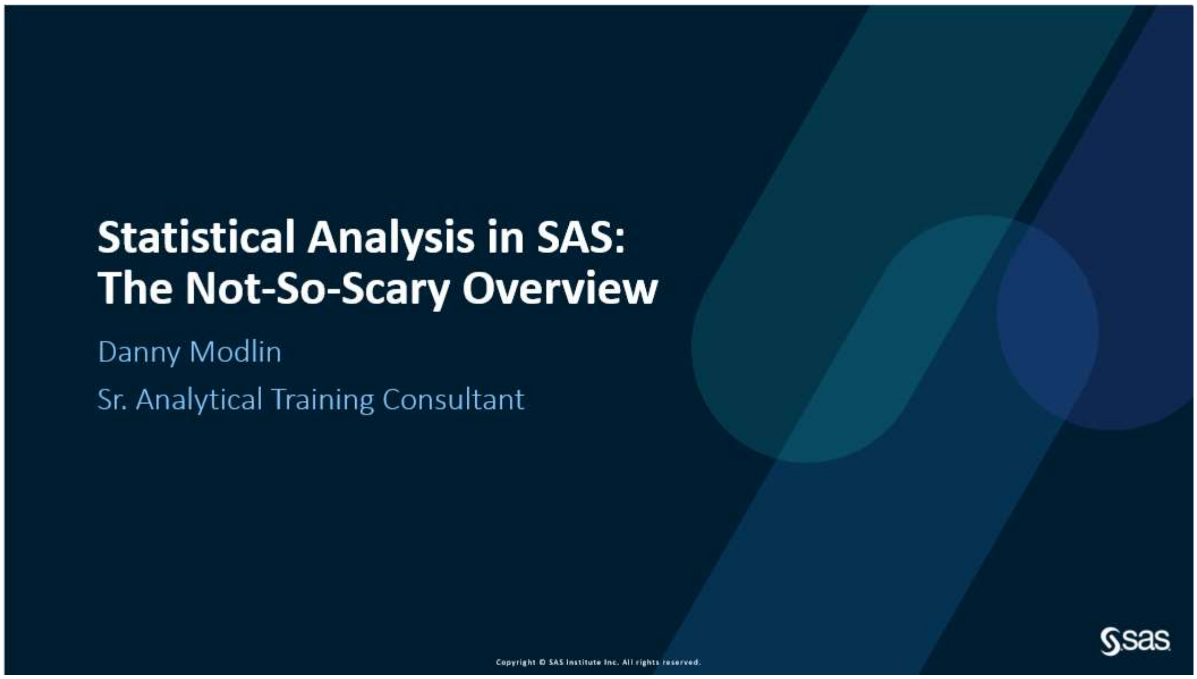
Statistical Analysis in SAS

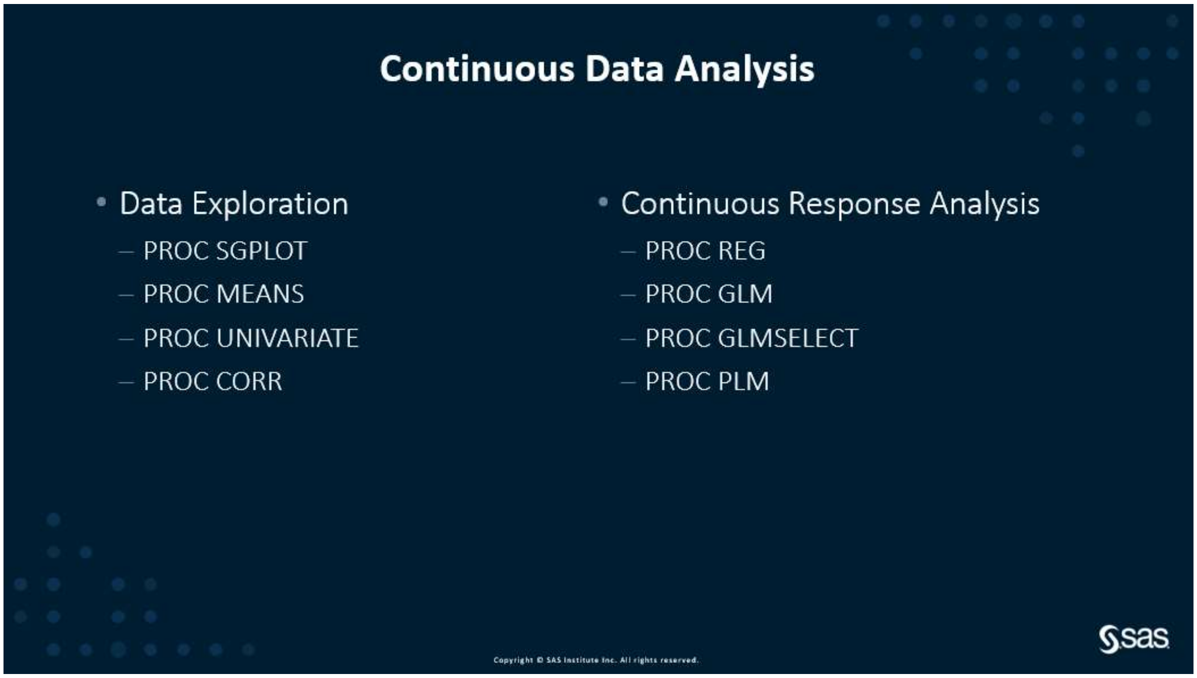
The Not So Scary Overview



Statistical Analysis in SAS:
The Not-So-Scary Overview

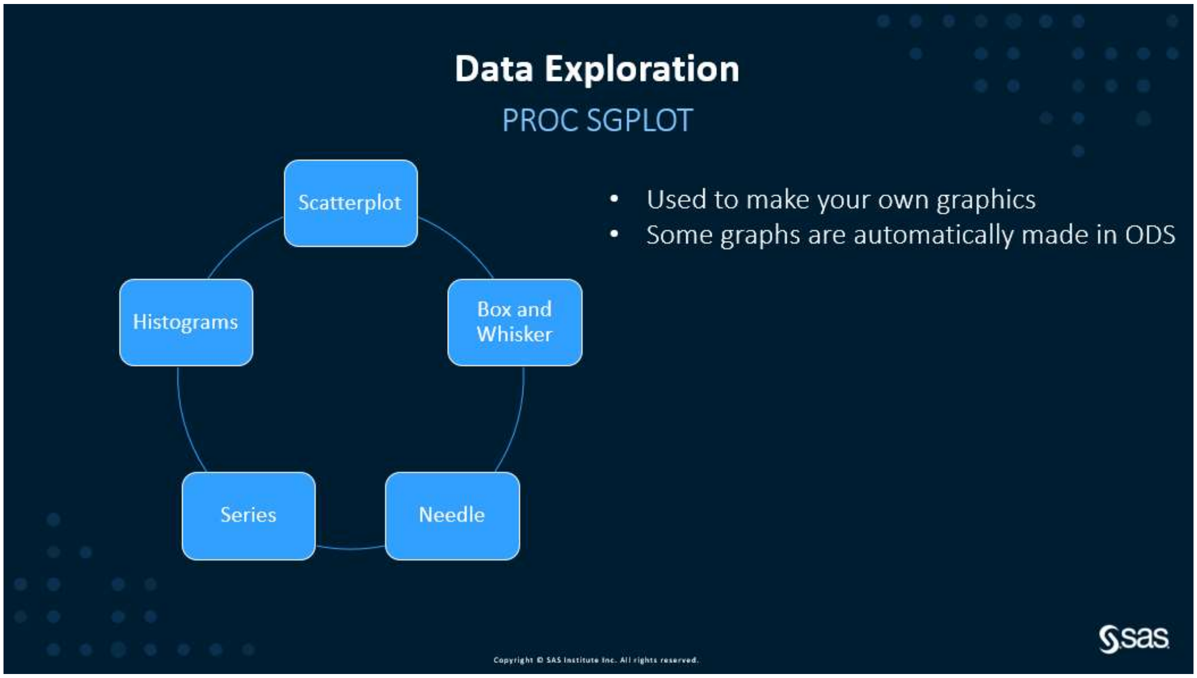
Danny Modlin

Sr. Analytical Training Consultant



Continuous Data Analysis

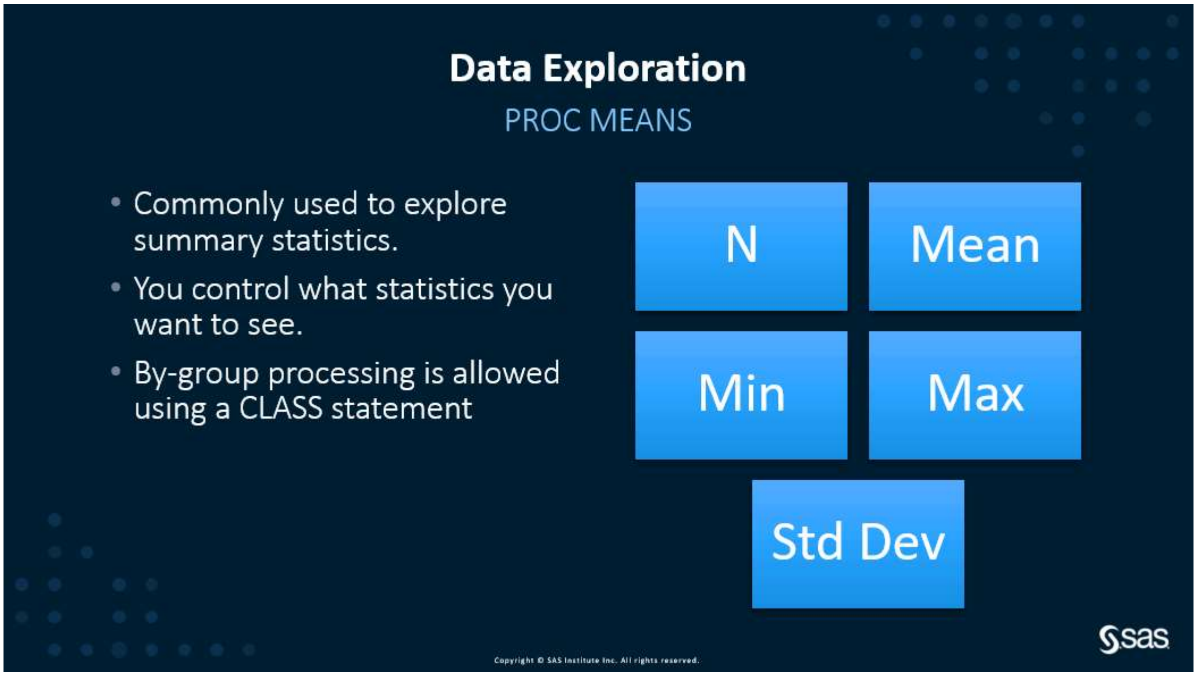
- Data Exploration
  - PROC SGPLOT
  - PROC MEANS
  - PROC UNIVARIATE
  - PROC CORR
- Continuous Response Analysis
  - PROC REG
  - PROC GLM
  - PROC GLMSELECT
  - PROC PLM



Data Exploration

PROC SGPLOT

- Used to make your own graphics
- Some graphs are automatically made in ODS



Data Exploration

PROC MEANS

- Commonly used to explore summary statistics.
- You control what statistics you want to see.
- By-group processing is allowed using a CLASS statement

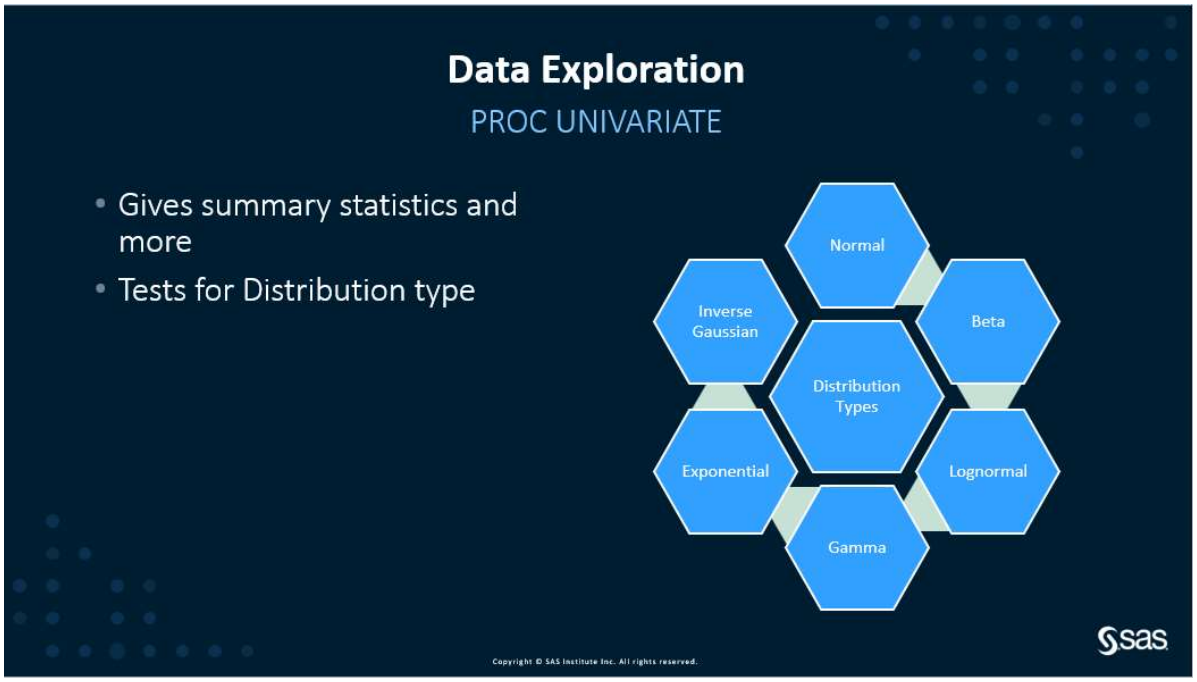
N

Mean

Min

Max

Std Dev



Data Exploration

PROC UNIVARIATE

- Gives summary statistics and more
- Tests for Distribution type

Normal

Inverse Gaussian

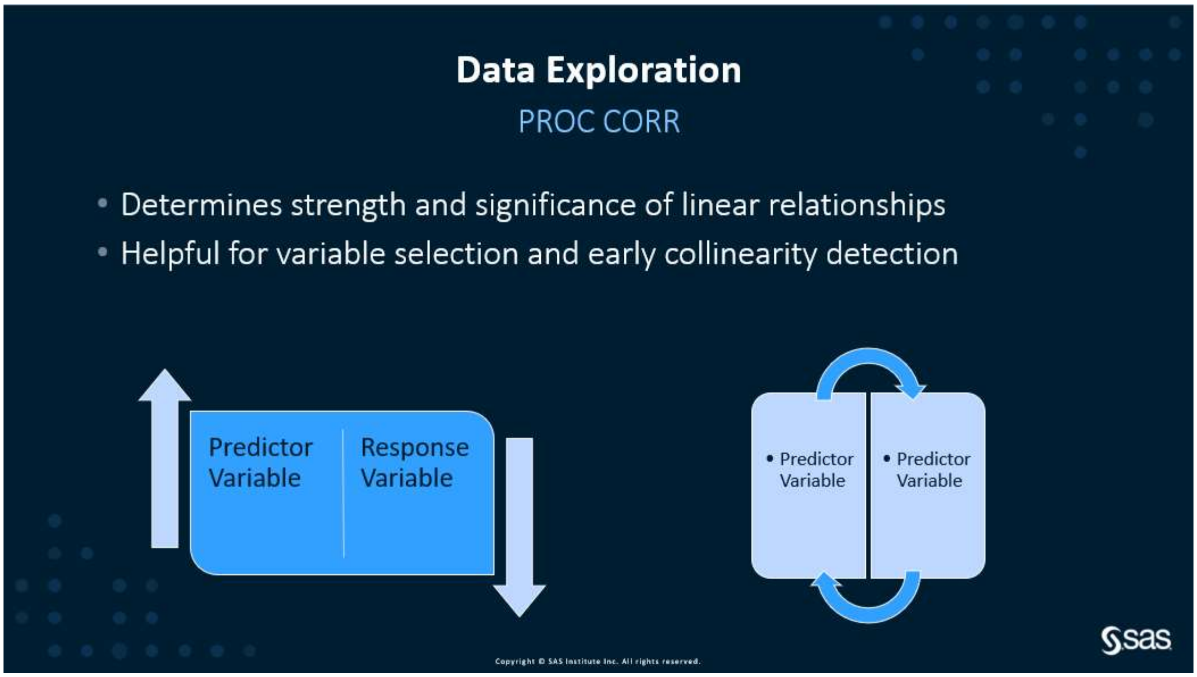
Beta

Distribution Types

Exponential

Lognormal

Gamma



Data Exploration

PROC CORR

- Determines strength and significance of linear relationships
- Helpful for variable selection and early collinearity detection

Predictor Variable

Response Variable

- Predictor Variable
- Predictor Variable

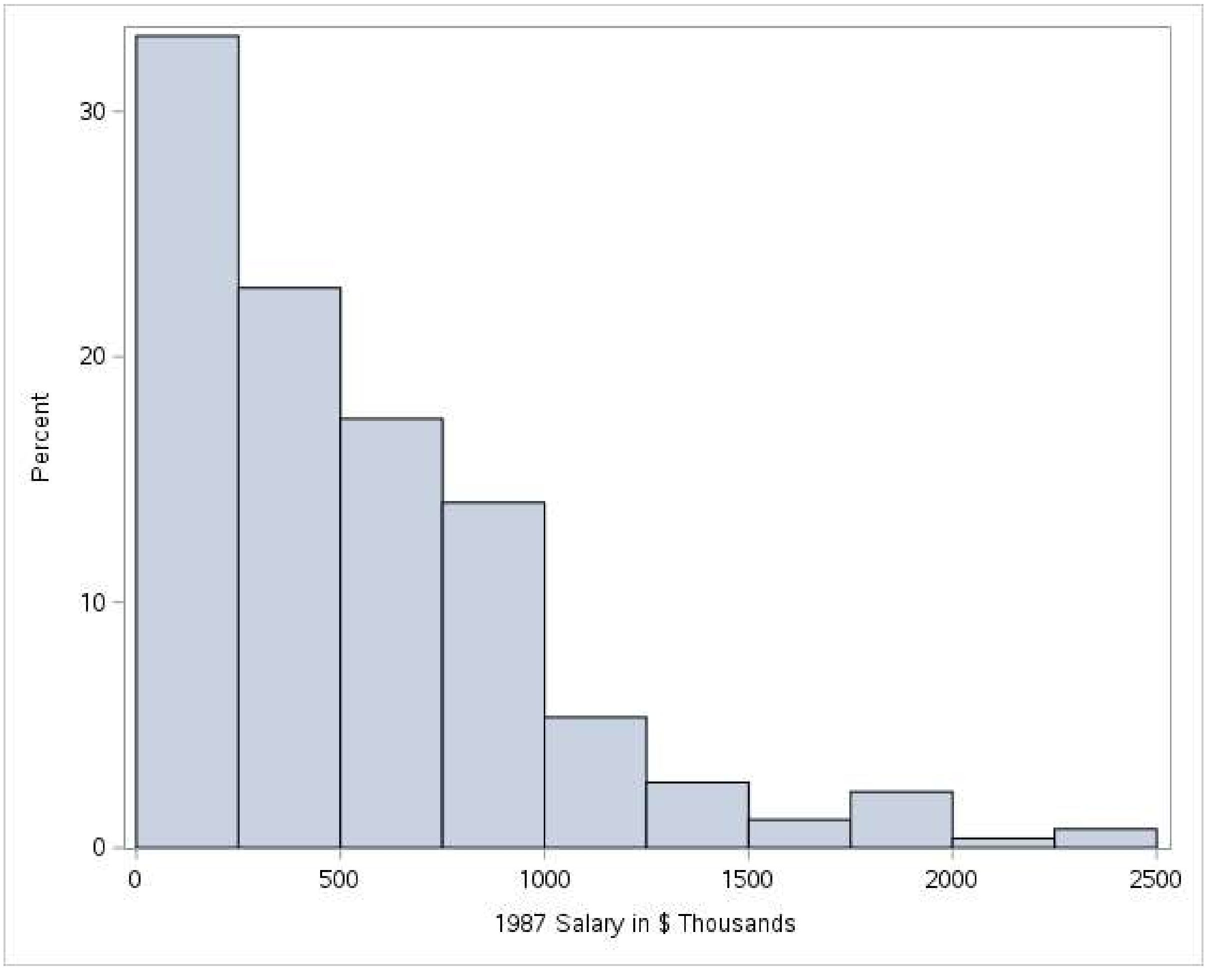
Demonstration Time

Copyright © SAS Institute Inc. All rights reserved.

PROC SGPLOT: Making your own graphics

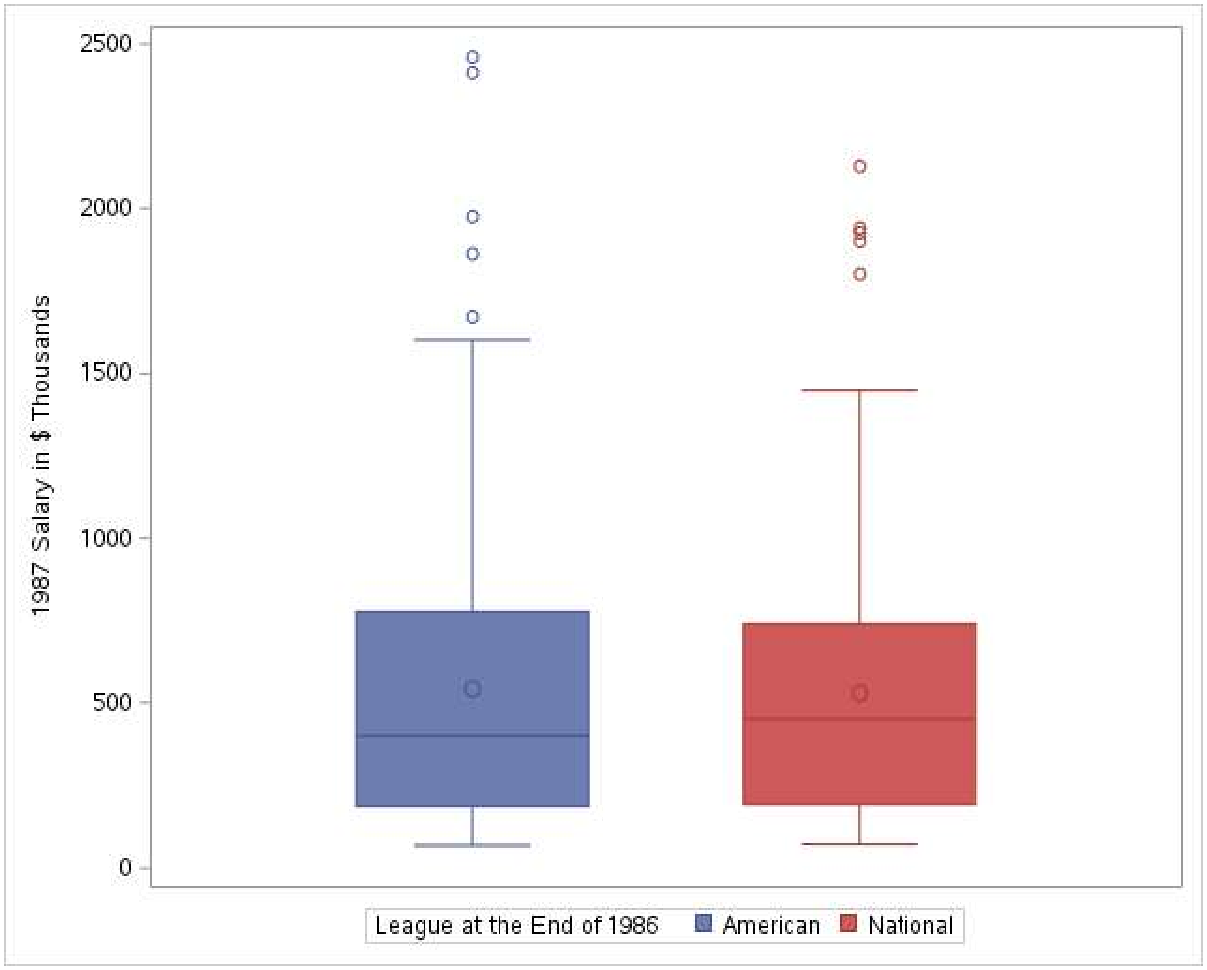
In [2]:

```
proc sgplot data=sashelp.baseball;
   histogram salary;
run;
```



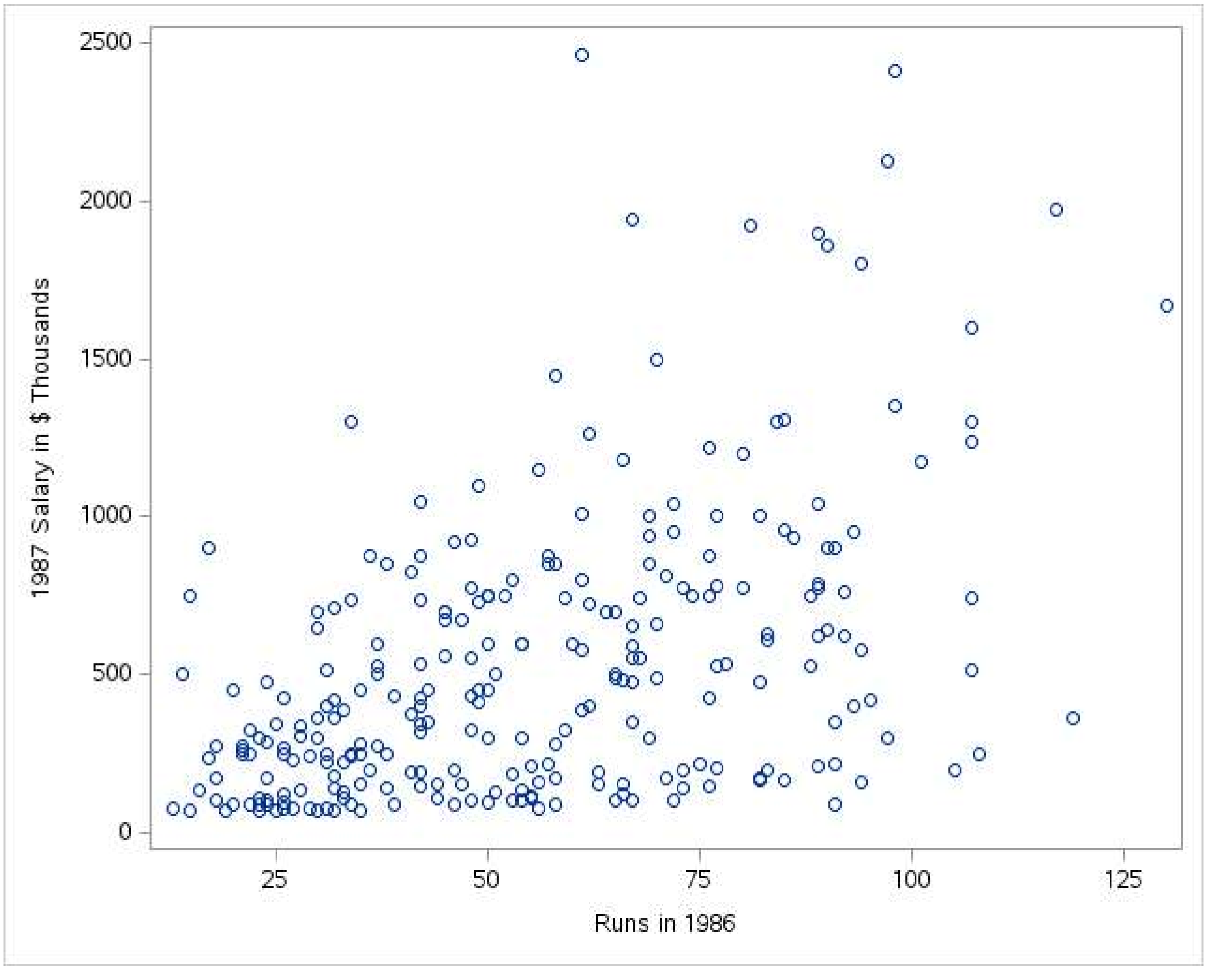
In [3]:

```
proc sgplot data=sashelp.baseball;
   vbox salary / group=league;
run;
```

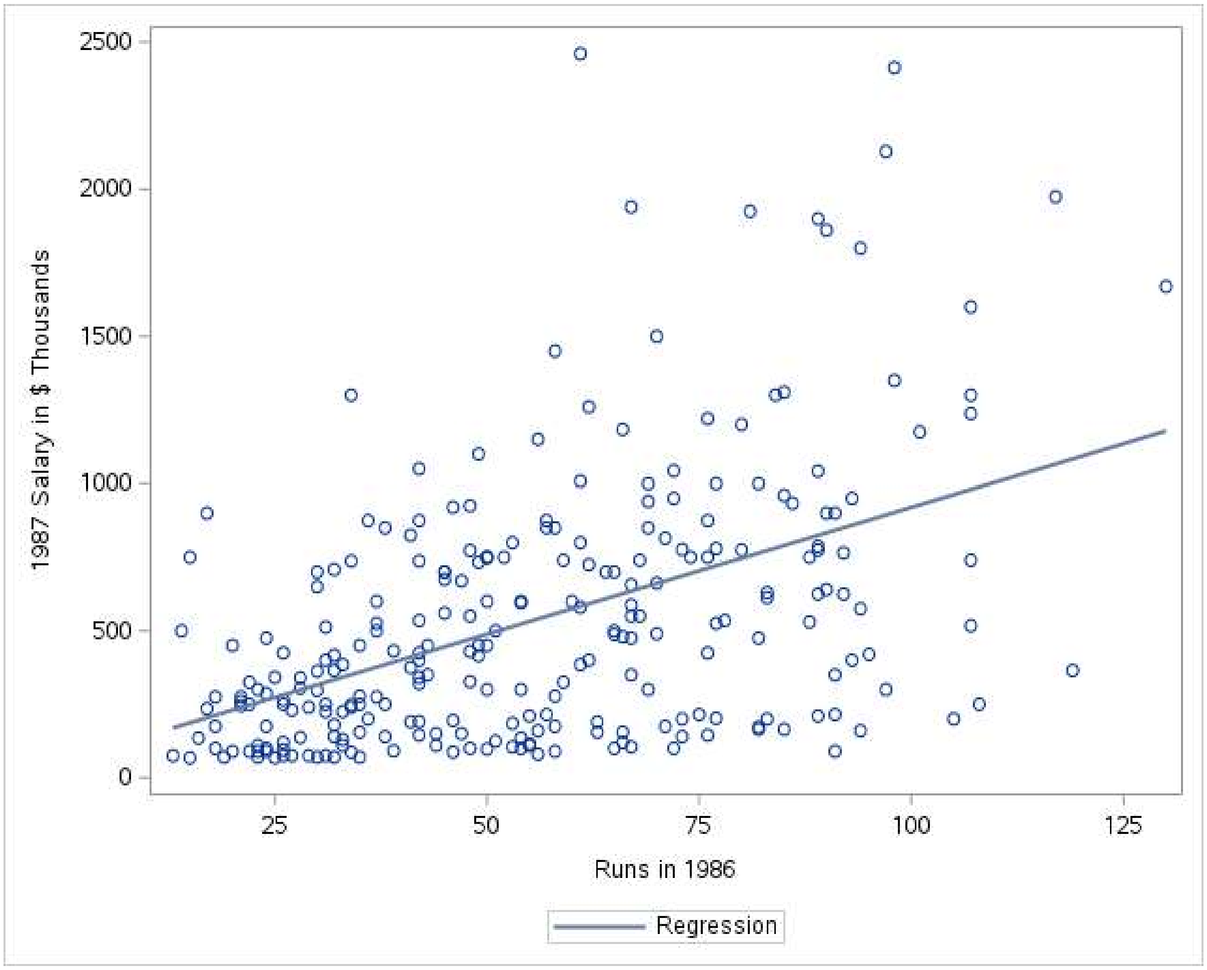


In [4]:

```
proc sgplot data=sashelp.baseball;
   scatter y=salary x=nRuns;
run;
```



In [5]:
```
proc sgplot data=sashelp.baseball;
   reg y=salary x=nRuns;
run;
```



PROC MEANS: Summary statistics tables

In [6]:

```
proc means data=sashelp.baseball;
    var salary;
run;
```

The SAS System

The MEANS Procedure

Analysis Variable : Salary 1987 Salary in $ Thousands				
N	Mean	Std Dev	Minimum	Maximum
263	535.9258821	451.1186807	67.5000000	2460.00

In [7]:

```
proc means data=sashelp.baseball median var qrange;
    var salary;
run;
```

The SAS System

The MEANS Procedure

Analysis Variable : Salary 1987 Salary in $ Thousands		
Median	Variance	Quartile Range
425.0000000	203508.06	560.0000000

In [8]:

```
proc means data=sashelp.baseball median n mean std min max;
    var salary;
run;
```

The SAS System

The MEANS Procedure

Analysis Variable : Salary 1987 Salary in $ Thousands					
Median	N	Mean	Std Dev	Minimum	Maximum
425.0000000	263	535.9258821	451.1186807	67.5000000	2460.00

In [9]:

```
proc means data=sashelp.baseball;
   class league;
   var salary nHits;
run;
```

The SAS System

The MEANS Procedure

League at the End of 1986	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
American	**175**	**Salary** **nHits**	**1987 Salary in $ Thousands** **Hits in 1986**	139 175	541.9995468 107.6857143	464.7827551 45.2275037	67.5000000 36.0000000	2460.00 238.0000000
National	**147**	**Salary** **nHits**	**1987 Salary in $ Thousands** **Hits in 1986**	124 147	529.1175000 98.2925170	437.0732479 42.4882274	70.0000000 31.0000000	2127.33 211.0000000

PROC UNIVARIATE: Summary statistics with distribution questions

In [10]:

```
proc univariate data=sashelp.baseball;
   var salary;
   histogram salary / kernel normal;
run;
```

The UNIVARIATE Procedure
Variable: Salary (1987 Salary in $ Thousands)

Moments			
N	263	Sum Weights	263
Mean	535.925882	Sum Observations	140948.507
Std Deviation	451.118681	Variance	203508.064
Skewness	1.58896735	Kurtosis	3.05896473
Uncorrected SS	128857066	Corrected SS	53319112.8
Coeff Variation	84.1755727	Std Error Mean	27.8171695

Basic Statistical Measures			
Location		Variability	
Mean	535.9259	Std Deviation	451.11868
Median	425.0000	Variance	203508
Mode	750.0000	Range	2393
		Interquartile Range	560.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	19.26601	Pr > |t|	<.0001
Sign	M	131.5	Pr >= |M|	<.0001
Signed Rank	S	17358	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	2460.00
99%	2127.33
95%	1350.00
90%	1050.00
75% Q3	750.00
50% Median	425.00
25% Q1	190.00
10%	100.00
5%	86.50
1%	70.00

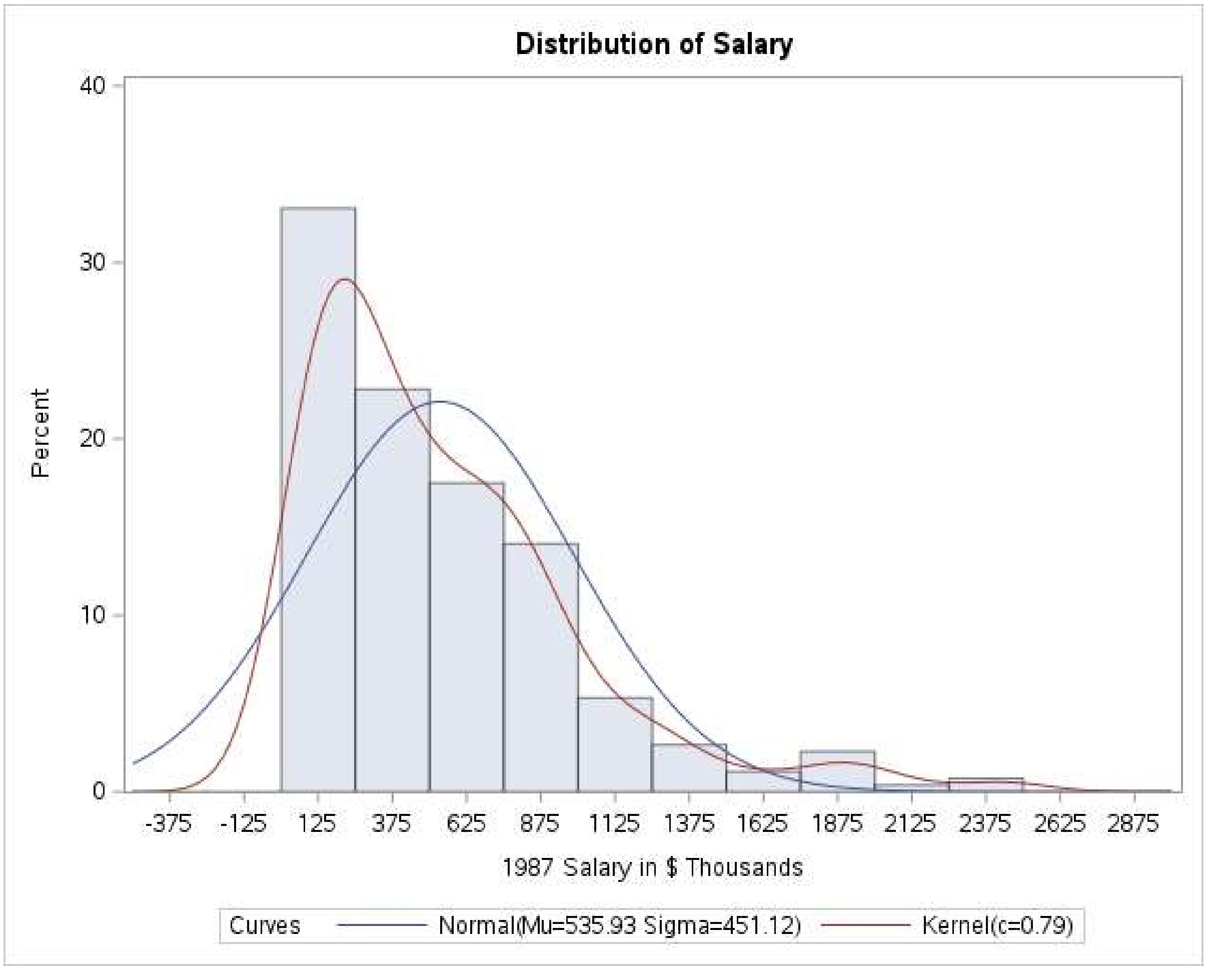
Quantiles (Definition 5)	
Level	Quantile
0% Min	67.50

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
67.5	41	1940.00	230
68.0	213	1975.00	83
70.0	260	2127.33	218
70.0	110	2412.50	164
70.0	93	2460.00	101

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	59	18.32	100.00

The SAS System

The UNIVARIATE Procedure



The SAS System

The UNIVARIATE Procedure
Fitted Normal Distribution for Salary (1987 Salary in $ Thousands)

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	535.9259
Std Dev	Sigma	451.1187

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.14955006	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	1.40541623	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	9.30772350	**Pr > A-Sq**	<0.005

Quantiles for Normal Distribution		
Percent	Quantile	
	Observed	Estimated
1.0	70.0000	-513.5331
5.0	86.5000	-206.0983
10.0	100.0000	-42.2060
25.0	190.0000	231.6510
50.0	425.0000	535.9259
75.0	750.0000	840.2008
90.0	1050.0000	1114.0577
95.0	1350.0000	1277.9501
99.0	2127.3330	1585.3849

In [11]:

```
proc univariate data=sashelp.baseball;
   var logsalary;
   histogram logsalary / kernel normal;
run;
```

The UNIVARIATE Procedure
Variable: logSalary (Log Salary)

Moments			
N	263	Sum Weights	263
Mean	5.92722154	Sum Observations	1558.85927
Std Deviation	0.88919239	Variance	0.7906631
Skewness	-0.1820065	Kurtosis	-0.8827516
Uncorrected SS	9446.85795	Corrected SS	207.153733
Coeff Variation	15.0018416	Std Error Mean	0.05482995

Basic Statistical Measures			
Location		Variability	
Mean	5.927222	Std Deviation	0.88919
Median	6.052089	Variance	0.79066
Mode	6.620073	Range	3.59579
		Interquartile Range	1.37305

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	108.1019	Pr > |t|	<.0001
Sign	M	131.5	Pr >= |M|	<.0001
Signed Rank	S	17358	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	7.80792
99%	7.66262
95%	7.20786
90%	6.95655
75% Q3	6.62007
50% Median	6.05209
25% Q1	5.24702
10%	4.60517
5%	4.46014
1%	4.24850

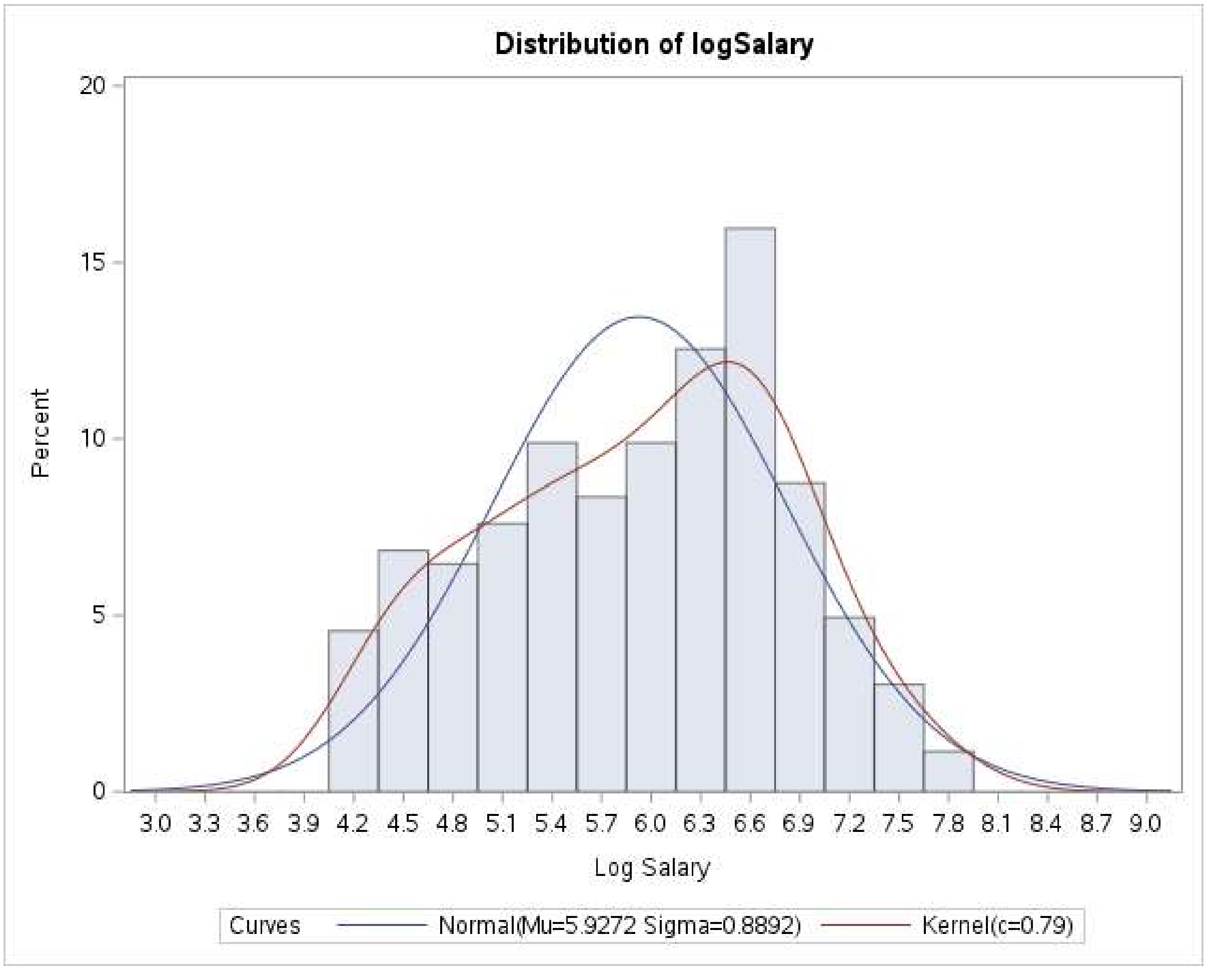
Quantiles (Definition 5)	
Level	Quantile
0% Min	4.21213

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
4.21213	41	7.57044	230
4.21951	213	7.58832	83
4.24850	260	7.66262	218
4.24850	110	7.78842	164
4.24850	93	7.80792	101

Missing Values			
Missing Value	Count	Percent Of	
		All Obs	Missing Obs
.	59	18.32	100.00

The SAS System

The UNIVARIATE Procedure



The SAS System

The UNIVARIATE Procedure
Fitted Normal Distribution for logSalary (Log Salary)

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	5.927222
Std Dev	Sigma	0.889192

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.07412552	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	0.36572122	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	2.24599244	**Pr > A-Sq**	<0.005

Quantiles for Normal Distribution		
Percent	Quantile	
	Observed	Estimated
1.0	4.24850	3.85865
5.0	4.46014	4.46463
10.0	4.60517	4.78768
25.0	5.24702	5.32747
50.0	6.05209	5.92722
75.0	6.62007	6.52697
90.0	6.95655	7.06677
95.0	7.20786	7.38981
99.0	7.66262	7.99579

PROC CORR: relationships among variable for multiple uses

In [12]:

```
proc corr data=sashelp.baseball plots=matrix;
   var nRBI nHome nRuns;
   with salary;
run;
```

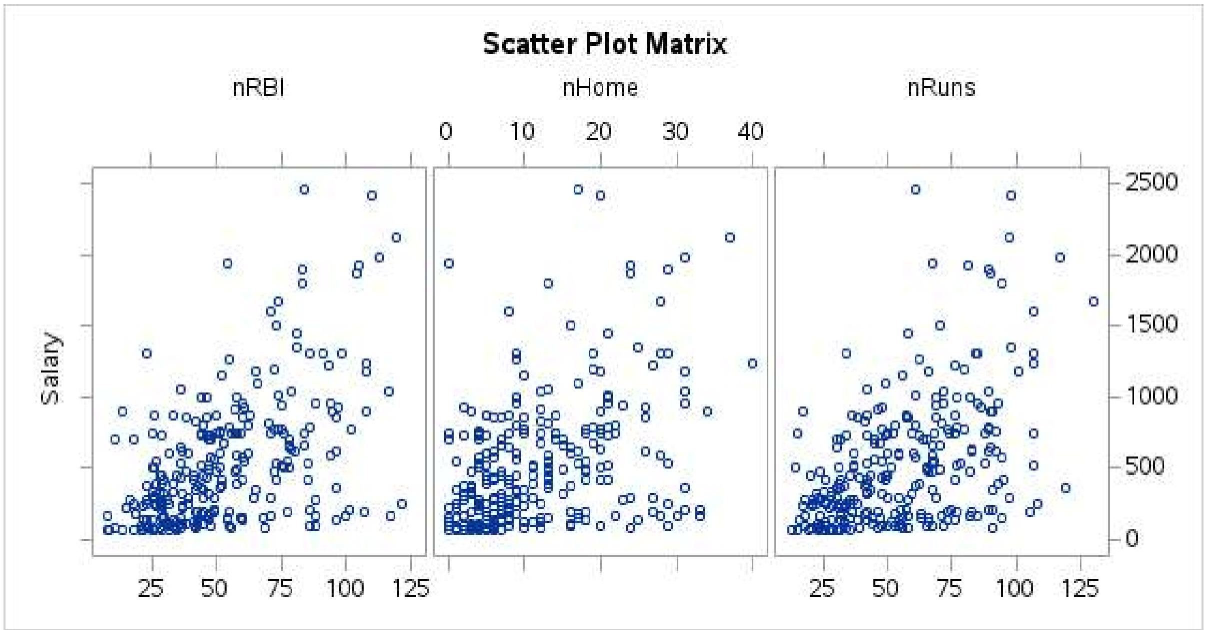
The SAS System

The CORR Procedure

1 With Variables:	Salary
3 Variables:	nRBI nHome nRuns

							Simple Statistics
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**	**Label**
Salary	263	535.92588	451.11868	140949	67.50000	2460	1987 Salary in $ Thousands
nRBI	322	49.37267	25.50116	15898	8.00000	121.00000	RBIs in 1986
nHome	322	11.10248	8.69877	3575	0	40.00000	Home Runs in 1986
nRuns	322	52.21739	25.05737	16814	12.00000	130.00000	Runs in 1986

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations			
	nRBI	**nHome**	**nRuns**
Salary 1987 Salary in $ Thousands	0.51723 <.0001 263	0.39885 <.0001 263	0.47903 <.0001 263



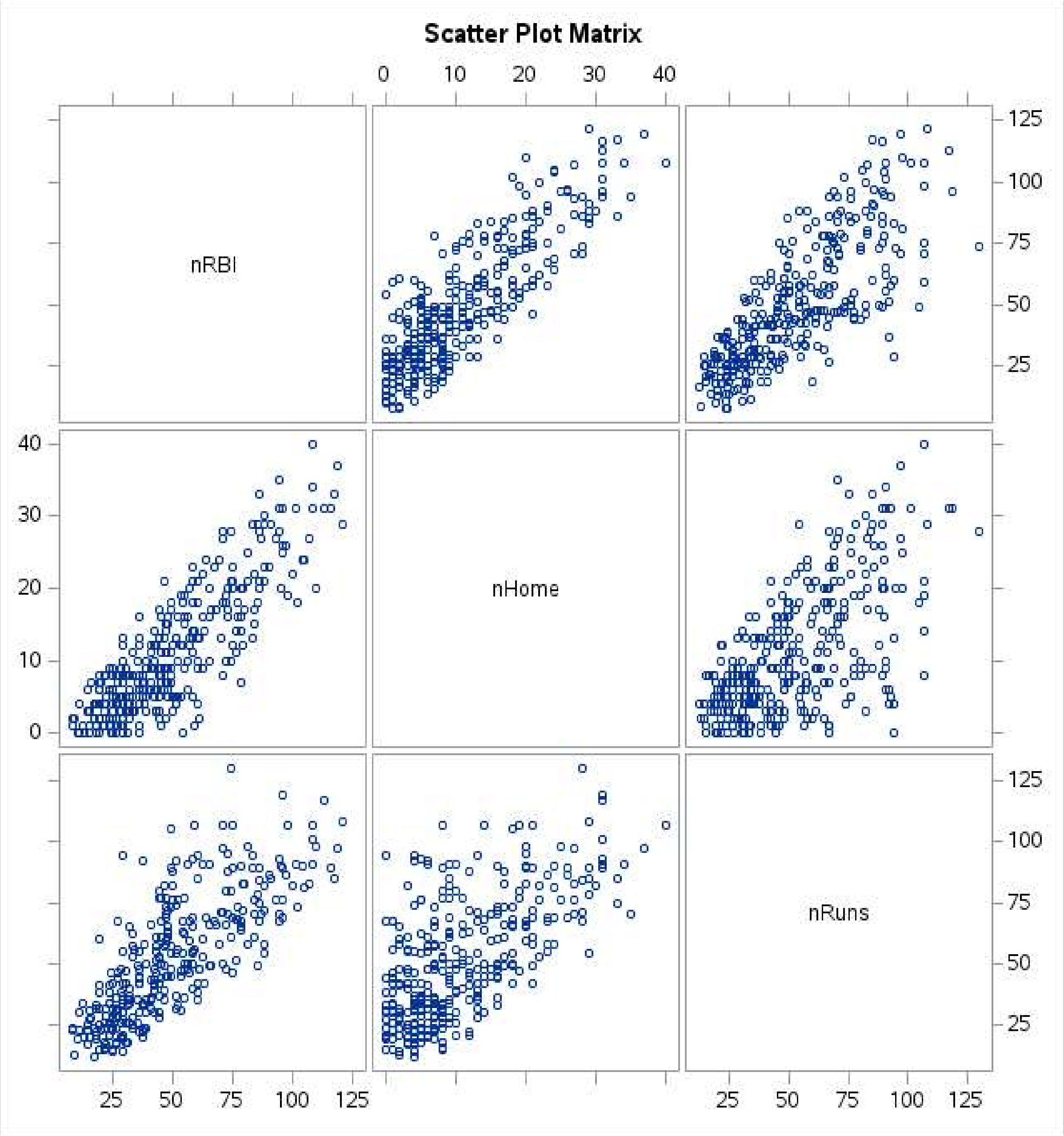
In [13]:
```
proc corr data=sashelp.baseball plots=matrix;
   var nRBI nHome nRuns;
run;
```

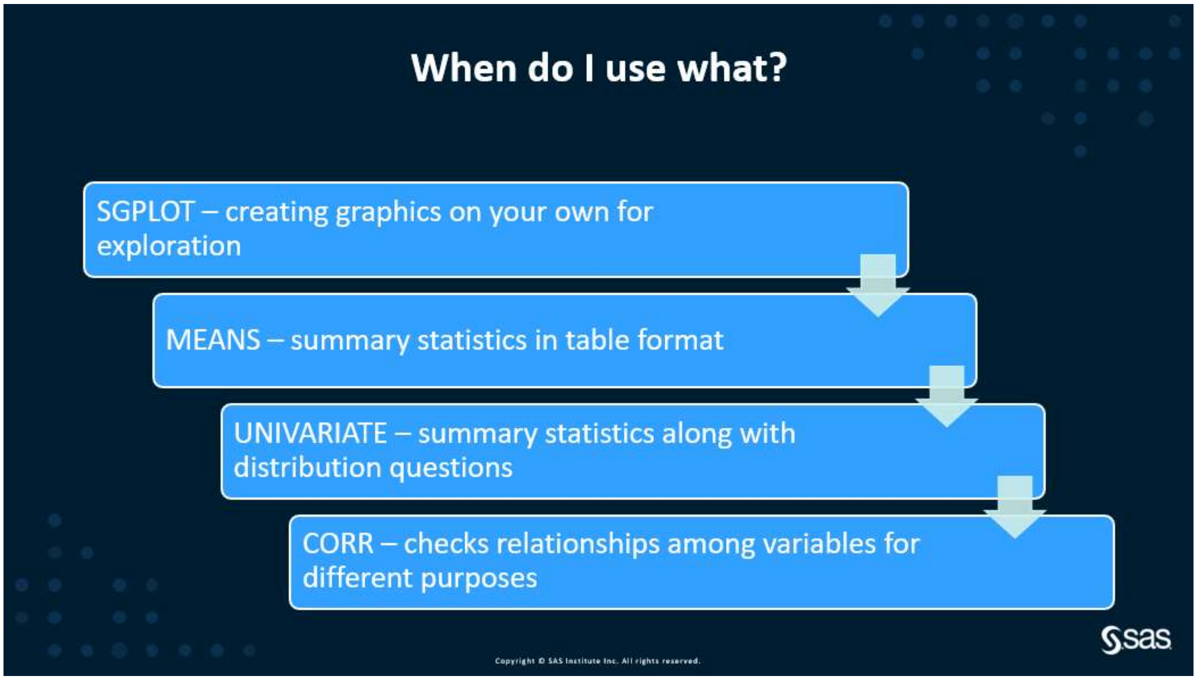
The CORR Procedure

3 Variables:	nRBI nHome nRuns

Simple Statistics							
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**	**Label**
nRBI	322	49.37267	25.50116	15898	8.00000	121.00000	RBIs in 1986
nHome	322	11.10248	8.69877	3575	0	40.00000	Home Runs in 1986
nRuns	322	52.21739	25.05737	16814	12.00000	130.00000	Runs in 1986

Pearson Correlation Coefficients, N = 322 Prob > |r| under H0: Rho=0			
	nRBI	**nHome**	**nRuns**
nRBI **RBIs in 1986**	1.00000	0.85394 <.0001	0.78053 <.0001
nHome **Home Runs in 1986**	0.85394 <.0001	1.00000	0.63965 <.0001
nRuns **Runs in 1986**	0.78053 <.0001	0.63965 <.0001	1.00000


Scatter Plot Matrix
0
10
20
30
40
125
100
75
50
25
nRBI
40
30
20
10
0
nHome
125
100
75
50
25
nRuns
25
50
75
100
125
25
50
75
100
125



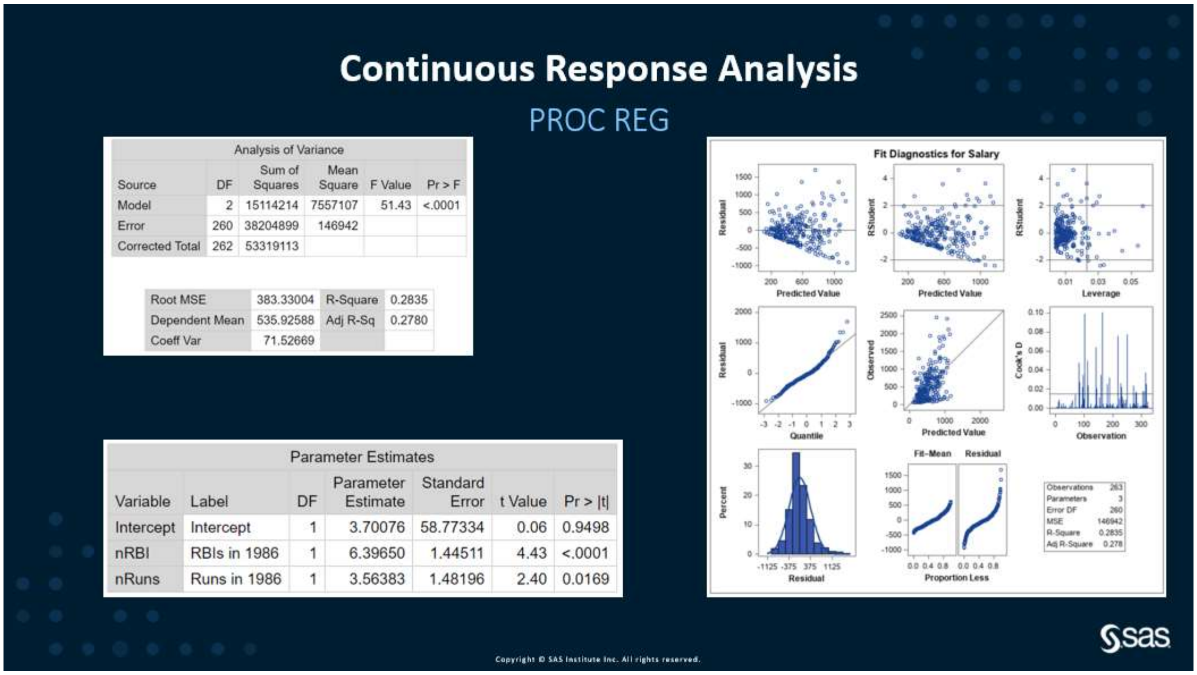
When do I use what?

SGPLOT – creating graphics on your own for exploration

MEANS – summary statistics in table format

UNIVARIATE – summary statistics along with distribution questions

CORR – checks relationships among variables for different purposes



Continuous Response Analysis

PROC REG

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	15114214	7557107	51.43	<.0001
Error	260	38204899	146942		
Corrected Total	262	53319113			

Root MSE	383.33004	R-Square	0.2835
Dependent Mean	535.92588	Adj R-Sq	0.2780
Coeff Var	71.52669		

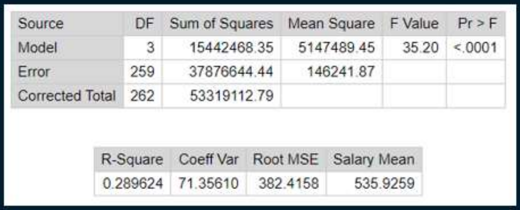
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.70076	58.77334	0.06	0.9498
nRBI	RBIs in 1986	1	6.39650	1.44511	4.43	<.0001
nRuns	Runs in 1986	1	3.56383	1.48196	2.40	0.0169

Fit Diagnostics for Salary

Observations	263
Parameters	3
Error DF	260
MSE	146942
R-Square	0.2835
Adj R-Square	0.278

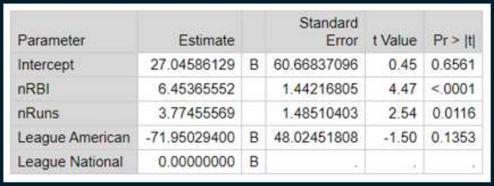
Continuous Response Analysis

PROC GLM

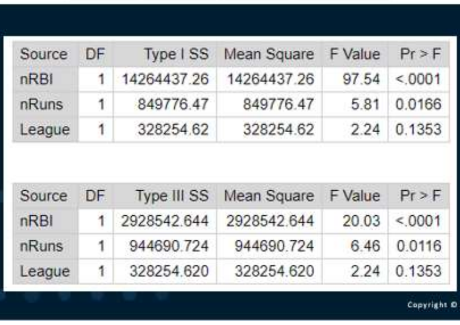


Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	15442468.35	5147489.45	35.20	<.0001
Error	259	37876644.44	146241.87		
Corrected Total	262	53319112.79			

R-Square	Coeff Var	Root MSE	Salary Mean
0.289624	71.35610	382.4158	535.9259



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	27.04586129	B	60.66837096	0.45	0.6561
nRBI	6.45365552		1.44216805	4.47	<.0001
nRuns	3.77455569		1.48510403	2.54	0.0116
League American	-71.95029400	B	48.02451808	-1.50	0.1353
League National	0.00000000	B	.	.	.

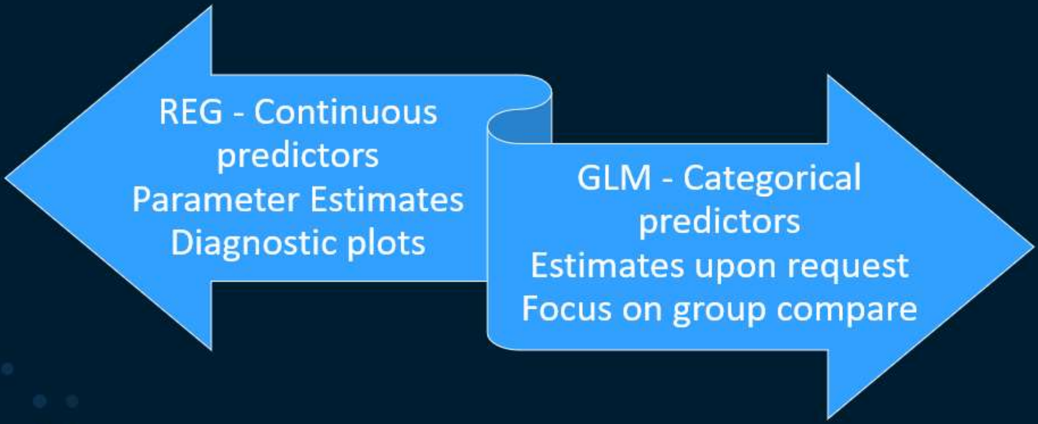


Source	DF	Type I SS	Mean Square	F Value	Pr > F
nRBI	1	14264437.26	14264437.26	97.54	<.0001
nRuns	1	849776.47	849776.47	5.81	0.0166
League	1	328254.62	328254.62	2.24	0.1353

Source	DF	Type III SS	Mean Square	F Value	Pr > F
nRBI	1	2928542.644	2928542.644	20.03	<.0001
nRuns	1	944690.724	944690.724	6.46	0.0116
League	1	328254.620	328254.620	2.24	0.1353

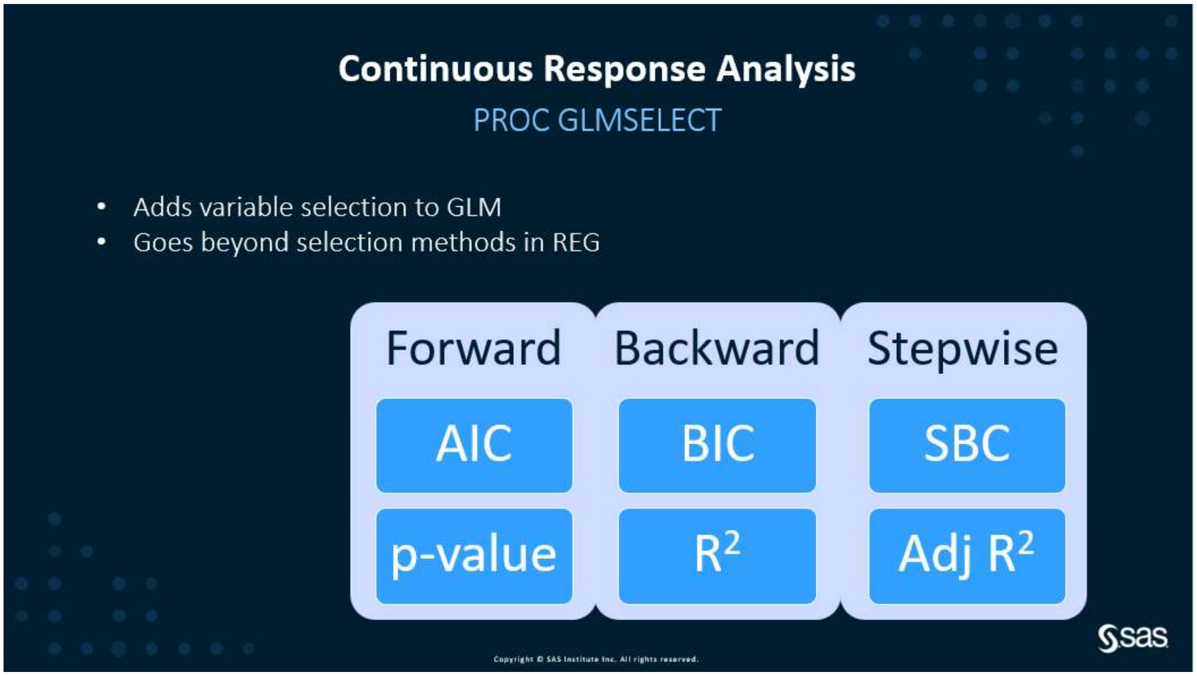
Continuous Response Analysis

Difference between REG and GLM



REG - Continuous predictors
Parameter Estimates
Diagnostic plots

GLM - Categorical predictors
Estimates upon request
Focus on group compare

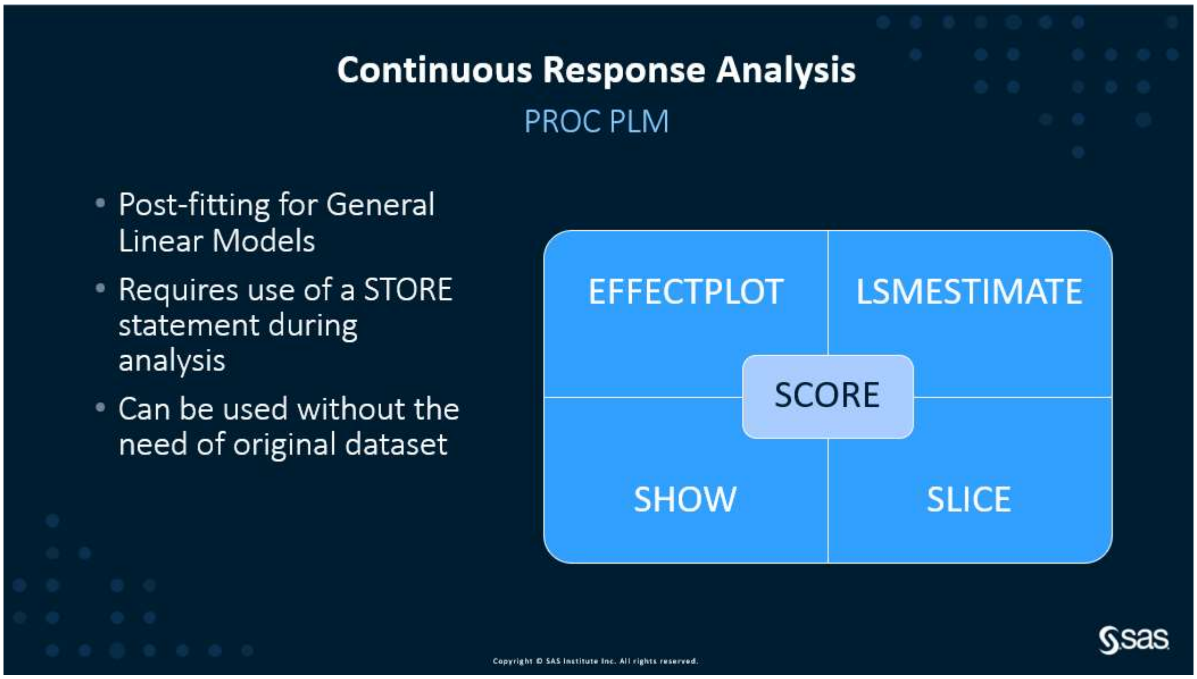


Continuous Response Analysis

PROC GLMSELECT

- Adds variable selection to GLM
- Goes beyond selection methods in REG

Forward	Backward	Stepwise
AIC	BIC	SBC
p-value	R^2	Adj R^2



Continuous Response Analysis

PROC PLM

- Post-fitting for General Linear Models
- Requires use of a STORE statement during analysis
- Can be used without the need of original dataset

EFFECTPLOT
LSMESTIMATE
SCORE
SHOW
SLICE

Demonstration Time

PROC REG: continuous predictors

In [14]:
```
proc reg data=sashelp.baseball;
   model salary = nRBI nHome nRuns;
run;
```

The SAS System

The REG Procedure

Model: MODEL1

Dependent Variable: Salary 1987 Salary in $ Thousands

Number of Observations Read	322
Number of Observations Used	263
Number of Observations with Missing Values	59

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	15366571	5122190	34.96	<.0001
Error	259	37952542	146535		
Corrected Total	262	53319113			

Root MSE	382.79879	R-Square	0.2882
Dependent Mean	535.92588	Adj R-Sq	0.2800
Coeff Var	71.42756		

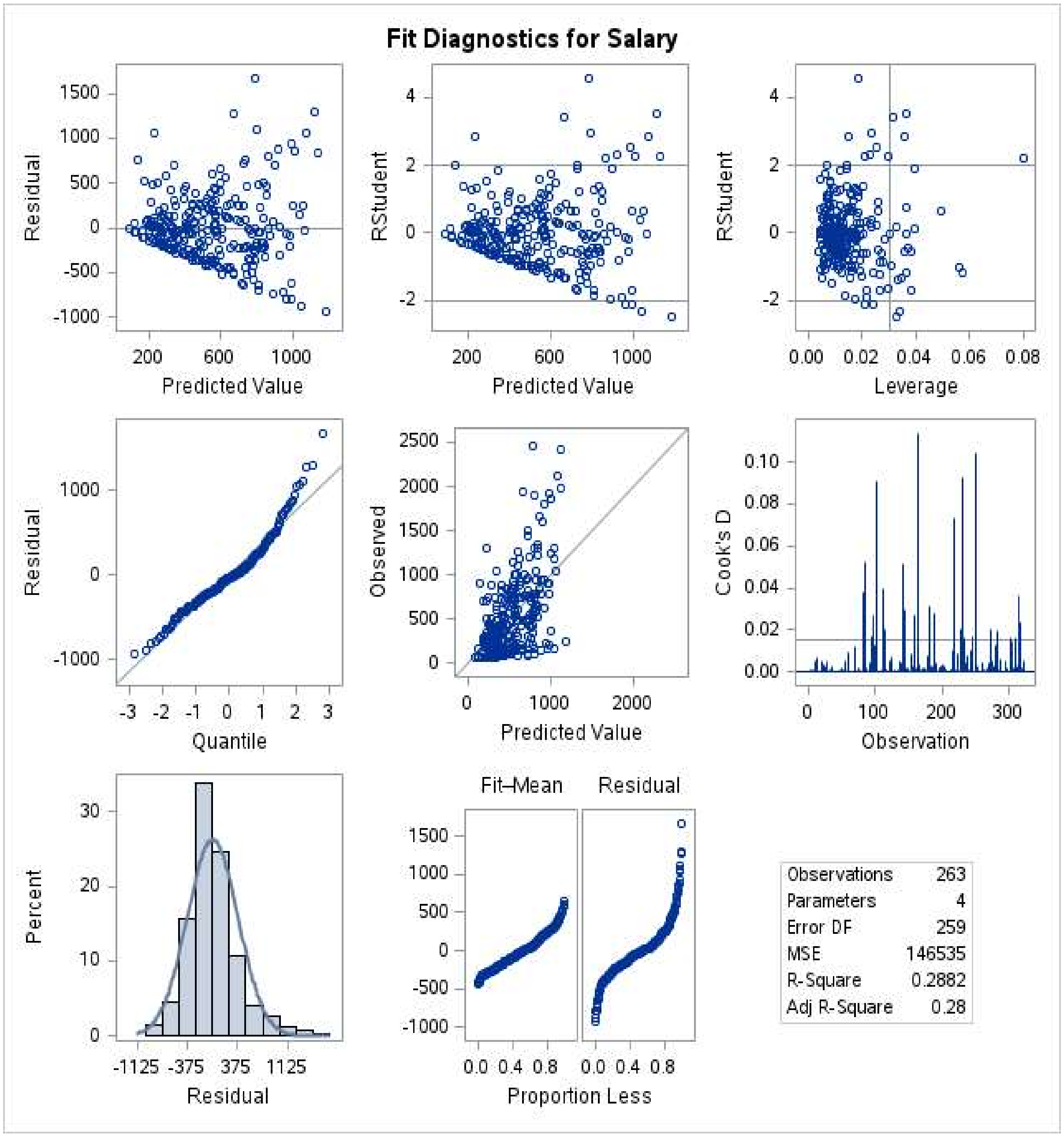
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-16.87956	60.75095	-0.28	0.7814
nRBI	RBIs in 1986	1	8.48984	2.15106	3.95	0.0001
nHome	Home Runs in 1986	1	-6.73206	5.12993	-1.31	0.1906
nRuns	Runs in 1986	1	3.39676	1.48537	2.29	0.0230

The SAS System

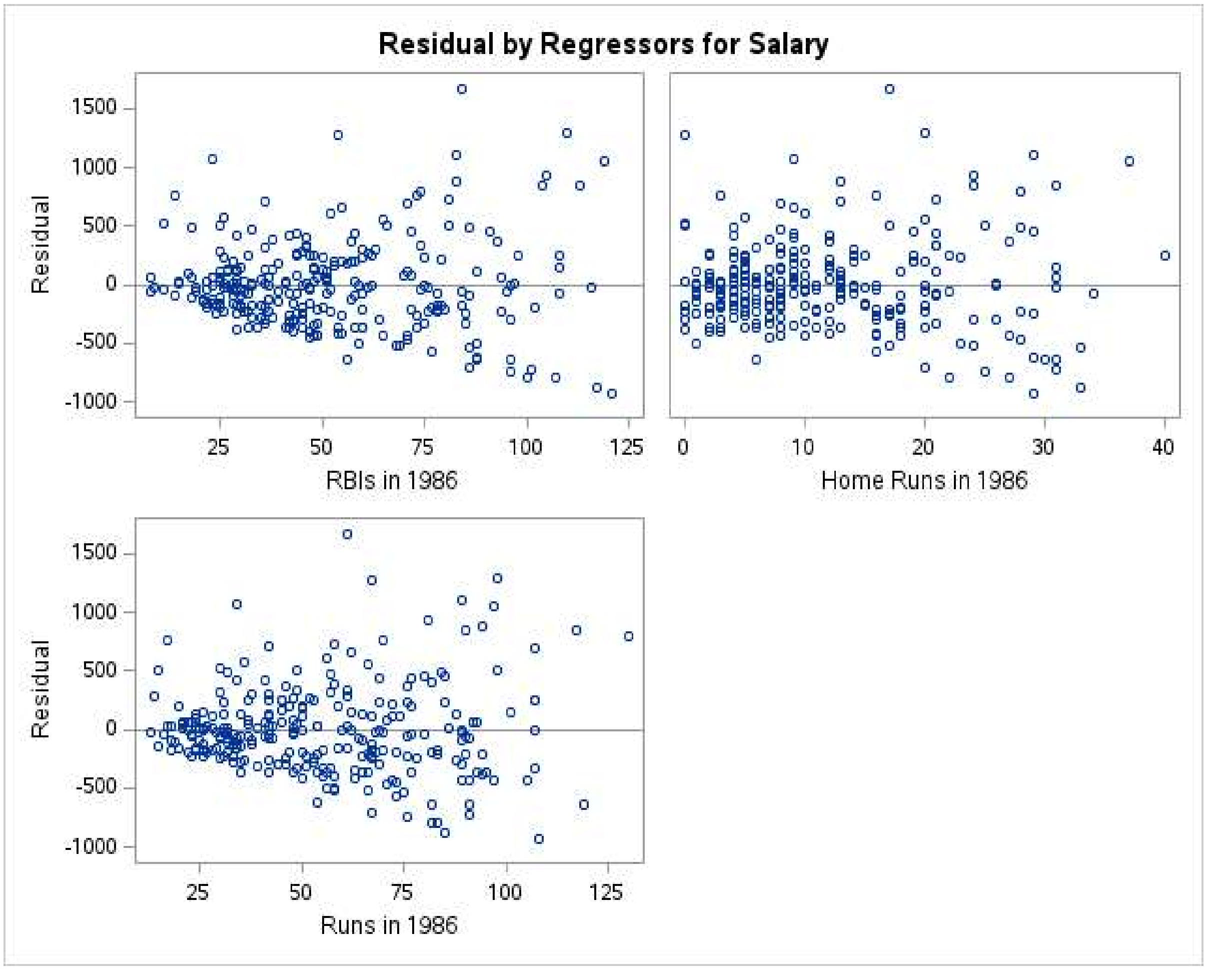
The REG Procedure

Model: MODEL1

Dependent Variable: Salary 1987 Salary in $ Thousands



Observations	263
Parameters	4
Error DF	259
MSE	146535
R-Square	0.2882
Adj R-Square	0.28



PROC GLM: categorical predictors

```
In [15]: proc glm data=sashelp.baseball;
            class league;
            model salary = nRBI nHome nRuns league;
         run;
```

The SAS System

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
League	2	American National

Number of Observations Read	322
Number of Observations Used	263

The SAS System

The GLM Procedure

Dependent Variable: Salary 1987 Salary in $ Thousands

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	15626224.51	3906556.13	26.74	<.0001
Error	258	37692888.28	146096.47		
Corrected Total	262	53319112.79			

R-Square	Coeff Var	Root MSE	Salary Mean
0.293070	71.32062	382.2257	535.9259

Source	DF	Type I SS	Mean Square	F Value	Pr > F
nRBI	1	14264437.26	14264437.26	97.64	<.0001
nHome	1	335833.23	335833.23	2.30	0.1307
nRuns	1	766300.24	766300.24	5.25	0.0228
League	1	259653.78	259653.78	1.78	0.1837

Source	DF	Type III SS	Mean Square	F Value	Pr > F
nRBI	1	2140929.139	2140929.139	14.65	0.0002
nHome	1	183756.160	183756.160	1.26	0.2631
nRuns	1	855238.624	855238.624	5.85	0.0162
League	1	259653.783	259653.783	1.78	0.1837

In [16]:

```
proc glm data=sashelp.baseball;
   class league;
   model salary = nRBI nHome nRuns league / solution;
run;
```

The SAS System

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
League	2	American National

Number of Observations Read	322
Number of Observations Used	263

The SAS System

The GLM Procedure

Dependent Variable: Salary 1987 Salary in $ Thousands

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	15626224.51	3906556.13	26.74	<.0001
Error	258	37692888.28	146096.47		
Corrected Total	262	53319112.79			

R-Square	Coeff Var	Root MSE	Salary Mean
0.293070	71.32062	382.2257	535.9259

Source	DF	Type I SS	Mean Square	F Value	Pr > F
nRBI	1	14264437.26	14264437.26	97.64	<.0001
nHome	1	335833.23	335833.23	2.30	0.1307
nRuns	1	766300.24	766300.24	5.25	0.0228
League	1	259653.78	259653.78	1.78	0.1837

Source	DF	Type III SS	Mean Square	F Value	Pr > F
nRBI	1	2140929.139	2140929.139	14.65	0.0002
nHome	1	183756.160	183756.160	1.26	0.2631
nRuns	1	855238.624	855238.624	5.85	0.0162
League	1	259653.783	259653.783	1.78	0.1837

Parameter	Estimate		Standard Error	t Value	Pr > |t|

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	6.93179332	B	63.23489614	0.11	0.9128
nRBI	8.25072642		2.15531572	3.83	0.0002
nHome	-5.79810409		5.16993294	-1.12	0.2631
nRuns	3.60910175		1.49167884	2.42	0.0162
League American	-64.58756513	B	48.44750777	-1.33	0.1837
League National	0.00000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

PROC GLMSELECT: GLM plus additional variable selection beyond REG

In [17]:

```
proc glmselect data=sashelp.baseball;
   class league;
   model salary = nRBI nHome nRuns league / selection=forward select=AIC;
run;
```

The SAS System

The GLMSELECT Procedure

Data Set	SASHELP.BASEBALL
Dependent Variable	Salary
Selection Method	Forward
Select Criterion	AIC
Stop Criterion	AIC
Effect Hierarchy Enforced	None

Number of Observations Read	322
Number of Observations Used	263

Class Level Information

Class	Levels	Values
League	2	American National

Dimensions

Number of Effects	5
Number of Parameters	6

The SAS System

The GLMSELECT Procedure

Forward Selection Summary

Step	Effect Entered	Number Effects In	Number Parms In	AIC
0	**Intercept**	1	1	3480.7683
1	**nRBI**	2	2	3400.8879
2	**nRuns**	3	3	3397.1022
3	**League**	4	4	3396.8328*

*** Optimal Value of Criterion**

Selection stopped at a local minimum of the AIC criterion.

Stop Details				
Candidate For	Effect	Candidate AIC		Compare AIC
Entry	nHome	3397.5538	>	3396.8328

The SAS System

The GLMSELECT Procedure
Selected Model

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

Effects:	Intercept nRBI nRuns League

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	15442468	5147489	35.20
Error	259	37876644	146242	
Corrected Total	262	53319113		

Root MSE	382.41583
Dependent Mean	535.92588
R-Square	0.2896
Adj R-Sq	0.2814
AIC	3396.83279
AICC	3397.06625
SBC	3146.12140

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	27.045861	60.668371	0.45
nRBI	1	6.453656	1.442168	4.47
nRuns	1	3.774556	1.485104	2.54
League American	1	-71.950294	48.024518	-1.50
League National	0	0	.	.

In [18]:

```
proc glmselect data=sashelp.baseball;
   class league;
   model salary = nRBI nHome nRuns league / selection=backward select=sl;
run;
```

The SAS System

The GLMSELECT Procedure

Data Set	SASHELP.BASEBALL
Dependent Variable	Salary
Selection Method	Backward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Stay Significance Level (SLS)	0.1
Effect Hierarchy Enforced	None

Number of Observations Read	322
Number of Observations Used	263

Class Level Information		
Class	**Levels**	**Values**
League	2	American National

Dimensions	
Number of Effects	5
Number of Parameters	6

The SAS System

The GLMSELECT Procedure

Backward Selection Summary					
Step	**Effect Removed**	**Number Effects In**	**Number Parms In**	**F Value**	**Pr > F**
0		5	5		
1	**nHome**	4	4	1.26	0.2631
2	**League**	3	3	2.24	0.1353

Selection stopped because the next candidate for removal has SLS < 0.1.

Stop Details					
Candidate For	**Effect**	**Candidate Significance**		**Compare Significance**	

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Removal	nRuns	0.0169	<	0.1000	(SLS)

The SAS System

The GLMSELECT Procedure

Selected Model

The selected model is the model at the last step (Step 2).

Effects:	Intercept nRBI nRuns

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	15114214	7557107	51.43
Error	260	38204899	146942	
Corrected Total	262	53319113		

Root MSE	383.33004
Dependent Mean	535.92588
R-Square	0.2835
Adj R-Sq	0.2780
AIC	3397.10223
AICC	3397.25727
SBC	3142.81870

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	3.700763	58.773344	0.06
nRBI	1	6.396502	1.445110	4.43
nRuns	1	3.563828	1.481963	2.40

Running PROC PLM after GLMSELECT: Slicing into the Interaction Term

In [20]:

```
proc glmselect data=sashelp.baseball;
   class league division;
   model salary = nRBI nHome nRuns league|division / selection=none showpvalues;
   store out=baseballitem;
run;

proc plm restore=baseballitem plots=all;
   slice league*division / sliceby=league;
run;
```

The SAS System

The GLMSELECT Procedure

Data Set	SASHELP.BASEBALL
Dependent Variable	Salary
Selection Method	None

Number of Observations Read	322
Number of Observations Used	263

Class Level Information		
Class	**Levels**	**Values**
League	2	American National
Division	2	East West

Dimensions	
Number of Effects	7
Number of Parameters	12

The SAS System

The GLMSELECT Procedure

Least Squares Summary				
Step	**Effect Entered**	**Number Effects In**	**Number Parms In**	**SBC**
0	**Intercept**	1	1	3219.3405
1	**nRBI**	2	2	3143.0322*
2	**nHome**	3	3	3146.3331
3	**nRuns**	4	4	3146.6479
4	**League**	5	5	3150.4145
5	**Division**	6	6	3150.2327
6	**League*Division**	7	7	3153.2216
*** Optimal Value of Criterion**				

The SAS System

The GLMSELECT Procedure

Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	16802369	2800395	19.63	<.0001
Error	256	36516744	142644		
Corrected Total	262	53319113			

Root MSE	377.68179
Dependent Mean	535.92588
R-Square	0.3151
Adj R-Sq	0.2991
AIC	3393.21651
AICC	3393.78344
SBC	3153.22159

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.942555	69.096788	0.14	0.8857
nRBI	1	8.091463	2.136544	3.79	0.0002
nHome	1	-5.267185	5.126548	-1.03	0.3052
nRuns	1	3.264631	1.478834	2.21	0.0282
League American	1	-134.717548	65.909080	-2.04	0.0420
League National	0	0	.	.	.
Division East	1	34.218084	68.217205	0.50	0.6164
Division West	0	0	.	.	.
League*Division American East	1	148.651987	93.514182	1.59	0.1132
League*Division American West	0	0	.	.	.
League*Division National East	0	0	.	.	.
League*Division National West	0	0	.	.	.

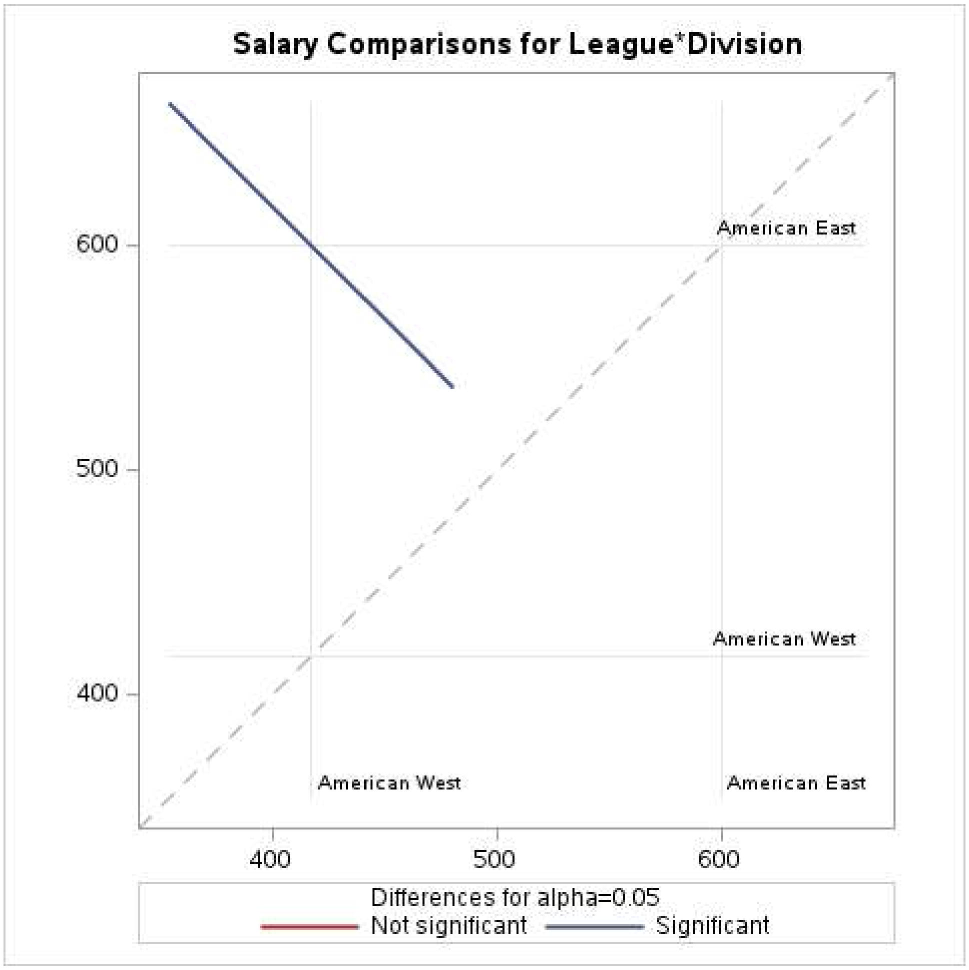
The SAS System

The PLM Procedure

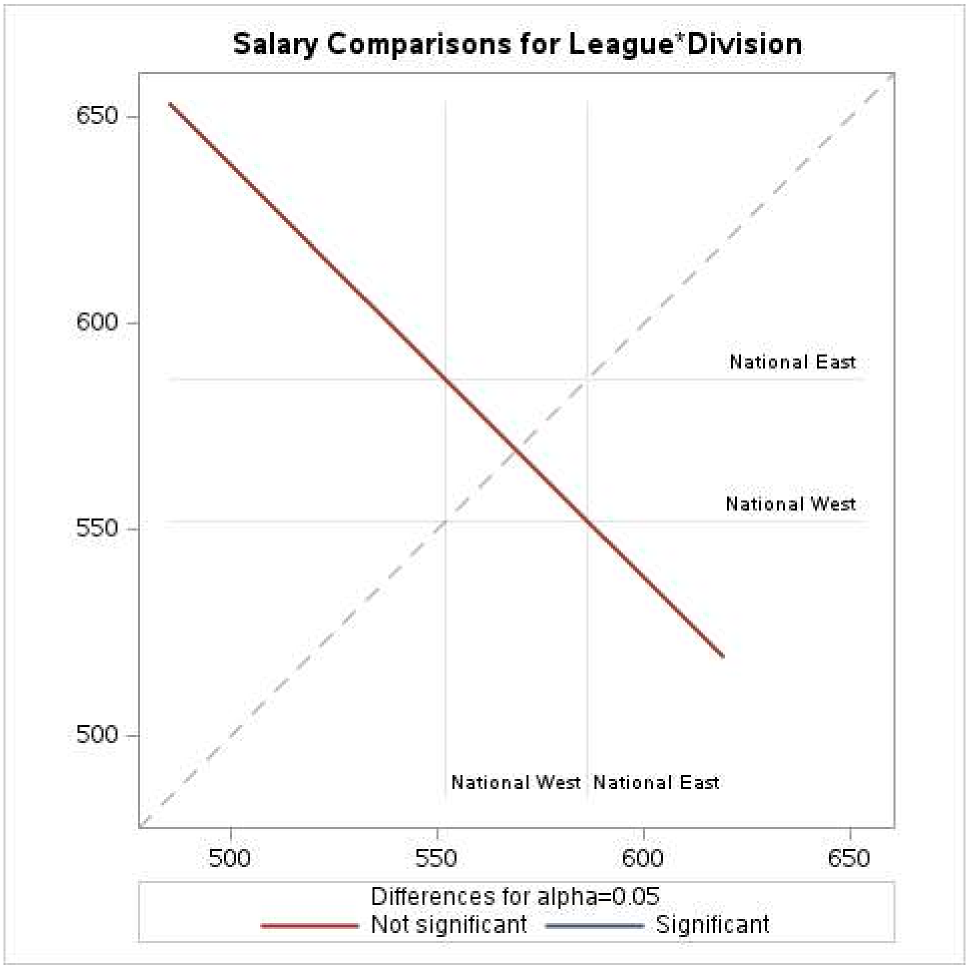
Store Information	
Item Store	WORK.BASEBALLITEM
Data Set Created From	SASHELP.BASEBALL
Created By	PROC GLMSELECT
Date Created	31JUL25:15:10:53
Response Variable	Salary
Class Variables	League Division
Model Effects	Intercept nRBI nHome nRuns League Division League*Division

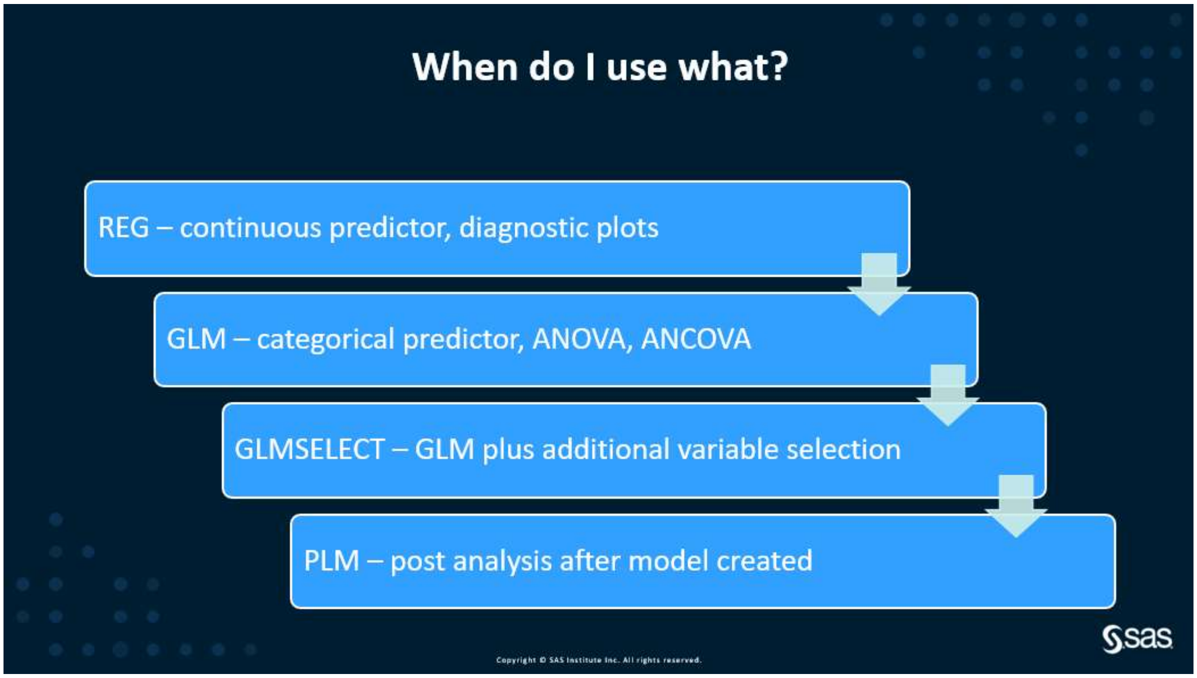
Class Level Information		
Class	Levels	Values
League	2	American National
Division	2	East West

F Test for League*Division Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
League American	1	256	8.02	0.0050



F Test for League*Division Least Squares Means Slice				
Slice	Num DF	Den DF	F Value	Pr > F
League National	1	256	0.25	0.6164





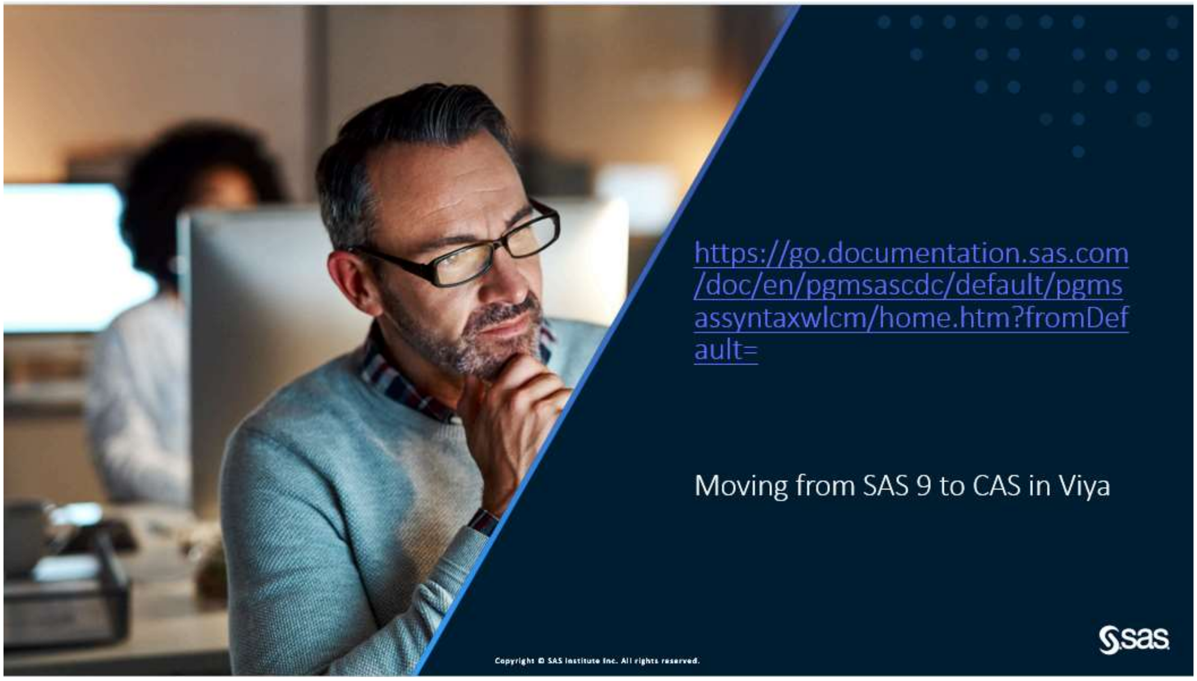
When do I use what?

REG – continuous predictor, diagnostic plots

GLM – categorical predictor, ANOVA, ANCOVA

GLMSELECT – GLM plus additional variable selection

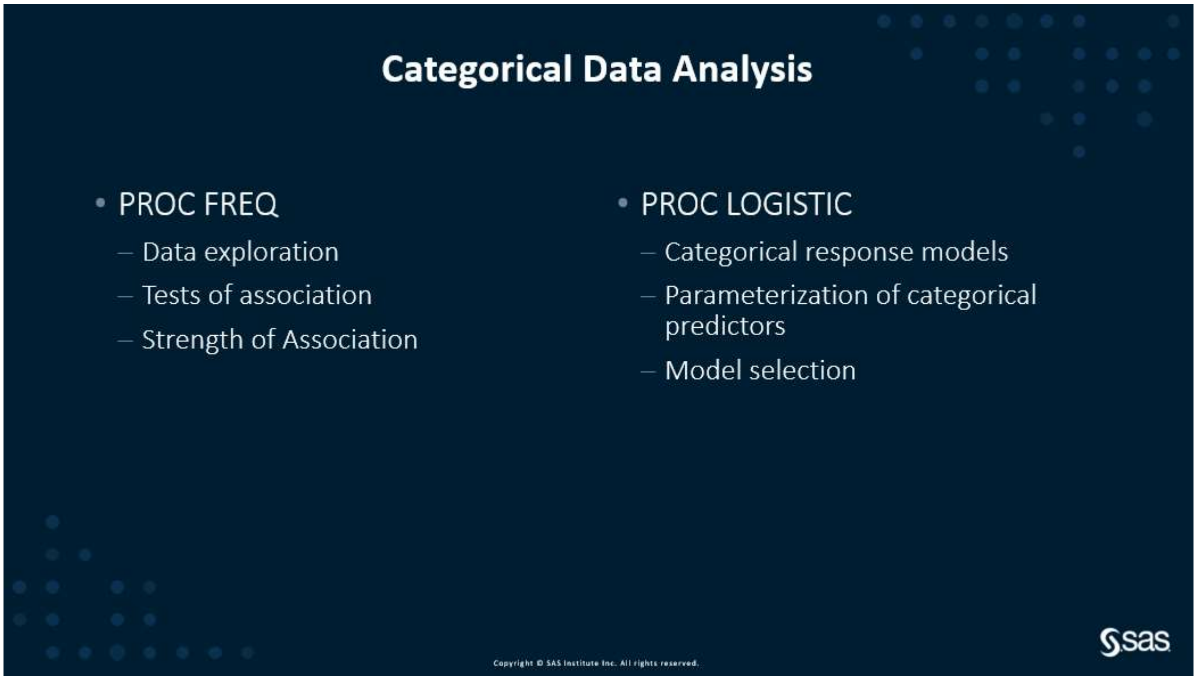
PLM – post analysis after model created



https://go.documentation.sas.com/doc/en/pgmsascdc/default/pgmsassyntaxwlcm/home.htm?fromDefault=

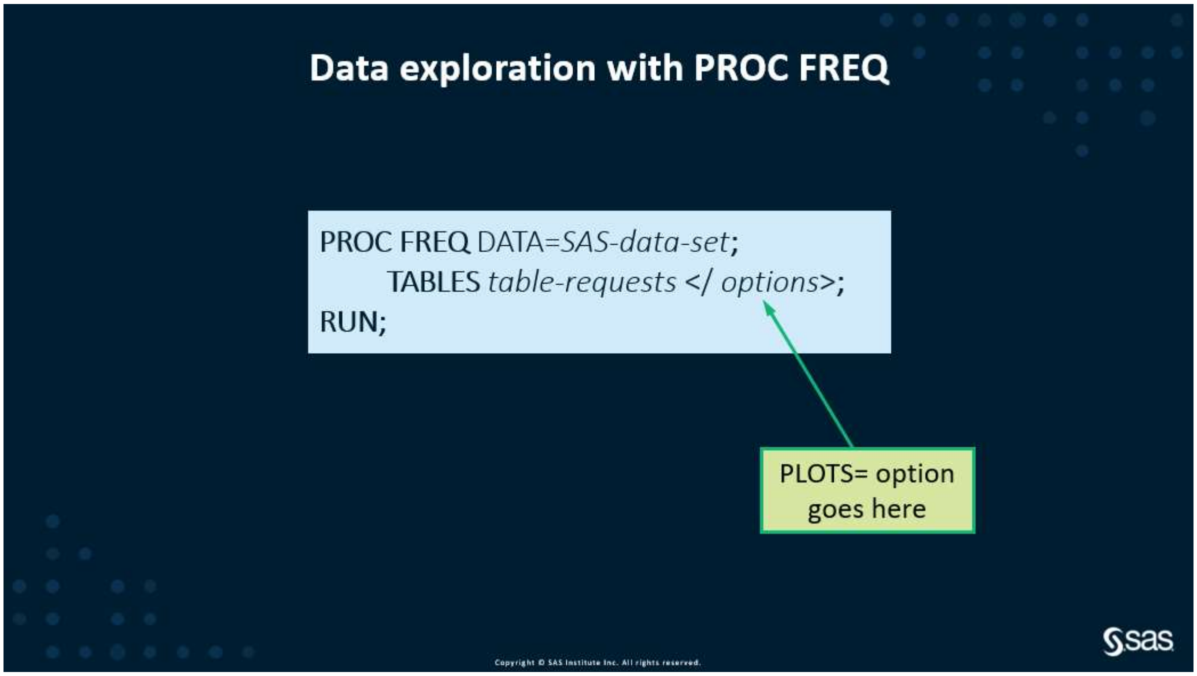
Moving from SAS 9 to CAS in Viya

https://go.documentation.sas.com/doc/en/pgmsascdc/default/pgmsassyntaxwlcm/home.htm?fromDefault=



Categorical Data Analysis

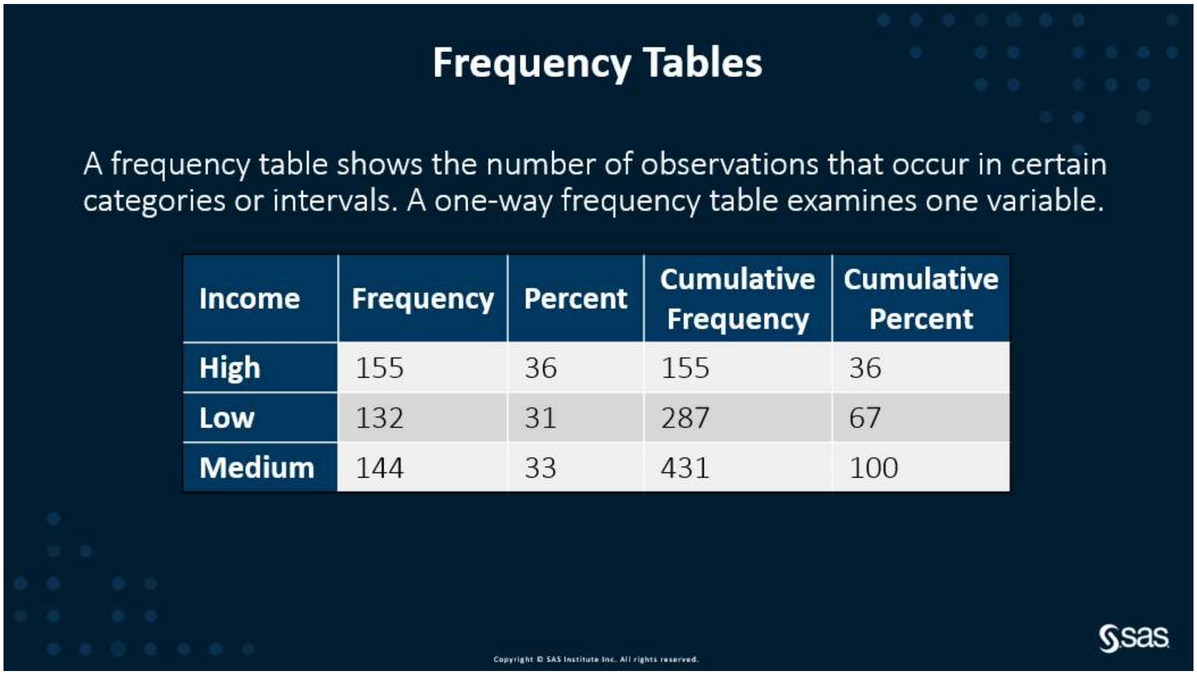
- PROC FREQ
  - Data exploration
  - Tests of association
  - Strength of Association
- PROC LOGISTIC
  - Categorical response models
  - Parameterization of categorical predictors
  - Model selection



Data exploration with PROC FREQ

```
PROC FREQ DATA=SAS-data-set;
      TABLES table-requests </ options>;
RUN;
```

PLOTS= option goes here



Frequency Tables

A frequency table shows the number of observations that occur in certain categories or intervals. A one-way frequency table examines one variable.

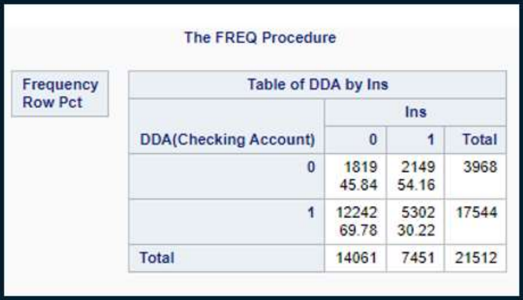
Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	155	36	155	36
Low	132	31	287	67
Medium	144	33	431	100

Crosstabulation Tables

A crosstabulation table shows the number of observations for each combination of the row and column variables.

	column 1	column 2	…	column c
row 1	cell$_{11}$	cell$_{12}$	…	cell$_{1c}$
row 2	cell$_{21}$	cell$_{22}$	…	cell$_{2c}$
…	…	…	…	…
row r	cell$_{r1}$	cell$_{r2}$	…	cell$_{rc}$

Evidence of Association

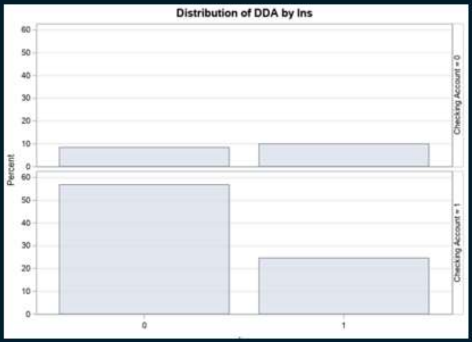


The FREQ Procedure

Frequency
Row Pct

Table of DDA by Ins

	Ins		
DDA(Checking Account)	0	1	Total
0	1819 45.84	2149 54.16	3968
1	12242 69.78	5302 30.22	17544
Total	14061	7451	21512



Demonstration Time

Running PROC FREQ to Explore Categorical Variables

In [2]:

```
proc freq data=sashelp.heart;
   table status BP_status Chol_status sex Weight_status;
run;
```

The FREQ Procedure

Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Alive	3218	61.78	3218	61.78
Dead	1991	38.22	5209	100.00

Blood Pressure Status				
BP_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	2267	43.52	2267	43.52
Normal	2143	41.14	4410	84.66
Optimal	799	15.34	5209	100.00

Cholesterol Status				
Chol_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Borderline	1861	36.80	1861	36.80
Desirable	1405	27.78	3266	64.58
High	1791	35.42	5057	100.00
Frequency Missing = 152				

Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	2873	55.15	2873	55.15
Male	2336	44.85	5209	100.00

Weight Status				
Weight_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Normal	1472	28.29	1472	28.29
Overweight	3550	68.23	5022	96.52
Underweight	181	3.48	5203	100.00
Frequency Missing = 6				

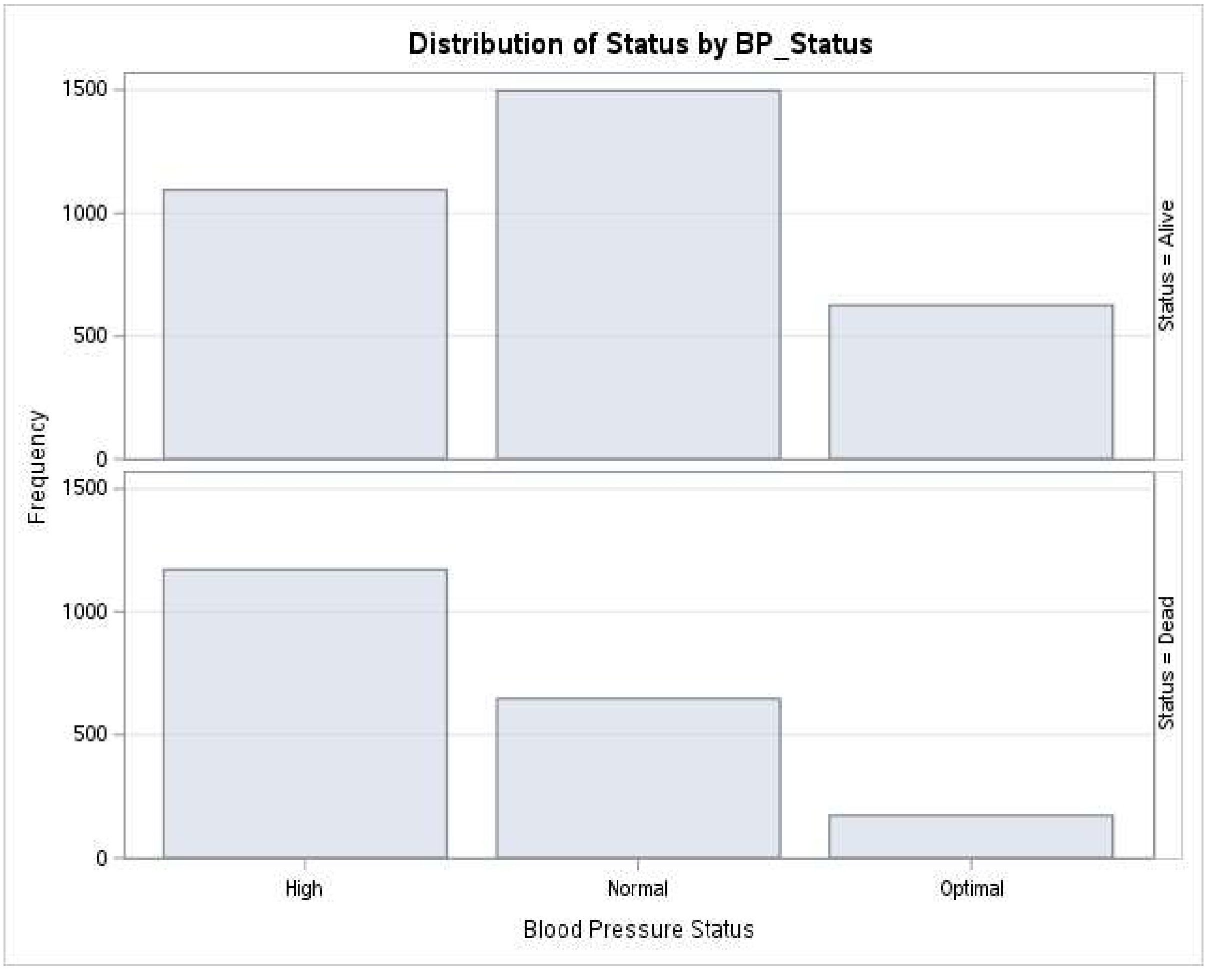
In [3]:

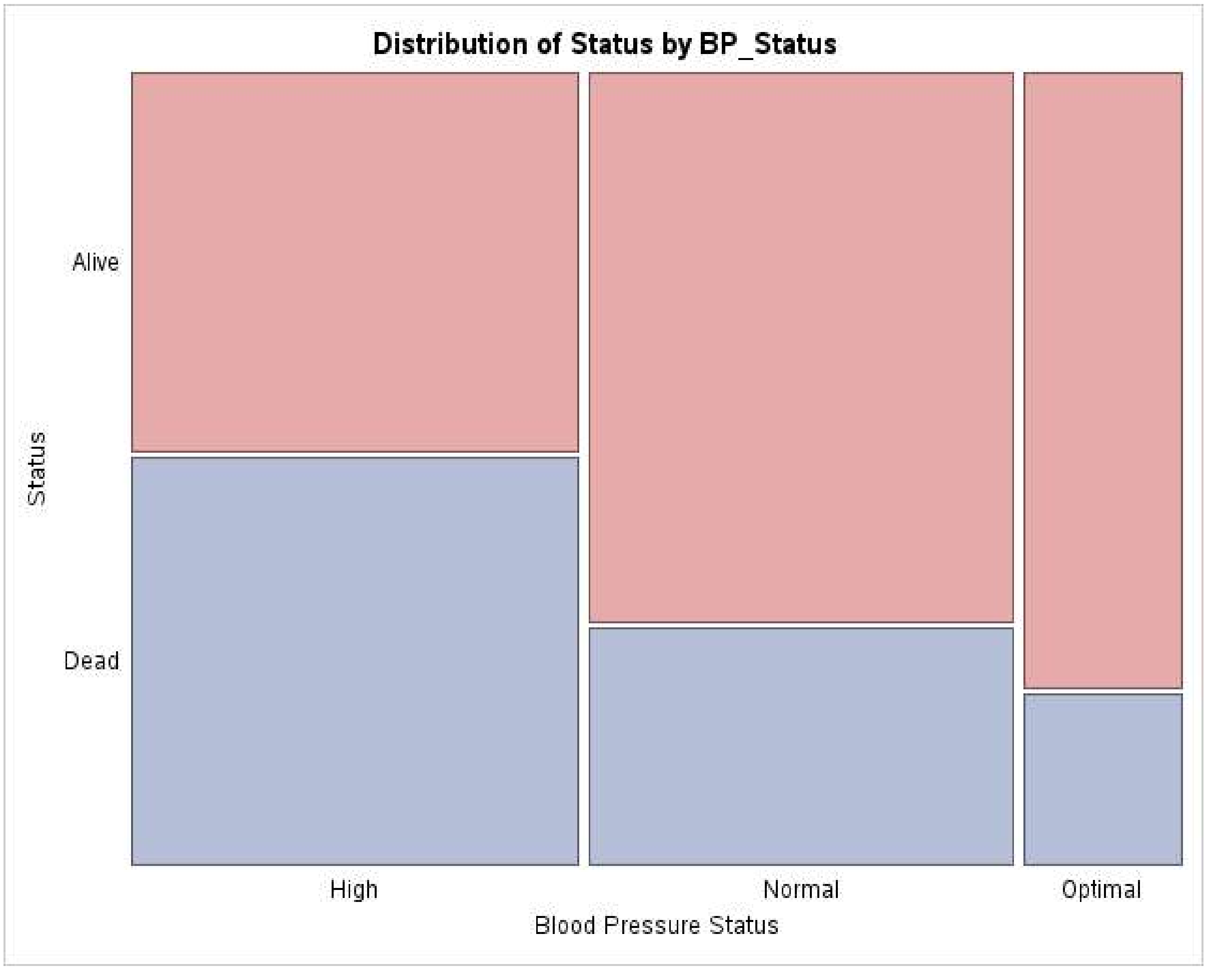
```
proc freq data=sashelp.heart;
   table status*(BP_status Chol_status sex Weight_status) / plots=all chisq;
run;
```

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Status by BP_Status				
Status	BP_Status(Blood Pressure Status)			
	High	Normal	Optimal	Total
Alive	1095 21.02 34.03 48.30	1497 28.74 46.52 69.86	626 12.02 19.45 78.35	3218 61.78
Dead	1172 22.50 58.86 51.70	646 12.40 32.45 30.14	173 3.32 8.69 21.65	1991 38.22
Total	2267 43.52	2143 41.14	799 15.34	5209 100.00





Statistics for Table of Status by BP_Status

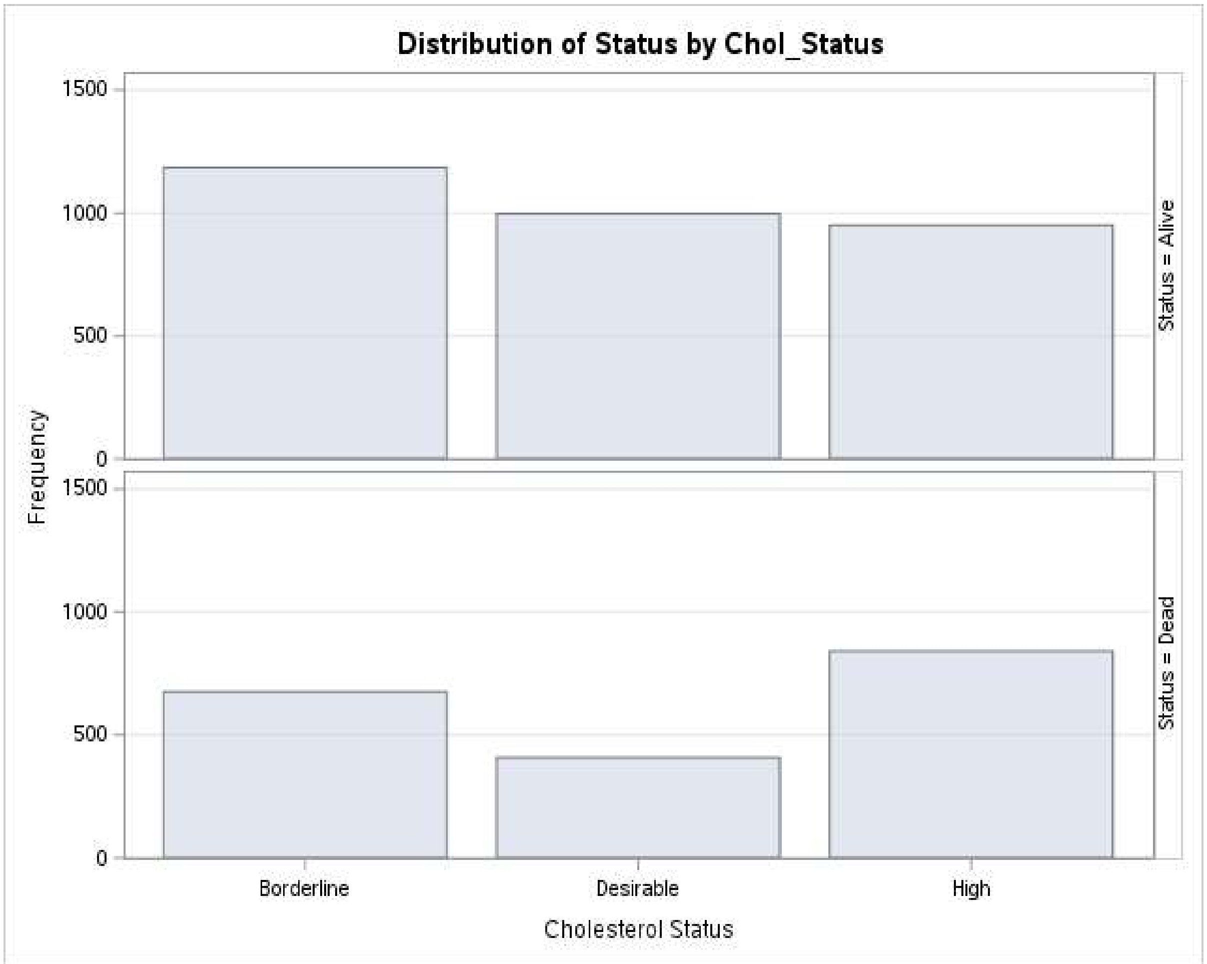
Statistic	DF	Value	Prob
Chi-Square	2	326.4757	<.0001
Likelihood Ratio Chi-Square	2	331.0338	<.0001
Mantel-Haenszel Chi-Square	1	306.1190	<.0001
Phi Coefficient		0.2504	
Contingency Coefficient		0.2429	
Cramer's V		0.2504	

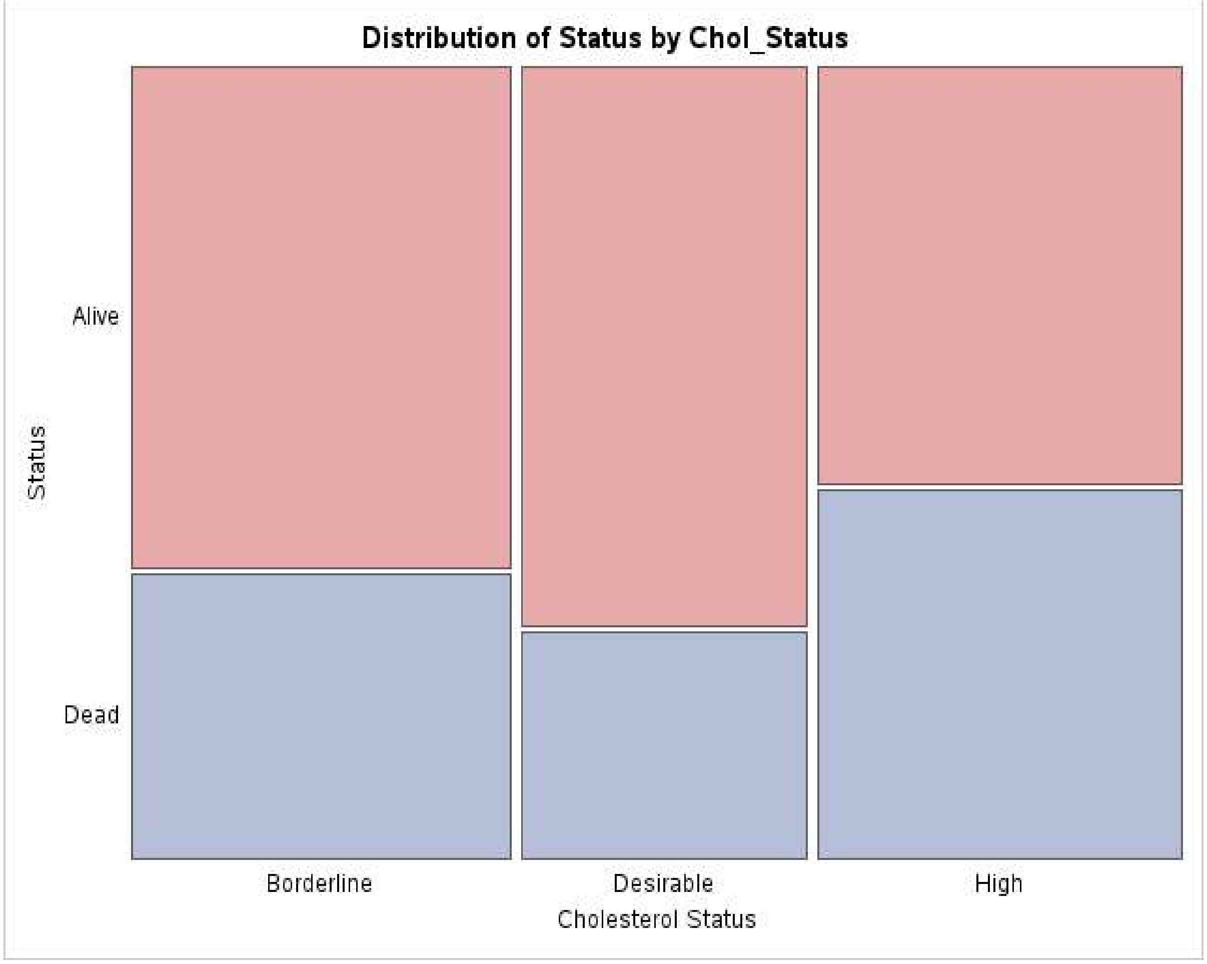
Sample Size = 5209

Frequency Percent Row Pct Col Pct

Table of Status by Chol_Status				
Status	Chol_Status(Cholesterol Status)			
	Borderline	Desirable	High	Total
Alive	1186 23.45 37.83 63.73	998 19.74 31.83 71.03	951 18.81 30.33 53.10	3135 61.99
Dead	675 13.35	407 8.05	840 16.61	1922 38.01

Table of Status by Chol_Status				
Status	Chol_Status(Cholesterol Status)			
	Borderline	Desirable	High	Total
	35.12 36.27	21.18 28.97	43.70 46.90	
Total	1861 36.80	1405 27.78	1791 35.42	5057 100.00
Frequency Missing = 152				





Statistics for Table of Status by Chol_Status

Statistic	DF	Value	Prob
Chi-Square	2	111.2331	<.0001
Likelihood Ratio Chi-Square	2	111.7053	<.0001
Mantel-Haenszel Chi-Square	1	42.6684	<.0001
Phi Coefficient		0.1483	
Contingency Coefficient		0.1467	
Cramer's V		0.1483	

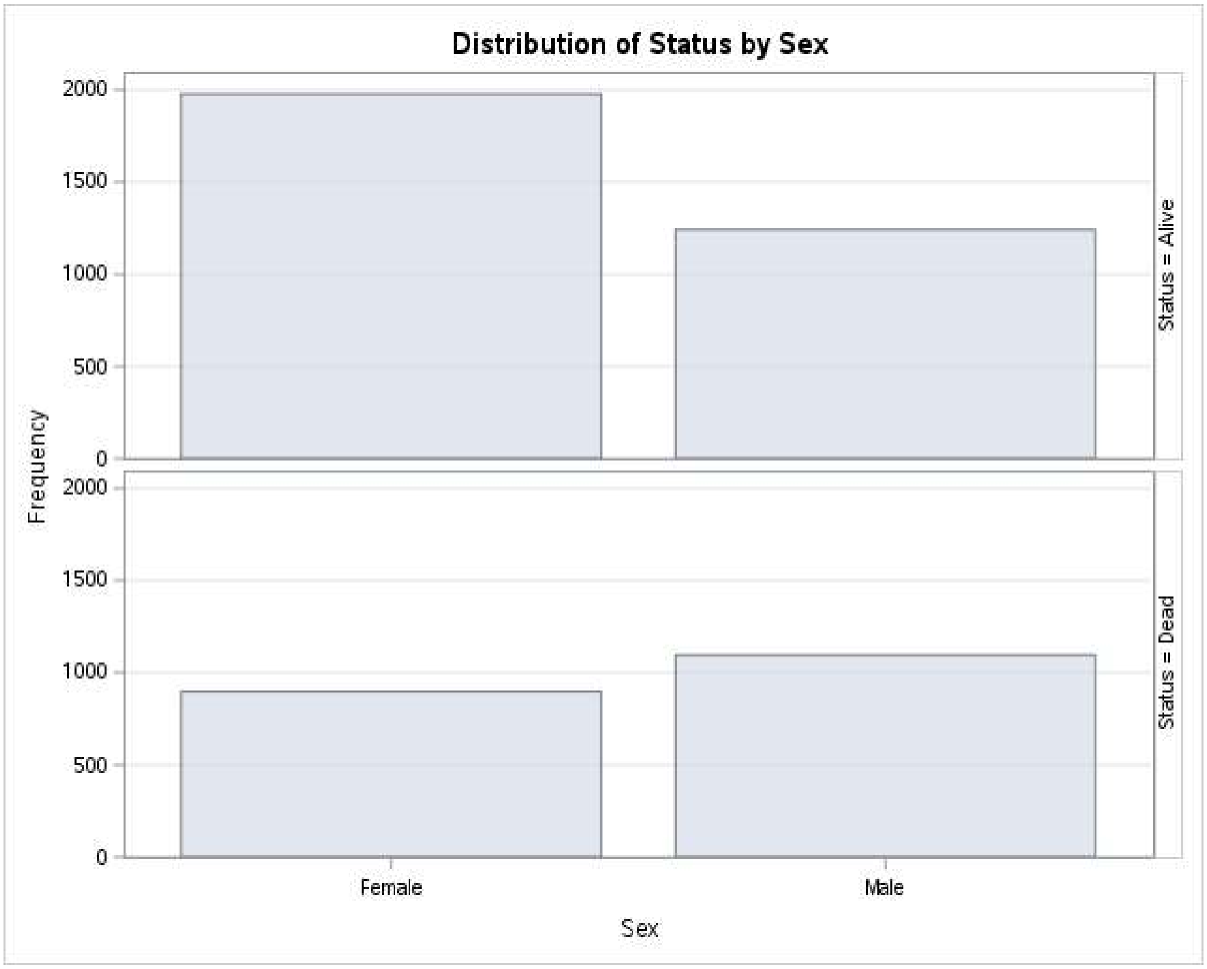
Sample Size = 5057

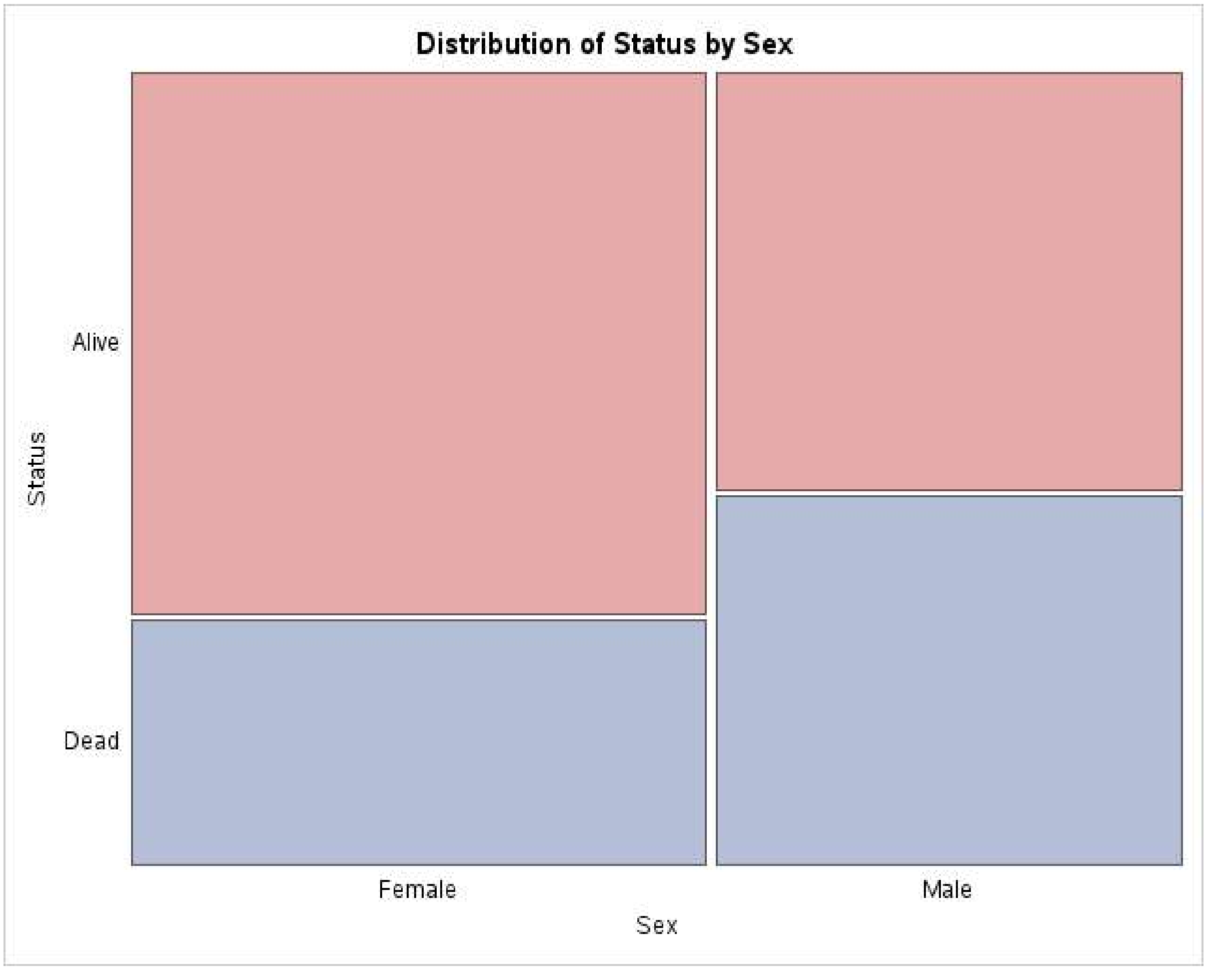
Frequency Missing = 152

Frequency Percent Row Pct Col Pct

Table of Status by Sex			
Status	Sex		
	Female	Male	Total
Alive	1977 37.95 61.44 68.81	1241 23.82 38.56 53.13	3218 61.78

Table of Status by Sex			
Status	Sex		
	Female	Male	Total
Dead	896 17.20 45.00 31.19	1095 21.02 55.00 46.88	1991 38.22
Total	2873 55.15	2336 44.85	5209 100.00





Statistics for Table of Status by Sex

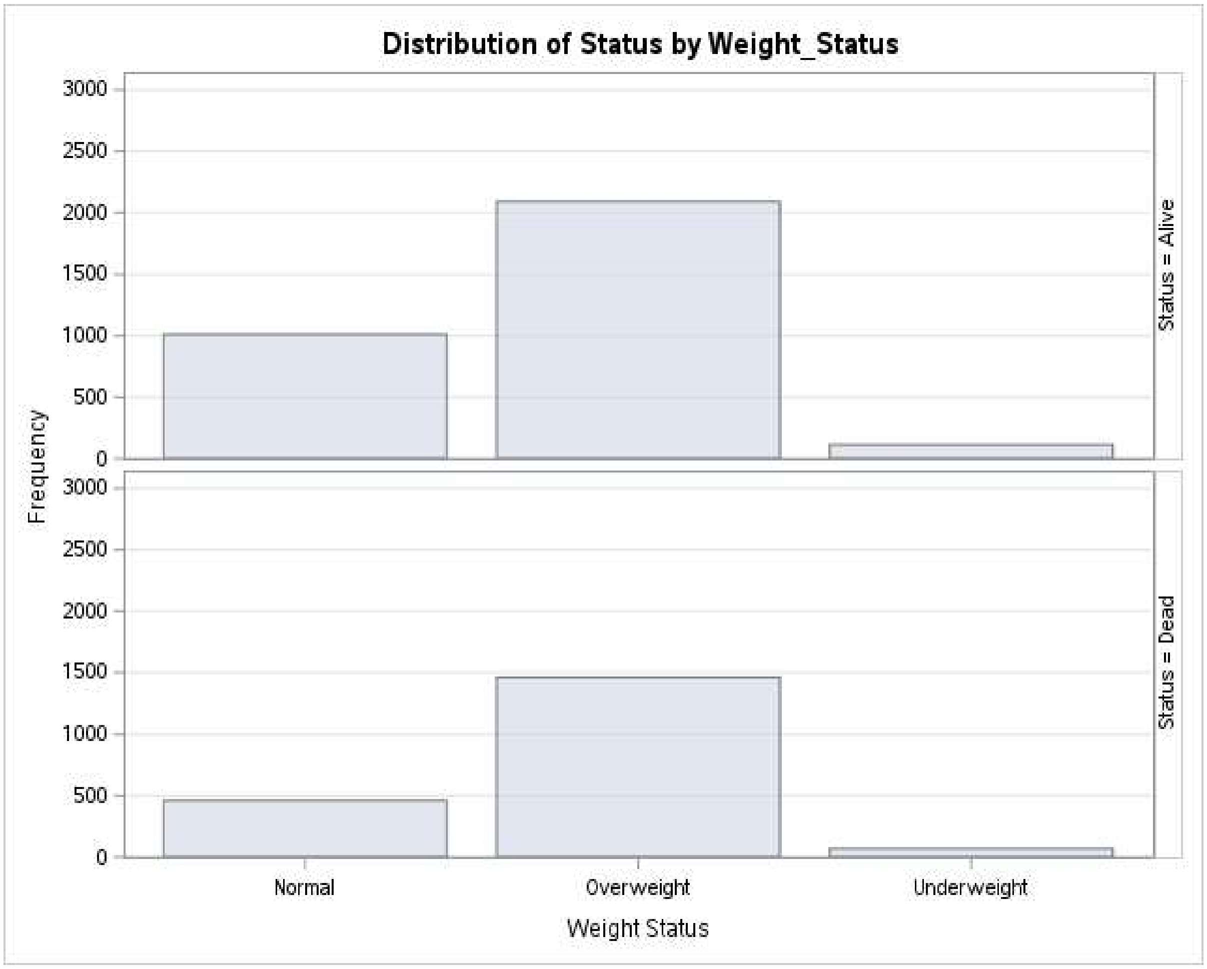
Statistic	DF	Value	Prob
Chi-Square	1	134.2906	<.0001
Likelihood Ratio Chi-Square	1	134.2973	<.0001
Continuity Adj. Chi-Square	1	133.6270	<.0001
Mantel-Haenszel Chi-Square	1	134.2648	<.0001
Phi Coefficient		0.1606	
Contingency Coefficient		0.1585	
Cramer's V		0.1606	

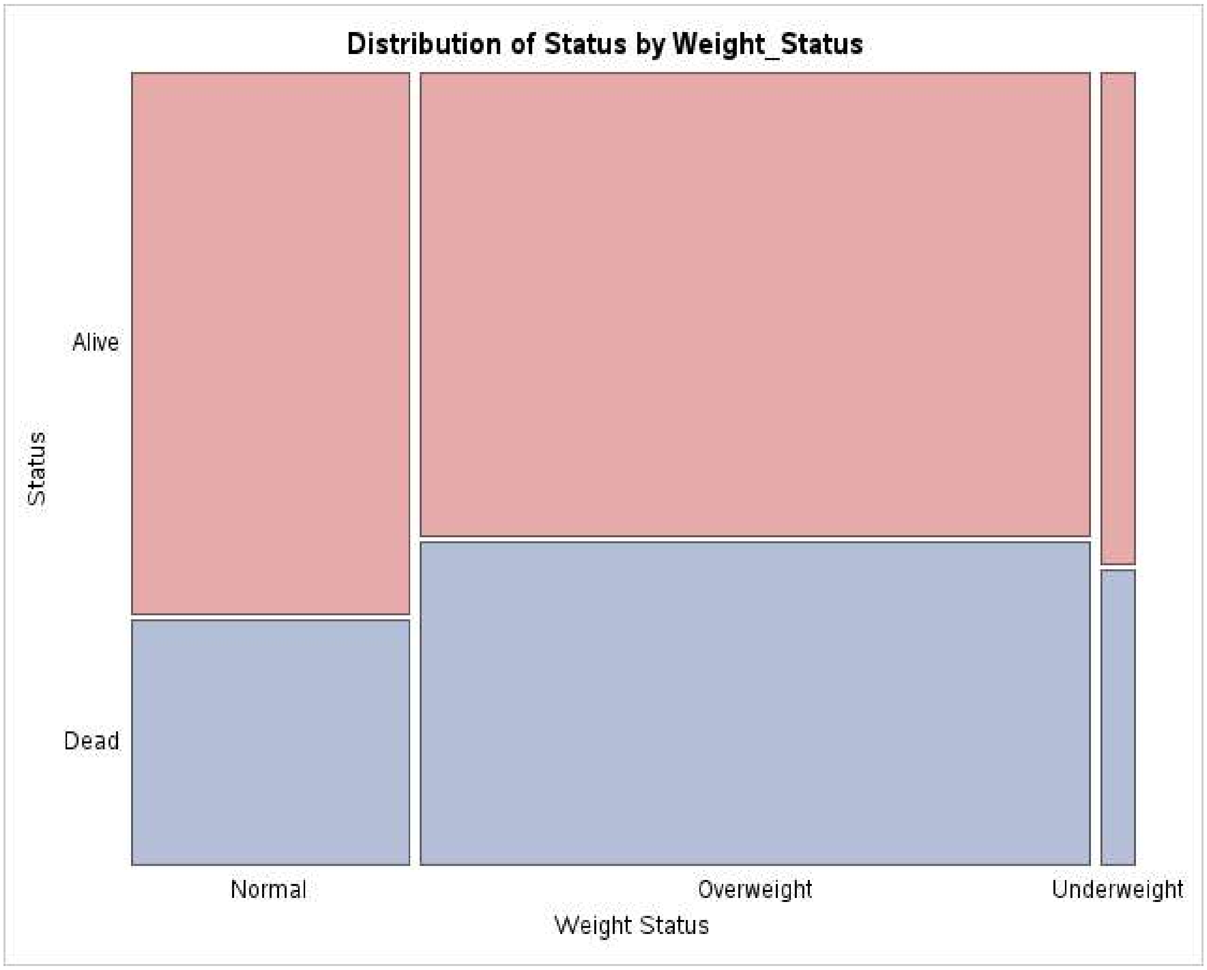
Fisher's Exact Test	
Cell (1,1) Frequency (F)	1977
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 5209

Frequency
Percent
Row Pct
Col Pct

Table of Status by Weight_Status				
Status	Weight_Status(Weight Status)			
	Normal	Overweight	Underweight	Total
Alive	1012 19.45 31.48 68.75	2090 40.17 65.01 58.87	113 2.17 3.51 62.43	3215 61.79
Dead	460 8.84 23.14 31.25	1460 28.06 73.44 41.13	68 1.31 3.42 37.57	1988 38.21
Total	1472 28.29	3550 68.23	181 3.48	5203 100.00
Frequency Missing = 6				



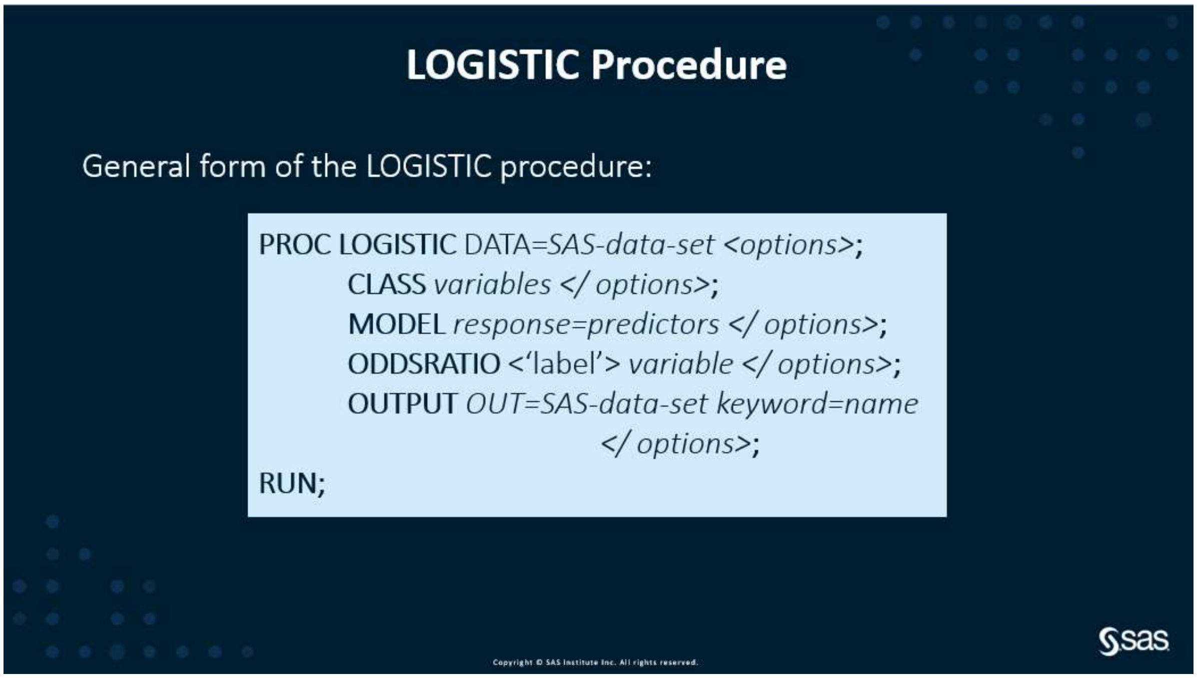


Statistics for Table of Status by Weight_Status

Statistic	DF	Value	Prob
Chi-Square	2	43.0256	<.0001
Likelihood Ratio Chi-Square	2	43.7488	<.0001
Mantel-Haenszel Chi-Square	1	32.5916	<.0001
Phi Coefficient		0.0909	
Contingency Coefficient		0.0906	
Cramer's V		0.0909	

Sample Size = 5203

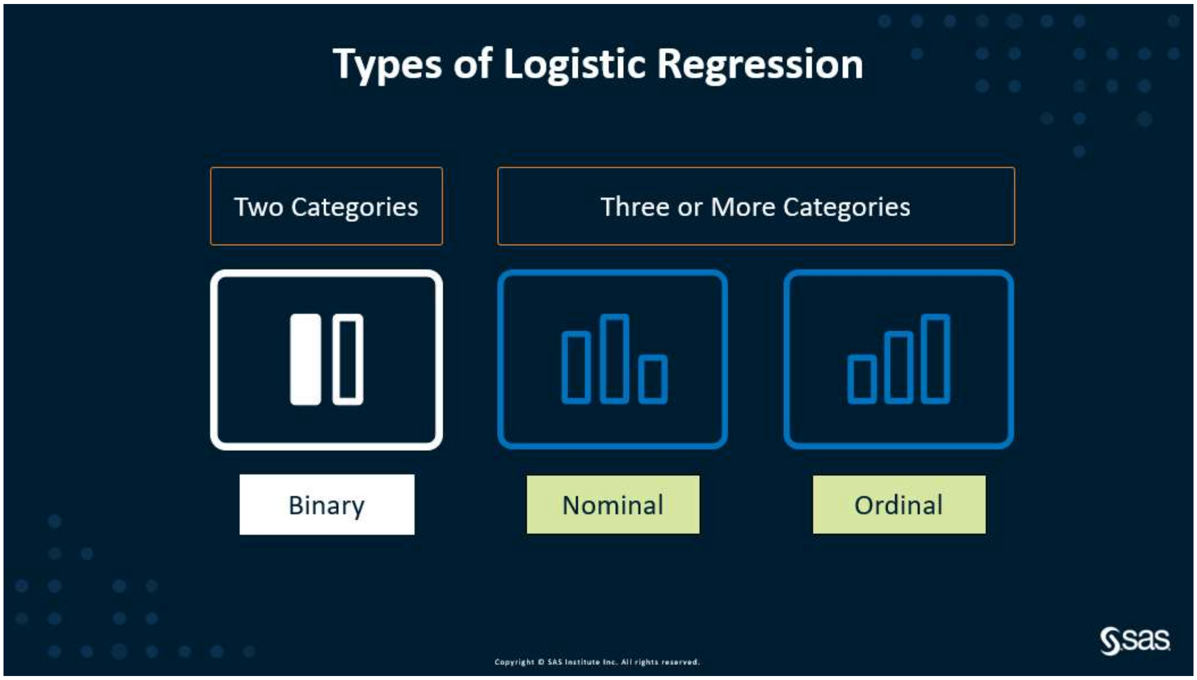
Frequency Missing = 6



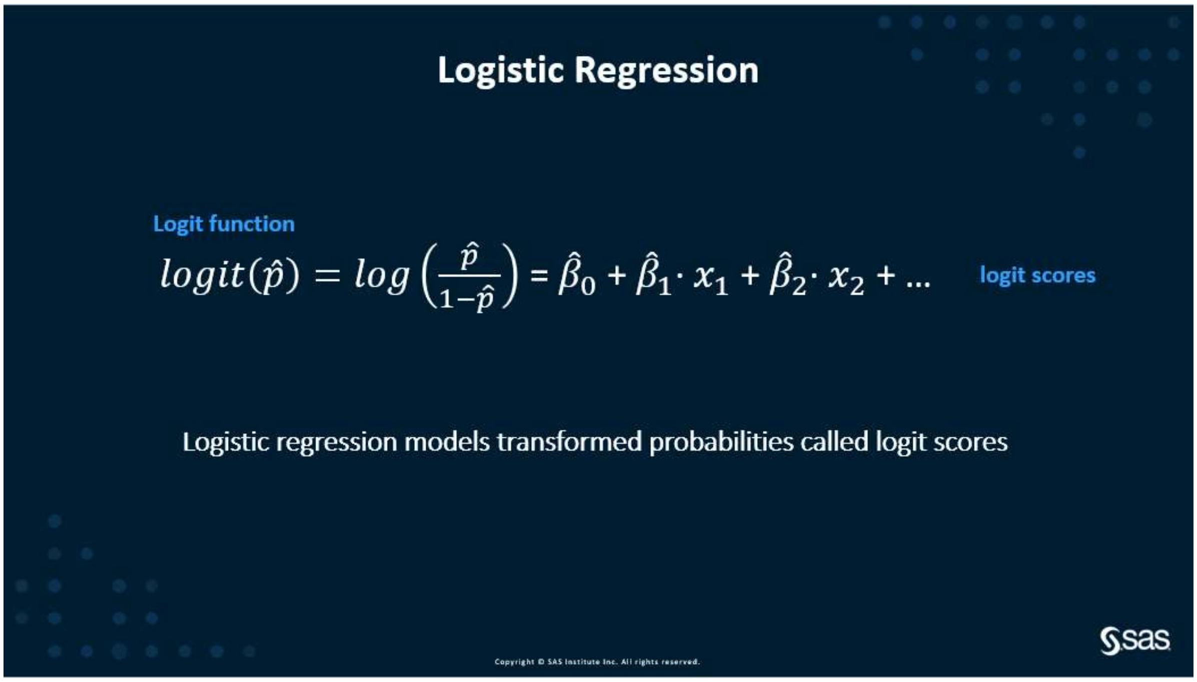
LOGISTIC Procedure

General form of the LOGISTIC procedure:

```
PROC LOGISTIC DATA=SAS-data-set <options>;
        CLASS variables </ options>;
        MODEL response=predictors </ options>;
        ODDSRATIO <'label'> variable </ options>;
        OUTPUT OUT=SAS-data-set keyword=name
                               </ options>;
RUN;
```



Types of Logistic Regression



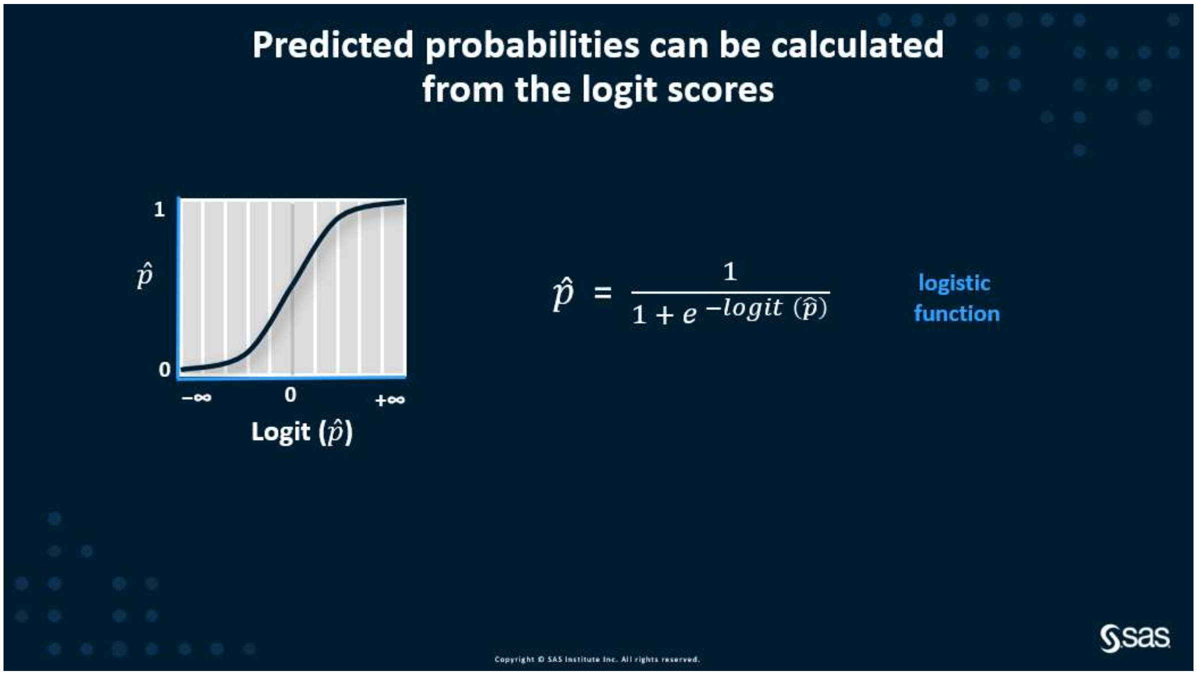
Logistic Regression

Logit function

$$logit(\hat{p}) = log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1\cdot x_1 + \hat{\beta}_2\cdot x_2 + \ldots$$

logit scores

Logistic regression models transformed probabilities called logit scores



Predicted probabilities can be calculated from the logit scores

$$\hat{p} = \frac{1}{1 + e^{-logit\ (\hat{p})}}$$

logistic function

Demonstration Time

Using PROC LOGISTIC to Model a Categorical Response

In [4]:

```
proc logistic data=sashelp.heart plots=all;
   class BP_status Chol_status sex Weight_status;
   model status(event='Alive') = BP_status Chol_status sex Weight_status height;
run;quit;
```

The LOGISTIC Procedure

Model Information		
Data Set	SASHELP.HEART	Framingham Heart Study
Response Variable	Status	
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	5209
Number of Observations Used	5047

Response Profile		
Ordered Value	**Status**	**Total Frequency**
1	Alive	3132
2	Dead	1915

Probability modeled is Status='Alive'.

Note: 162 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information			
Class	**Value**	**Design Variables**	
BP_Status	**High**	1	0
	Normal	0	1
	Optimal	-1	-1
Chol_Status	**Borderline**	1	0
	Desirable	0	1
	High	-1	-1
Sex	**Female**	1	
	Male	-1	
Weight_Status	**Normal**	1	0
	Overweight	0	1
	Underweight	-1	-1

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

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	6702.256	6188.615
SC	6708.783	6247.354
-2 Log L	6700.256	6170.615

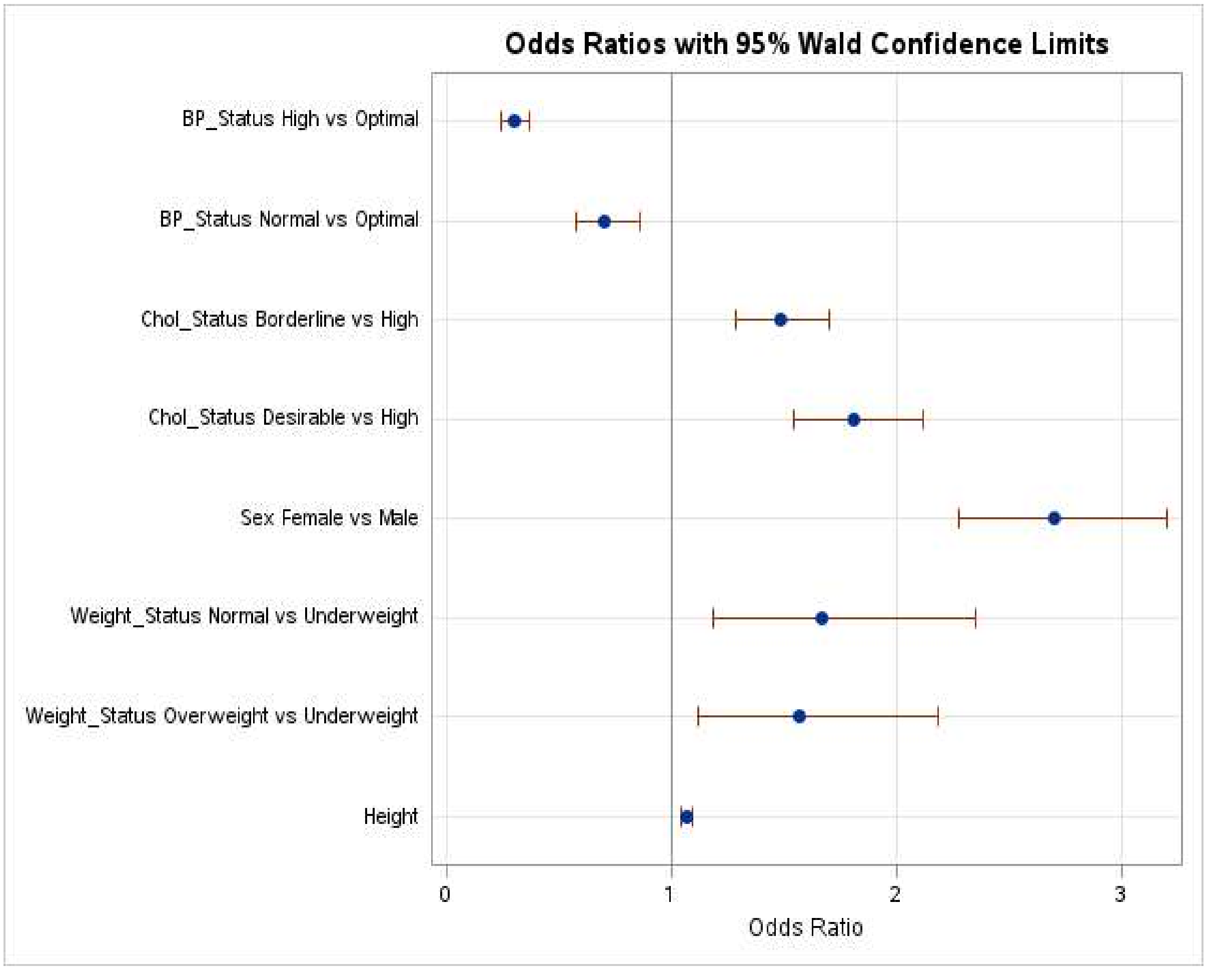
Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	529.6415	8	<.0001
Score	508.6632	8	<.0001
Wald	466.7464	8	<.0001

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
BP_Status	2	220.0289	<.0001
Chol_Status	2	60.6296	<.0001
Sex	1	129.7312	<.0001
Weight_Status	2	8.6593	0.0132
Height	1	29.4169	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-3.6764	0.7846	21.9541	<.0001
BP_Status	**High**	1	-0.6761	0.0466	210.7309	<.0001
BP_Status	**Normal**	1	0.1612	0.0467	11.9255	0.0006
Chol_Status	**Borderline**	1	0.0649	0.0427	2.3070	0.1288
Chol_Status	**Desirable**	1	0.2633	0.0479	30.2284	<.0001
Sex	**Female**	1	0.4963	0.0436	129.7312	<.0001
Weight_Status	**Normal**	1	0.1923	0.0687	7.8364	0.0051
Weight_Status	**Overweight**	1	0.1286	0.0650	3.9105	0.0480
Height		1	0.0650	0.0120	29.4169	<.0001

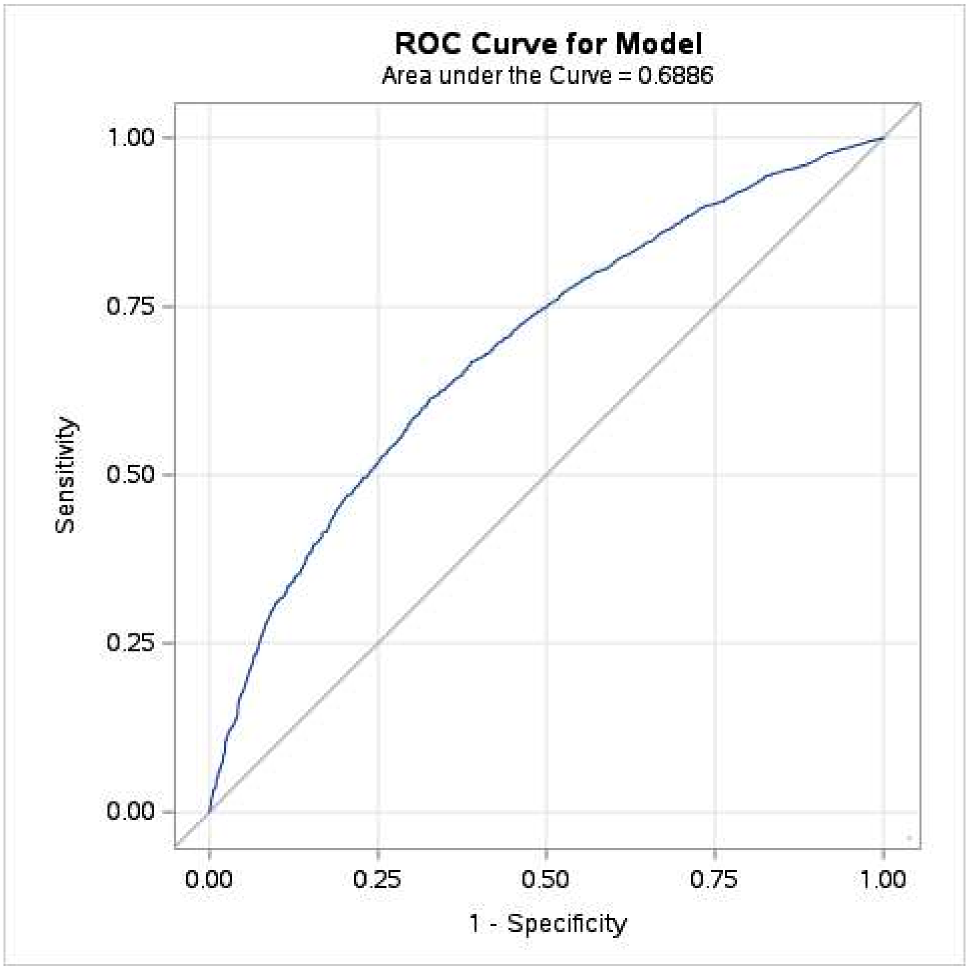
Odds Ratio Estimates

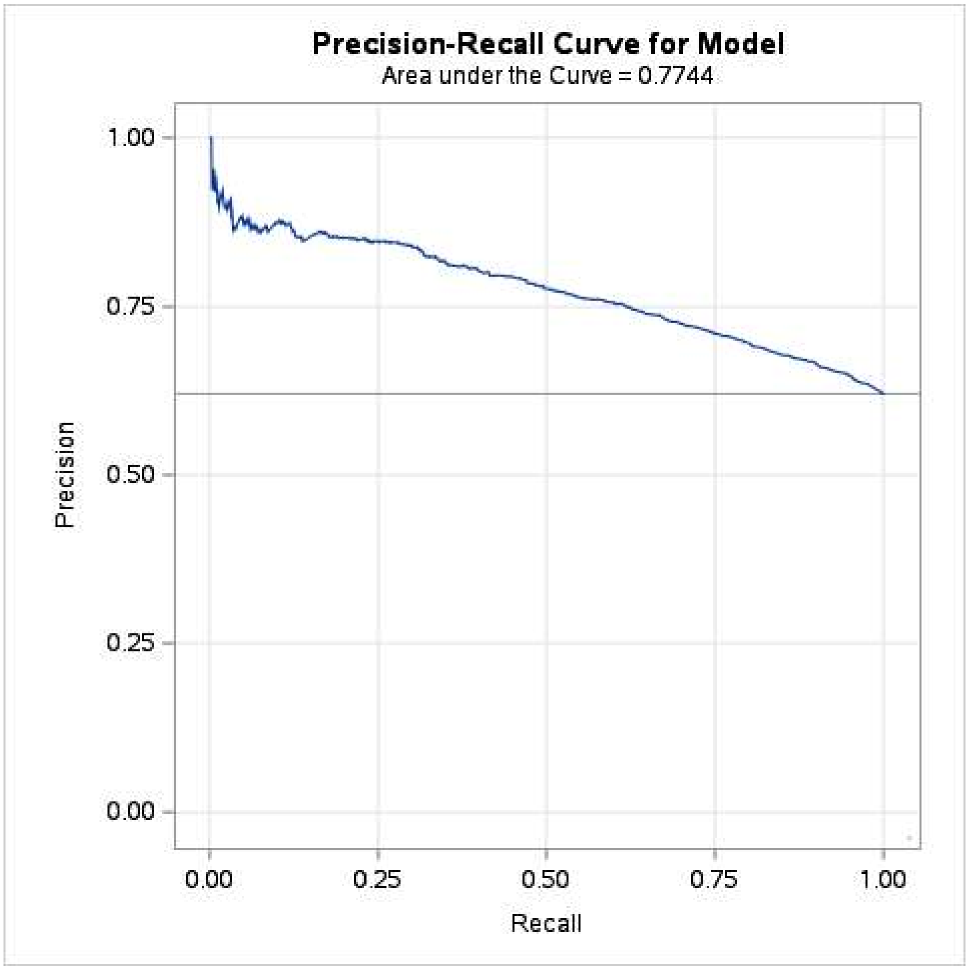
Effect	Point Estimate	95% Wald Confidence Limits	
BP_Status High vs Optimal	0.304	0.248	0.372
BP_Status Normal vs Optimal	0.702	0.574	0.859
Chol_Status Borderline vs High	1.482	1.289	1.703
Chol_Status Desirable vs High	1.807	1.544	2.115
Sex Female vs Male	2.698	2.275	3.201
Weight_Status Normal vs Underweight	1.671	1.187	2.352
Weight_Status Overweight vs Underweight	1.567	1.123	2.188
Height	1.067	1.042	1.093

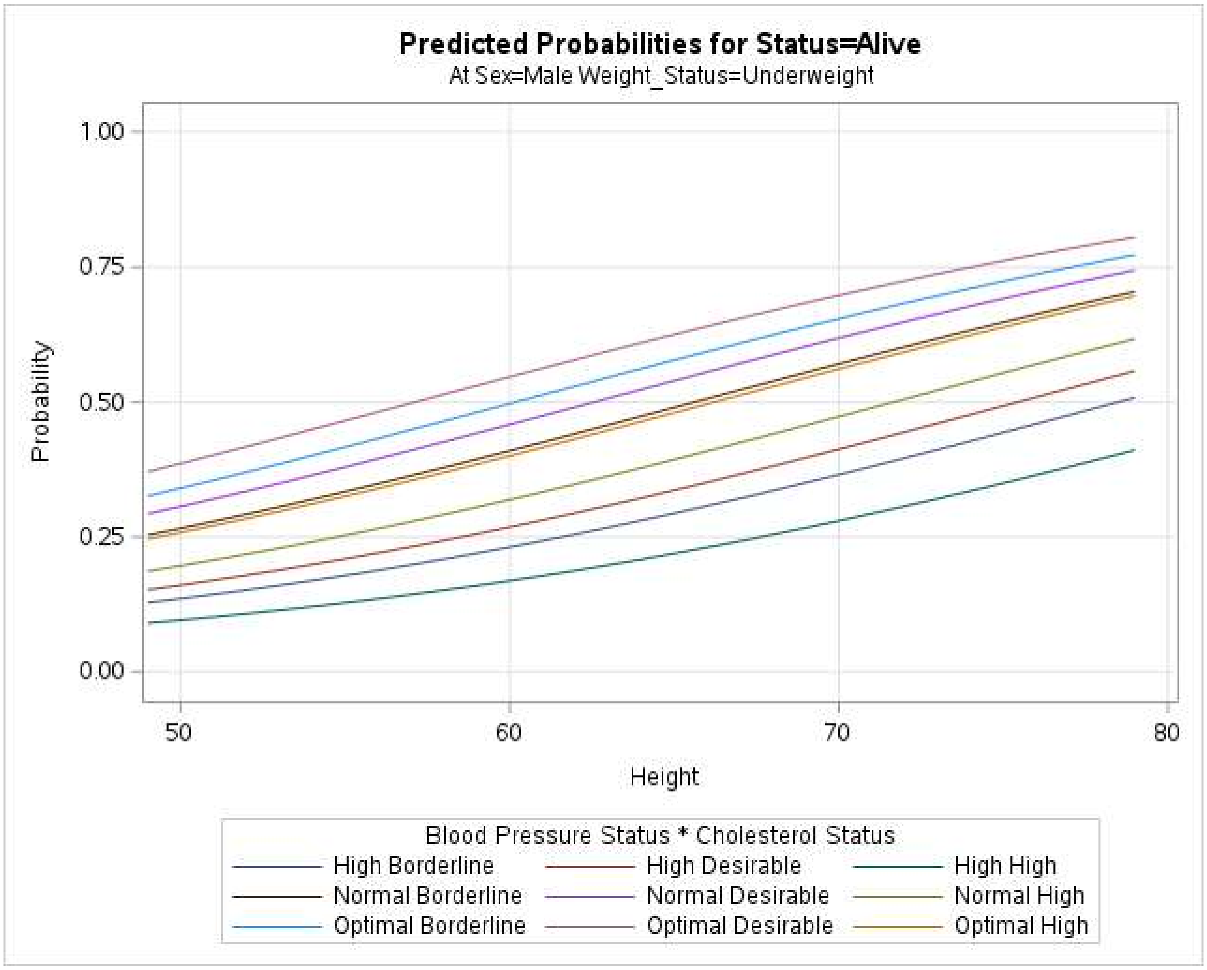


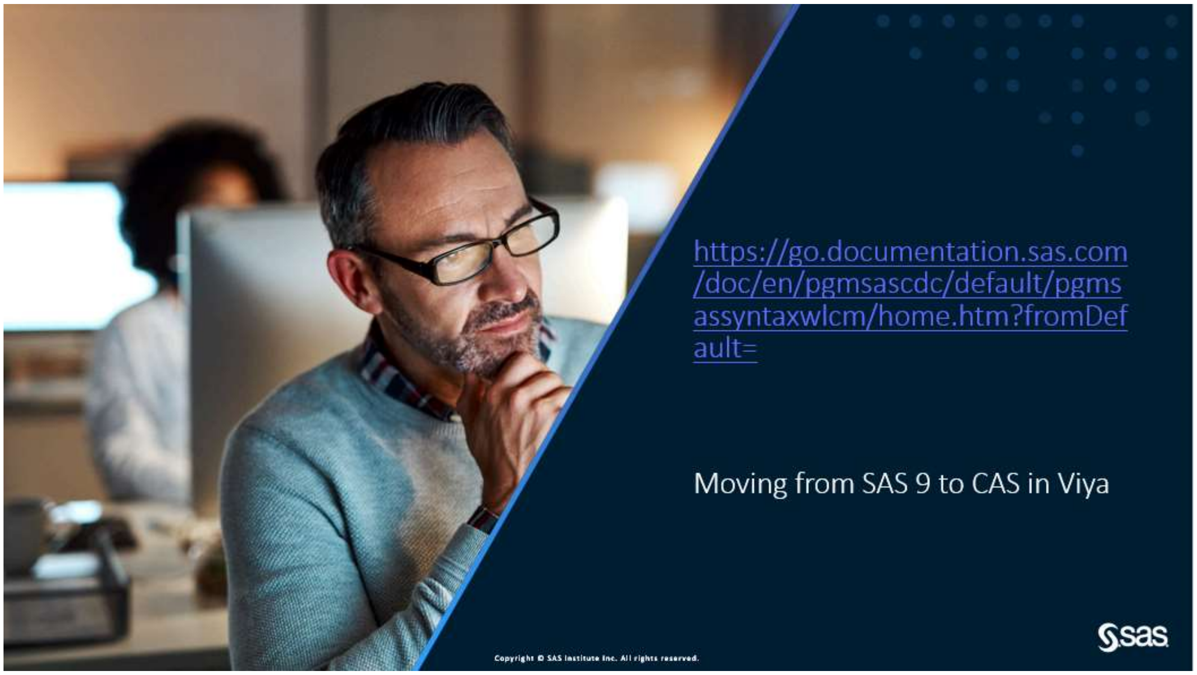
Association of Predicted Probabilities and Observed Responses

Percent Concordant	68.8	Somers' D	0.377
Percent Discordant	31.1	Gamma	0.378
Percent Tied	0.1	Tau-a	0.178
Pairs	5997780	c	0.689


ROC Curve for Model
Area under the Curve = 0.6886
Sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - Specificity


Precision-Recall Curve for Model
Area under the Curve = 0.7744
Precision
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Recall





https://go.documentation.sas.com/doc/en/pgmsascdc/default/pgmsassyntaxwlcm/home.htm?fromDefault=

Moving from SAS 9 to CAS in Viya

https://go.documentation.sas.com/doc/en/pgmsascdc/default/pgmsassyntaxwlcm/home.htm?fromDefault=

Questions?

In [ ]: