B2B-02

Tailoring Proc Summary for More Efficient Summarizations

Bill Parman, Cigna®, Chattanooga, TN, USA

ABSTRACT

When using Proc Summary many who are new to SAS® programming sort their data and then summarize it. Although there are many summarization techniques possible with Proc Summary, the objective of this paper is the presentation of a fundamental technique, showing how to eliminate the perceived need to pre-sort Proc Summary input data sets and how to tailor Proc Summary to produce only the exact summaries required.

KEY WORDS

NWAY, CLASS, TYPES, DROP, WHERE, _TYPE_, and _FREQ_

INTRODUCTION

Proc Summary is a versatile, powerful, and time-saving procedure. It is one of the author's favorite SAS® procedures. To demonstrate how to summarize SAS® data sets without using Proc Sort and how to code Proc Summary to achieve exactly and only the summaries required, the data set SASHELP.PRDSAL2, containing 23,040 observations, is used in all code examples. All code examples can be run in SAS® as coded. The following scenario is used to demonstrate the objectives of this paper: Suppose you are asked to summarize actual sales in SASHELP.PRDSAL2 in two ways: 1) State within Year and 2) Country within Year.

SOLUTION-1: USING PROC SORT AND PROC SUMMARY

```
Proc Sort
   Data=SASHELP.PRDSAL2
   Out =WORK.PRDSAL2_SRTD
   ;
   By Year Country;
Run;
Proc Summary
   Data=WORK.PRDSAL2_SRTD
   ;
   By  Year Country;
   Var Actual;
   Output

Out=Country_Smry(Drop=_Type_
     Rename=(_Freq_=Count))
     Sum= / Autoname;
Run;
```

```
Proc Sort
   Data=SASHELP.PRDSAL2
   Out =WORK.PRDSAL2_SRTD
   ;
   By Year State;
Run;
Proc Summary
   Data=WORK.PRDSAL2_SRTD
   ;
   By  Year State;
   Var Actual;
   Output

Out=State_Smry(Drop=_Type_
      Rename=(_Freq_=Count))
      Sum= / Autoname;
Run;
```

1995	Canada	1152	$712,440.80
1995	Mexico	1152	$465,915.20
1995	U.S.A.	3456	$2112784.28
1996	Canada	1152	$708,818.40
1996	Mexico	1152	$472,192.80
1996	U.S.A.	3456	$2198151.78
1997	Canada	1152	$890,551.00
1997	Mexico	1152	$582,394.00
1997	U.S.A.	3456	$2640980.35
1998	Canada	1152	$886,023.00
1998	Mexico	1152	$590,241.00
1998	U.S.A.	3456	$2747689.72

Year	State	Count	Actual_Sum
1995	Baja California Norte	288	$116,479.20
1995	British Columbia	288	$177,009.60
	<rows left out>		
1995	Texas	288	$349,455.60
1995	Washington	288	$227,792.80
1996	Baja California Norte	288	$115,099.20
1996	British Columbia	288	$179,060.00
	<rows left out>		
1996	Texas	288	$368,937.60
1996	Washington	288	$229,609.60
1997	Baja California Norte	288	$145,599.00
1997	British Columbia	288	$221,262.00
	<rows left out>		
1997	Texas	288	$436,819.50
1997	Washington	288	$284,741.00
1998	Baja California Norte	288	$143,874.00
1998	British Columbia	288	$223,825.00
	<rows left out>		
1998	Texas	288	$461,172.00
1998	Washington	288	$287,012.00

Output 1: Results for Solution-1

Solution 1 produces exactly what is required: 1) Actual sales summarized by Country within Year and 2) Actual sales summarized by State within Year. However, consider the I/O to achieve the results:

I/O Summary of Solution-1			
Steps	**Country Summary**	**State Summary**	**Total**
Proc Sort – In	23,040	23,040	46,080
Proc Sort – Out	23,040	23,040	46,080
Proc Summary - In	23,040	23,040	46,080
Proc Summary – Out	12	64	76
Total			138,316

Table 1: I/O Summary for Solution-1

As one can see, it takes 4 steps and 138,316 input/output operations to create the two required summarizes. Is it possible to create both summaries in **one** step and reduce the input/output operations? To answer this question, consider some background information about Proc Summary.

PROC SUMMARY BACKGROUND

The CLASS statement and NWAY vs. No NWAY

When using the CLASS statement with Proc Summary, 2^n summarizations are possible, where n is the number of variables listed in the CLASS statement. The inclusion of, or the exclusion of, the NWAY proc option in the Proc Summary statement determines how many of the possible summarizations Proc Summary produces. When NWAY is excluded from the Proc Summary proc options, 2^n summaries are produced. When NWAY is included as a Proc Summary option, only the highest type of summary is produced.

In our scenario, there are 3 variables involved with the two summaries: Year, Country, and State. If these 3 variables are listed in the CLASS statement in Proc Summary and NWAY is excluded as a Proc Summary proc option, 2^3 or 8 summaries are produced in the data set specified in the OUTPUT statement. Additionally, no Proc Sort is required. The summarization combinations in our scenario are:

1. Summarization across ALL rows (_Type_ = 0)
2. Summarization by State (_Type_ = 1)
3. Summarization by Country (_Type_ = 2)
4. Summarization by State within Country (_Type_ = 3)
5. Summarization by Year (_Type_ = 4)
6. Summarization by State within Year (_Type_ = 5)
7. Summarization by Country within Year (_Type_ = 6)
8. Summarization by State within Country within Year (_Type_ = 7)

Notice that the summaries highlighted in red are exactly the ones specified in our scenario.

If NWAY is included as a Proc Summary proc option and the same CLASS statement described above is used, only one of the 8 possible summaries is produced, namely, “Summarization by State within Country within Year (_Type_ = 7).”

In summary (no pun intended), using the CLASS statement with Proc Summary while excluding the NWAY proc option, more summaries are produced than required. On the other hand, using the CLASS statement with Proc Summary while including the NWAY proc option, neither of the required summaries is produced (See the code in Figures 2 and 3 and their output in Outputs 2 and 3).

Is there another Proc Summary statement that will limit its output to exactly what our scenario requires? See section “Solution 2: Using Proc Summary Only” for the answer.

```
Proc Summary
   Data = SASHELP.PRDSAL2
   ;
   CLASS
      Year
      Country
      State
      ;
   VAR
      Actual
      ;
   OUTPUT
      OUT = Smry_Results Sum= / Autoname;
Run;
```

Year	Country	State	_Type_	_Freq_	Actual_Sum
.			0	23040	15008182.32
.		Baja California Norte	1	1152	521,051.40
.		British Columbia	1	1152	801,156.60
<rows left out>					
.	Canada		2	4608	3197833.20
.	Mexico		2	4608	2110743.00
.	U.S.A.		2	13824	9699606.12
.	Canada	British Columbia	3	1152	801,156.60
.	Canada	Ontario	3	1152	780,046.20
<rows left out>					
.	U.S.A.	Texas	3	1152	1616384.70
.	U.S.A.	Washington	3	1152	1029155.40
1995			4	5760	3291140.28
1996			4	5760	3379162.98
1997			4	5760	4113925.35
1998			4	5760	4223953.72
1995		Baja California Norte	5	288	116,479.20
1995		British Columbia	5	288	177,009.60
<rows left out>					
1998		Texas	5	288	461,172.00
1998		Washington	5	288	287,012.00
1995	Canada		6	1152	712,440.80
1995	Mexico		6	1152	465,915.20
<rows left out>					
1998	Mexico		6	1152	590,241.00
1998	U.S.A.		6	3456	2747689.72
1995	Canada	British Columbia	7	288	177,009.60
1995	Canada	Ontario	7	288	171,178.40
<rows left out>					
1998	U.S.A.	Texas	7	288	461,172.00
1998	U.S.A.	Washington	7	288	287,012.00

Output 2: Results for Proc Summary with the CLASS statement while not using NWAY

```
Proc Summary
   Data = SASHELP.PRDSAL2
   NWAY
   ;
   CLASS Year Country State;
   VAR   Actual;
   OUTPUT
      OUT = Smry_Results Sum= / Autoname;
Run;
```

Year	Country	State	_Type_	_Freq_	Actual_Sum
1995	**Canada**	**British Columbia**	**7**	**288**	**177,009.60**
1995	**Canada**	**Ontario**	**7**	**288**	**171,178.40**
<rows left out>					
1998	**U.S.A.**	**Texas**	**7**	**288**	**461,172.00**
1998	**U.S.A.**	**Washington**	**7**	**288**	**287,012.00**

Output 3: Results for Proc Summary with the CLASS statement while using NWAY

SOLUTION-2: USING PROC SUMMARY ONLY

By adding the TYPES statement to the one-step Proc Summary solution, not only will the Proc Sort steps be eliminated, but Proc Summary will be limited to ONLY the two summaries required in our scenario.

Of the possible summaries determined by the number of variables listed in the CLASS statement, the TYPES statement makes possible the specification of exactly which ones Proc Summary produces.

Recall from the previous section, the results of the Proc Summary with the CLASS statement while excluding NWAY. The output data set included the two summaries required by our scenario (_TYPE_ = 5 and _TYPE_=6, highlighted in red), but additional summaries NOT required were produced as well. By adding the TYPES statement, an additional OUTPUT statement, and sub-setting the summary results on the _TYPE_ variable, our precise solution is possible in one step.

During development, to determine the WHERE= and DROP= data set options in each OUTPUT statement do the following:

1. Run the following PROC SUMMARY without any data set options applied to the two OUTPUT files.
    a. Both summaries are written to each file
    b. Observe the two files, noting the values in the _TYPE_ column (5 and 6)
    c. For each _TYPE_ value, observe which column is blank (State or Country)
2. With the information from step-1, code the appropriate data set options, selecting the correct *_TYPE_* value for each summary and dropping the column that does not apply (Country or State).

```
PROC SUMMARY DATA = SASHELP.PRDSAL2;
   CLASS Year Country State;
   TYPES Year * (Country State);
   VAR Actual;
   OUTPUT
      OUT=Country_Smry(Drop=State Where=(_Type_=6)) Sum= / Autoname;
   OUTPUT
      OUT=State_Smry(Drop=Country Where=(_Type_=5)) Sum= / Autoname;
Run;
```

In the TYPES statement think of Year * (Country State) as Year * Country and Year * State, where the asterisk (*) symbolically represents the word "By". In other words Year * (Country State) symbolically represents the summaries: Country within Year and State within Year.

Remember, the summary combinations specified in the TYPES statement must be within the possibilities determined by the CLASS statement.

Year	State	_Type_	_Freq_	Actual_Sum
1995	Baja California Norte	5	288	116,479.20
1995	British Columbia	5	288	177,009.60
<rows left out>				
1995	Texas	5	288	349,455.60
1995	Washington	5	288	227,792.80
1996	Baja California Norte	5	288	115,099.20
1996	British Columbia	5	288	179,060.00
<rows left out>				
1996	Texas	5	288	368,937.60
1996	Washington	5	288	229,609.60
1997	Baja California Norte	5	288	145,599.00
1997	British Columbia	5	288	221,262.00
<rows left out>				
1997	Texas	5	288	436,819.50
1997	Washington	5	288	284,741.00
1998	Baja California Norte	5	288	143,874.00
1998	British Columbia	5	288	223,825.00
<rows left out>				
1998	Texas	5	288	461,172.00
1998	Washington	5	288	287,012.00

Output 4: Results for Solution-2 - State within Year

Year	Country	_Type_	_Freq_	Actual_Sum
1995	Canada	6	1152	712,440.80
1995	Mexico	6	1152	465,915.20
1995	U.S.A.	6	3456	2112784.28
1996	Canada	6	1152	708,818.40
1996	Mexico	6	1152	472,192.80
1996	U.S.A.	6	3456	2198151.78
1997	Canada	6	1152	890,551.00
1997	Mexico	6	1152	582,394.00
1997	U.S.A.	6	3456	2640980.35
1998	Canada	6	1152	886,023.00
1998	Mexico	6	1152	590,241.00
1998	U.S.A.	6	3456	2747689.72

Output 5: Results for Solution-2 - Country within Year

SUMMARY

By now it should be obvious that the one-step solution for the given scenario is distinctly more efficient. Beyond the reduction of steps, with it comes a significant reduction in I/O, which improves overall performance. For the given scenario, the ratio of Solution-1 I/O to Solution-2 I/O is 6:1. Often times, more than anything else, it is I/O, caused by over-processed data, that "kills" the performance of SAS® programs. Solution-2 illustrates an easy way to improve program performance when using Proc Summary to produce specific summary files.

I/O Comparison of Solution-1 and Solution-2			
		Solution-1	**Solution-2**
Proc Sort	In	23,040	---
	Out	23,040	---
Proc Summary	In	23,040	---
	Out	64	---
Proc Sort	In	23,040	---
	Out	23,040	---
Proc Summary	In	23,040	23,040
	Out	12	76
	Total	**138,316**	**23,116**

Table 2: I/O Comparison Summary for Solution-1 and Solution-2

NOTE: There is a close relationship between Proc Summary and Proc Means. Read the documentation to understand this relationship and do not be surprised when opening an example while in the Proc Summary documentation only to find that Proc Means has been used to illustrate the example's concept.

RESOURCES

Base SAS® 9.3 Procedures Guide
http://support.sas.com/documentation/cdl/en/proc/65145/HTML/default/viewer.htm#p0aq3hsvflztfzn1xa2wt6s35oy6.htm

CONTACT INFORMATION

Name: Bill Parman
Enterprise: Cigna®
Address: 401 Chestnut St
City, State ZIP: Chattanooga, TN 37402
E-mail: Francis.Parman@Cigna.Com