Paper 2992-2019

Library Datasets Summary Macro %DATA_SPECS

Jeffrey Meyers, Mayo Clinic, Rochester, MN

ABSTRACT

The field of clinical research often involves sharing data with other research groups and receiving data from other research groups. This creates the need to have a quick and concise way to summarize incoming or outgoing data that allows the user to get a grasp of the number of datasets, number of variables, and number of observations included in the library as well as the specifics of each variable within each dataset. The CONTENTS procedure can fulfill this role to an extent, but the DATA_SPECS macro uses the REPORT procedure along with the Excel Output Delivery System (ODS) destination to create a report that is fine tuned to summarize a library. The macro produces a one page overview of the datasets included in the specified library, and then creates a new worksheet for each dataset that lists all of the variables within that dataset along with their labels, formats, and a short distribution summary that varies depending on variable type. This gives the user an overview of the data that can be used in documents such as data dictionaries, and demonstrates an example of the powerful reports that can be generated with the ODS Excel destination.

INTRODUCTION

The field of clinical research often involves transferring data to and receiving from other research organizations, and this leads to the need for an app or program that can quickly summarize multiple datasets in a streamlined format. The CONTENTS procedure can give a quick summary of the variables contained within one or more datasets, but does not give any information about the values contained within those variables. The DATA_SPECS macro was developed to not only show the important metadata of the datasets within a library, but also gives distribution information that changes depending on the variable type. The macro also checks for variables that exist in multiple datasets which is important to determine if information potentially changes from one dataset to another. The DATA_SPECS macro then outputs all of this information into an EXCEL document using the Output Delivery System (ODS) EXCEL destination to give a streamlined report.

DATA_SPECS MACRO DESCRIPTION

The DATA_SPECS macro is simple to use and is very thoroughly developed for sharing. The macro has customizability with few parameters, full documentation, and full error checking on the parameters. The program will run without changing the user's settings, leaves behind no temporary datasets, and does not put any output to the log or results windows.

MACRO PARAMETERS

The DATA_SPECS macro only has two required parameters and five optional parameters. The required parameters are the only options that require inputs when calling the macro as the optional parameters all have default settings.

Required Parameters

LIBN

The *LIBN* parameter refers to the libname of interest to be summarized by the macro. The library must exist and the libname must be established prior to running the macro.

OUT

The *OUT* parameter refers to the outputted Excel file that is created at the end of the macro that will contain the library summary. The file path should include the .XLSX at the end of the filename and requires the full path.

Optional Parameters

INDEX

The *INDEX* parameter allows the user to specify any potential patient identifying variables that will be used by the macro to determine how many unique patients or subjects are within each dataset if the variable(s) are within the dataset. Multiple variables can be specified in two different ways. If there are multiple index variables, such as patient ID and study center ID, they can be specified in a list separated by spaces. The macro will search through the list and find all of the unique combinations of the listed variable as a new combined INDEX variable. The second scenario is that different datasets have different patient ID variables. This often happens when multiple data transfers happen from the same group, but by different programmers. This can be accommodated by the macro by separating lists with a | (pipe) symbol. The macro will search each dataset for the first of the separated lists that exists and use that to create the index variable.

CAT_THRESHOLD

The *CAT_THRESHOLD* parameter determines how many unique values a numeric variable can have before it is considered continuous. When the number of distinct values is less than the threshold then a frequency distribution will be shown in the summary. Otherwise continuous statistics such as median and range are shown in the summary.

WHERE

The *WHERE* parameter allows the user to pass a WHERE statement into the dataset produced by the CONTENTS procedure within the macro. This can be useful if certain variables or datasets are not wanted within the summary.

FORMAT

The *FORMAT* parameter determines whether the variables within each dataset's summary are shown in a long or wide format. The long format will list all of the variables vertically, and the wide format will list all of the variables horizontally.

ORDER

The *ORDER* parameter determines whether the variables within each dataset are shown alphabetically or by the variable order within the original dataset.

ERROR CHECKING

The DATA_SPECS macro has full error checking programmed in order to improve the user experience. The macro will check the following list of issues:

1. If the library exists and has been assigned
2. If the *OUT* and *LIBN* parameters are missing in the macro call or set to null
3. If the *CAT_THRESHOLD* parameter is not a number greater than 0
4. If the *FORMAT* parameter is not set to a value of LONG or WIDE
5. If the *ORDER* parameter is not set to a value of VARNUM or ALPHA
6. If the current session's SAS® version is not at least 9.4 or greater

If any issues are found the macro will do the following:

1. Stop the macro and restore any changed options
2. Send an error message to the log that describes the issue and gives potential solutions. For example if using the wrong value for the *FORMAT* parameter the macro will provide the list of acceptable values in the log.

REPORT EXAMPLES

The macro outputs an Excel workbook in XLSX format that contains one worksheet with an overview summary of the library and one worksheet for each dataset included in the library. The created file is generated using the ODS Excel tagsets which is exclusive to SAS 9.4 or later. Unlike previous tag sets ODS Excel creates a fully-fledged XLSX file instead of an XML file with large file sizes and compatibility issues.

LIBRARY DATASET SUMMARY TAB

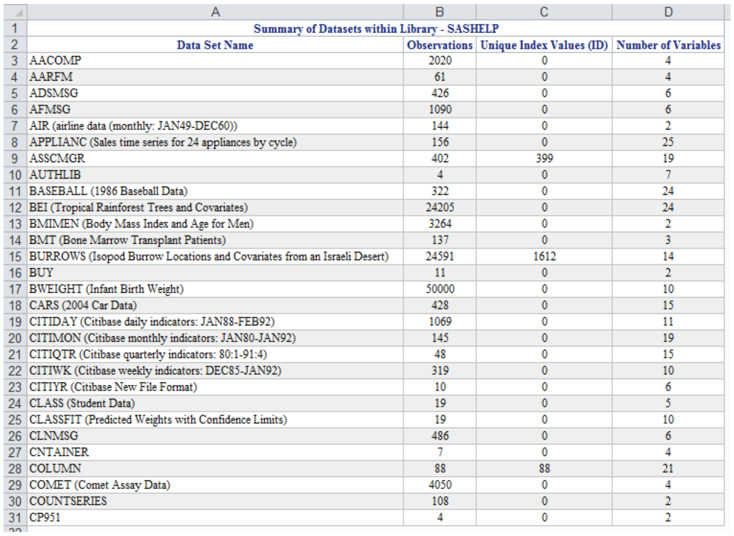
The dataset summary worksheet contains up to two tables. The first table always occurs, and lists each dataset's name, dataset label, number of observations, number of unique index values, and number of variables. The second table only occurs if the macro detects variables that exist in multiple datasets (excluding the index variables).

Table of Datasets within the Library

The table of datasets contains each dataset's name, label, number of observations, number of unique index values, and number of variables. This is designed to allow the user to determine the following at a quick glance:

3. The length and width of each dataset
4. How many datasets are contained within the library
5. Whether each dataset contains the same number of patient records (unique index values)

Figure 1 is a screenshot of the table of several datasets within the library SASHELP



Summary of Datasets within Library - SASHELP

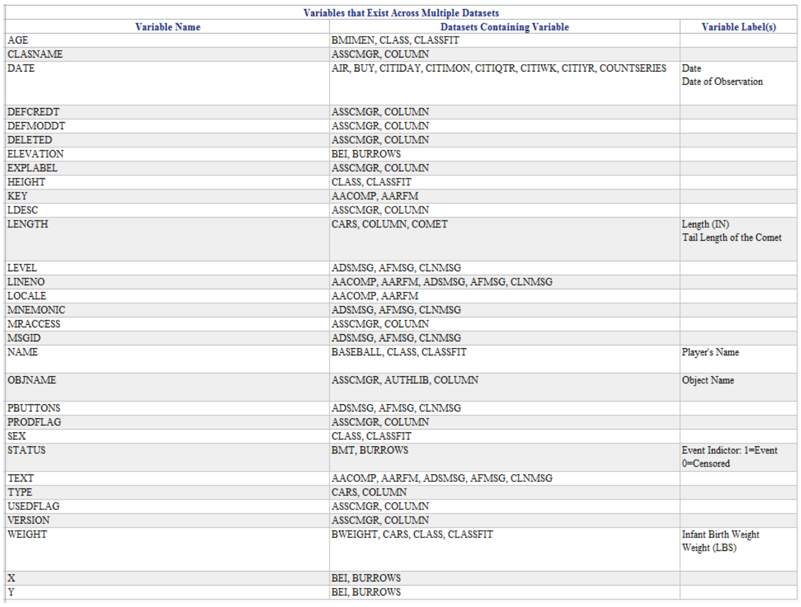
Data Set Name	Observations	Unique Index Values (ID)	Number of Variables
AACOMP	2020	0	4
AARFM	61	0	4
ADSMSG	426	0	6
AFMSG	1090	0	6
AIR (airline data (monthly: JAN49-DEC60))	144	0	2
APPLIANC (Sales time series for 24 appliances by cycle)	156	0	25
ASSCMGR	402	399	19
AUTHLIB	4	0	7
BASEBALL (1986 Baseball Data)	322	0	24
BEI (Tropical Rainforest Trees and Covariates)	24205	0	24
BMIMEN (Body Mass Index and Age for Men)	3264	0	2
BMT (Bone Marrow Transplant Patients)	137	0	3
BURROWS (Isopod Burrow Locations and Covariates from an Israeli Desert)	24591	1612	14
BUY	11	0	2
BWEIGHT (Infant Birth Weight)	50000	0	10
CARS (2004 Car Data)	428	0	15
CITIDAY (Citibase daily indicators: JAN88-FEB92)	1069	0	11
CITIMON (Citibase monthly indicators: JAN80-JAN92)	145	0	19
CITIQTR (Citibase quarterly indicators: 80:1-91:4)	48	0	15
CITIWK (Citibase weekly indicators: DEC85-JAN92)	319	0	10
CITIYR (Citibase New File Format)	10	0	6
CLASS (Student Data)	19	0	5
CLASSFIT (Predicted Weights with Confidence Limits)	19	0	10
CLNMSG	486	0	6
CNTAINER	7	0	4
COLUMN	88	88	21
COMET (Comet Assay Data)	4050	0	4
COUNTSERIES	108	0	2
CP951	4	0	2

Figure 1. The datasets starting with the letters A, B, and C are displayed from the library SASHELP.

Table of Variables that Exist Across Datasets within the Library

The second table shows any variable except those designated by the *INDEX* parameter that exists across multiple datasets, and also displays all unique labels for the variables. This gives the user an indication that the variables should be compared to see if the values are different for the same patient in different datasets for variables such as age.

Figure 2 is a screenshot of the table of variables that exist across the datasets from figure 1 within the library SASHELP



Variables that Exist Across Multiple Datasets		
Variable Name	**Datasets Containing Variable**	**Variable Label(s)**
AGE	BMIMEN, CLASS, CLASSFIT	
CLASNAME	ASSCMGR, COLUMN	
DATE	AIR, BUY, CITIDAY, CITIMON, CITIQTR, CITIWK, CITIYR, COUNTSERIES	Date Date of Observation
DEFCREDT	ASSCMGR, COLUMN	
DEFMODDT	ASSCMGR, COLUMN	
DELETED	ASSCMGR, COLUMN	
ELEVATION	BEI, BURROWS	
EXPLABEL	ASSCMGR, COLUMN	
HEIGHT	CLASS, CLASSFIT	
KEY	AACOMP, AARFM	
LDESC	ASSCMGR, COLUMN	
LENGTH	CARS, COLUMN, COMET	Length (IN) Tail Length of the Comet
LEVEL	ADSMSG, AFMSG, CLNMSG	
LINENO	AACOMP, AARFM, ADSMSG, AFMSG, CLNMSG	
LOCALE	AACOMP, AARFM	
MNEMONIC	ADSMSG, AFMSG, CLNMSG	
MRACCESS	ASSCMGR, COLUMN	
MSGID	ADSMSG, AFMSG, CLNMSG	
NAME	BASEBALL, CLASS, CLASSFIT	Player's Name
OBJNAME	ASSCMGR, AUTHLIB, COLUMN	Object Name
PBUTTONS	ADSMSG, AFMSG, CLNMSG	
PRODFLAG	ASSCMGR, COLUMN	
SEX	CLASS, CLASSFIT	
STATUS	BMT, BURROWS	Event Indictor: 1=Event 0=Censored
TEXT	AACOMP, AARFM, ADSMSG, AFMSG, CLNMSG	
TYPE	CARS, COLUMN	
USEDFLAG	ASSCMGR, COLUMN	
VERSION	ASSCMGR, COLUMN	
WEIGHT	BWEIGHT, CARS, CLASS, CLASSFIT	Infant Birth Weight Weight (LBS)
X	BEI, BURROWS	
Y	BEI, BURROWS	

Figure 2. Variables that exist across the datasets listed in figure 1. Variables with multiple labels can be seen in the third column.

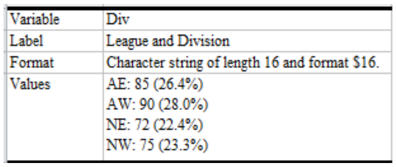
DATASET SPECIFIC TABS

The DATA_SPECS macro will create one worksheet for each dataset contained in the library that contains each variable's name, label, type, format, and a short data description. The data description will differ depending on the variable type and the parameter *CAT_THRESHOLD*. Character variables will always be considered discrete, numeric variables that have distinct values less than or equal to the number specified by *CAT_THRESHOLD* will be considered discrete, and numeric variables that have distinct values greater than the number specified by *CAT_THRESHOLD* will be considered continuous.

Discrete Variables

The number of distinct values that a discrete variable contains will determine which type of data description is listed. For variables that have distinct values less than or equal to the number specified by *CAT_THRESHOLD* the distinct values will be listed in a frequency table.

Figure 3 is a screenshot of the summary of an unformatted character discrete variable with number of distinct values less than or equal to *CAT_THRESHOLD*

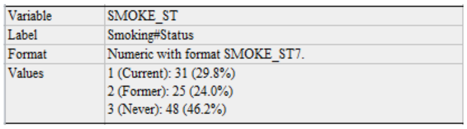


Variable	Div
Label	League and Division
Format	Character string of length 16 and format $16.
Values	AE: 85 (26.4%) AW: 90 (28.0%) NE: 72 (22.4%) NW: 75 (23.3%)

Figure 3. Summary of the variable DIV from the SASHELP.BASEBALL dataset.

The distribution will be different if the discrete variable is formatted versus unformatted. If the discrete variable is formatted then the unformatted values will be listed with the formatted values following in parentheses. If the formatted value matches the unformatted values then the macro does not display the formatted values in parentheses.

Figure 4 is a screenshot of the summary of a formatted numeric discrete variable with number of distinct values less than or equal to *CAT_THRESHOLD*

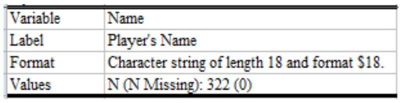


Variable	SMOKE_ST
Label	Smoking#Status
Format	Numeric with format SMOKE_ST7.
Values	1 (Current): 31 (29.8%) 2 (Former): 25 (24.0%) 3 (Never): 48 (46.2%)

Figure 4. Summary of the variable SMOKE_ST from an internal study dataset.

Variables with distinct values above the *CAT_THRESHOLD* will be shown with a simple number of non-missing and missing counts.

Figure 5 is a screenshot of the summary of a discrete variable with number of distinct values greater than *CAT_THRESHOLD*



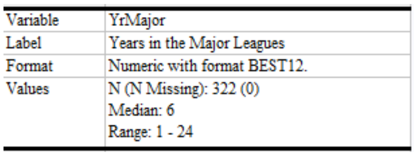
Variable	Name
Label	Player's Name
Format	Character string of length 18 and format $18.
Values	N (N Missing): 322 (0)

Figure 5. Summary of the variable NAME from the SASHELP.BASEBALL dataset.

Continuous Variables

Continuous variables will be displayed with a distribution of values that includes number of non-missing values, number of missing values, median and range. The values are displayed using the format of the variable which allows dates to be shown as dates and variables with decimals to be shown with decimals.

Figure 6 is a screenshot of the summary of a continuous variable.



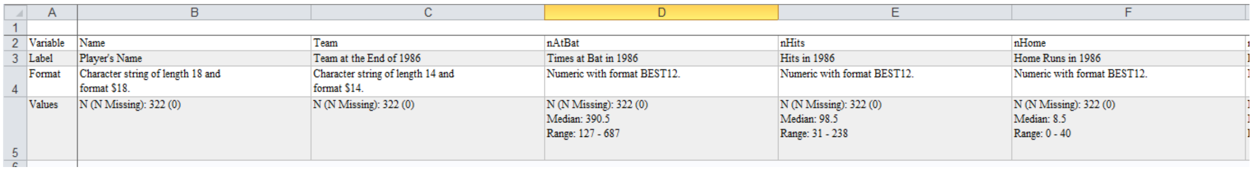
Variable	YrMajor
Label	Years in the Major Leagues
Format	Numeric with format BEST12.
Values	N (N Missing): 322 (0) Median: 6 Range: 1 - 24

Figure 6. Summary of the variable YRMAJOR from the SASHELP.BASEBALL dataset.

Format of Table when FORMAT=WIDE

Setting the *FORMAT* parameter to long will display the variables horizontally in a column. Each row will be shaded every other to make reading easier. The first row and the first column are frozen so that they are always visible.

Figure 7 is a screenshot of the table when *FORMAT*=WIDE



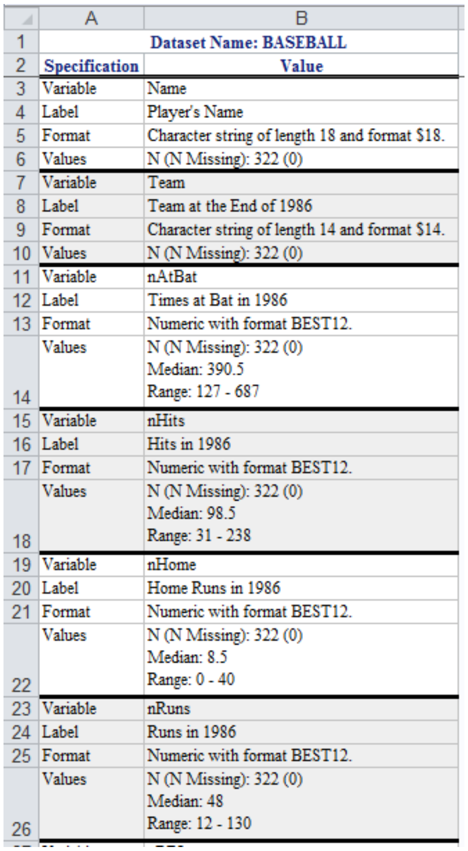
	A	B	C	D	E	F
1						
2	Variable	Name	Team	nAtBat	nHits	nHome
3	Label	Player's Name	Team at the End of 1986	Times at Bat in 1986	Hits in 1986	Home Runs in 1986
4	Format	Character string of length 18 and format $18.	Character string of length 14 and format $14.	Numeric with format BEST12.	Numeric with format BEST12.	Numeric with format BEST12.
5	Values	N (N Missing): 322 (0)	N (N Missing): 322 (0)	N (N Missing): 322 (0) Median: 390.5 Range: 127 - 687	N (N Missing): 322 (0) Median: 98.5 Range: 31 - 238	N (N Missing): 322 (0) Median: 8.5 Range: 0 - 40

Figure 7. Worksheet tab for the dataset SASHELP.BASEBALL when *FORMAT*=WIDE

Format of Table when *FORMAT*=LONG

Setting the *FORMAT* parameter to long will display the variables vertically in a column. Variable sections will be shaded every other to make reading easier. The first two rows are frozen so that they are always visible.

Figure 8. is a screenshot of the table when *FORMAT*=LONG



	A	B
1		Dataset Name: BASEBALL
2	Specification	Value
3	Variable	Name
4	Label	Player's Name
5	Format	Character string of length 18 and format $18.
6	Values	N (N Missing): 322 (0)
7	Variable	Team
8	Label	Team at the End of 1986
9	Format	Character string of length 14 and format $14.
10	Values	N (N Missing): 322 (0)
11	Variable	nAtBat
12	Label	Times at Bat in 1986
13	Format	Numeric with format BEST12.
14	Values	N (N Missing): 322 (0) Median: 390.5 Range: 127 - 687
15	Variable	nHits
16	Label	Hits in 1986
17	Format	Numeric with format BEST12.
18	Values	N (N Missing): 322 (0) Median: 98.5 Range: 31 - 238
19	Variable	nHome
20	Label	Home Runs in 1986
21	Format	Numeric with format BEST12.
22	Values	N (N Missing): 322 (0) Median: 8.5 Range: 0 - 40
23	Variable	nRuns
24	Label	Runs in 1986
25	Format	Numeric with format BEST12.
26	Values	N (N Missing): 322 (0) Median: 48 Range: 12 - 130

Figure 8. Worksheet tab for the dataset SASHELP.BASEBALL when *FORMAT*=LONG

CONCLUSION

The DATA_SPECS macro is a powerful tool for quickly summarizing a library of datasets. The ability to quickly determine the number and size of datasets as well as have a quick glance at the content of the variables within each dataset will increase the efficiency of understanding and integrating new or shared data. The DATA_SPECS macro is also an excellent example of the versatile reports that can be created using the Excel output destination. Being able to create tables within separate tabs allows for quick navigation, and not having vertical or horizontal space limits allows for more information to be presented within a single table. The Excel output destination still has the useful options that TAGSETS.EXCELXP did, but with the benefits of smaller file sizes and more flexible customizations.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Jeffrey Meyers
Enterprise: Mayo Clinic
Address: 200 First St. SW
City, State ZIP: Rochester, MN 55905
E-mail: Meyers.jeffrey@mayo.edu / jpmeyers.spa@gmail.com
Web: https://communities.sas.com/t5/SAS-Communities-Library/Library-Datasets-Summary-Macro-DATA-SPECS/ta-p/544757