

Explore, Analyze, and Summarize Your Data with SAS Software: Selecting the Best Power Tool from a Rich Portfolio

LeRoy Bessler PhD
bessler@execpc.com

LeRoy Bessler

The Power To Show™

bessler@execpc.com

SAS is the most widely installed
analytical tool on mainframes.
I don't know the situation for
midrange and PCs.

LeRoy Bessler

The Power To Show™

bessler@execpc.com

My Focus for SAS Tools Here

- Data Exploration
- Simple Analyses
- Data Summarization

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Expectations

- I have 120 slides
- I may have to skip some of them
- All slides are in Proceedings (black / white)
- Full-color .ppt file available via email
- Verbose slides meant to be self-explanatory
(standard wisdom is keep slides very brief)

LeRoy Bessler

The Power To Show™

bessler@execpc.com

About this Color Presentation

- The hardcopy is black-and-white only
- For **COLOR** slides, ask bessler@execpc.com
- You might also request:
"Communication-Effective Use of Color"
by the same author
(I recommend all of his work.)

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Agenda

- The PROCs
 - Functional Reasons for Selecting a PROC
 - Default Examples of Using Each
 - Deeper Examination of Some
 - Using ODS to Customize "Function"
- SQL vs MEANS Processing Efficiency
(specific to my sample data & applications)

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Scope

- Some features / options of the PROCs
- Statistics
 - simple descriptive statistics
 - only ones that I can explain
 - informed comments are welcome
- I will try to take questions as we go

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Common, But Out Of Scope

- BY processing
- WEIGHT statement
- FREQ statement (not the PROC)

LeRoy Bessler

The Power To Show™

bessler@execpc.com

For Significance of Statistics Not Covered Here, but Available

Please see Appendix 1

in SAS Procedures Manual:

- how to interpret statistics
- Base SAS procedures, not SAS/STAT
- examples included

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Procedures Covered Today

- FREQ
- MEANS / SUMMARY
- UNIVARIATE
- RANK
- CLUSTER (SAS/STAT, not Base SAS)
- CORR (was not advertised in my abstract)
- SQL

LeRoy Bessler

The Power To Show™

bessler@execpc.com

FREQ

- Counts, Percent of Total Count
- other statistics not discussed here
- 2-Way Table

LeRoy Bessler

The Power To Show™

bessler@execpc.com

MEANS / SUMMARY

- Means
- Sums
- LARGE collection of other statistics

LeRoy Bessler

The Power To Show™

bessler@execpc.com

UNIVARIATE

- Everything in MEANS / SUMMARY
- A Bit More
- Default listing output much richer than MEANS / SUMMARY

LeRoy Bessler

The Power To Show™

bessler@execpc.com

RANK

- Rank numeric values
- or
- Assign values to (ranked) quartiles, deciles, percentiles, etc.

LeRoy Bessler

The Power To Show™

bessler@execpc.com

CORR

- Identify related pairs of numeric variables within the same data set
- Or is any pair of variables independent?

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Reasons to Use SQL

- User prefers SQL language over SAS
- Wants to get data from relational data base and let source pre-process data delivered
- Some function that only SQL can do (e.g., cartesian product) or can do faster
- Programmer obsessed with brevity, possibly at the expense of faster results

LeRoy Bessler

The Power To Show™

bessler@execpc.com

PROCS' Simple Descriptive Statistics Ranked from Richest to Poorest

- UNIVARIATE
- MEANS / SUMMARY
- TABULATE
- SQL
- REPORT
- CORR*

*strictly speaking, not "Simple Desc. Statistics"

LeRoy Bessler

The Power To Show™

bessler@execpc.com

OK, so what are on Bessler's List for "Simple Descriptive Statistics"?

- N (count), MIN, MAX, MEAN
 - SUM, MODE
 - MEDIAN, RANGE, EXTREME VALUES
 - PERCENTILES, DECILES, QUARTILES
 - STD (assumes a Normal Distribution)
- RED = Default Output MEANS / SUMMARY

LeRoy Bessler

The Power To Show™

bessler@execpc.com

What about PROC FREQ?

- Delivers Counts (Frequencies)
- No "Simple Descriptive Statistics"
- Can Deliver Huge Volume of Other Statistics, none of which are presented here

LeRoy Bessler

The Power To Show™

bessler@execpc.com

For Demo, Using SASHELP.CLASS

```
options nocenter nodate nonumber;  
title ' ';  
proc print data=sashelp.class;  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Alice	F	13	56.5	84.0
3	Barbara	F	13	65.3	98.0
4	Carol	F	14	62.8	102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83.0
7	Jane	F	12	59.8	84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	84.0
10	John	M	12	59.0	99.5
11	Joyce	F	11	51.3	50.5
12	Judy	F	14	64.3	90.0
13	Louise	F	12	56.3	77.0
14	Mary	F	15	66.5	112.0
15	Philip	M	16	72.0	150.0
16	Robert	M	12	64.8	128.0
17	Ronald	M	15	67.0	133.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

PROC FREQ

```
proc freq data=sashelp.class;  
tables Age;  
run;
```

/* If no TABLES statement,
listing for every var in data= */

LeRoy Bessler

The Power To Show™

bessler@execpc.com

The FREQ Procedure

Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
11	2	10.53	2	10.53
12	5	26.32	7	36.84
13	3	15.79	10	52.63
14	4	21.05	14	73.68
15	4	21.05	18	94.74
16	1	5.26	19	100.00

LeRoy Bessler

The Power To Show™

bessler@execpc.com

PROC FREQ

```
proc freq data=sashelp.class;  
tables Age*Sex;  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

The FREQ Procedure
Table of Age by Sex

Age	Sex		
	F	M	Total
Frequency Percent Row Pct Col Pct			
11	1 5.26 50.00 11.11	1 5.26 50.00 10.00	2 10.53
12	2 10.53 40.00 22.22	3 15.79 60.00 30.00	5 26.32
13	2 10.53 66.67 22.22	1 5.26 33.33 10.00	3 15.79
14	2 10.53 50.00 22.22	2 10.53 50.00 20.00	4 21.05
15	2 10.53 50.00 22.22	2 10.53 50.00 20.00	4 21.05
16	0 0.00 0.00 0.00	1 5.26 100.00 10.00	1 5.26
Total	9 47.37	10 52.63	19 100.00

LeRoy Bessler bessler@execpc.com

Age Sex

Frequency Percent Row Pct Col Pct	F	M	Total
11	1 5.26 50.00 11.11	1 5.26 50.00 10.00	2 10.53
15	2 10.53 50.00 22.22	2 10.53 50.00 20.00	4 21.05
16	0 0.00 0.00 0.00	1 5.26 100.00 10.00	1 5.26
Total	9 47.37	10 52.63	19 100.00

LeRoy Bessler bessler@execpc.com

PROC FREQ

```
proc freq data=sashelp.class;
tables Sex*Age;
run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

The FREQ Procedure
Table of Sex by Age

Sex	Age						Total
	11	12	13	14	15	16	
Frequency Percent Row Pct Col Pct							
F	1 5.26 11.11 50.00	2 10.53 22.22 40.00	2 10.53 22.22 66.67	2 10.53 22.22 50.00	2 10.53 22.22 50.00	0 0.00 0.00 0.00	9 47.37
M	1 5.26 10.00 50.00	3 15.79 30.00 60.00	1 5.26 10.00 33.33	2 10.53 20.00 50.00	2 10.53 20.00 50.00	1 5.26 10.00 100.00	10 52.63
Total	2 10.53	5 26.32	3 15.79	4 21.05	4 21.05	1 5.26	19 100.00

LeRoy Bessler The Power To Show™ bessler@execpc.com

Titles from PROC FREQ

- I supplied a blank user title (title ' ');
- PROC FREQ forces:
 - "The FREQ Procedure" subtitle
removable only if using ODS
 - "Table of Var1 by Var2" subtitle
based on 'tables Var1*Var2;'
var labels are used in row/column labels

LeRoy Bessler The Power To Show™ bessler@execpc.com

PROC FREQ

```
proc freq data=sashelp.class;
tables Sex*Age*Height; run;
```

Result: Two CrossTabulations
Table 1 of Age by Height
Controlling for Sex=F
(and a Table 2 for Sex=M)

LeRoy Bessler The Power To Show™ bessler@execpc.com

List Form, Not Cross Tabulation

- Simpler, Easier-to-use
- Counts, Percents of Grand Total

```
proc freq data=sashelp.class;  
tables Sex*Age / list;  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

The FREQ Procedure					
Sex	Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	11	1	5.26	1	5.26
F	12	2	10.53	3	15.79
F	13	2	10.53	5	26.32
F	14	2	10.53	7	36.84
F	15	2	10.53	9	47.37
M	11	1	5.26	10	52.63
M	12	3	15.79	13	68.42
M	13	1	5.26	14	73.68
M	14	2	10.53	16	84.21
M	15	2	10.53	18	94.74
M	16	1	5.26	19	100.00

LeRoy Bessler

The Power To Show™

bessler@execpc.com

PROC FREQ – Output Data Set

```
proc freq data=sashelp.class noprint;  
tables Age / out=OutputFromFREQ;  
run;  
title 'Output Data Set from PROC FREQ';  
title2 'NOTE: No Cumulative Statistics';  
proc print data=OutputFromFREQ  
label noobs; run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Output Data Set from PROC FREQ NOTE: No Cumulative Statistics

Age	Frequency Count	Percent of Total Frequency
11	2	10.5263
12	5	26.3158
13	3	15.7895
14	4	21.0526
15	4	21.0526
16	1	5.2632

Using ODS for More Output Data

```
ods trace on;  
proc freq data=sashelp.class;  
tables Age;  
run;  
ods trace off;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

SAS Log:

```
1 ods trace on;  
2 proc freq data=sashelp.class;  
3 tables Age;  
4 run;
```

Output Added:

```
-----  
Name: OneWayFreqs  
Label: One-Way Frequencies  
Template: Base.Freq.OneWayFreqs  
Path: Freq.Age.OneWayFreqs  
-----
```

NOTE: There were 19 observations
read from the data set SASHELP.CLASS.

```
5 ods trace off;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Get ODS Output Data

```
ods listing close;  
ods output OneWayFreqs=WithCums;  
proc freq data=sashelp.class;  
tables Age;  
run;  
ods listing;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Print ODS Output Data Set

```
proc print data=WithCums  
noobs  
label;  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Table	Age	Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Age	11	11	2	10.53	2	10.53
Age	12	12	5	26.32	7	36.84
Age	13	13	3	15.79	10	52.63
Age	14	14	4	21.05	14	73.68
Age	15	15	4	21.05	18	94.74
Age	16	16	1	5.26	19	100.00

LeRoy Bessler

The Power To Show™

bessler@execpc.com

ODS Output Data Set Contents

```
proc contents data=WithCums  
varnum;  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

-----Variables Ordered by Position-----

#	Variable	Type	Len	Format	Label
1	Table	Char	32		
2	F_Age	Char	3		Age
3	Age	Num	8		
4	Frequency	Num	8	BEST8.	
5	Percent	Num	8	6.2	
6	CumFrequency	Num	8	BEST8.	Cumulative Frequency
7	CumPercent	Num	8	6.2	Cumulative Percent

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Keep Desired ODS Output Data

```
data KeepFREQs;  
set WithCums(drop=Table F_Age);  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Print the Kept Data

```
proc print data=KeepFREQs  
    noobs  
    label;  
  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
11	2	10.53	2	10.53
12	5	26.32	7	36.84
13	3	15.79	10	52.63
14	4	21.05	14	73.68
15	4	21.05	18	94.74
16	1	5.26	19	100.00

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Kept Data Set Contents

```
proc contents data=KeepFREQs  
    varnum;  
  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

-----Variables Ordered by Position-----					
#	Variable	Type	Len	Format	Label
1	Age	Num	8		
2	Frequency	Num	8	BEST8.	
3	Percent	Num	8	6.2	
4	CumFrequency	Num	8	BEST8.	Cumulative Frequency
5	CumPercent	Num	8	6.2	Cumulative Percent

LeRoy Bessler

The Power To Show™

bessler@execpc.com

MEANS / SUMMARY

Only Two Differences

- MEANS default is PRINT,
SUMMARY default is NOPRINT
- If VAR statement omitted,
SUMMARY only counts observations,
MEANS analyzes all numeric variables
not listed on some other statement

LeRoy Bessler

The Power To Show™

bessler@execpc.com

PROC MEANS / PROC SUMMARY

```
proc means data=sashelp.class;  
run;  
  
proc summary data=sashelp.class  
    print;  
  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

The MEANS Procedure					
Variable	N	Mean	Std Dev	Minimum	Maximum
Age	19	13.3157895	1.4926722	11.0000000	16.0000000
Height	19	62.3368421	5.1270752	51.3000000	72.0000000
Weight	19	100.0263158	22.7739335	50.5000000	150.0000000

The SUMMARY Procedure	
	N
Obs	19

LeRoy Bessler The Power To Show™ bessler@execpc.com

PROC SUMMARY

```
proc summary data=sashelp.class
    print;
    var Age Height Weight;
run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

The SUMMARY Procedure					
Variable	N	Mean	Std Dev	Minimum	Maximum
Age	19	13.3157895	1.4926722	11.0000000	16.0000000
Height	19	62.3368421	5.1270752	51.3000000	72.0000000
Weight	19	100.0263158	22.7739335	50.5000000	150.0000000

LeRoy Bessler The Power To Show™ bessler@execpc.com

Listing Output (PRINT option)

No control

- format of variables
- labels for columns
- report layout

Solution: Procedure Output Data Set

LeRoy Bessler The Power To Show™ bessler@execpc.com

SUMMARY Output Data Set

```
proc summary data=sashelp.class;
var Age Height Weight;
output out=SUMMARY;
run;
proc print data=SUMMARY
    noobs label;
run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

TYPE	_FREQ_	_STAT_	Age	Height	Weight
0	19	N	19.0000	19.0000	19.000
0	19	MIN	11.0000	51.3000	50.500
0	19	MAX	16.0000	72.0000	150.000
0	19	MEAN	13.3158	62.3368	100.026
0	19	STD	1.4927	5.1271	22.774

LeRoy Bessler The Power To Show™ bessler@execpc.com

Trim the SUMMARY Output

```
proc summary data=sashelp.class;
var Age Height Weight;
output out=SUMMARY
      (drop=_type_ _freq_);
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Transpose the SUMMARY Output

```
proc transpose data=SUMMARY
      out=TransposedSUMMARY;
run;
proc print
      data=TransposedSUMMARY
      noobs;
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

NAME	COL1	COL2	COL3	COL4	COL5
Age	19	11.0	16	13.316	1.4927
Height	19	51.3	72	62.337	5.1271
Weight	19	50.5	150	100.026	22.7739

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Label & Format the Transpose

```
data Finished;
set TransposedSUMMARY;
rename _NAME_=Var COL1=N
      COL2=Min COL3=Max
      COL4=Mean COL5=StdDev;
format COL2-COL5 5.1; run;
proc print data=Finished noobs; run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Var	N	Min	Max	Mean	Std Dev
Age	19	11.0	16.0	13.3	1.5
Height	19	51.3	72.0	62.3	5.1
Weight	19	50.5	150.0	100.0	22.8

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Print of SUMMARY by CLASS

```
proc summary data=sashelp.class
      print maxdec=1 fw=7
      min max mean std;
class Sex Age;
var Height;
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

The SUMMARY Procedure
Analysis Variable : Height

Sex	Age	N Obs	Minimum	Maximum	Mean	Std Dev
F	11	1	51.3	51.3	51.3	.
	12	2	56.3	59.8	58.1	2.5
	13	2	56.5	65.3	60.9	6.2
	14	2	62.8	64.3	63.6	1.1
	15	2	62.5	66.5	64.5	2.8
M	11	1	57.5	57.5	57.5	.
	12	3	57.3	64.8	60.4	3.9
	13	1	62.5	62.5	62.5	.
	14	2	63.5	69.0	66.3	3.9
	15	2	66.5	67.0	66.8	0.4
	16	1	72.0	72.0	72.0	.

**Output SUM by CLASS
with Subtotals & Grand Total**

```
proc summary data=sashelp.class;
class Sex Age;
var Weight;
output out=SUMbyCLASS sum=;
run;
proc print data= SUMbyCLASS; run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

Sex	Age	_TYPE_	_FREQ_	Weight
	.	0	19	1900.5
	11	1	2	135.5
	12	1	5	472.0
	13	1	3	266.0
	14	1	4	407.5
	15	1	4	469.5
	16	1	1	150.0
F	.	2	9	811.0
M	.	2	10	1089.5
F	11	3	1	50.5
F	12	3	2	161.5
F	13	3	2	182.0
F	14	3	2	192.5
F	15	3	2	224.5
M	11	3	1	85.0
M	12	3	3	310.5
M	13	3	1	84.0
M	14	3	2	215.0
M	15	3	2	245.0
M	16	3	1	150.0

SUM only each CLASS level

```
proc summary
data=sashelp.class nway;
class Sex Age; var Weight;
output out=SUMeachCLASS
(drop=_type_
rename=( _freq_=Count))
sum=; run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

Sex	Age	Count	Weight
F	11	1	50.5
F	12	2	161.5
F	13	2	182.0
F	14	2	192.5
F	15	2	224.5
M	11	1	85.0
M	12	3	310.5
M	13	1	84.0
M	14	2	215.0
M	15	2	245.0
M	16	1	150.0

If not NWAY SUMMARY / MEANS

- You can do some snazzy things
- Out of scope to present here
- For how to easily decode _TYPE_:

“A Format to Make the _TYPE_ Field
of PROC MEANS Easier to Interpret”
Matt.Pettis@thomson.com

Please tell him that I referred you

LeRoy Bessler The Power To Show™ bessler@execpc.com

Why PROC UNIVARIATE instead of MEANS / SUMMARY?

- MODE is available (as option)
- Other added Statistics options
- Very Rich Default Output Listing

LeRoy Bessler

The Power To Show™

bessler@execpc.com

PROC UNIVARIATE

```
proc univariate data=sashelp.class;  
run;
```

* Default Output from UNIVARIATE *;

LeRoy Bessler

The Power To Show™

bessler@execpc.com

PROC UNIVARIATE

```
ods trace on;  
proc univariate data=sashelp.class;  
run;  
ods trace off;  
* ODS trace Output *;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Select one UNIVARIATE section from the six per variable

ods select

Univariate.Height.ExtremeObs;

```
proc univariate data=sashelp.class;  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

The UNIVARIATE Procedure Variable: Height

Extreme Observations

----Lowest----		----Highest----	
Value	Obs	Value	Obs
51.3	11	66.5	14
56.3	13	66.5	19
56.5	2	67.0	17
57.3	6	69.0	1
57.5	18	72.0	15

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Default PROC RANK Output

```
proc rank data=sashelp.class  
out=rankings;  
run;  
proc print data=rankings label noobs;  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Name	Sex	Values of Age Were Replaced by Ranks	Values of Height Were Replaced by Ranks	Values of Weight Were Replaced by Ranks
Alfred	M	12.5	18.0	15.5
Alice	F	9.0	3.0	4.5
Barbara	F	9.0	14.0	9.0
Carol	F	12.5	10.0	11.5
Henry	M	12.5	11.0	11.5
James	M	5.0	4.0	3.0
Janet	F	5.0	7.0	6.0
Jeffrey	M	16.5	9.5	15.5
John	M	9.0	8.5	4.5
Joyce	F	5.0	6.0	10.0
Judy	F	1.5	1.0	1.0
Louise	F	12.5	12.0	8.0
Mary	F	5.0	2.0	2.0
Philip	M	16.5	15.5	13.5
Robert	M	19.0	19.0	19.0
Ronald	M	5.0	13.0	17.0
Thomas	M	16.5	17.0	18.0
William	M	1.5	5.0	7.0
William	M	16.5	15.5	13.5

LeRoy Bessler The Power To Show™ bessler@execpc.com

Rank By Descending Height

```
proc rank data=sashelp.class
    descending;
var Height;
ranks RankFromTallestToShortest;
run;
proc print;
run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

NOTE: Ties end with .5			
Obs	Name	Height	RankFrom Tallest To Shortest
1	Alfred	69.0	2.0
2	Alice	56.5	17.0
3	Barbara	65.3	6.0
4	Carol	62.8	10.0
5	Henry	63.5	9.0
6	James	57.3	16.0
7	Jane	59.8	13.0
8	Janet	62.5	11.5
9	Jeffrey	62.5	11.5
10	John	59.0	14.0
11	Joyce	51.3	19.0
12	Judy	64.3	8.0
13	Louise	56.3	18.0
14	Mary	66.5	4.5
15	Philip	72.0	1.0
16	Robert	64.8	7.0
17	Ronald	67.0	3.0
18	Thomas	57.5	15.0
19	William	66.5	4.5

LeRoy pc.com

Rank Descending with Ties=Low

```
proc rank data=sashelp.class
    descending ties=low;
var Height;
ranks RankDescWithTiesEqualLow;
run;
proc print;
run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

TIES=LOW refers to lower numbered (i.e., higher) rank			
Obs	Name	Height	Rank Descending WithTies EqualLow
1	Alfred	69.0	2
2	Alice	56.5	17
3	Barbara	65.3	6
4	Carol	62.8	10
5	Henry	63.5	9
6	James	57.3	16
7	Jane	59.8	13
8	Janet	62.5	11
9	Jeffrey	62.5	11
10	John	59.0	14
11	Joyce	51.3	19
12	Judy	64.3	8
13	Louise	56.3	18
14	Mary	66.5	4
15	Philip	72.0	1
16	Robert	64.8	7
17	Ronald	67.0	3
18	Thomas	57.5	15
19	William	66.5	4

Assigning Percentiles By Height

```
proc rank data=sashelp.class
    groups=100;
var Height;
ranks PercentileByHeight;
run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

Sort & Print Percentiles

```
proc sort;  
by descending PercentileByHeight;  
run;  
  
proc print noobs;  
var Name Height PercentileByHeight;  
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

For Percentiles From Tallest to Shortest,
DO NOT use DESCENDING option
NOTE: Output from PROC RANK has been sorted

Name	Height	Percentile ByHeight
Philip	72.0	95
Alfred	69.0	90
Ronald	67.0	85
Mary	66.5	77
William	66.5	77
Barbara	65.3	70
Robert	64.8	65
Judy	64.3	60
Henry	63.5	55
Carol	62.8	50
Janet	62.5	42
Jeffrey	62.5	42
Jane	59.8	35
John	59.0	30
Thomas	57.5	25
James	57.3	20
Alice	56.5	15
Louise	56.3	10
Joyce	51.3	5

LeR

com

General Problem

- You have a data set with numeric measurement variable
- For bar chart or pie chart, how to assign variable to (sub)ranges?
- Reason: You do not want one bar / slice per observation

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Possible Solutions

- there may be natural break points based on the nature of the variable
- quartiles, quintiles, deciles
- mean and +/- standard deviation
- median and 20th & 80th percentiles
- etc.
- clusters

LeRoy Bessler

The Power To Show™

bessler@execpc.com

PROC CLUSTER (SAS/STAT)

- Can observations be formed into groups based on closeness?
- Eleven methods to measure closeness (I like METHOD=CENTROID)
- Can also be applied to categorical data (but not demonstrated here)
- For very large data sets, see manual

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Why Use PROC CLUSTER

- Example here is similar to my own first use of it (which does illustrate what it does)
- Serious statisticians use it for “real” applications

LeRoy Bessler

The Power To Show™

bessler@execpc.com

PROC CLUSTER Minimal Output

```
proc cluster data=sashelp.class
    method=centroid;
```

```
var weight;
```

```
run;
```

➤ Output not informative, not useful
(for me)

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	518.652047		1.0000	1.0000

Root-Mean-Square Total-Sample Standard Deviation = 22.77393
Root-Mean-Square Distance Between Observations = 32.20721

NCL	--Clusters Joined--	FREQ	Nom Cent Dist	T
18	OB4 OB5	2	0	T
17	OB1 OB8	2	0	T
16	OB2 OB9	2	0	T
15	OB14 OB19	2	0	T
14	CL16 OB7	3	0.0155	
13	CL17 CL15	4	0.0155	T
12	CL14 OB18	4	0.0259	
11	CL12 OB6	5	0.0427	
10	OB3 OB10	2	0.0466	
9	CL10 CL18	4	0.1164	
8	OB16 OB17	2	0.1552	
7	CL11 OB12	6	0.1832	
6	CL7 OB13	7	0.251	
5	CL13 CL9	8	0.3609	
4	OB15 CL8	3	0.6055	
3	CL5 CL6	15	0.6989	
2	CL3 CL4	18	1.2751	
1	CL2 OB11	19	1.6282	

PROC CLUSTER

```
proc cluster data=sashelp.class
    method=centroid
```

```
    noprint outtree=clusters;
```

```
var weight;
```

```
id name;
```

```
copy sex age height;
```

```
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

PROC CLUSTER (Continued)

```
proc print data=clusters;
```

```
var name sex age height weight
```

```
    _parent_ _ncl_ _freq_;
```

```
run;
```

* Output from PROC CLUSTER *;

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Need PROC TREE To Use Clusters

```
proc tree data=clusters noprint
```

```
    out=tree
```

```
    nclusters=5; /* max = orig nobs */
```

```
id name;
```

```
copy weight sex age height;
```

```
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Using PROC TREE Output

```
proc sort data=tree out=SortedTree;
```

```
by cluster weight;
```

```
run;
```

```
proc print data=SortedTree noobs;
```

```
var name weight cluster;
```

```
run;
```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

5 Clusters			8 Clusters		
Name	Height	CLUSTER	Name	Height	CLUSTER
Barbara	98.0	1	Barbara	98.0	1
John	99.5	1	John	99.5	1
Carol	102.5	1	Carol	102.5	1
Henry	102.5	1	Henry	102.5	1
Mary	112.0	1	Mary	112.0	2
William	112.0	1	William	112.0	2
Alfred	112.5	1	Alfred	112.5	2
Janet	112.5	1	Janet	112.5	2
Louise	77.0	2	James	83.0	3
James	83.0	2	Alice	84.0	3
Alice	84.0	2	Jeffrey	84.0	3
Jeffrey	84.0	2	Jane	84.5	3
Jane	84.5	2	Thomas	85.0	3
Thomas	85.0	2	Robert	128.0	4
Judy	90.0	2	Ronald	133.0	4
Robert	128.0	3	Judy	90.0	5
Ronald	133.0	3	Louise	77.0	6
Philip	150.0	4	Philip	150.0	7
Joyce	50.5	5	Joyce	50.5	8

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Two-Way Clusters

```

proc cluster data=YourData noprint
    method=centroid
    outtree=clusters2way;
var Var1 Var2 ;
id IdVar ;
copy OtherVarsIfWanted ;
run;

```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Two-Way Clusters (Continued)

```

proc tree data=clusters2way
    noprint
out=tree nclusters=HowMany ;
id IdVar ;
copy Var1 Var2
    OtherVarsIfWanted ;
run;

```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

Two-Way Clusters (Continued)

```

proc sort data=tree;
by cluster Var1 Var2;
run;

proc print data=tree noobs;
var IdVar Var1 Var2 cluster;
run;

```

LeRoy Bessler

The Power To Show™

bessler@execpc.com

8 Clusters				8 Clusters			
Name	Height	Age	CLUSTER	Name	Height	Age	CLUSTER
Barbara	98.0	13	1	Mary	66.5	15	1
John	99.5	12	1	William	66.5	15	1
Carol	102.5	14	1	Ronald	67.0	15	1
Henry	102.5	14	1	Jeffrey	62.5	13	2
James	83.0	12	2	Janet	62.5	15	2
Alice	84.0	13	2	Carol	62.8	14	2
Jeffrey	84.0	13	2	Henry	63.5	14	2
Jane	84.5	12	2	Judy	64.3	14	2
Thomas	85.0	11	2	John	59.0	12	3
Mary	112.0	15	3	Jane	59.8	12	3
William	112.0	15	3	Louise	56.3	12	4
Alfred	112.5	14	3	Alice	56.5	13	4
Janet	112.5	15	3	James	57.3	12	4
Robert	128.0	12	4	Thomas	57.5	11	4
Ronald	133.0	15	4	Robert	64.8	12	5
Judy	90.0	14	5	Barbara	65.3	13	5
Louise	77.0	12	6	Alfred	69.0	14	6
Philip	150.0	16	7	Joyce	51.3	11	7
Joyce	50.5	11	8	Philip	72.0	16	8

LeRoy Bessler

The Power To Show™

bessler@execpc.com

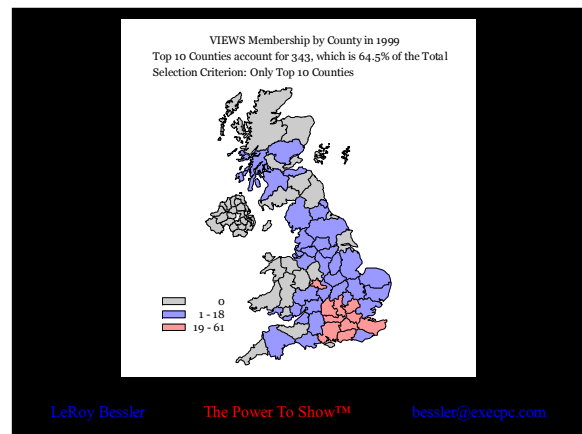
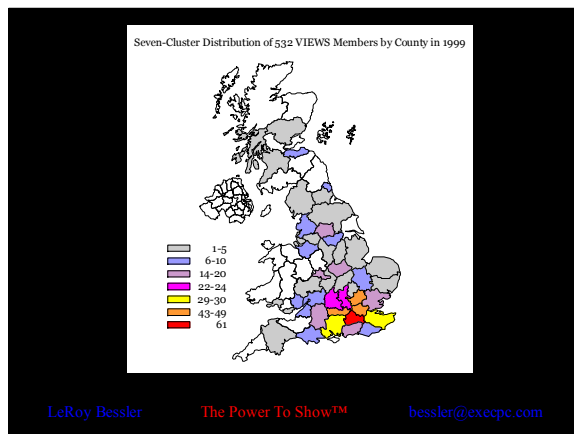
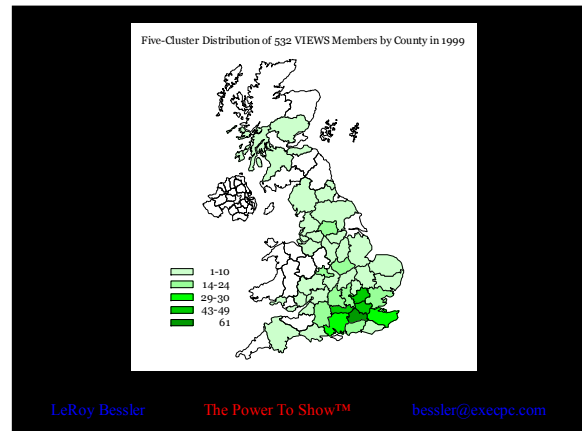
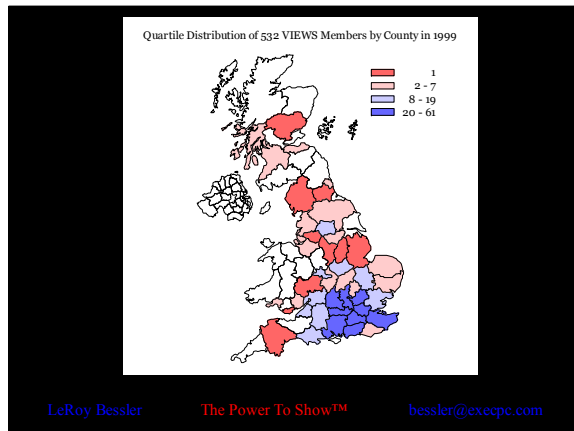
How I got into using PROC CLUSTER

- Not for analytical purposes
- Wanted a different way to assign subranges to a statistical map
- Some other choices – break points at:
 - percentiles 20, 50, 80
 - quartiles
 - mean and +/- 1 std dev of the mean

LeRoy Bessler

The Power To Show™

bessler@execpc.com



Problem

- Are values of one numeric variable affected by another numeric variable in the same data set?
- Possible application:
Identify independent variables for a multi-variable regression or other mathematical model

SOLUTION: PROC CORR

```
proc corr data=sashelp.class;
run;

proc corr data=sashelp.shoes
nosimple;
run;
```

sashelp.class

The CORR Procedure

3 Variables: Age Height Weight

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Age	19	13.31579	1.49267	253.00000	11.00000	16.00000
Height	19	62.33694	5.12708	1184	51.30000	72.00000
Weight	19	100.92632	22.77393	1901	50.50000	150.00000

Pearson Correlation Coefficients, N = 19
Prob > |r| under H0: Rho=0

	Age	Height	Weight
Age	1.00000	0.81143 <.0001	0.74089 <.0001
Height	0.81143 <.0001	1.00000	0.87779 <.0001
Weight	0.74089 <.0001	0.87779 <.0001	1.00000

LeRoy Bessler The Power To Show™ bessler@execpc.com

sashelp.shoes

The CORR Procedure

4 Variables: Stores Sales Inventory Returns

Pearson Correlation Coefficients, N = 395
Prob > |r| under H0: Rho=0

	Stores	Sales	Inventory	Returns
Stores	1.00000	0.42374 <.0001	0.47891 <.0001	0.40611 <.0001
Number of Stores				
Sales	0.42374 <.0001	1.00000	0.95444 <.0001	0.96434 <.0001
Total Sales				
Inventory	0.47891 <.0001	0.95444 <.0001	1.00000	0.89850 <.0001
Total Inventory				
Returns	0.40611 <.0001	0.96434 <.0001	0.89850 <.0001	1.00000
Total Returns				

LeRoy Bessler The Power To Show™ bessler@execpc.com

PROC CORR

- Correlation Measures
 - Pearson (default)
 - Hoeffding
 - Kendall
 - Spearman
- NOMISS option – see manual for benefits

LeRoy Bessler The Power To Show™ bessler@execpc.com

Some Options

```
proc corr data=sashelp.shoes
  nosimple
  rank
  noprob;
run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

Rank the Correlation Coefficients, Suppress the Significance Probabilities

The CORR Procedure

4 Variables: Stores Sales Inventory Returns

Pearson Correlation Coefficients, N = 395

	Stores	Sales	Inventory	Returns
Stores	1.00000	0.47891	0.42374	0.40611
Number of Stores				
Sales	0.47891	1.00000	0.95444	0.96434
Total Sales				
Inventory	0.42374	0.95444	1.00000	0.89850
Total Inventory				
Returns	0.40611	0.96434	0.89850	1.00000
Total Returns				

LeRoy Bessler The Power To Show™ bessler@execpc.com

Pearson Coeff. Output Data Set

```
proc corr data=sashelp.shoes
  noprint
  outp=OutputFromCORR;
run;
proc print data=OutputFromCORR
  noobs;
run;
```

LeRoy Bessler The Power To Show™ bessler@execpc.com

listing of OUTP= data					
TYPE	_NAME_	Stores	Sales	Inventory	Returns
MEAN		11.648	85700.17	250898.86	2967.32
STD		8.874	129107.23	351514.63	4611.74
N		395.000	395.00	395.00	395.00
CORR	Stores	1.000	0.42	0.48	0.41
CORR	Sales	0.424	1.00	0.95	0.96
CORR	Inventory	0.479	0.95	1.00	0.90
CORR	Returns	0.406	0.96	0.90	1.00

LeRoy Bessler

The Power To Show™

bessler@excecpc.com

```
proc sql; /* to emulate other PROCs */
select Sex, Age,
count(*) as Count, sum(Height) as Sum,
mean(Height) as Mean,
std(Height) as StdDev,
min(Height) as Min, max(Height) as Max
from sashelp.class
group by Sex, Age; quit;
```

LeRoy Bessler The Power To Show™ bessler@excecpc.com

Sex	Age	Count	Sum	Mean	StdDev	Minimum	Maximum
F	11	1	51.3	51.3	.	51.3	51.3
F	12	2	116.1	58.05	2.474874	56.3	59.8
F	13	2	121.8	60.9	6.22254	56.5	65.3
F	14	2	127.1	63.55	1.06066	62.8	64.3
F	15	2	129	64.5	2.828427	62.5	66.5
M	11	1	57.5	57.5	.	57.5	57.5
M	12	3	181.1	60.36667	3.932345	57.3	64.8
M	13	1	62.5	62.5	.	62.5	62.5
M	14	2	132.5	66.25	3.889887	63.5	69
M	15	2	133.5	66.75	0.353553	66.5	67
M	16	1	72	72	.	72	72

LeRoy Bessler The Power To Show™ bessler@excecpc.com

PROC MEANS Equivalent

```
proc means data=sashelp.class
sum mean std min max
/* default stats */
maxdec=1 fw=7;

class Sex Age;
var Height;
run;
```

LeRoy Bessler The Power To Show™ bessler@excecpc.com

The MEANS Procedure							
Analysis Variable : Height							
Sex	Age	N Obs	Sum	Mean	Std Dev	Minimum	Maximum
F	11	1	51.3	51.3	.	51.3	51.3
	12	2	116.1	58.1	2.5	56.3	59.8
	13	2	121.8	60.9	6.2	56.5	65.3
	14	2	127.1	63.6	1.1	62.8	64.3
	15	2	129.0	64.5	2.8	62.5	66.5
M	11	1	57.5	57.5	.	57.5	57.5
	12	3	181.1	60.4	3.9	57.3	64.8
	13	1	62.5	62.5	.	62.5	62.5
	14	2	132.5	66.3	3.9	63.5	69.0
	15	2	133.5	66.8	0.4	66.5	67.0
	16	1	72.0	72.0	.	72.0	72.0

LeRoy Bessler The Power To Show™ bessler@excecpc.com

```
proc sql; /* to create output data set */
create table SQLout
as
select Sex, Age,
list of statistics vars is here
from sashelp.class
group by Sex, Age; quit;
proc print noobs data=SQLout; run;
```

LeRoy Bessler The Power To Show™ bessler@excecpc.com

Sex	Age	Count	Sum	Mean	StdDev	Minimum	Maximum
F	11	1	51.3	51.3000	.	51.3	51.3
F	12	2	116.1	58.0500	2.47487	56.3	59.8
F	13	2	121.8	60.9000	6.22254	56.5	65.3
F	14	2	127.1	63.5500	1.06066	62.8	64.3
F	15	2	129.0	64.5000	2.82843	62.5	66.5
M	11	1	57.5	57.5000	.	57.5	57.5
M	12	3	181.1	60.3667	3.93234	57.3	64.8
M	13	1	62.5	62.5000	.	62.5	62.5
M	14	2	132.5	66.2500	3.88909	63.5	69.0
M	15	2	133.5	66.7500	0.35355	66.5	67.0
M	16	1	72.0	72.0000	.	72.0	72.0

LeRoy Bessler The Power To Show™ bessler@execpc.com

PROC SQL vs PROC MEANS*

for each var1-var2 combination, provide
count sum mean minimum maximum for var3:

obs -	<u>cpu time</u>	<u>elapsed</u>
100 -	SQL 1.00X MEANS	SQL 1.00X MEANS
1,000 -	SQL 1.00X MEANS	SQL 1.00X MEANS
10,000 -	SQL 1.70X MEANS	SQL 1.20X MEANS
100,000 -	SQL 1.57X MEANS	SQL 1.46X MEANS
1,000,000 -	SQL 1.58X MEANS	SQL 1.80X MEANS
10,000,000 -	SQL 1.29X MEANS	SQL 1.15X MEANS
50,000,000 -	SQL 1.20X MEANS	SQL 1.30X MEANS

* Results may vary. Even reruns varied for me.

LeRoy Bessler The Power To Show™ bessler@execpc.com

References on SQL

Introduction to PROC SQL
by kronk@sys-seminar.com (Katie Minten Ronk)

Improve Your Queries: Hints and Tips for Using SQL
by Marje.Fecht@prowerk.com
and Linda.Mitterling@sas.com

Efficiency Techniques for Beginning PROC SQL Users
by KirkLafler@cs.com

Top Ten Reasons to Use PROC SQL
by Weiming.R.Hu@kpchr.org

Please tell them that I referred you

LeRoy Bessler The Power To Show™ bessler@execpc.com

Other References

DATA: "Getting to Know You; Getting to Know All About You"
(The scope is PROC UNIVARIATE.)
by [Larry & Sue Douglass \(ldouglas.umd.edu\)](mailto:Larry & Sue Douglass (ldouglas.umd.edu))

Steps to Success with PROC MEANS
(How can you NOT want to read this?)
by [Andrew Karp \(SierraInfo@aol.com\)](mailto:Andrew Karp (SierraInfo@aol.com))

PROC FREQ—What's it really good for?
by Theresa.L.Gordon@census.gov
and Monique.E.Eleby@census.gov

Please tell them that I referred you

LeRoy Bessler The Power To Show™ bessler@execpc.com

Notice

SAS & SAS/GRAPH are trademarks or registered trademarks of SAS Institute Inc. in the USA and other countries.
® denotes USA registration.

The Power to Show is a trademark of LeRoy Bessler PhD.

LeRoy Bessler The Power To Show™ bessler@execpc.com

Citation:

LeRoy Bessler, "Explore, Analyze, and Summarize Your Data with SAS Software: Selecting the Best Power Tool from a Rich Portfolio", Proceedings of 2004 Wisconsin Illinois SAS Users Conference (Published by Software User Services, PO Box 96, Milwaukee, WI 53201), 2004.

LeRoy Bessler The Power To Show™ bessler@execpc.com