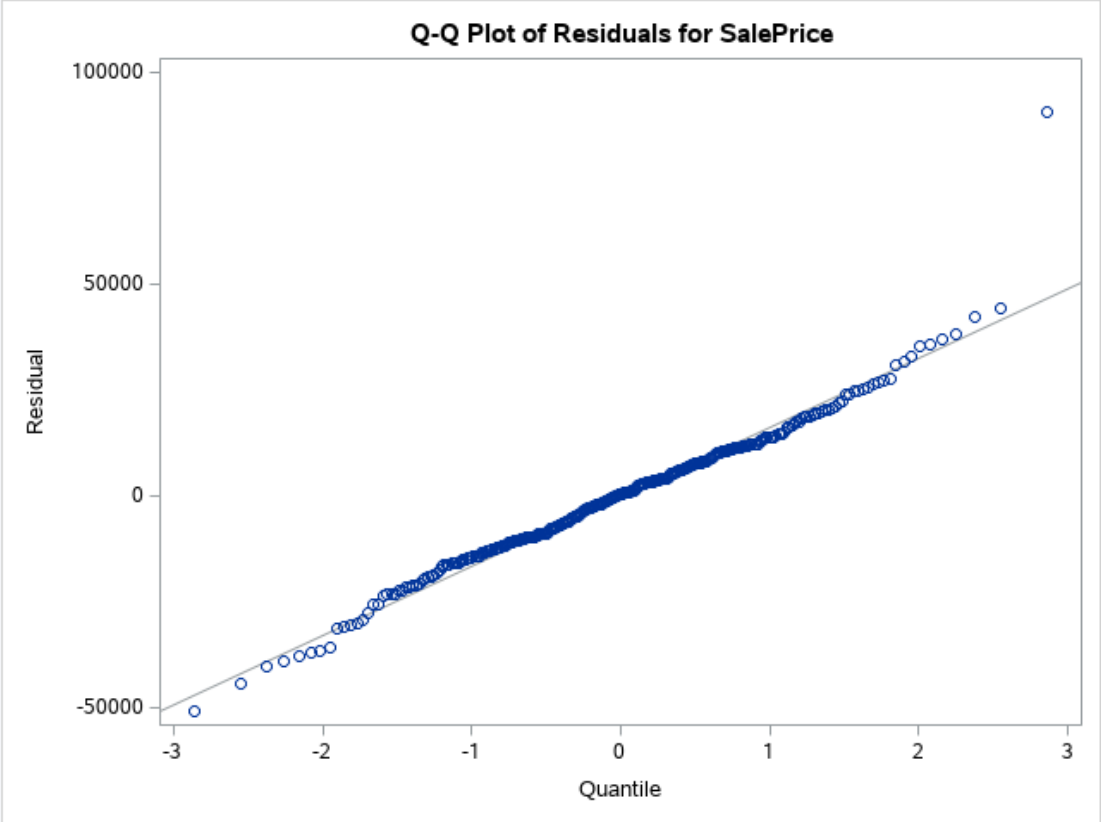
Root MSE	16590	R-Square	0.8108
Dependent Mean	137525	Adj R-Sq	0.8056
Coeff Var	12.06328		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	44347	6191.27194	7.16	<.0001
Gr_Liv_Area	Above grade (ground) living area square feet	1	63.19776	5.58574	11.31	<.0001
Basement_Area	Basement area in square feet	1	28.69218	3.41703	8.40	<.0001
Garage_Area	Size of garage in square feet	1	35.75419	6.44584	5.55	<.0001
Deck_Porch_Area	Total area of decks and porches in square feet	1	31.37054	7.95944	3.94	0.0001
Lot_Area	Lot size in square feet	1	0.69950	0.31676	2.21	0.0280
Age_Sold	Age of house when sold, in years	1	-420.81504	44.21914	-9.52	<.0001
Bedroom_AbvGr	Bedrooms above grade	1	-4834.84875	1688.85823	-2.86	0.0045
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3022.12472	1920.83907	1.57	0.1167

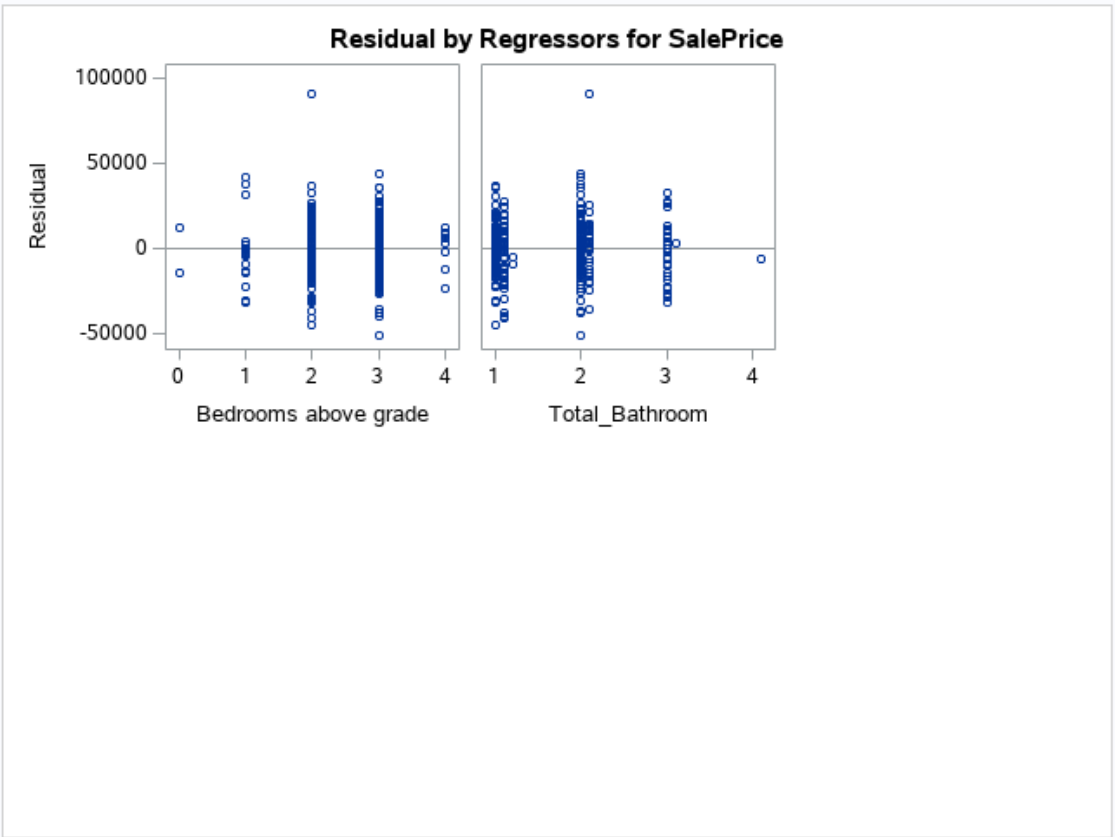
SalePrice Model - Plots of Diagnostic Statistics

The REG Procedure
Model: CONTINUOUS
Dependent Variable: SalePrice Sale price in dollars


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


Q-Q Plot of Residuals for SalePrice
100000
50000
0
-50000
Residual
-3
-2
-1
0
1
2
3
Quantile


Residual by Regressors for SalePrice
Residual
100000
50000
0
-50000
500 750 1000 1250 1500
Gr_Liv_Area
0 500 1000 1500
Basement area in square feet
0 200 400 600 800
Size of garage in square feet
0 200 400 600 800
Deck_Porch_Area
0 10000 20000
Lot size in square feet
0 25 50 75 100 125
Age_Sold


Residual by Regressors for SalePrice
Residual
100000
50000
0
-50000
0 1 2 3 4
Bedrooms above grade
1 2 3 4
Total_Bathroom

/* We can add the option `diagnostics(unpack)` to obtain full-size separate plots for the diagnostic plots */

```
proc reg data=STAT1.ameshousing3
         plots(only)=(      QQ      RESIDUALBYPREDICTED      RESIDUALS
diagnostics(unpack) );
    CONTINUOUS: model SalePrice
                  = &interval;
    title 'SalePrice Model - Plots of Diagnostic Statistics';
run;
quit;
```

15. Identifying outliers and influential observations

We can use the following statistical procedures for each scenario:

outliers	influential observations
STUDENT	Cook's *D*
	RSTUDENT
	DFFITS
	DFBETAS

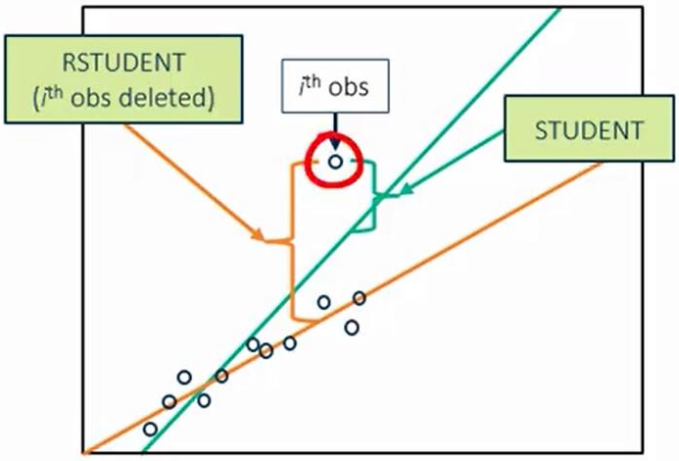
outliers

STUDENT residuals

< 2	occur by chance
2 - 3	infrequent
> 3	rare – investigate!

influential observations

RSTUDENT residuals



influential observations

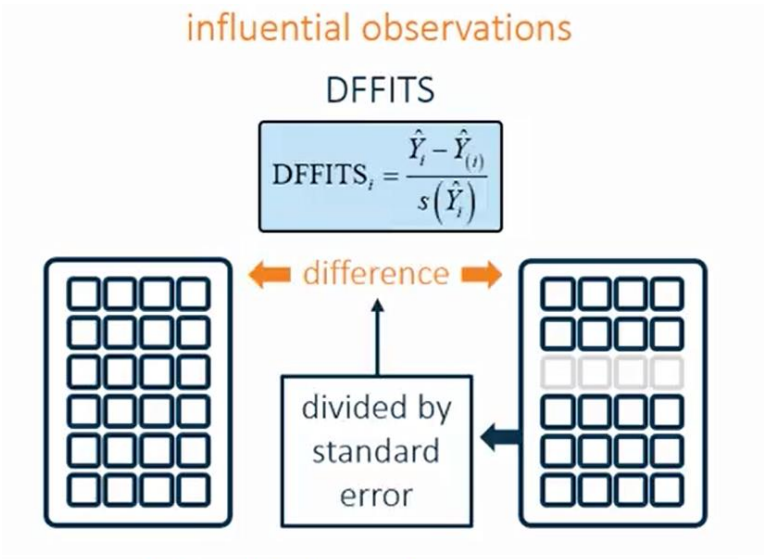
Cook's *D* statistic

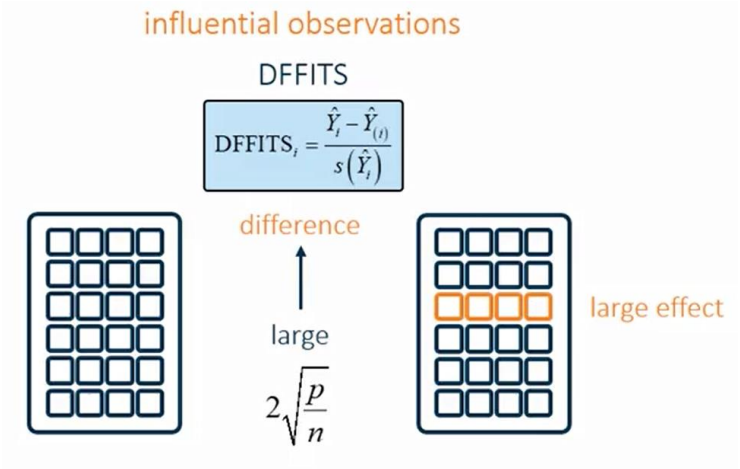
explanatory models
for parameter estimation

influential observations

Cook's *D* statistic

influential → Cook's $D_i > \frac{4}{n}$

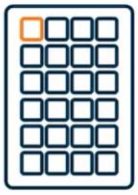




DFBETAS
difference in betas

general cutoff:
2

size-adjusted cutoff:
$2\sqrt{\frac{1}{n}}$



large DFBETA → influential observation

```
/* save the list of interval variables in a macro */
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;


/* */
ods select none;
proc glmselect data=STAT1.ameshousing3 plots=all;
     STEPWISE: model SalePrice = &interval / selection=stepwise
details=steps select=SL slentry=0.05 slstay=0.05;
     title "Stepwise Model Selection for SalePrice - SL 0.05";
run;
quit;
ods select all;


ods graphics on;
ods output RSTUDENTBYPREDICTED=Rstud
           COOKSDPLOT=Cook
           DFFITSPLOT=Dffits
           DFBETASPANEL=Dfbs;
proc reg data=STAT1.ameshousing3
         plots(only label)=
              (RSTUDENTBYPREDICTED
               COOKSD
               DFFITS
               DFBETAS);
    SigLimit: model SalePrice = &_GLSIND;
    title 'SigLimit Model - Plots of Diagnostic Statistics';
run;
quit;
```

SigLimit Model - Plots of Diagnostic Statistics

The REG Procedure
Model: SigLimit
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	300
Number of Observations Used	300

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	3.424508E11	48921543221	176.86	<.0001
Error	292	80772716963	276618894		
Corrected Total	299	4.232235E11			

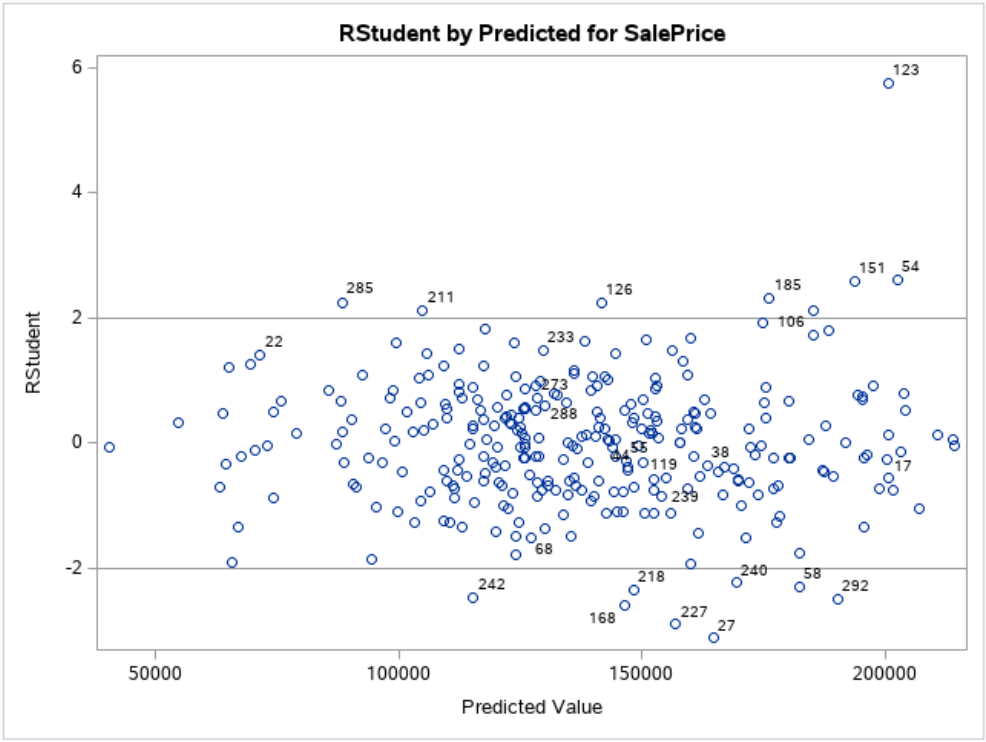
Root MSE	16632	R-Square	0.8091
Dependent Mean	137525	Adj R-Sq	0.8046
Coeff Var	12.09371		

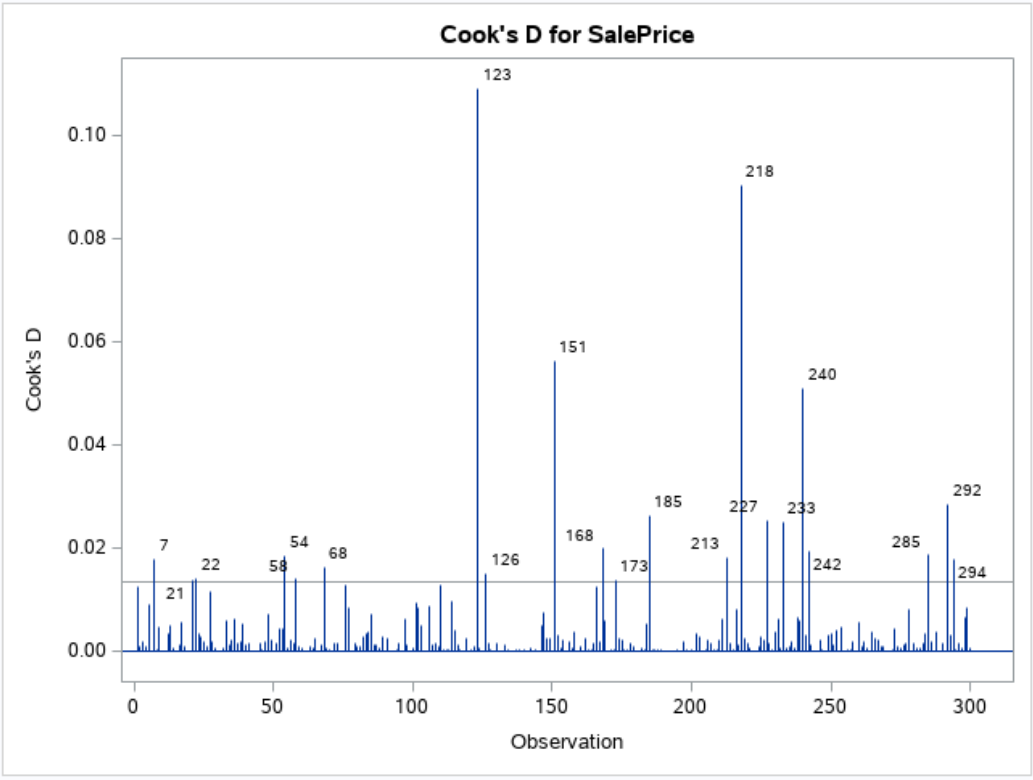
Parameter Estimates

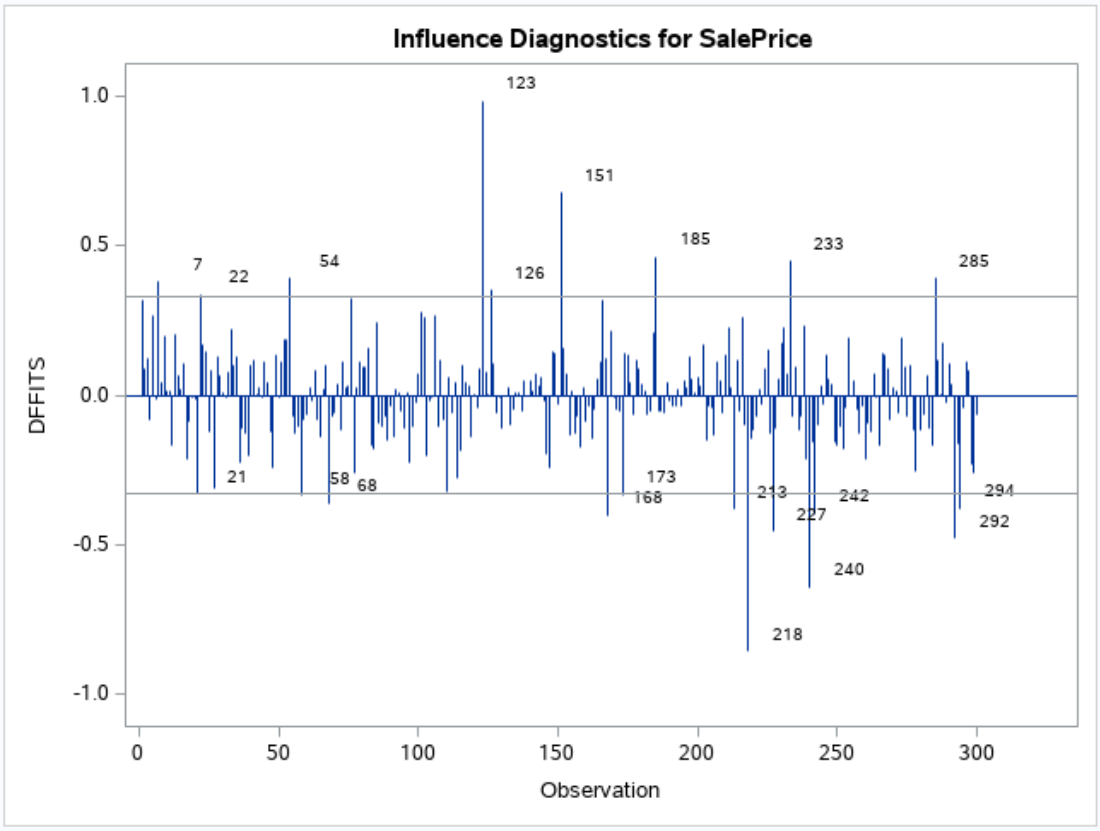
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	47463	5880.67404	8.07	<.0001
Gr_Liv_Area	Above grade (ground) living area square feet	1	65.30372	5.43667	12.01	<.0001
Basement_Area	Basement area in square feet	1	29.84908	3.34540	8.92	<.0001
Garage_Area	Size of garage in square feet	1	36.30961	6.45241	5.63	<.0001
Deck_Porch_Area	Total area of decks and porches in square feet	1	32.05255	7.96768	4.02	<.0001
Lot_Area	Lot size in square feet	1	0.70813	0.31751	2.23	0.0265
Age_Sold	Age of house when sold, in years	1	-447.19868	41.01931	-10.90	<.0001
Bedroom_AbvGr	Bedrooms above grade	1	-5042.76650	1687.92817	-2.99	0.0031

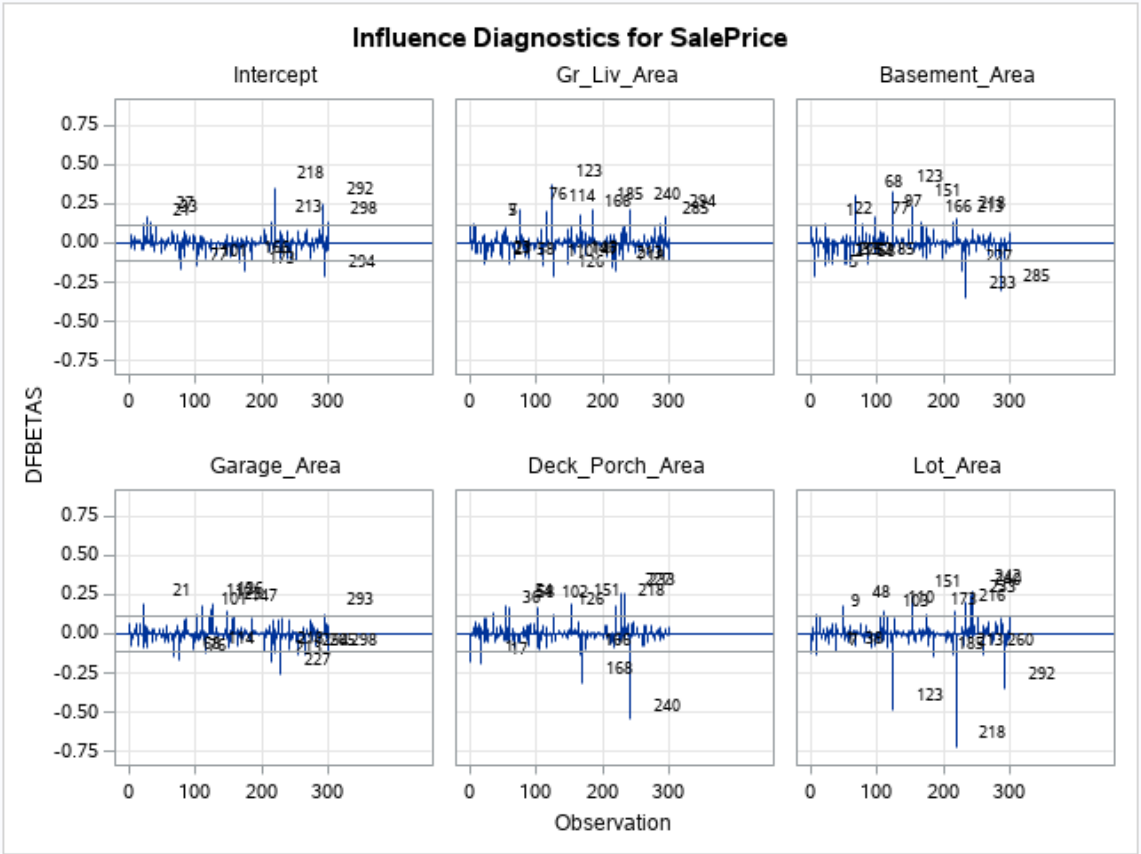
SigLimit Model - Plots of Diagnostic Statistics

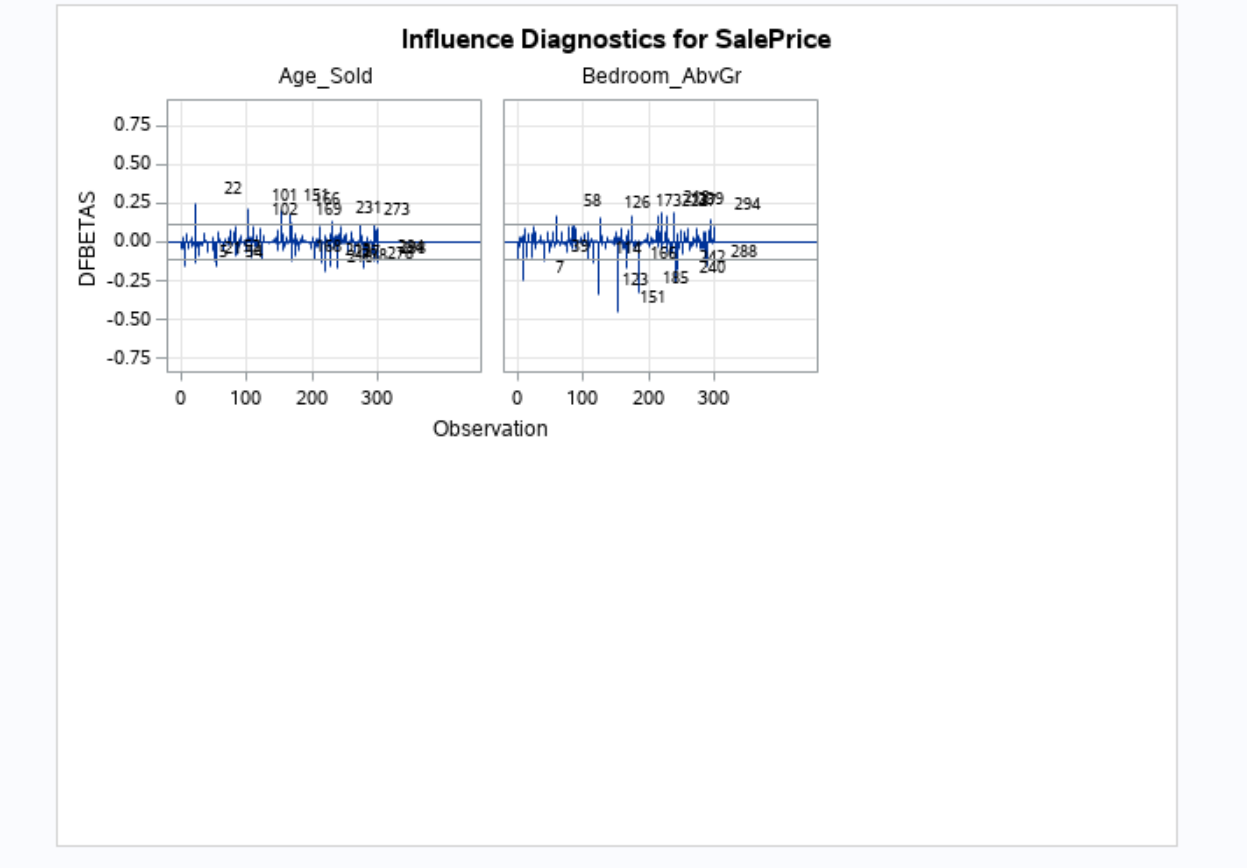
The REG Procedure
Model: SigLimit
Dependent Variable: SalePrice Sale price in dollars






Influence Diagnostics for SalePrice
DFFITS
Observation
1.0
0.5
0.0
-0.5
-1.0
0
50
100
150
200
250
300
123
151
185
233
285
7
22
54
126
21
58
68
173
168
213
227
242
294
292
240
218


Influence Diagnostics for SalePrice
Intercept
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
DFBETAS
Observation
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
0
100
200
300
218
292
213
298
294
123
76
114
185
240
68
151
122
166
233
285
21
293
227
102
168
9
48
260
292
123
218



/* Let's examine the influential observations using PROC PRINT statements */

```
title;
proc print data=Rstud;
run;
```

here's a sample of the output:

Obs	Model	Dependent	RStudent	PredictedValue	outLevLabel	Observation
1	SigLimit	SalePrice	1.73092	185283.46	.	1
2	SigLimit	SalePrice	0.67964	180284.34	.	2
3	SigLimit	SalePrice	0.63948	104541.46	.	3
4	SigLimit	SalePrice	-0.58261	169597.56	.	4
5	SigLimit	SalePrice	1.32153	158490.53	.	5
6	SigLimit	SalePrice	-0.05738	125932.40	.	6
7	SigLimit	SalePrice	1.92473	174736.56	.	7
8	SigLimit	SalePrice	0.36232	153008.35	.	8
9	SigLimit	SalePrice	1.47568	156222.61	.	9
10	SigLimit	SalePrice	0.10202	140446.83	.	10
11	SigLimit	SalePrice	0.15571	125421.52	.	11
12	SigLimit	SalePrice	-1.12570	150512.33	.	12
13	SigLimit	SalePrice	1.65684	150725.00	.	13
14	SigLimit	SalePrice	0.54343	126013.31	.	14
15	SigLimit	SalePrice	0.15339	151460.95	.	15
16	SigLimit	SalePrice	0.55394	125741.85	.	16
17	SigLimit	SalePrice	-0.56164	200736.13	17	17
18	SigLimit	SalePrice	-0.64482	90640.01	.	18
19	SigLimit	SalePrice	-0.06749	120117.08	.	19
20	SigLimit	SalePrice	-0.05783	124845.09	.	20
21	SigLimit	SalePrice	-1.47822	123941.30	.	21
22	SigLimit	SalePrice	1.40038	71388.82	22	22
23	SigLimit	SalePrice	1.24923	117326.83	.	23
24	SigLimit	SalePrice	0.62756	147826.80	.	24
25	SigLimit	SalePrice	-0.82201	173773.56	.	25
26	SigLimit	SalePrice	0.65376	175204.37	.	26
27	SigLimit	SalePrice	-3.10785	164685.93	27	27
28	SigLimit	SalePrice	1.15239	135960.27	.	28
29	SigLimit	SalePrice	0.46690	151302.69	.	29
30	SigLimit	SalePrice	0.05430	142857.51	.	30
31	SigLimit	SalePrice	-0.05717	143939.38	.	31
32	SigLimit	SalePrice	0.53249	204219.38	.	32
33	SigLimit	SalePrice	1.21513	65033.63	.	33
34	SigLimit	SalePrice	0.81924	112472.69	.	34
35	SigLimit	SalePrice	0.88479	175434.69	.	35

```
proc print data=Cook;
run;
```

here's a sample of the output:

Obs	Model	Dependent	CooksD	Observation	CooksDLabel
1	SigLimit	SalePrice	0.01260	1	.
2	SigLimit	SalePrice	0.00102	2	.
3	SigLimit	SalePrice	0.00186	3	.
4	SigLimit	SalePrice	0.00092	4	.
5	SigLimit	SalePrice	0.00904	5	.
6	SigLimit	SalePrice	0.00002	6	.
7	SigLimit	SalePrice	0.01782	7	7
8	SigLimit	SalePrice	0.00024	8	.
9	SigLimit	SalePrice	0.00486	9	.
10	SigLimit	SalePrice	0.00003	10	.
11	SigLimit	SalePrice	0.00004	11	.
12	SigLimit	SalePrice	0.00347	12	.
13	SigLimit	SalePrice	0.00501	13	.
14	SigLimit	SalePrice	0.00052	14	.
15	SigLimit	SalePrice	0.00004	15	.
16	SigLimit	SalePrice	0.00142	16	.
17	SigLimit	SalePrice	0.00577	17	.
18	SigLimit	SalePrice	0.00094	18	.
19	SigLimit	SalePrice	0.00001	19	.
20	SigLimit	SalePrice	0.00002	20	.
21	SigLimit	SalePrice	0.01368	21	21
22	SigLimit	SalePrice	0.01406	22	22
23	SigLimit	SalePrice	0.00350	23	.
24	SigLimit	SalePrice	0.00270	24	.
25	SigLimit	SalePrice	0.00196	25	.
26	SigLimit	SalePrice	0.00088	26	.
27	SigLimit	SalePrice	0.01176	27	.
28	SigLimit	SalePrice	0.00202	28	.
29	SigLimit	SalePrice	0.00056	29	.
30	SigLimit	SalePrice	0.00001	30	.
31	SigLimit	SalePrice	0.00001	31	.
32	SigLimit	SalePrice	0.00070	32	.
33	SigLimit	SalePrice	0.00612	33	.
34	SigLimit	SalePrice	0.00132	34	.
35	SigLimit	SalePrice	0.00211	35	.

```
proc print data=Dffits;
run;
```

Obs	Model	Dependent	Observation	DFFITS	DFFITSOUT
1	SigLimit	SalePrice	1	0.31861	.
2	SigLimit	SalePrice	2	0.09029	.
3	SigLimit	SalePrice	3	0.12177	.
4	SigLimit	SalePrice	4	-0.08573	.
5	SigLimit	SalePrice	5	0.26928	.
6	SigLimit	SalePrice	6	-0.01301	.
7	SigLimit	SalePrice	7	.	0.37928
8	SigLimit	SalePrice	8	0.04366	.
9	SigLimit	SalePrice	9	0.19752	.
10	SigLimit	SalePrice	10	0.01636	.
11	SigLimit	SalePrice	11	0.01720	.
12	SigLimit	SalePrice	12	-0.16660	.
13	SigLimit	SalePrice	13	0.20071	.
14	SigLimit	SalePrice	14	0.06417	.
15	SigLimit	SalePrice	15	0.01742	.
16	SigLimit	SalePrice	16	0.10651	.
17	SigLimit	SalePrice	17	-0.21458	.
18	SigLimit	SalePrice	18	-0.08653	.
19	SigLimit	SalePrice	19	-0.00777	.
20	SigLimit	SalePrice	20	-0.01146	.
21	SigLimit	SalePrice	21	.	-0.33145
22	SigLimit	SalePrice	22	.	0.33593
23	SigLimit	SalePrice	23	0.16745	.
24	SigLimit	SalePrice	24	0.14693	.
25	SigLimit	SalePrice	25	-0.12508	.
26	SigLimit	SalePrice	26	0.08382	.
27	SigLimit	SalePrice	27	-0.31129	.
28	SigLimit	SalePrice	28	0.12730	.
29	SigLimit	SalePrice	29	0.06685	.
30	SigLimit	SalePrice	30	0.00901	.
31	SigLimit	SalePrice	31	-0.00958	.
32	SigLimit	SalePrice	32	0.07492	.
33	SigLimit	SalePrice	33	0.22141	.
34	SigLimit	SalePrice	34	0.10270	.
35	SigLimit	SalePrice	35	0.12977	.

```
proc print data=Dfbs;
run;
```

Obs	Model	Dependent	Observation	_DFBETAS1	_DFBETASOUT1	_DFBETAS2	_DFBETASOUT2	_DFBETAS3	_DFBETASOUT3	_DFBETAS4	_DFBETASOUT4	_DFBETAS5	_DFBETASOUT5	_DFBETAS6	_DFBETASOUT6	_DFBETAS7	_DFBETASOUT7	_DFBETAS8	_DFBETASOUT8
1	SigLimit	SalePrice	1	-0.00567	.	0.10783	.	.	0.11744	0.07141	.	.	-0.18180	.	-0.12067	.	.	.	.
2	SigLimit	SalePrice	2	0.00677	.	0.02128	.	0.03237	.	0.00993	.	0.00943	.	-0.04517	.	.	.	.	.
3	SigLimit	SalePrice	3	0.05218	.	-0.04403	.	0.01625	.	-0.08221	.	-0.01416	.	0.03778	.	.	.	.	.
4	SigLimit	SalePrice	4	-0.03621	.	0.00052	.	-0.01769	.	0.00315	.	-0.01452	.	0.05082	.	.	.	.	.
5	SigLimit	SalePrice	5	0.02059	.	.	0.12008	.	-0.21199	-0.03160	.	0.04749	.	0.01234	.	.	.	.	.
6	SigLimit	SalePrice	6	-0.00736	.	0.00672	.	0.00198	.	0.00249	.	-0.01051	.	-0.00022	.	.	.	.	.
7	SigLimit	SalePrice	7	0.04779	.	.	0.11635	0.10354	.	0.01939	.	0.08053	.	.	-0.13126	.	.	.	.
8	SigLimit	SalePrice	8	-0.01381	.	-0.00500	.	0.02086	.	0.01649	.	-0.02605	.	0.00021	.	.	.	.	.
9	SigLimit	SalePrice	9	-0.04188	.	-0.07039	.	0.02326	.	0.03276	.	0.06244	.	.	0.12030	.	.	.	.
10	SigLimit	SalePrice	10	0.00188	.	-0.00134	.	0.00327	.	-0.00808	.	0.01138	.	0.00328	.	.	.	.	.
11	SigLimit	SalePrice	11	-0.00046	.	-0.00547	.	0.00843	.	-0.00308	.	-0.00783	.	0.00258	.	.	.	.	.
12	SigLimit	SalePrice	12	0.01953	.	-0.07213	.	-0.09553	.	0.06919	.	0.04227	.	0.00780	.	.	.	.	.
13	SigLimit	SalePrice	13	0.02733	.	0.01455	.	0.03310	.	-0.07737	.	0.06917	.	0.10997	.	.	.	.	.
14	SigLimit	SalePrice	14	0.00592	.	0.01107	.	0.00854	.	-0.00881	.	-0.03227	.	0.03467	.	.	.	.	.
15	SigLimit	SalePrice	15	-0.00374	.	-0.00114	.	0.00949	.	-0.00466	.	0.00569	.	-0.00594	.	.	.	.	.
16	SigLimit	SalePrice	16	0.00424	.	-0.07299	.	0.00809	.	0.06251	.	0.04623	.	-0.03030	.	.	.	.	.
17	SigLimit	SalePrice	17	0.01850	.	-0.00557	.	-0.00275	.	0.02321	.	.	-0.19555	-0.00443	.	.	.	.	.
18	SigLimit	SalePrice	18	-0.04906	.	0.04644	.	-0.03161	.	0.00269	.	0.01831	.	0.03320	.	.	.	.	.
19	SigLimit	SalePrice	19	0.00237	.	-0.00275	.	-0.00116	.	0.00153	.	0.00324	.	0.00233	.	.	.	.	.
20	SigLimit	SalePrice	20	0.00036	.	0.00153	.	0.00350	.	-0.00743	.	-0.00589	.	0.00253	.	.	.	.	.
21	SigLimit	SalePrice	21	.	0.12113	.	-0.13573	.	-0.14695	.	0.18720	0.09741	.	-0.02495	.	.	.	.	.
22	SigLimit	SalePrice	22	-0.04997	.	0.05380	.	.	0.12409	-0.08686	.	-0.05163	.	-0.04014	.	.	.	.	.
23	SigLimit	SalePrice	23	0.04449	.	.	-0.11834	0.02931	.	0.05758	.	-0.04330	.	-0.02803	.	.	.	.	.
24	SigLimit	SalePrice	24	-0.01389	.	-0.00640	.	-0.06488	.	-0.02513	.	0.10459	.	-0.00082	.	.	.	.	.
25	SigLimit	SalePrice	25	0.04655	.	-0.02173	.	-0.05301	.	-0.05341	.	0.07340	.	0.04282	.	.	.	.	.
26	SigLimit	SalePrice	26	0.01347	.	-0.01421	.	0.00902	.	-0.00776	.	-0.01001	.	0.01773	.	.	.	.	.
27	SigLimit	SalePrice	27	.	0.16637	-0.00213	.	.	-0.12998	-0.09471	.	0.03141	.	-0.06843	.	.	.	.	.
28	SigLimit	SalePrice	28	0.00883	.	-0.05857	.	0.03687	.	0.03968	.	-0.05989	.	-0.00858	.	.	.	.	.
29	SigLimit	SalePrice	29	0.02643	.	-0.01986	.	-0.01107	.	0.02597	.	-0.01415	.	0.02973	.	.	.	.	.
30	SigLimit	SalePrice	30	0.00897	.	-0.00145	.	0.00017	.	0.00014	.	0.00013	.	-0.00160	.	.	.	.	.
31	SigLimit	SalePrice	31	-0.00744	.	0.00164	.	-0.00003	.	-0.00006	.	-0.00089	.	0.00169	.	.	.	.	.
32	SigLimit	SalePrice	32	-0.02439	.	0.01977	.	0.02466	.	0.01600	.	-0.00981	.	-0.01729	.	.	.	.	.
33	SigLimit	SalePrice	33	.	0.13368	-0.08691	.	.	-0.13475	0.01229	.	0.00667	.	0.06850	.	.	.	.	.
34	SigLimit	SalePrice	34	0.02450	.	-0.05775	.	0.02738	.	0.03622	.	0.02790	.	-0.00636	.	.	.	.	.
35	SigLimit	SalePrice	35	-0.02052	.	-0.02248	.	0.01275	.	0.00427	.	0.10276	.	-0.00402	.	.	.	.	.
36	SigLimit	SalePrice	36	0.06119	.	-0.05003	.	-0.07012	.	-0.05777	.	.	0.13675	-0.00389	.	.	.	.	.
37	SigLimit	SalePrice	37	-0.05498	.	0.03200	.	-0.02248	.	0.02312	.	-0.05841	.	0.06573	.	.	.	.	.
38	SigLimit	SalePrice	38	0.01489	.	0.00953	.	0.00890	.	0.03255	.	-0.02709	.	.	-0.11852	.	.	.	.
39	SigLimit	SalePrice	39	0.10779	.	0.01382	.	-0.08020	.	-0.04580	.	0.03985	.	0.03038	.	.	.	.	.
40	SigLimit	SalePrice	40	0.00131	.	0.05138	.	-0.08089	.	-0.01319	.	0.01907	.	0.01138	.	.	.	.	.

/* save the first 300 observations in this dataset */

```
data Dfbs01;
      set Dfbs (obs=300);
run;
```

/* save the remaining observations in this dataset */

```
data Dfbs02;
      set Dfbs (firstobs=301);
run;
```

/* combine both datasets */

```
data Dfbs2;
      update Dfbs01 Dfbs02;
      by Observation;
run;
```

```
/*  Merge datasets from above.*/
data influential;
    merge Rstud
          Cook
          Dffits
             Dfbs2;
    by observation;

/*  Flag observations that have exceeded at least one cutpoint;*/
    if (ABS(Rstudent)>3) or (Cooksdlabel ne ' ') or Dffitsout then
flag=1;
    array dfbetas{*} _dfbetasout: ;
    do i=2 to dim(dfbetas);
         if dfbetas{i} then flag=1;
    end;

/*  Set to missing values of influence statistics for those*/
/*  that have not exceeded cutpoints;*/
    if ABS(Rstudent)<=3 then RStudent=.;
    if Cooksdlabel eq ' ' then CooksD=.;

/*  Subset only observations that have been flagged.*/
    if flag=1;
    drop i flag;
run;

title;
proc print data=influential;
    id observation;
    var Rstudent CooksD Dffitsout _dfbetasout:;
run;
```

Observation	RStudent	CooksD	DFFITSOUT	_DFBETASOUT1	_DFBETASOUT2	_DFBETASOUT3	_DFBETASOUT4	_DFBETASOUT5	_DFBETASOUT6	_DFBETASOUT7	_DFBETASOUT8
1	.	.	.	.	.	0.11744	.	-0.18180	-0.12067	.	.
5	.	.	.	.	0.12008	-0.21199	.	.	.	-0.15614	.
7	.	0.01782	0.37928	.	0.11635	.	.	.	-0.13126	.	-0.25391
9	.	.	.	.	.	.	.	.	0.12030	.	.
17	.	.	.	.	.	.	.	-0.19555	.	.	.
21	.	0.01368	-0.33145	0.12113	-0.13573	-0.14695	0.18720	.	.	-0.13114	.
22	.	0.01406	0.33593	.	.	0.12409	.	.	.	0.25385	.
23	.	.	.	.	-0.11834	.	.	.	.	.	.
27	-3.10785	.	.	0.16637	.	-0.12996	.	.	.	.	.
33	.	.	.	0.13368	.	-0.13475	.	.	.	.	.
36	.	.	.	.	.	.	.	0.13675	.	.	.
38	.	.	.	.	.	.	.	.	-0.11852	.	.
39	.	.	.	.	.	.	.	.	.	.	-0.12765
48	.	.	.	.	.	.	.	.	0.17608	.	.
52	.	.	.	.	.	-0.13187	.	.	.	-0.12227	.
53	.	.	.	.	.	-0.13121	.	.	.	.	.
54	.	0.01861	0.38966	.	.	.	.	0.18598	.	-0.15989	.
58	.	0.01402	-0.33731	.	-0.13499	-0.14035	.	0.16875	.	.	0.16723
68	.	0.01642	-0.36316	.	.	0.30122	-0.15249	.	.	.	.
76	.	.	.	.	0.21746	.	-0.16784	.	.	.	.
77	.	.	.	-0.16620	.	0.12547	.	.	.	.	.
85	.	.	.	.	.	-0.13503	.	.	.	.	.
97	.	.	.	.	.	0.17242	.	.	.	.	.
101	.	.	.	-0.14484	.	.	0.12467	.	.	0.21130	.
102	.	.	.	.	.	.	.	0.17062	.	0.11704	.
103	.	.	.	.	.	.	.	.	0.11796	.	.
110	.	.	.	.	-0.14851	.	0.18387	.	0.14328	.	.
114	.	.	.	.	0.20243	.	-0.11961	.	.	.	-0.13709
123	5.74803	0.10918	0.98452	.	0.37049	0.33210	0.16428	.	-0.48778	.	-0.33805
126	.	0.01518	0.35093	.	-0.21801	.	0.19091	0.12337	.	.	0.16108
146	.	.	.	.	-0.11679	.	.	.	.	.	.
147	.	.	.	.	-0.13364	.	0.14746	.	.	.	.
151	.	0.05626	0.67740	.	.	0.24219	.	0.18916	0.24197	0.20531	-0.45423
166	.	.	.	-0.12630	0.17790	0.14104	.	-0.13577	.	0.18060	-0.17280
168	.	0.01995	-0.40345	.	.	.	.	-0.31234	.	-0.12216	.
169	.	.	.	.	.	.	.	.	.	0.12081	.
173	.	0.01381	-0.33360	-0.18701	.	.	.	.	0.12808	.	0.16800
185	.	0.02643	0.46326	.	0.21312	.	.	.	-0.15273	.	-0.33146
213	.	0.01817	-0.38308	0.14101	-0.16098	0.13936	-0.17657	.	-0.13738	-0.14126	0.17538
216	.	.	.	.	.	.	.	.	0.14638	.	.
218	.	0.09031	-0.85644	0.35641	-0.18290	0.15837	-0.12556	0.18696	-0.72465	-0.19044	0.19164
227	.	0.02530	-0.45550	.	.	-0.18188	-0.26007	0.25544	.	-0.16114	0.17296
231	.	.	.	.	.	.	.	.	.	0.13129	.
233	.	0.02512	0.44925	.	.	-0.34640	.	0.25701	0.20069	.	.
238	.	.	.	.	.	.	.	.	.	-0.17478	.
239	.	.	.	.	.	.	.	.	.	.	0.18814
240	.	0.05107	-0.64349	.	0.21804	.	.	-0.54816	0.25496	.	-0.26231
242	.	0.01928	-0.39614	.	.	.	.	.	0.27169	.	-0.18557
254	.	.	.	.	.	.	-0.13397	.	.	.	.
260	.	.	.	.	.	.	.	.	-0.13327	.	.
265	.	.	.	.	.	.	-0.13113	.	.	.	.
273	.	.	.	.	.	.	.	.	.	0.11745	.
278	.	.	.	.	.	.	.	.	.	-0.16699	.
285	.	0.01871	0.38963	.	0.12842	-0.30732	.	.	.	.	.
288	.	.	.	.	.	.	.	.	.	.	-0.16036
292	.	0.02839	-0.48084	0.25018	.	.	.	.	-0.34680	.	.
293	.	.	.	.	.	.	0.12390	.	.	.	.
294	.	0.01773	-0.37825	-0.21136	0.17478	.	.	.	.	-0.12004	0.14720
298	.	.	.	0.13082	.	.	-0.13526	.	.	-0.13275	.

16. Dealing with collinearity

We'll merge a new dataset into the existing one.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/* first, we must sort both the existing and new datasets, before we can merge them */
proc sort data=STAT1.ameshousing3 out=STAT1.ames_sorted;
    by PID;
run;
proc sort data=STAT1.amesaltuse;
    by PID;
run;

/* next, we can merge the two datasets */
data amescombined;
    merge STAT1.ames_sorted STAT1.amesaltuse;
    by PID;
run;

/* now let's create a correlations matrix for the variables in the combined datasets */
title;
proc corr data=amescombined nosimple;
    var &interval;
    with score;
run;
```

The CORR Procedure

1 With Variables:	score
8 Variables:	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Gr_Liv_Area	Basement_Area	Garage_Area	Deck_Porch_Area	Lot_Area	Age_Sold	Bedroom_AbvGr	Total_Bathroom
score	-0.61394 <.0001 300	-0.97894 <.0001 300	-0.38872 <.0001 300	-0.35979 <.0001 300	-0.29249 <.0001 300	0.39125 <.0001 300	-0.28357 <.0001 300	-0.51877 <.0001 300

/* There's collinearity issue for any variable for which VIF >= 10 */

```
proc reg data=amescombined;
    model SalePrice = &interval score / vif;
    title 'Collinearity Diagnostics';
run;
quit;
```

Collinearity Diagnostics

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	2930
Number of Observations Used	300
Number of Observations with Missing Values	2630

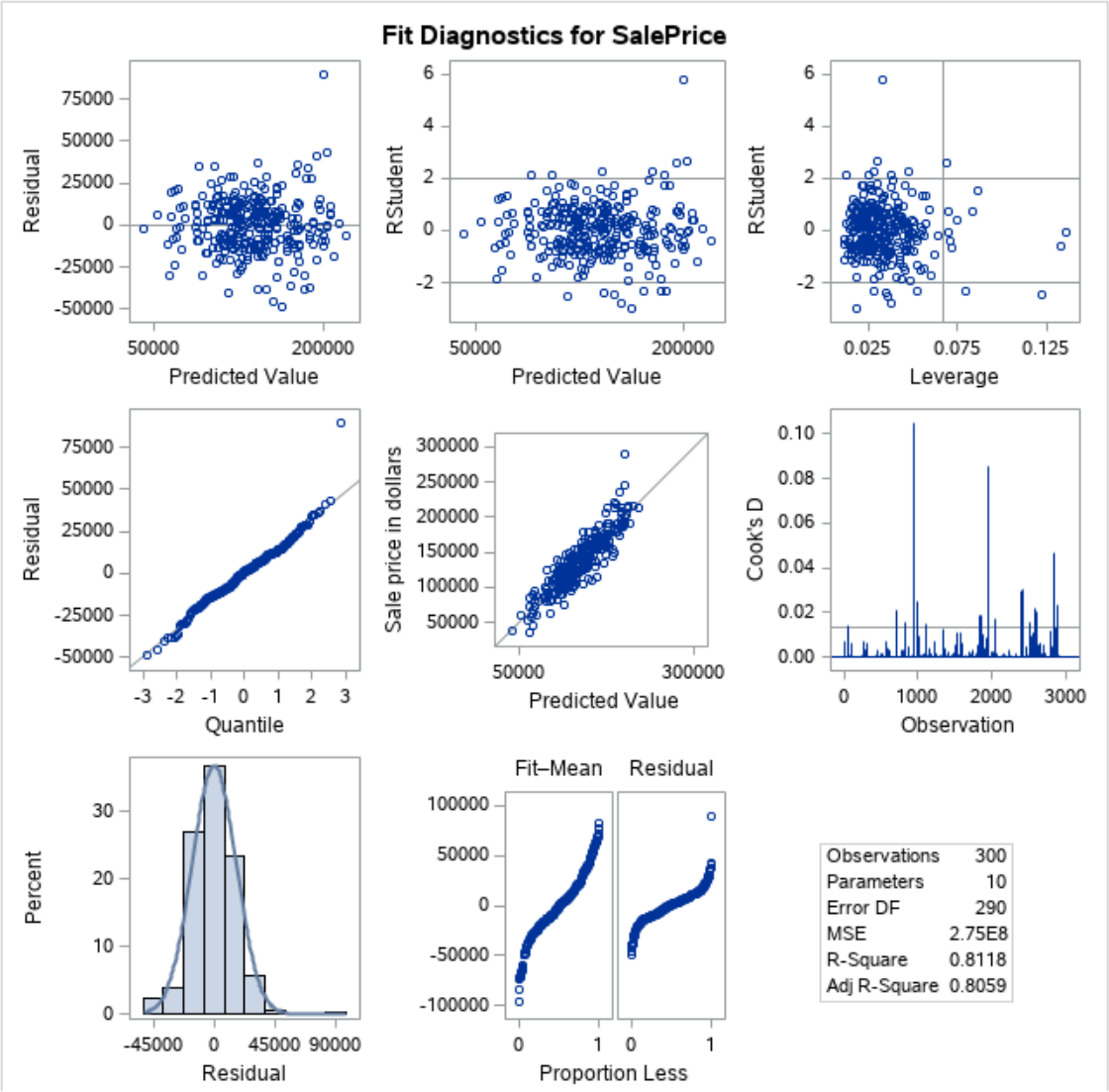
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	3.435663E11	38174038666	138.98	<.0001
Error	290	79657171514	274679902		
Corrected Total	299	4.232235E11			

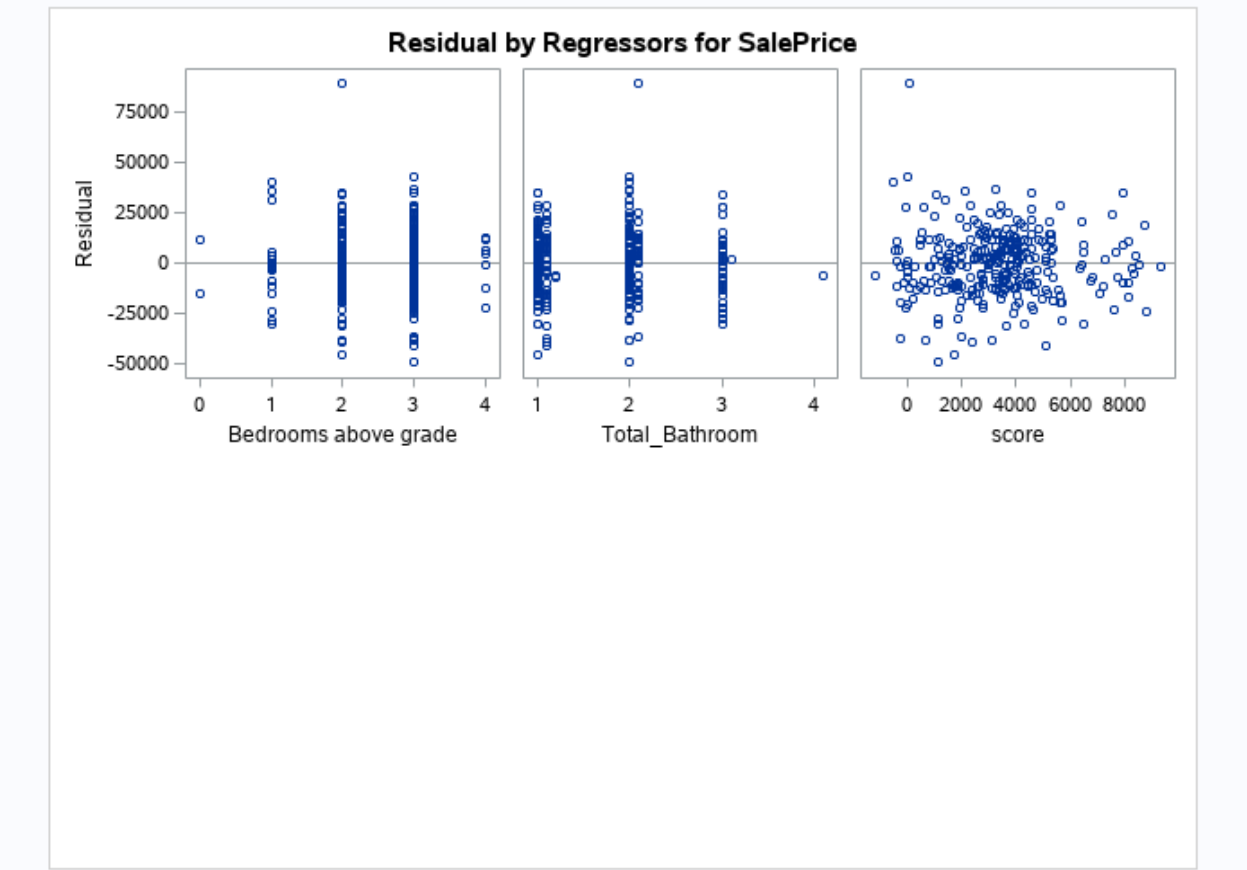
Root MSE	16573	R-Square	0.8118
Dependent Mean	137525	Adj R-Sq	0.8059
Coeff Var	12.05125		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-4254871	3419274	-1.24	0.2144	0
Gr_Liv_Area	Above grade (ground) living area square feet	1	923.25717	684.04818	1.35	0.1782	27569
Basement_Area	Basement area in square feet	1	2178.35638	1709.68175	1.27	0.2036	411868
Garage_Area	Size of garage in square feet	1	35.01213	6.46640	5.41	<.0001	1.41398
Deck_Porch_Area	Total area of decks and porches in square feet	1	30.64725	7.97228	3.84	0.0001	1.21667
Lot_Area	Lot size in square feet	1	0.69964	0.31644	2.21	0.0278	1.20422
Age_Sold	Age of house when sold, in years	1	-422.21228	44.18905	-9.55	<.0001	1.60476
Bedroom_AbvGr	Bedrooms above grade	1	-4888.35244	1687.71153	-2.90	0.0041	1.48233
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3047.94315	1919.03449	1.59	0.1133	1.73073
score		1	429.97552	341.96962	1.26	0.2096	533085

Collinearity Diagnostics

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice Sale price in dollars





/* We remove SCORE from the variables list and reran the VIF test */

```
proc reg data=amescombined;
    NOSCORE: model SalePrice = &interval / vif;
    title2 'Removing Score';
run;
quit;
```

Removing Score

The REG Procedure
Model: NOSCORE
Dependent Variable: SalePrice Sale price in dollars

Number of Observations Read	2930
Number of Observations Used	300
Number of Observations with Missing Values	2630

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	3.431321E11	42891512314	155.84	<.0001
Error	291	80091420996	275228251		
Corrected Total	299	4.232235E11			

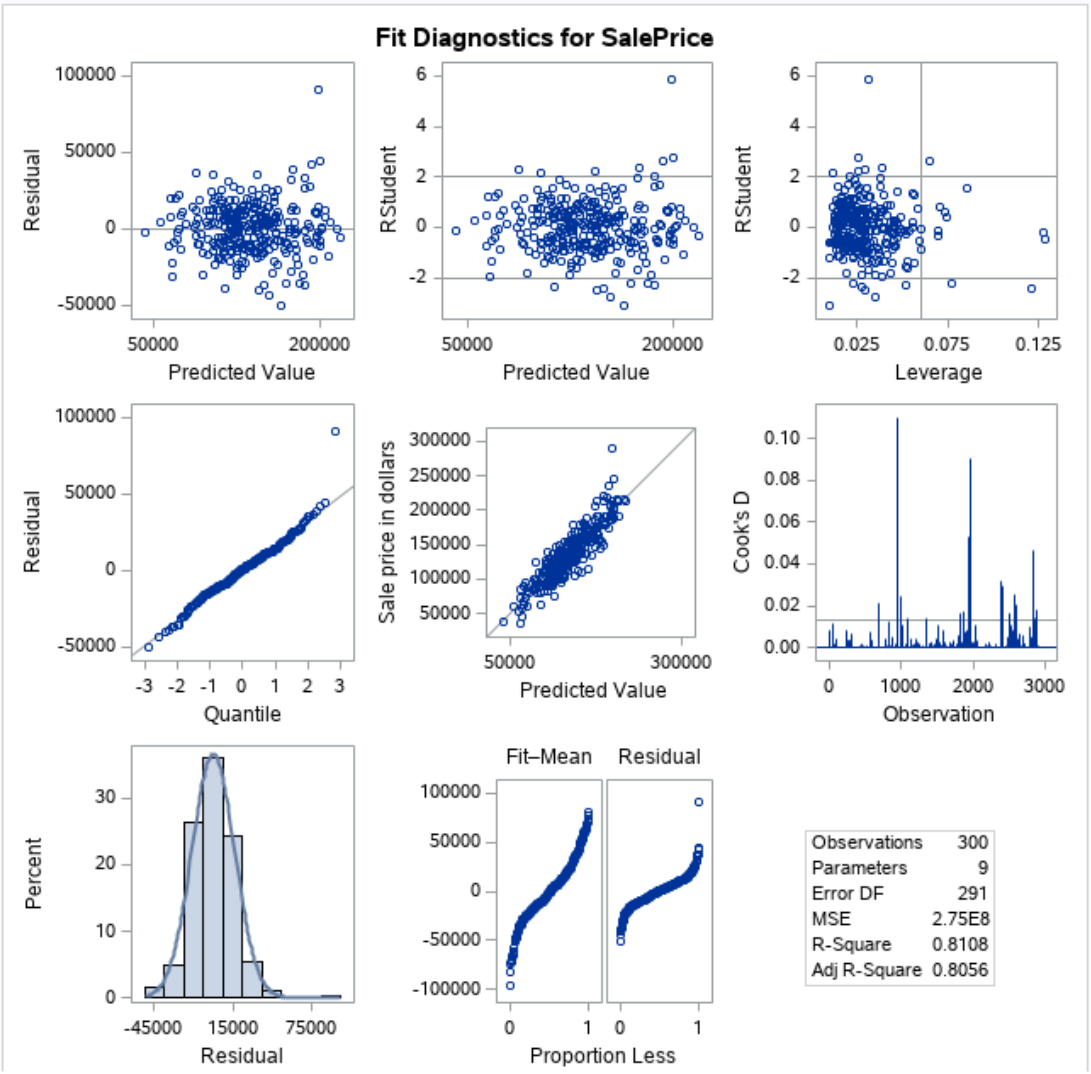
Root MSE	16590	R-Square	0.8108
Dependent Mean	137525	Adj R-Sq	0.8056
Coeff Var	12.06328		

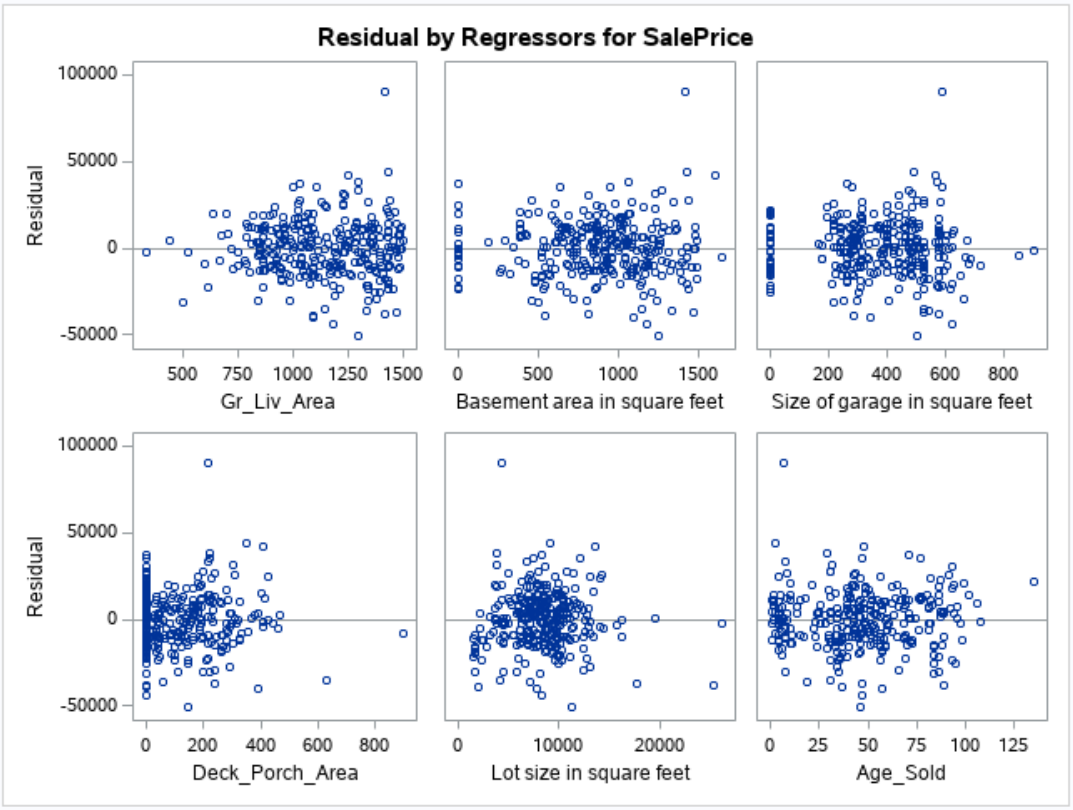
Parameter Estimates

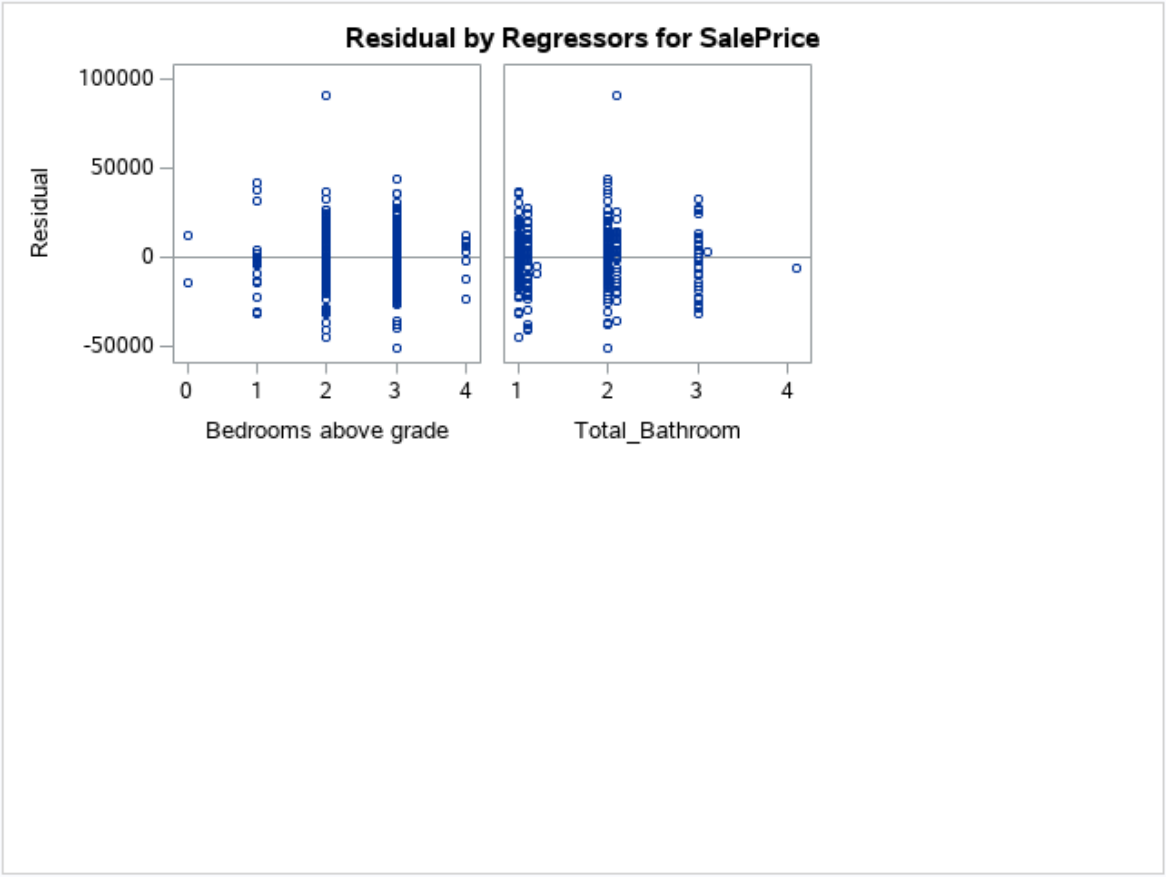
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	44347	6191.27194	7.16	<.0001	0
Gr_Liv_Area	Above grade (ground) living area square feet	1	63.19776	5.58574	11.31	<.0001	1.83461
Basement_Area	Basement area in square feet	1	28.69218	3.41703	8.40	<.0001	1.64195
Garage_Area	Size of garage in square feet	1	35.75419	6.44584	5.55	<.0001	1.40220
Deck_Porch_Area	Total area of decks and porches in square feet	1	31.37054	7.95944	3.94	0.0001	1.21034
Lot_Area	Lot size in square feet	1	0.69950	0.31676	2.21	0.0280	1.20422
Age_Sold	Age of house when sold, in years	1	-420.81504	44.21914	-9.52	<.0001	1.60375
Bedroom_AbvGr	Bedrooms above grade	1	-4834.84875	1688.85823	-2.86	0.0045	1.48138
Total_Bathroom	Total number of bathrooms (half bathrooms counted 10%)	1	3022.12472	1920.83907	1.57	0.1167	1.73053

Removing Score

The REG Procedure
Model: NOSCORE
Dependent Variable: SalePrice Sale price in dollars




Residual by Regressors for SalePrice
Residual
100000
50000
0
-50000
500 750 1000 1250 1500
Gr_Liv_Area
0 500 1000 1500
Basement area in square feet
0 200 400 600 800
Size of garage in square feet
0 200 400 600 800
Deck_Porch_Area
0 10000 20000
Lot size in square feet
0 25 50 75 100 125
Age_Sold


Residual by Regressors for SalePrice
Residual
100000
50000
0
-50000
0 1 2 3 4
Bedrooms above grade
1 2 3 4
Total_Bathroom

17. Building a predictive model

We use PROC GLMSELECT for this purpose.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
        Season_Sold Garage_Type_2 Foundation_2 Heating_QC
        Masonry_Veneer Lot_Shape_2 Central_Air;

ods graphics;

proc glmselect data=STAT1.ameshousing3
              plots=all
              valdata=STAT1.ameshousing4;
    class &categorical / param=glm ref=first;
    model SalePrice=&categorical &interval /
              selection=backward
              select=sbc
              choose=validate
              showpvalues;
    store out=STAT1.amesstore;
    title "Selecting the Best Model using Honest Assessment";
run;
```

Selecting the Best Model using Honest Assessment

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Validation Data Set	STAT1.AMESHOUSING4
Dependent Variable	SalePrice
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	SBC
Choose Criterion	Validation ASE
Effect Hierarchy Enforced	None

Observation Profile for Analysis Data	
Number of Observations Read	300
Number of Observations Used	294
Number of Observations Used for Training	294

Observation Profile for Validation Data	
Number of Observations Read	300
Number of Observations Used	293

Class Level Information		
Class	Levels	Values
House_Style2	5	1Story 2Story SFoyer SLvl 1.5Fin
Overall_Qual2	3	5 6 4
Overall_Cond2	3	5 6 4
Fireplaces	3	1 2 0
Season_Sold	4	2 3 4 1
Garage_Type_2	3	Detached NA Attached
Foundation_2	3	Cinder Block Concrete/Slab Brick/Tile/Stone
Heating_QC	4	Fa Gd TA Ex
Masonry_Veneer	2	Y N
Lot_Shape_2	2	Regular Irregular
Central_Air	2	Y N

Dimensions	
Number of Effects	20
Number of Parameters	43

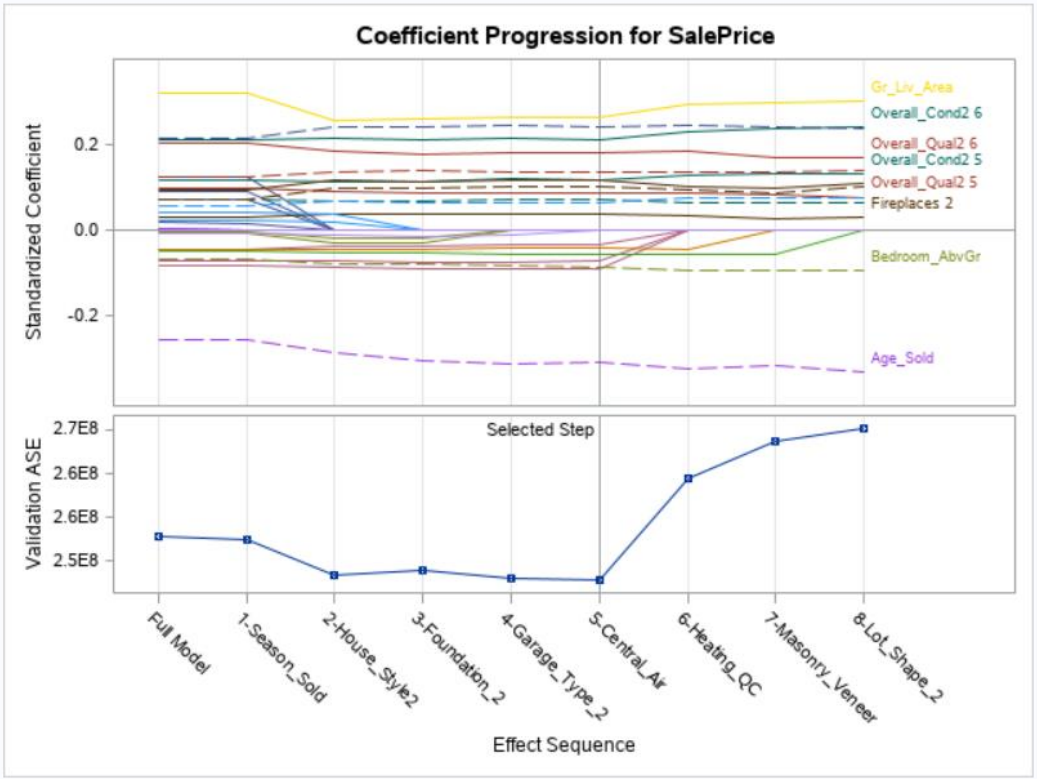
Selecting the Best Model using Honest Assessment

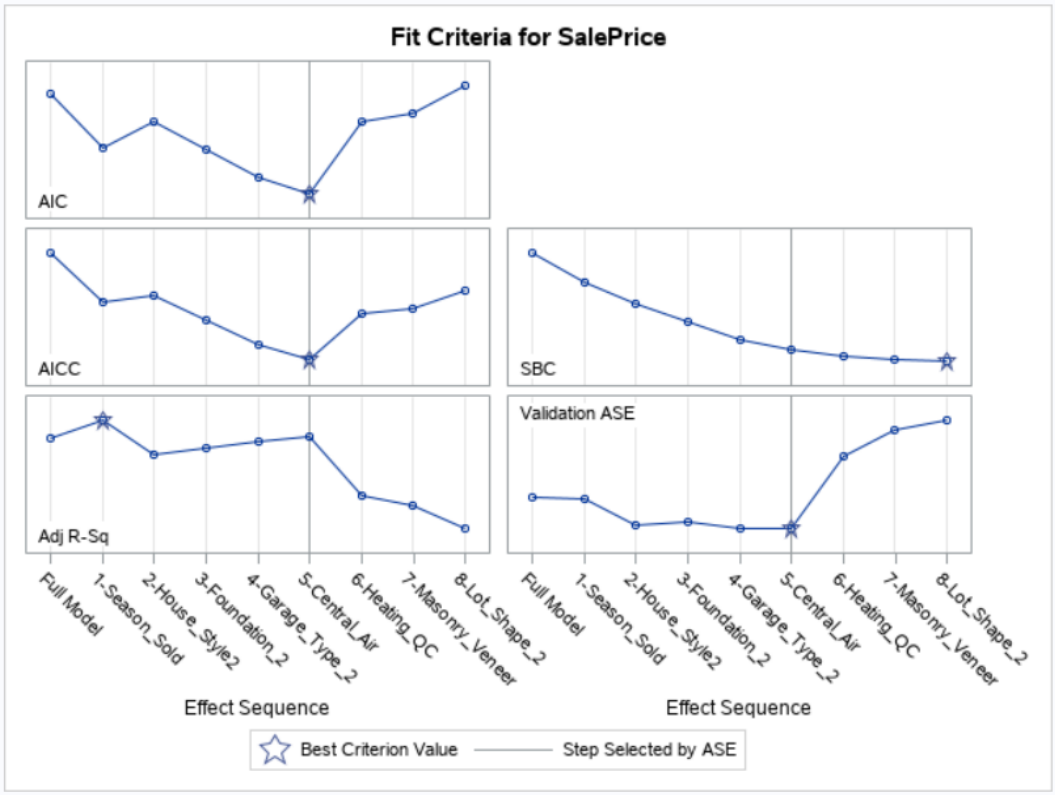
The GLMSELECT Procedure

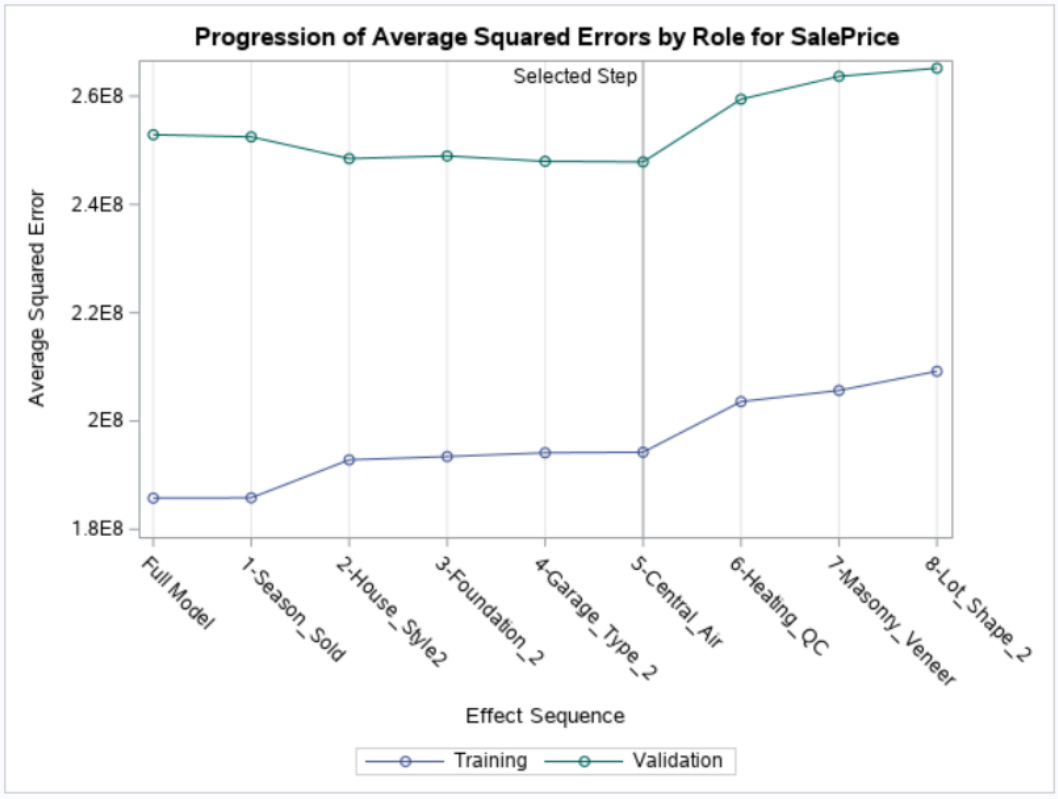
Backward Selection Summary						
Step	Effect Removed	Number Effects In	Number Parms In	SBC	ASE	Validation ASE
0		20	32	5779.6460	185773538	252878776
1	Season_Sold	19	29	5762.6753	185824120	252480746
2	House_Style2	18	25	5750.8247	192832172	248469026
3	Foundation_2	17	23	5740.3830	193440101	248951925
4	Garage_Type_2	16	21	5730.0735	194137231	247966687
5	Central_Air	15	20	5724.5490	194242334	247854963*
6	Heating_QC	14	17	5721.3123	203586891	259432895
7	Masonry_Veneer	13	16	5718.5873	205646000	263660934
8	Lot_Shape_2	12	15	5717.9317*	209193215	265159474
* Optimal Value of Criterion						

Selection stopped at a local minimum of the SBC criterion.

Stop Details				
Candidate For	Effect	Candidate SBC		Compare SBC
Removal	Deck_Porch_Area	5718.6683	>	5717.9317




Fit Criteria for SalePrice
AIC
AICC
SBC
Validation ASE
Adj R-Sq
Full Model
1-Season_Sold
2-House_Style2
3-Foundation_2
4-Garage_Type_2
5-Central_Air
6-Heating_QC
7-Masonry_Veneer
8-Lot_Shape_2
Effect Sequence
Effect Sequence
Best Criterion Value
Step Selected by ASE


Progression of Average Squared Errors by Role for SalePrice
Selected Step
Average Squared Error
2.6E8
2.4E8
2.2E8
2E8
1.8E8
Full Model
1-Season_Sold
2-House_Style2
3-Foundation_2
4-Garage_Type_2
5-Central_Air
6-Heating_QC
7-Masonry_Veneer
8-Lot_Shape_2
Effect Sequence
Training
Validation

Selecting the Best Model using Honest Assessment

The GLMSELECT Procedure
Selected Model

The selected model, based on Validation ASE, is the model at Step 5.

Effects:	Intercept Overall_Qual2 Overall_Cond2 Fireplaces Heating_QC Masonry_Veneer Lot_Shape_2 Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	19	3.566452E11	18770797693	90.06
Error	274	57107246191	208420607	
Corrected Total	293	4.137524E11		

Root MSE	14437
Dependent Mean	137179
R-Square	0.8620
Adj R-Sq	0.8524
AIC	5946.87742
AICC	5950.27448
SBC	5724.54902
ASE (Train)	194242334
ASE (Validate)	247854963

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	51207	7079.121457	7.23	<.0001
Overall_Qual2 5	1	6782.080263	3104.469941	2.18	0.0298
Overall_Qual2 6	1	13659	3414.565419	4.00	<.0001
Overall_Qual2 4	0	0	.	.	.
Overall_Cond2 5	1	8996.618020	4137.937302	2.17	0.0305
Overall_Cond2 6	1	15909	4025.283609	3.95	<.0001
Overall_Cond2 4	0	0	.	.	.
Fireplaces 1	1	9716.205925	2044.560791	4.75	<.0001
Fireplaces 2	1	7235.661619	4540.159269	1.59	0.1122
Fireplaces 0	0	0	.	.	.
Heating_QC Fa	1	-11668	4315.812370	-2.70	0.0073
Heating_QC Gd	1	-3178.918390	2496.841385	-1.27	0.2040
Heating_QC TA	1	-6689.247126	2133.424223	-3.14	0.0019
Heating_QC Ex	0	0	.	.	.
Masonry_Veneer Y	1	-3369.652622	2079.343731	-1.62	0.1063
Masonry_Veneer N	0	0	.	.	.
Lot_Shape_2 Regular	1	-4507.715447	2036.544994	-2.21	0.0277
Lot_Shape_2 Irregular	0	0	.	.	.
Gr_Liv_Area	1	42.972194	5.709351	7.53	<.0001
Basement_Area	1	25.491273	3.170869	8.04	<.0001
Garage_Area	1	29.698556	5.913131	5.02	<.0001
Deck_Porch_Area	1	20.952561	7.235245	2.90	0.0041

Lot_Area	1	1.199858	0.307660	3.90	0.0001
Age_Sold	1	-422.187733	47.675825	-8.86	<.0001
Bedroom_AbvGr	1	-4541.124997	1523.500120	-2.98	0.0031
Total_Bathroom	1	3806.351237	1714.333548	2.22	0.0272

18. Scoring new data

For this exercise, we use the validation dataset for scoring: 'ameshousing4'. In the business context, scoring takes place with newly unseen data.

/* First approach: using a PROC PLM for re-scoring */

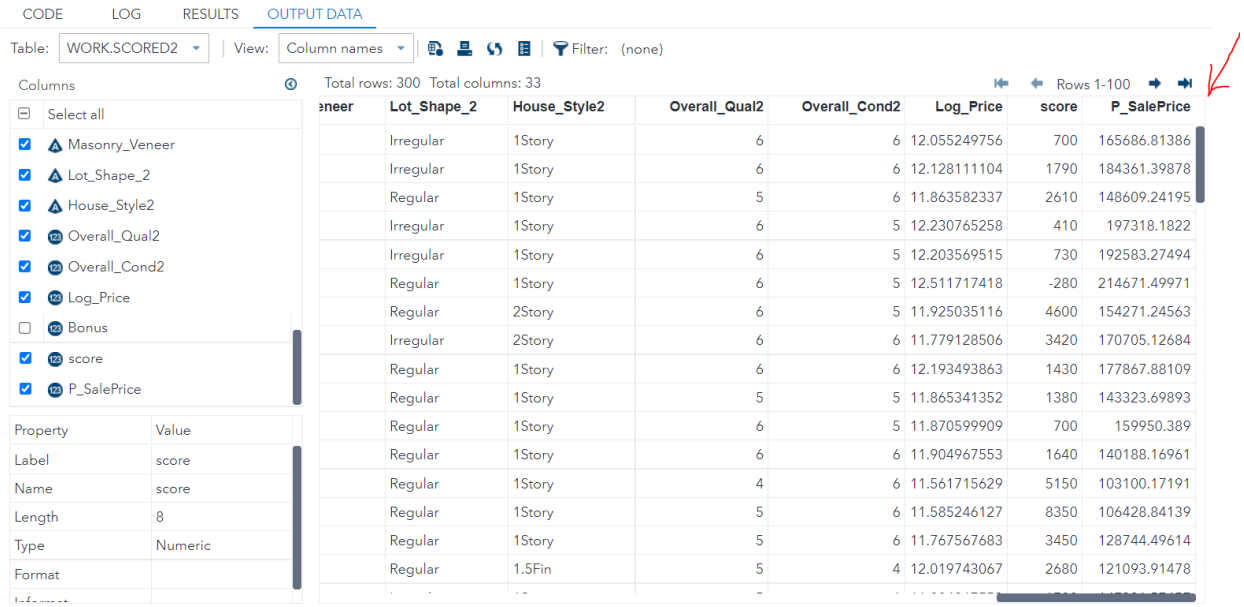
```
proc plm restore=STAT1.amesstore;
    score data=STAT1.ameshousing4 out=scored;
    code file="&homefolder\scoring.sas";
run;
```

The PLM Procedure

Store Information	
Item Store	STAT1.AMESSTORE
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLMSELECT
Date Created	22AUG21:00:30:01
Response Variable	SalePrice
Class Variables	House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces Season_Sold Garage_Type_2 Foundation_2 ...
Model Effects	Intercept Overall_Qual2 Overall_Cond2 Fireplaces Heating_QC Masonry_Veneer Lot_Shape_2 Gr_Liv_Are..

/* Second approach: using a DATA STEP for re-scoring */

```
data scored2;
    set STAT1.ameshousing4;
    %include "&homefolder\scoring.sas";
run;
```



/* Comparing the outcomes of both scoring approaches */

```
proc compare base=scored compare=scored2 criterion=0.0001;
    var Predicted;
    with P_SalePrice;
run;
```

The COMPARE Procedure
Comparison of WORK.SCORED with WORK.SCORED2
(Method=RELATIVE(2.22E-10), Criterion=0.0001)

Data Set Summary

Dataset	Created	Modified	NVar	NObs	Label
WORK.SCORED	22AUG21:00:43:28	22AUG21:00:43:28	33	300	Scoring Results for DATA=STAT1.AMESHOUSING4
WORK.SCORED2	22AUG21:00:45:47	22AUG21:00:45:47	33	300	

Variables Summary

Number of Variables in Common: 32.
Number of Variables in WORK.SCORED but not in WORK.SCORED2: 1.
Number of Variables in WORK.SCORED2 but not in WORK.SCORED: 1.
Number of VAR Statement Variables: 1.
Number of WITH Statement Variables: 1.

Observation Summary

Observation	Base	Compare
First Obs	1	1
Last Obs	300	300

Number of Observations in Common: 300.
Total Number of Observations Read from WORK.SCORED: 300.
Total Number of Observations Read from WORK.SCORED2: 300.

Number of Observations with Some Compared Variables Unequal: 0.
Number of Observations with All Compared Variables Equal: 300.

Values Comparison Summary

Number of Variables Compared with All Observations Equal: 1.
Number of Variables Compared with Some Observations Unequal: 0.
Total Number of Values which Compare Unequal: 0.
Total Number of Values not EXACTLY Equal: 297.
Maximum Difference Criterion Value: 2.3062E-15.

18. Categorical data analysis

Let's examine the distribution of bonus in the Ames housing dataset, by two of its predictors: fireplace, lot shape.

/* Let's format the bonus variable */

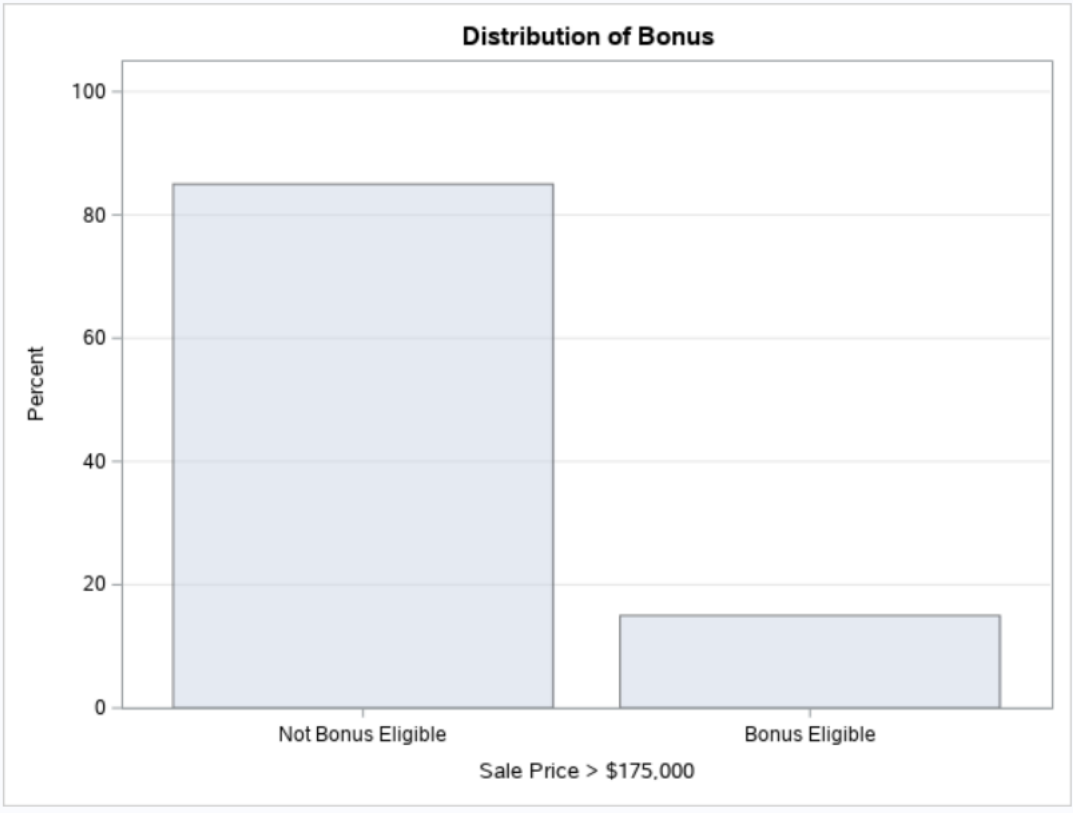
```
title;
proc format;
    value bonusfmt 1 = "Bonus Eligible"
                   0 = "Not Bonus Eligible"
                  ;
run;
```

/ * Let's create frequency tables */

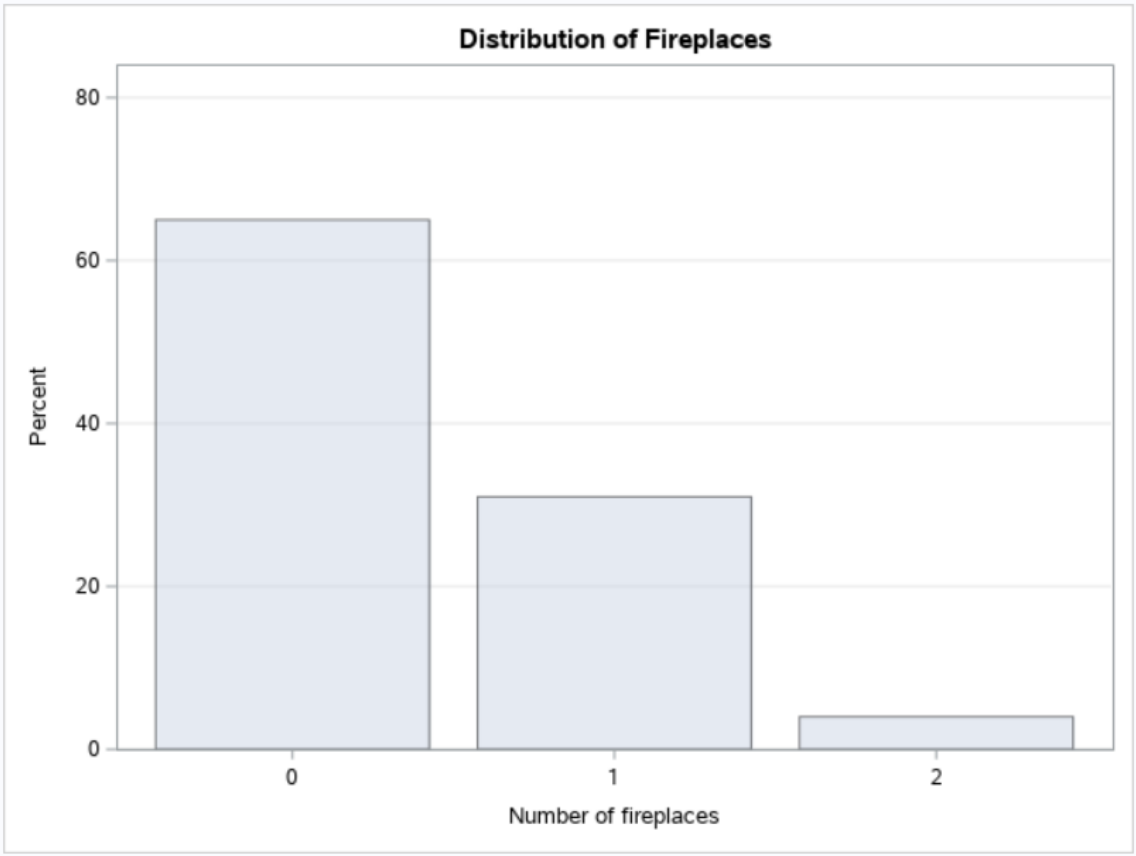
```
proc freq data=STAT1.ameshousing3;
    tables Bonus Fireplaces Lot_Shape_2
           Fireplaces*Bonus Lot_Shape_2*Bonus/
           plots(only)=freqplot(scale=percent);
    format Bonus bonusfmt.;
run;
```

The FREQ Procedure

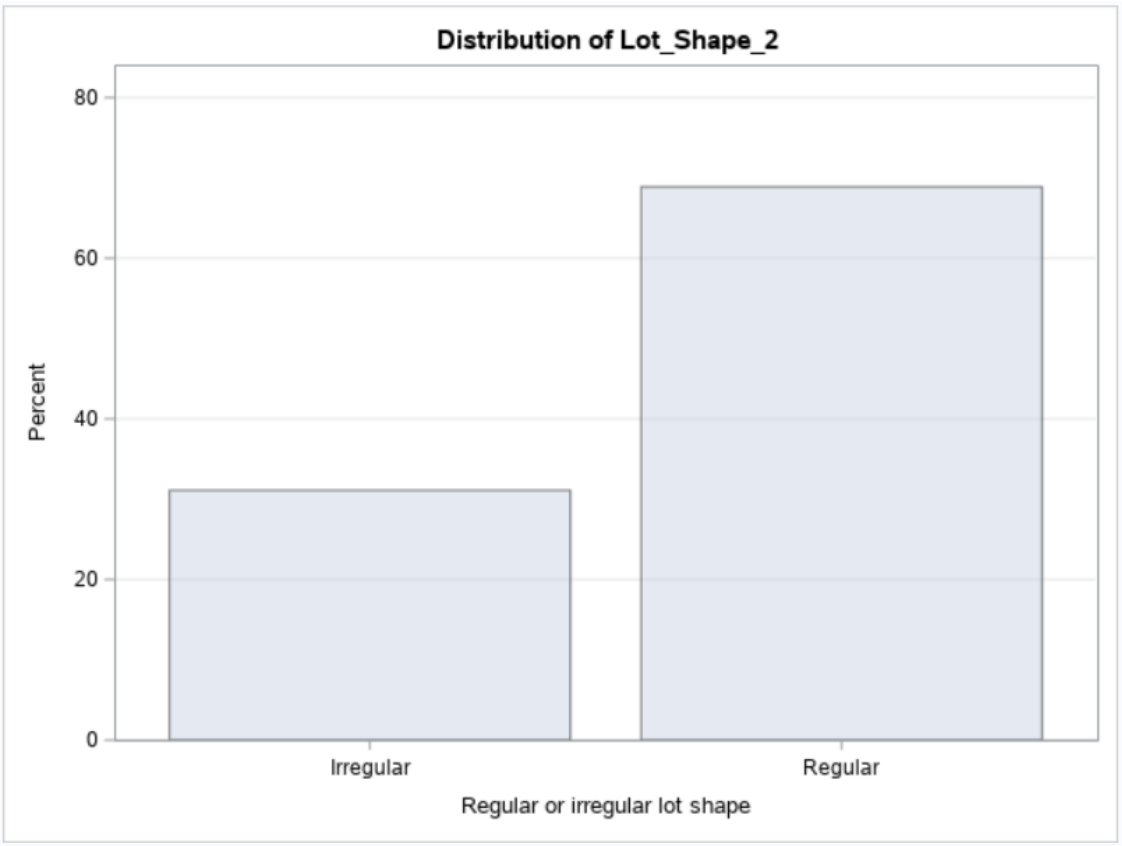
Sale Price > $175,000				
Bonus	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not Bonus Eligible	255	85.00	255	85.00
Bonus Eligible	45	15.00	300	100.00



Number of fireplaces				
Fireplaces	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	195	65.00	195	65.00
1	93	31.00	288	96.00
2	12	4.00	300	100.00

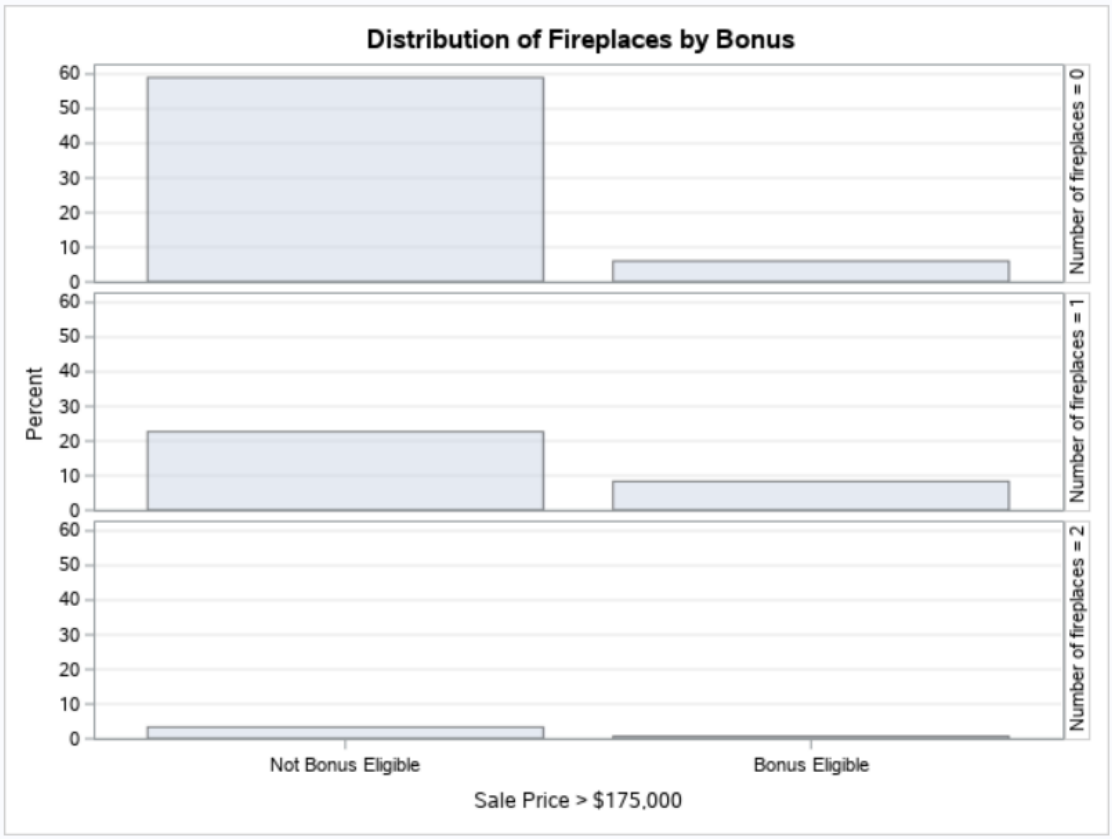


Regular or irregular lot shape				
Lot_Shape_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Irregular	93	31.10	93	31.10
Regular	206	68.90	299	100.00
Frequency Missing = 1				



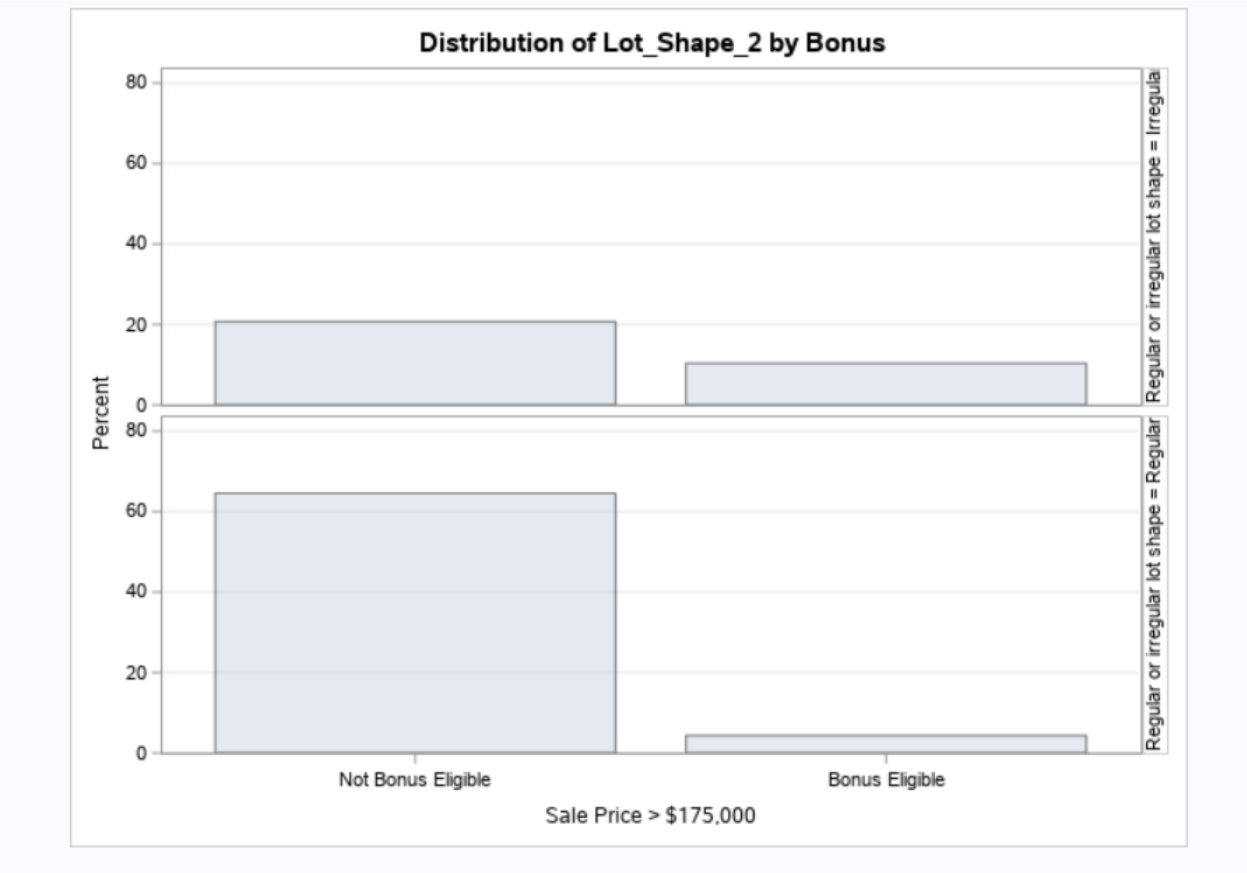
Frequency
Percent
Row Pct
Col Pct

Table of Fireplaces by Bonus			
	Bonus(Sale Price > $175,000)		
Fireplaces(Number of fireplaces)	Not Bonus Eligible	Bonus Eligible	Total
0	177 59.00 90.77 69.41	18 6.00 9.23 40.00	195 65.00
1	68 22.67 73.12 26.67	25 8.33 26.88 55.56	93 31.00
2	10 3.33 83.33 3.92	2 0.67 16.67 4.44	12 4.00
Total	255 85.00	45 15.00	300 100.00



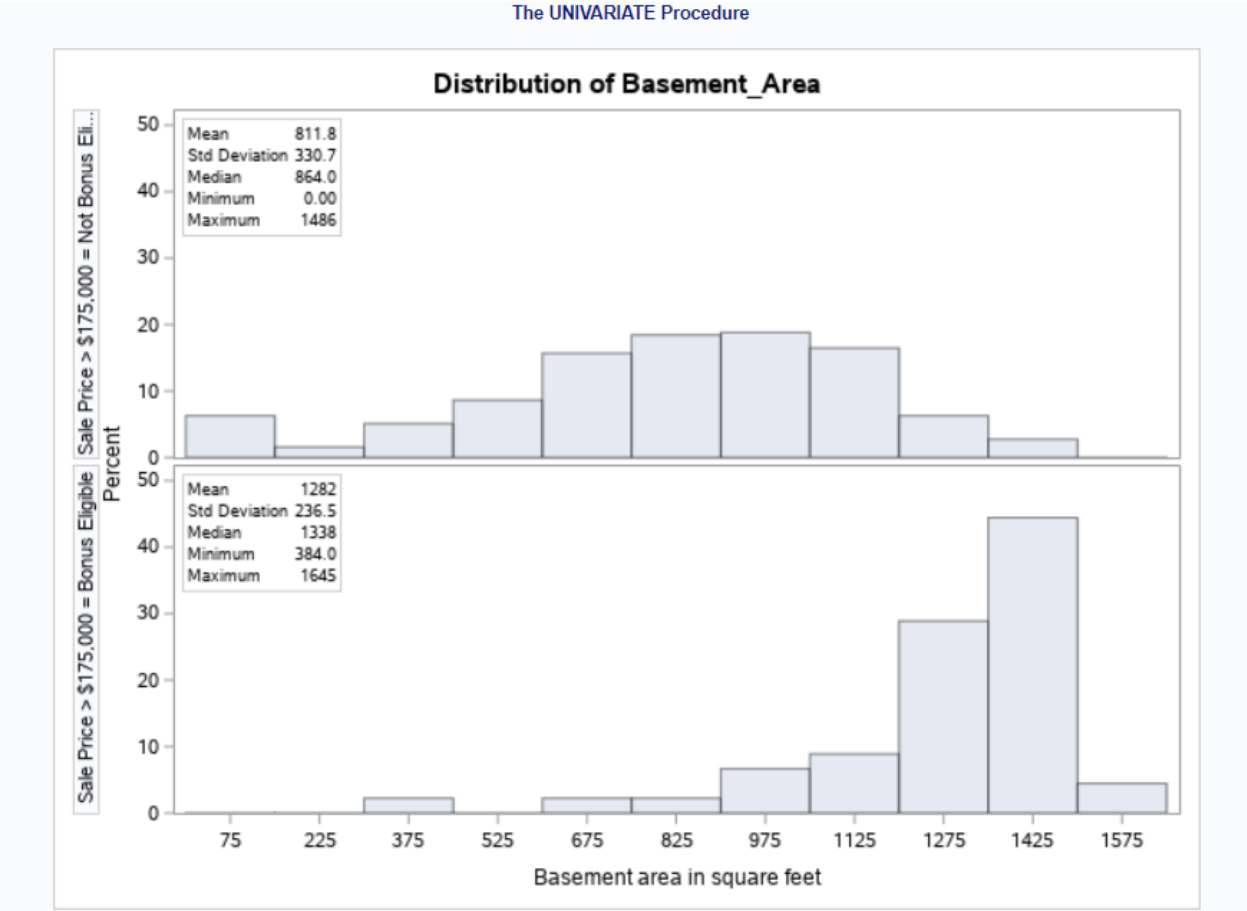
Frequency
Percent
Row Pct
Col Pct

Table of Lot_Shape_2 by Bonus			
	Bonus(Sale Price > $175,000)		
Lot_Shape_2(Regular or irregular lot shape)	Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 20.74 66.67 24.31	31 10.37 33.33 70.45	93 31.10
Regular	193 64.55 93.69 75.69	13 4.35 6.31 29.55	206 68.90
Total	255 85.28	44 14.72	299 100.00
Frequency Missing = 1			



/* create histograms of bonus */

```
proc univariate data=STAT1.ameshousing3 noprint;
    class Bonus;
    var Basement_Area ;
    histogram Basement_Area;
    inset mean std median min max / format=5.2 position=nw;
    format Bonus bonusfmt.;
run;
```



19.Performing a Chi-square Test of Association

We use PROC FREQ for this purpose.

```
ods graphics off;
proc freq data=STAT1.ameshousing3;
    tables (Lot_Shape_2 Fireplaces)*Bonus
          / chisq expected cellchi2 nocol nopercent
            relrisk;
    format Bonus bonusfmt.;
    title 'Associations with Bonus';
run;

ods graphics on;
```

Associations with Bonus

The FREQ Procedure

Frequency Expected Cell Chi-Square Row Pct

Table of Lot_Shape_2 by Bonus			
	Bonus(Sale Price > $175,000)		
Lot_Shape_2(Regular or irregular lot shape)	Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 79.314 3.7797 66.67	31 13.686 21.905 33.33	93
Regular	193 175.69 1.7064 93.69	13 30.314 9.8893 6.31	206
Total	255	44	299
Frequency Missing = 1			

Statistics for Table of Lot_Shape_2 by Bonus

Statistic	DF	Value	Prob
Chi-Square	1	37.2807	<.0001
Likelihood Ratio Chi-Square	1	34.4226	<.0001
Continuity Adj. Chi-Square	1	35.1587	<.0001
Mantel-Haenszel Chi-Square	1	37.1561	<.0001
Phi Coefficient		-0.3531	
Contingency Coefficient		0.3330	
Cramer's V		-0.3531	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	62
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.1347	0.0664	0.2735
Relative Risk (Column 1)	0.7116	0.6137	0.8251
Relative Risk (Column 2)	5.2821	2.9002	9.6202

Sample Size = 299
Frequency Missing = 1

Frequency Expected Cell Chi-Square Row Pct	Table of Fireplaces by Bonus		
	Bonus(Sale Price > $175,000)		
Fireplaces(Number of fireplaces)	Not Bonus Eligible	Bonus Eligible	Total
0	177 165.75 0.7636 90.77	18 29.25 4.3269 9.23	195
1	68 79.05 1.5446 73.12	25 13.95 8.7529 26.88	93
2	10 10.2 0.0039 83.33	2 1.8 0.0222 16.67	12
Total	255	45	300

Statistics for Table of Fireplaces by Bonus

Statistic	DF	Value	Prob
Chi-Square	2	15.4141	0.0004
Likelihood Ratio Chi-Square	2	14.4859	0.0007
Mantel-Haenszel Chi-Square	1	10.7458	0.0010
Phi Coefficient		0.2267	
Contingency Coefficient		0.2211	
Cramer's V		0.2267	

Sample Size = 300

/* Checking if bonus and fireplaces have a significant ordinal association */

/* The Mantel-Haenszel chi-square test is appropriate in this case */

```
ods graphics off;
proc freq data=STAT1.ameshousing3;
    tables Fireplaces*Bonus / chisq measures cl;
    format Bonus bonusfmt.;
    title 'Ordinal Association between FIREPLACES and BONUS?';
run;

ods graphics on;
```

Ordinal Association between FIREPLACES and BONUS?

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Fireplaces by Bonus			
	Bonus(Sale Price > $175,000)		
Fireplaces(Number of fireplaces)	Not Bonus Eligible	Bonus Eligible	Total
0	177 59.00 90.77 69.41	18 6.00 9.23 40.00	195 65.00
1	68 22.67 73.12 26.67	25 8.33 26.88 55.56	93 31.00
2	10 3.33 83.33 3.92	2 0.67 16.67 4.44	12 4.00
Total	255 85.00	45 15.00	300 100.00

Statistics for Table of Fireplaces by Bonus

Statistic	DF	Value	Prob
Chi-Square	2	15.4141	0.0004
Likelihood Ratio Chi-Square	2	14.4859	0.0007
Mantel-Haenszel Chi-Square	1	10.7458	0.0010
Phi Coefficient		0.2267	
Contingency Coefficient		0.2211	
Cramer's V		0.2267	

Statistic	Value	ASE	95% Confidence Limits	
Gamma	0.4964	0.1111	0.2786	0.7143
Kendall's Tau-b	0.2072	0.0585	0.0926	0.3218
Stuart's Tau-c	0.1449	0.0433	0.0600	0.2298
Somers' D C|R	0.1510	0.0451	0.0626	0.2395
Somers' D R|C	0.2842	0.0786	0.1301	0.4383
Pearson Correlation	0.1896	0.0591	0.0737	0.3054
Spearman Correlation	0.2107	0.0594	0.0943	0.3272
Lambda Asymmetric C|R	0.0000	0.0000	0.0000	0.0000
Lambda Asymmetric R|C	0.0667	0.0603	0.0000	0.1849
Lambda Symmetric	0.0467	0.0424	0.0000	0.1298
Uncertainty Coefficient C|R	0.0571	0.0298	0.0000	0.1156
Uncertainty Coefficient R|C	0.0313	0.0167	0.0000	0.0640
Uncertainty Coefficient Symmetric	0.0404	0.0213	0.0000	0.0823

Sample Size = 300

20. Simple logistic regression model

Let's assess the relationship between the probability of a bonus eligible home and one predictor: basement area.

```
ods graphics on;
proc logistic data=STAT1.ameshousing3 alpha=0.05
              plots(only)=(effect oddsratio);
    model Bonus(event='1')=Basement_Area / clodds=pl;
    title 'LOGISTIC MODEL (1):Bonus=Basement_Area';
run;
```

LOGISTIC MODEL (1):Bonus=Basement_Area

The LOGISTIC Procedure

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > $175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	300
Number of Observations Used	300

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	45

Probability modeled is Bonus='1'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

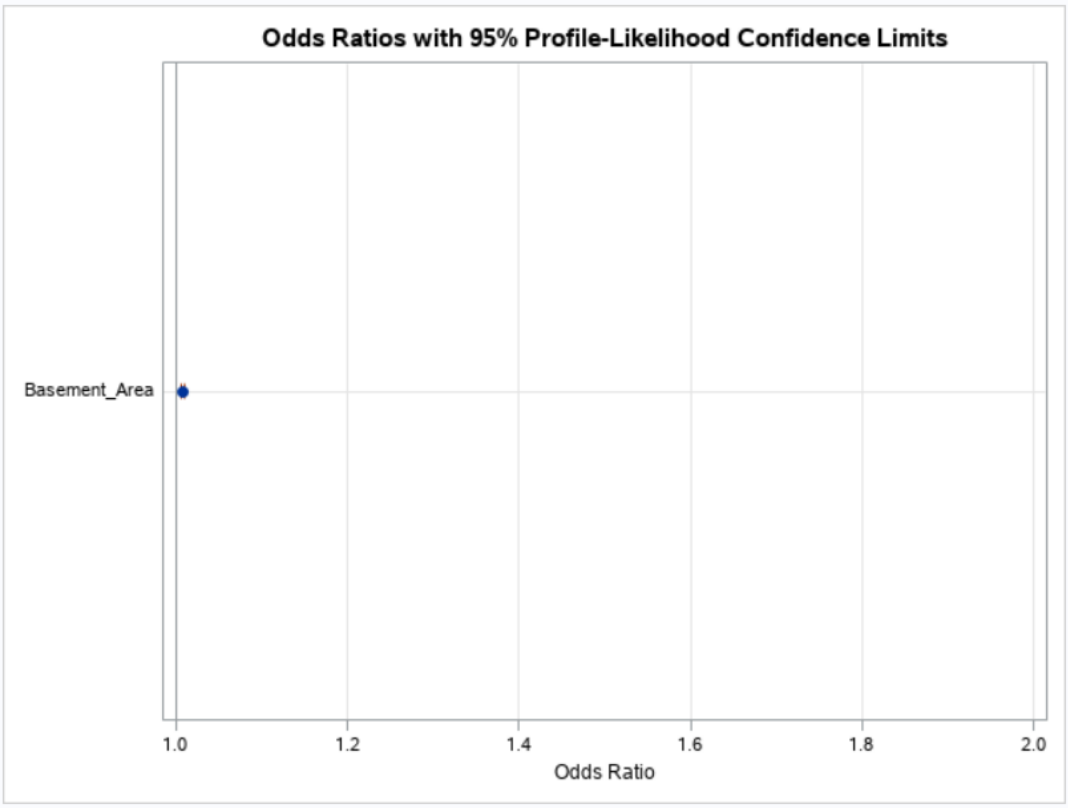
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	255.625	161.838
SC	259.329	169.246
-2 Log L	253.625	157.838

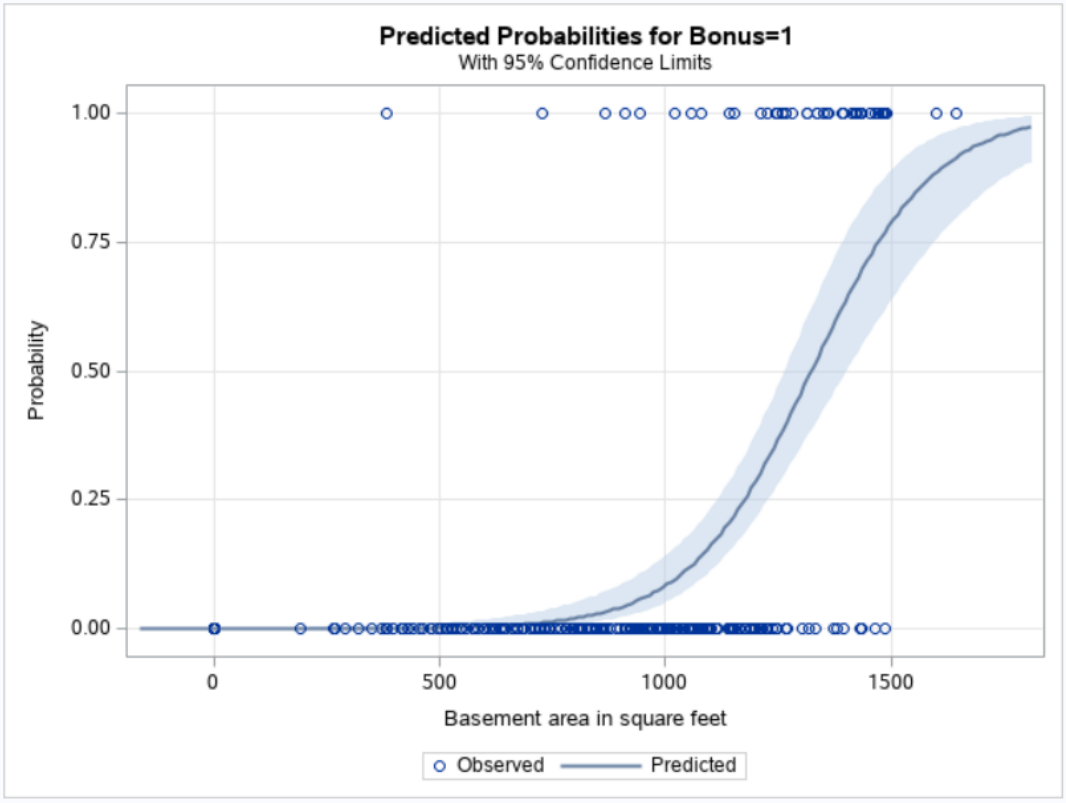
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	95.7870	1	<.0001
Score	65.5624	1	<.0001
Wald	48.0617	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-9.7854	1.2896	57.5758	<.0001
Basement_Area	1	0.00739	0.00107	48.0617	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	Somers' D	0.791
Percent Discordant	10.4	Gamma	0.792
Percent Tied	0.1	Tau-a	0.202
Pairs	11475	c	0.896

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	1.0000	1.007	1.005	1.010





Additional goodness of fit measures from PROC LOGISTIC:

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	Somers' D	0.791
Percent Discordant	10.4	Gamma	0.792
Percent Tied	0.1	Tau-a	0.202
Pairs	11475	c	0.896

21. Simple logistic regression model

Let's assess the relationship between the probability of a bonus eligible home and three predictors: basement area, number of fireplaces and lot shape.

```
ods graphics on;
proc logistic data=STAT1.ameshousing3 plots(only)=(effect
oddsratio);
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') /
param=ref;
    model Bonus(event='1')=Basement_Area Fireplaces Lot_Shape_2
/ clodds=pl;
    units Basement_Area=100;
    title 'LOGISTIC MODEL (2):Bonus= Basement_Area Fireplaces
Lot_Shape_2';
run;
```

LOGISTIC MODEL (2):Bonus= Basement_Area Fireplaces Lot_Shape_2

The LOGISTIC Procedure

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > $175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	300
Number of Observations Used	299

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	44

Probability modeled is Bonus='1'.

Note: 1 observation was deleted due to missing values for the response or explanatory variables.

Class Level Information			
Class	Value	Design Variables	
Fireplaces	0	0	0
	1	1	0
	2	0	1
Lot_Shape_2	Irregular	1	
	Regular	0	

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	140.499
SC	255.513	159.001
-2 Log L	249.812	130.499

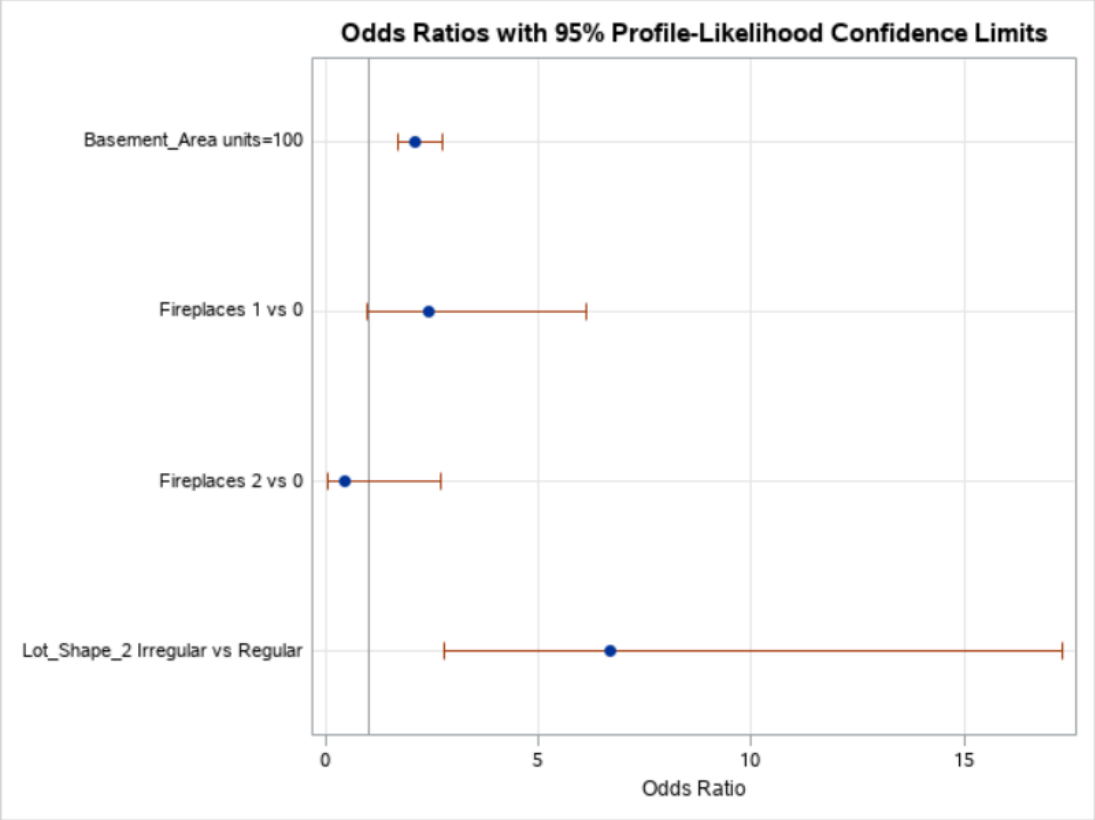
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	119.3133	4	<.0001
Score	91.7250	4	<.0001
Wald	49.8671	4	<.0001

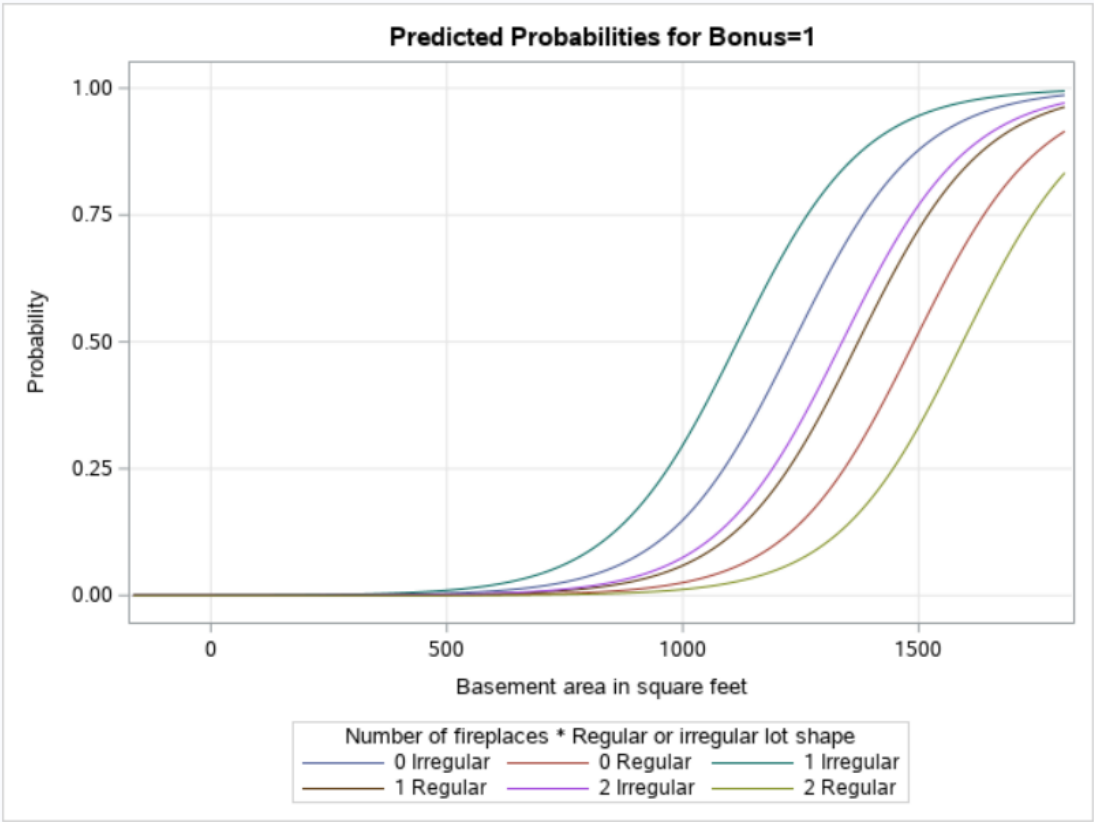
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Basement_Area	1	38.1356	<.0001
Fireplaces	2	5.2060	0.0741
Lot_Shape_2	1	16.9421	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-11.0882	1.5384	51.9467	<.0001
Basement_Area		1	0.00744	0.00120	38.1356	<.0001
Fireplaces	1	1	0.8810	0.4658	3.5770	0.0586
Fireplaces	2	1	-0.7683	0.9654	0.6335	0.4261
Lot_Shape_2	Irregular	1	1.9025	0.4622	16.9421	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	92.9	Somers' D	0.859
Percent Discordant	7.0	Gamma	0.860
Percent Tied	0.1	Tau-a	0.216
Pairs	11220	c	0.930

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	100.0	2.105	1.696	2.727
Fireplaces 1 vs 0	1.0000	2.413	0.973	6.127
Fireplaces 2 vs 0	1.0000	0.464	0.054	2.703
Lot_Shape_2 Irregular vs Regular	1.0000	6.703	2.786	17.301


Odds Ratios with 95% Profile-Likelihood Confidence Limits
Basement_Area units=100
Fireplaces 1 vs 0
Fireplaces 2 vs 0
Lot_Shape_2 Irregular vs Regular
0
5
10
15
Odds Ratio


Predicted Probabilities for Bonus=1
1.00
0.75
0.50
0.25
0.00
Probability
0
500
1000
1500
Basement area in square feet
Number of fireplaces * Regular or irregular lot shape
0 Irregular
0 Regular
1 Irregular
1 Regular
2 Irregular
2 Regular

22. Multiple logistic regression model: with interactions

Let's assess the relationship between the probability of a bonus eligible home and several predictors, with interactions between those predictors.

```
proc logistic data=STAT1.ameshousing3 plots(only)=(effect oddsratio);
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
    model Bonus(event='1')=Basement_Area|Fireplaces|Lot_Shape_2 @2 /
          selection=backward clodds=pl slstay=0.10;
    units Basement_Area=100;
    title 'LOGISTIC MODEL (3): Backward Elimination '
           'Bonus=Basement_Area|Fireplaces|Lot_Shape_2';
run;
```

LOGISTIC MODEL (3): Backward Elimination Bonus=Basement_Area|Fireplaces|Lot_Shape_2

The LOGISTIC Procedure

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > $175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	300
Number of Observations Used	299

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	44

Probability modeled is Bonus='1'.

Note: 1 observation was deleted due to missing values for the response or explanatory variables.

Backward Elimination Procedure

Class Level Information			
Class	Value	Design Variables	
Fireplaces	0	0	0
	1	1	0
	2	0	1
Lot_Shape_2	Irregular	1	
	Regular	0	

Step 0. The following effects were entered:

Intercept Basement_Area Fireplaces Basement_*Fireplaces Lot_Shape_2 Basement_*Lot_Shape_ Fireplace*Lot_Shape_

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	141.737
SC	255.513	178.741
-2 Log L	249.812	121.737

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	128.0756	9	<.0001
Score	109.4005	9	<.0001
Wald	40.8304	9	<.0001

Step 1. Effect Fireplace*Lot_Shape_ is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	141.166
SC	255.513	170.769
-2 Log L	249.812	125.166

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	124.6462	7	<.0001
Score	106.9810	7	<.0001
Wald	42.3266	7	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
3.4592	2	0.1774

Step 2. Effect Basement_*Fireplaces is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	138.872
SC	255.513	161.074
-2 Log L	249.812	126.872

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	122.9405	5	<.0001
Score	102.6370	5	<.0001
Wald	42.9826	5	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
5.5364	4	0.2365

Note: No (additional) effects met the 0.1 significance level for removal from the model.

Summary of Backward Elimination						
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Fireplace*Lot_Shape_	2	5	3.2305	0.1988	
2	Basement_*Fireplaces	2	4	1.7237	0.4224	

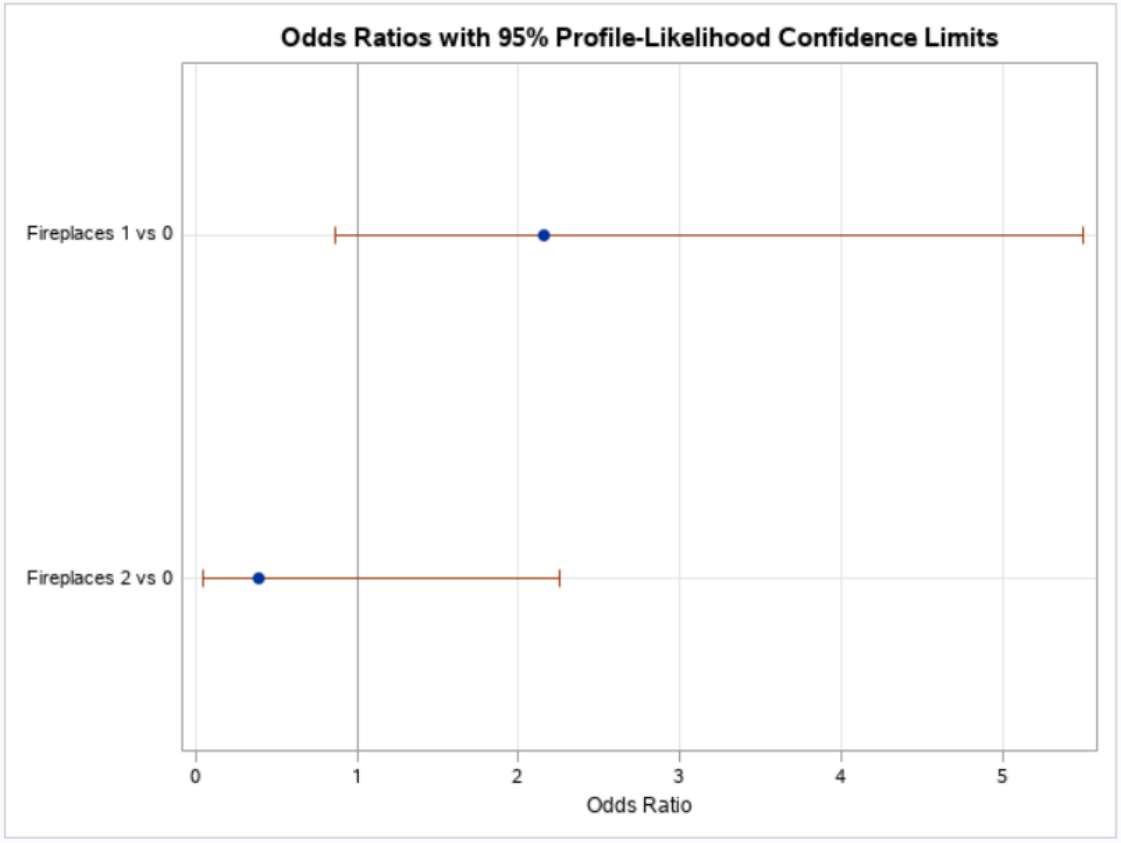
Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Basement_Area	1	18.2896	<.0001
Fireplaces	2	4.7171	0.0946
Lot_Shape_2	1	5.0247	0.0250
Basement_*Lot_Shape_	1	3.1127	0.0777

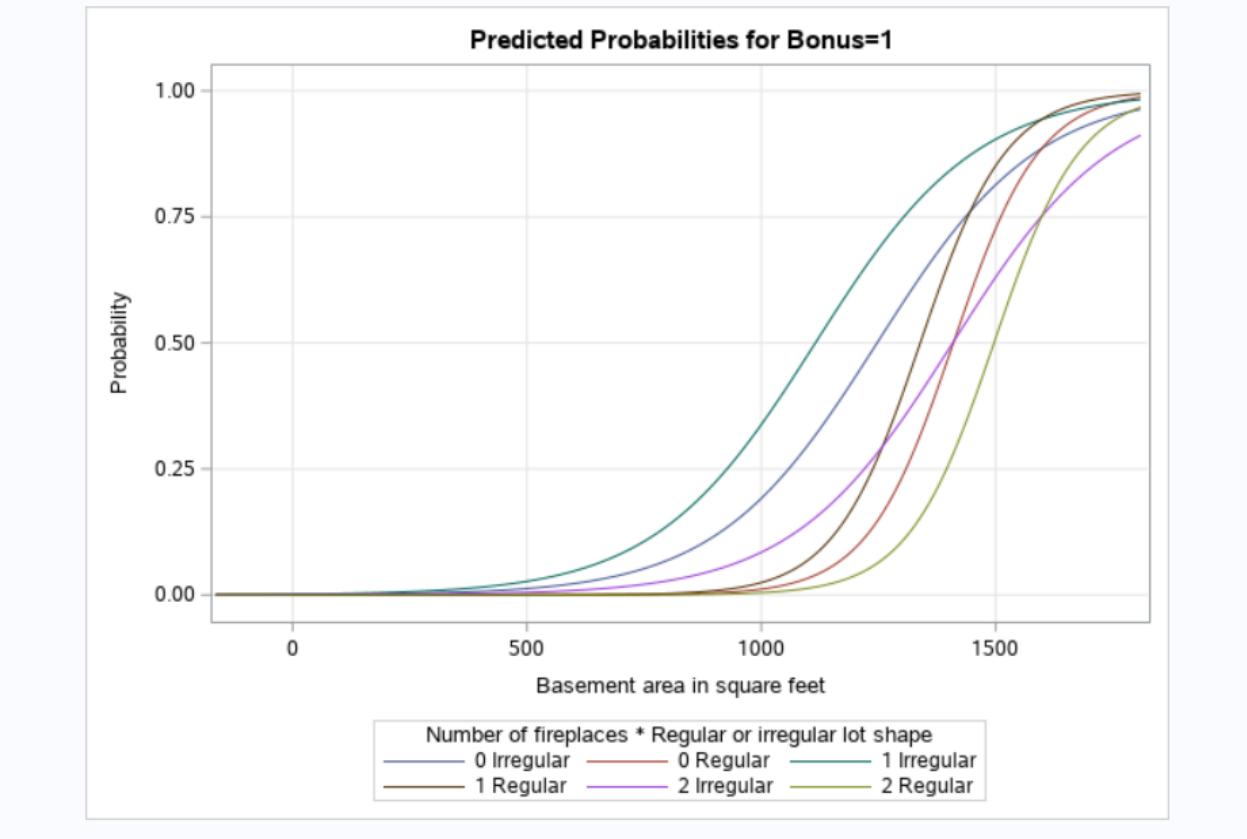
Note: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests. The joint test for an effect is a test that all the parameters associated with that effect are zero. Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-15.3017	3.2407	22.2952	<.0001
Basement_Area		1	0.0109	0.00254	18.2896	<.0001
Fireplaces	1	1	0.7671	0.4687	2.6781	0.1017
Fireplaces	2	1	-0.9405	0.9503	0.9795	0.3223
Lot_Shape_2	Irregular	1	8.0362	3.5850	5.0247	0.0250
Basement_*Lot_Shape_	Irregular	1	-0.00503	0.00285	3.1127	0.0777

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	93.8	Somers' D	0.876
Percent Discordant	6.2	Gamma	0.876
Percent Tied	0.1	Tau-a	0.221
Pairs	11220	c	0.938

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Fireplaces 1 vs 0	1.0000	2.153	0.865	5.500
Fireplaces 2 vs 0	1.0000	0.390	0.047	2.251





/* Let's re-estimate the multiple logistic regression with more detailed specifications */

```
proc logistic data=STAT1.ameshousing3
              plots(only)=oddsratio(range=clip);
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
    model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
    units Basement_Area=100;
    oddsratio Basement_Area / at (Lot_Shape_2=ALL) cl=pl;
    oddsratio Lot_Shape_2 / at (Basement_Area=1000 1500) cl=pl;
    title 'LOGISTIC MODEL (3.1): Bonus=Basement_Area|Lot_Shape_2
Fireplaces';
run;
```

LOGISTIC MODEL (3.1): Bonus=Basement_Area|Lot_Shape_2 Fireplaces

The LOGISTIC Procedure

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > $175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	300
Number of Observations Used	299

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	44

Probability modeled is Bonus='1'.

Note: 1 observation was deleted due to missing values for the response or explanatory variables.

Class Level Information			
Class	Value	Design Variables	
Fireplaces	0	0	0
	1	1	0
	2	0	1
Lot_Shape_2	Irregular	1	
	Regular	0	

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	138.872
SC	255.513	161.074
-2 Log L	249.812	126.872

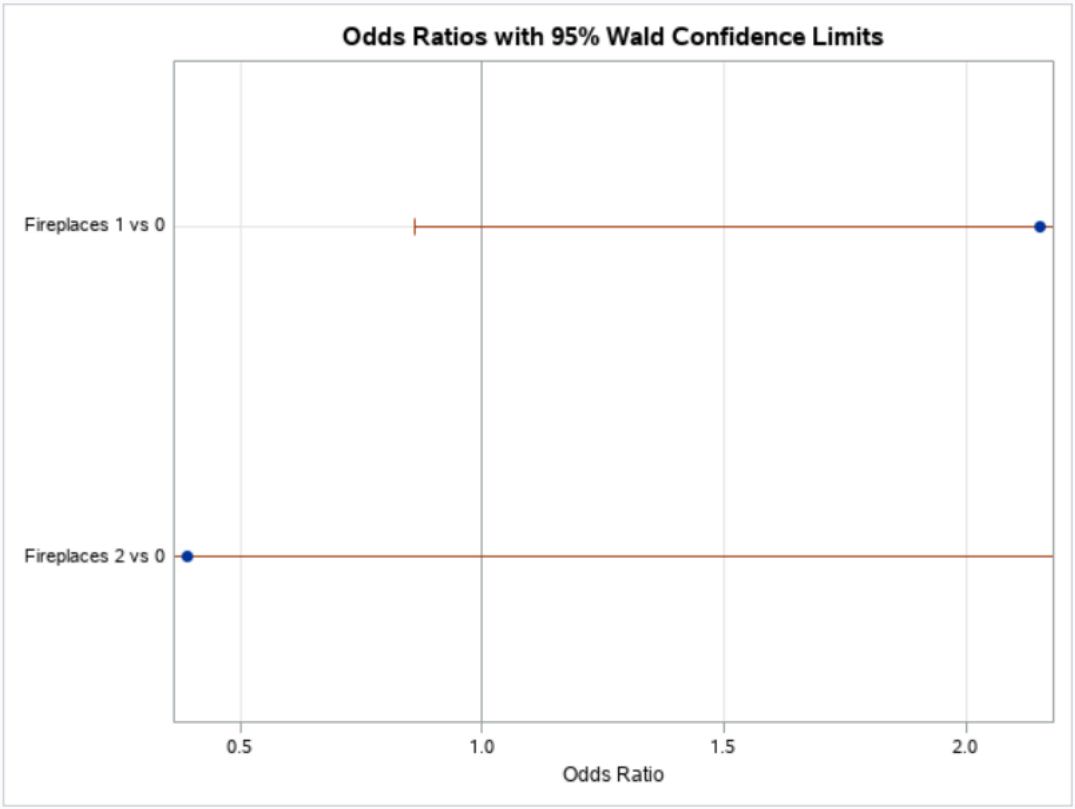
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	122.9405	5	<.0001
Score	102.6370	5	<.0001
Wald	42.9826	5	<.0001

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Basement_Area	1	18.2896	<.0001
Lot_Shape_2	1	5.0247	0.0250
Basement_*Lot_Shape_	1	3.1127	0.0777
Fireplaces	2	4.7171	0.0946

Note: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests. The joint test for an effect is a test that all the parameters associated with that effect are zero. Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

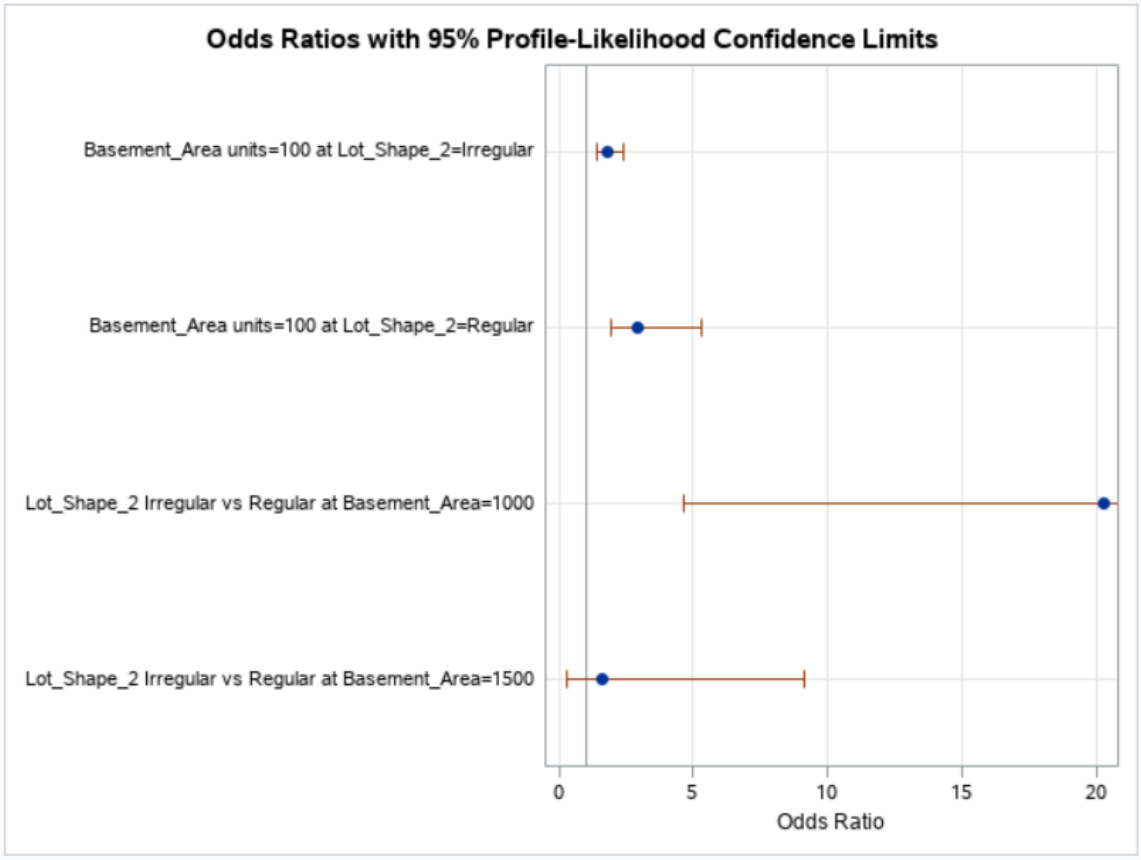
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-15.3017	3.2407	22.2952	<.0001
Basement_Area		1	0.0109	0.00254	18.2896	<.0001
Lot_Shape_2	Irregular	1	8.0362	3.5850	5.0247	0.0250
Basement_*Lot_Shape_	Irregular	1	-0.00503	0.00285	3.1127	0.0777
Fireplaces	1	1	0.7671	0.4687	2.6781	0.1017
Fireplaces	2	1	-0.9405	0.9503	0.9795	0.3223

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Fireplaces 1 vs 0	2.153	0.859	5.396
Fireplaces 2 vs 0	0.390	0.061	2.514



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	93.8	Somers' D	0.876
Percent Discordant	6.2	Gamma	0.876
Percent Tied	0.1	Tau-a	0.221
Pairs	11220	c	0.938

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
Basement_Area units=100 at Lot_Shape_2=Irregular	1.791	1.421	2.396
Basement_Area units=100 at Lot_Shape_2=Regular	2.960	1.932	5.315
Lot_Shape_2 Irregular vs Regular at Basement_Area=1000	20.278	4.623	146.987
Lot_Shape_2 Irregular vs Regular at Basement_Area=1500	1.643	0.283	9.145



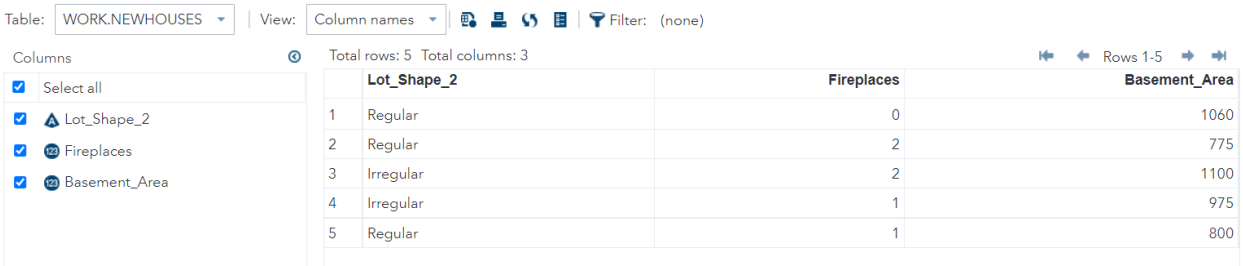
/* re-run the model above, suppress the output, but save it to be used in the PROC PLM procedure */

```
ods select none;
proc logistic data=STAT1.ameshousing3;
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
```

```
    model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
    units Basement_Area=100;
      store out=isbonus;
run;
ods select all;
```

/* Generate a new dataset with 5 entries, to use for predictions */

```
data newhouses;
      length Lot_Shape_2 $9;
      input Fireplaces Lot_Shape_2 $ Basement_Area;
      datalines;
      0  Regular    1060
      2  Regular     775
      2  Irregular  1100
      1  Irregular   975
      1  Regular     800
      ;
run;
```



	Lot_Shape_2	Fireplaces	Basement_Area
1	Regular	0	1060
2	Regular	2	775
3	Irregular	2	1100
4	Irregular	1	975
5	Regular	1	800

/* make predictions using the new dataset and estimated model parameters */

```
proc plm restore=isbonus;
      score data=newhouses out=scored_houses / ILINK;
      title 'Predictions using PROC PLM';
run;
```

```
proc print data=scored_houses;
run;
```

Predictions using PROC PLM

The PLM Procedure

Store Information	
Item Store	WORK.ISBONUS
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC LOGISTIC
Date Created	22AUG21:03:47:39
Response Variable	Bonus
Link Function	Logit
Distribution	Binary
Class Variables	Fireplaces Lot_Shape_2 Bonus
Model Effects	Intercept Basement_Area Lot_Shape_2 Basement_*Lot_Shape_ Fireplaces

Predictions using PROC PLM

Obs	Lot_Shape_2	Fireplaces	Basement_Area	Predicted
1	Regular	0	1060	0.02192
2	Regular	2	775	0.00040
3	Irregular	2	1100	0.14210
4	Irregular	1	975	0.30608
5	Regular	1	800	0.00286

Reference:

Course title: Statistics with SAS

Instructor: Jordan Bakerman

Weblink: https://www.coursera.org/learn/sas-statistics?#syllabus

********** The End **********