Paper 016-30

Datasets and Variables and Labels, Oh My!

Myra A. Oltsik, NOP World

ABSTRACT

It might seem a daunting task when one is faced with reading in several files containing hundreds of variables, all of which need to be renamed. Imagine, then, how frightening it might be to received 48 files with over 20,000 variables, all of which need to be renamed with meaningful names! There's no reason to be afraid with good documentation, organization, and a robust macro. This paper will help you set up a program which will let you read in multiple files with large numbers of variables in an efficient, quick manner.

INTRODUCTION

I received 48 SPSS® datasets that needed to be converted to SAS® datasets. In addition to the respondent ID, the number of variables in each file ranged from as few as 33 to as many as 1,650. In each file, only the Respondent ID had a "usable" variable name. The remaining fields were named as v1 to v33, for example. Though the labels for each variable were explicit and detailed, the variables needed to be renamed to make it easier for analysis.

DOCUMENTATION

Most of the information needed by the macro should be completed before the macro is ever invoked. The basic documentation needed for reading in multiple files are:

- A list of all the files: The list should include the input file name, the output dataset name, and the number of variables in each file. I created a spreadsheet from this list that will be used as input into the macro. (see Appendix, item 1)
- A list of the variables in each file: This list will need to relate the variable [number] with its label and variable name, which you may have to provide. I created macros for each file from this list to help in the renaming. (see Appendix, items 2 & 3)

When I assigned new names for each variable, I used the subject from the beginning of the label, and used the count of the number of variables associated with that subject. For example, there were 18 variables associate with the Car Rental business. In the order they are found in the file, the variables were renamed to Car_Rental_0001-Car_Rental_0018. [1]

[1] See my paper, "SUGI 28 Paper 251 - Show and Tell" for instructions on using Microsoft Excel to help you write code.

THE PROGRAM

After importing the Excel spreadsheet [2] with the list of files (see Appendix, item 4, line 8), macro variables are associated with each file (see Appendix, item 4, line 18) before the macro is initiated.

The macro is simple and straightforward. Through looping, for each file, the macro converts it from an SPSS® portable file (see Appendix, item 4, line 44) to a temporary SAS® dataset. Another temporary output dataset name is given (see Appendix, item 4, line 47), and the correct renaming macro is called (see Appendix, item 4, line 65). Since all the non-identifying variables are "binary", having the value of either 0 or 1, the macro also makes sure these variables have the minimum length allowed for the operating system (see Appendix, item 4, line 52). Finally, the temporary dataset is sorted on the Respondent ID variable to a permanent dataset (see Appendix, item 4, line 73).

The remaining lines of code in the macro produce information on the datasets' contents, and print the minimum and maximum values for each variable along with their means and sums. The temporary datasets are then deleted. A put statement to the log lets the user know which file has finished processing.

CONCLUSION

With careful setup and good documentation, it is relatively easy to read in multiple files and rename thousands of variables when the need presents itself. A well organized macro will take care of it for you.

CONTACT INFORMATION

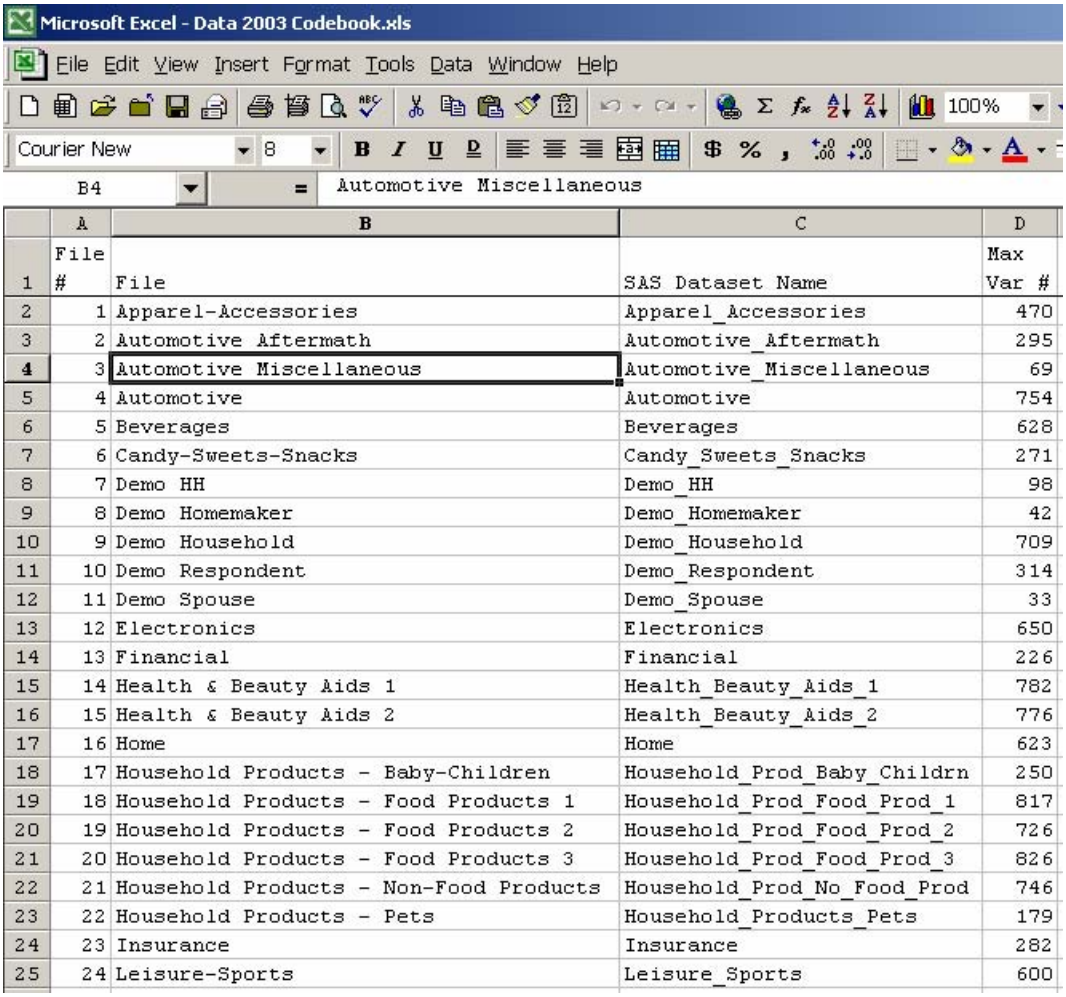
Myra A. Oltsik
Manager, Analytic Programming and Databases
Database Marketing/DemandMatch CoE
NOP World
500 Mamaroneck Ave, Suite 103
Harrison, NY 10528-1600
moltsik@nopworld.com

[2] SAS® 6.12 can import and export Microsoft Excel versions 5, 95, and 97. SAS® 8.2 can also handle Microsoft Excel 2000. SAS® 9 can also handle Microsoft Excel 2002.

Paper 016-30

APPENDIX

(1) Portion of spreadsheet with list of files:



	A	B	C	D
1	File #	File	SAS Dataset Name	Max Var #
2	1	Apparel-Accessories	Apparel_Accessories	470
3	2	Automotive Aftermath	Automotive_Aftermath	295
4	3	Automotive Miscellaneous	Automotive_Miscellaneous	69
5	4	Automotive	Automotive	754
6	5	Beverages	Beverages	628
7	6	Candy-Sweets-Snacks	Candy_Sweets_Snacks	271
8	7	Demo HH	Demo_HH	98
9	8	Demo Homemaker	Demo_Homemaker	42
10	9	Demo Household	Demo_Household	709
11	10	Demo Respondent	Demo_Respondent	314
12	11	Demo Spouse	Demo_Spouse	33
13	12	Electronics	Electronics	650
14	13	Financial	Financial	226
15	14	Health & Beauty Aids 1	Health_Beauty_Aids_1	782
16	15	Health & Beauty Aids 2	Health_Beauty_Aids_2	776
17	16	Home	Home	623
18	17	Household Products - Baby-Children	Household_Prod_Baby_Childrn	250
19	18	Household Products - Food Products 1	Household_Prod_Food_Prod_1	817
20	19	Household Products - Food Products 2	Household_Prod_Food_Prod_2	726
21	20	Household Products - Food Products 3	Household_Prod_Food_Prod_3	826
22	21	Household Products - Non-Food Products	Household_Prod_No_Food_Prod	746
23	22	Household Products - Pets	Household_Products_Pets	179
24	23	Insurance	Insurance	282
25	24	Leisure-Sports	Leisure_Sports	600

Paper 016-30

(2) Spreadsheet of Variables for File 3 – Automotive Miscellaneous:

	A	B	C	F
1	SPSS Var #	Label	File #	New Field Name
2	0	RespID	3	
3	1	wgtpop	3	Pop_Weight
4	2	All	3	
5	3	Car Rental - Business Use: Alamo	3	Car_Rental_0001
6	4	Car Rental - Business Use: Avis	3	Car_Rental_0002
7	5	Car Rental - Business Use: Budget	3	Car_Rental_0003
8	6	Car Rental - Business Use: Dollar	3	Car_Rental_0004
9	7	Car Rental - Business Use: Enterprise	3	Car_Rental_0005
10	8	Car Rental - Business Use: Hertz	3	Car_Rental_0006
11	9	Car Rental - Business Use: National	3	Car_Rental_0007
12	10	Car Rental - Business Use: Thrifty	3	Car_Rental_0008
13	11	Car Rental - Business Use: Other	3	Car_Rental_0009
14	12	Car Rental - Personal Use: Alamo	3	Car_Rental_0010
15	13	Car Rental - Personal Use: Avis	3	Car_Rental_0011
16	14	Car Rental - Personal Use: Budget	3	Car_Rental_0012
17	15	Car Rental - Personal Use: Dollar	3	Car_Rental_0013
18	16	Car Rental - Personal Use: Enterprise	3	Car_Rental_0014
19	17	Car Rental - Personal Use: Hertz	3	Car_Rental_0015
20	18	Car Rental - Personal Use: National	3	Car_Rental_0016
21	19	Car Rental - Personal Use: Thrifty	3	Car_Rental_0017
22	20	Car Rental - Personal Use: Other	3	Car_Rental_0018
23	21	Auto Clubs: Belong to an auto club	3	Auto_Clubs_0001
24	22	Auto Clubs: Member of AAA	3	Auto_Clubs_0002
25	23	Auto Clubs: Member of AARP	3	Auto_Clubs_0003
26	24	Auto Clubs: Member of Allstate	3	Auto_Clubs_0004
27	25	Auto Clubs: Member of Amoco	3	Auto_Clubs_0005
28	26	Auto Clubs: Member of Car dealer	3	Auto_Clubs_0006
29	27	Auto Clubs: Member of other	3	Auto_Clubs_0007
30	28	Driving: Have valid drivers license	3	Driving_0001
31	29	Driving: Drive automobile	3	Driving_0002

(3) Rename Macro for File 3 – Automotive Miscellaneous:

```
%macro rename3;
   v1    = Pop_Weight
   v3    = Car_Rental_0001
   v4    = Car_Rental_0002
   v5    = Car_Rental_0003
   v6    = Car_Rental_0004
   v7    = Car_Rental_0005
   v8    = Car_Rental_0006
   v9    = Car_Rental_0007
   v10   = Car_Rental_0008
   v11   = Car_Rental_0009
   v12   = Car_Rental_0010
   v13   = Car_Rental_0011
   v14   = Car_Rental_0012
   v15   = Car_Rental_0013
   v16   = Car_Rental_0014
   v17   = Car_Rental_0015
   v18   = Car_Rental_0016
```

```
   v19   = Car_Rental_0017
   v20   = Car_Rental_0018
   v21   = Auto_Clubs_0001
   v22   = Auto_Clubs_0002
. . .
   v49   = Mileage_0010
   v50   = Mileage_0011
   v51   = Mileage_0012
   v52   = Miles_Traveled_in_Town_City_0001
   v53   = Miles_Traveled_in_Town_City_0002
. . .
   v69   = Miles_Traveled_in_Town_City_0018
%mend;
```

(4) SAS® Program:

```
  1   libname lbnm "C:\TEST\data\sas";
  2
  3   %include "C:\TEST\programs\Rename Macro.sas";
  4
  5 /*****************************************************************/
  6 /* READ IN LIST OF ALL FILES WITH NUMBER OF VARIABLES IN EACH FILE. */
  7 /*****************************************************************/
  8   proc import out=data_files
  9      datafile= "C:\TEST\documentation\DATA 2003 codebook.xls"
 10      dbms=excel2000 replace;
 11      range="'Max Var Num by File$'";
 12      getnames=yes;
 13   run;
 14
 15 /*****************************************************************/
 16 /* MAKE MACRO VARIABLES FROM INFORMATION ABOUT EACH FILE.          */
 17 /*****************************************************************/
 18   data _null_;
 19      set data_files;
 20      call symput('file'||left(put(_n_,8.)),File);
 21      call symput('vars'||left(put(_n_,8.)),left(put(Max_Var__,8.)));
 22      call symput('dsnm'||left(put(_n_,8.)),SAS_Dataset_Name);
 23   run;
 24
 25 /*****************************************************************/
 26 /* COUNT NUMBER OF FILES TO BE READ.                              */
 27 /*****************************************************************/
 28   proc sql noprint;
 29      select count(*) into :dcount
 30      from data_files
 31      ;
 32   quit;
 33
 34   %macro multiple_read;
 35
 36      %do z = 1 %to &dcount;
 37
 38         %let file   = &&file&z;
 39         %let vars   = &&vars&z;
 40         %let dsnm   = &&dsnm&z;
```

Paper 016-30

```
41
42          filename spss "C:\TEST\Data\Received\Data 2003 &file..por";
43
44          proc convert spss=spss out=&dsnm._sp;
45          run;
46
47          data &dsnm;
48             length
49                RESPONDENT_ID $10.
50                WAVE          $2.
51                V1            8.
52                V3-V&vars     3.
53             ;
54             set &dsnm._sp(drop=V2);
55             RESPONDENT_ID = put(RESPID,10.);
56             WAVE = substr(RESPONDENT_ID,1,2);
57                   label
58                      V1            = 'Population Weight'
59                      RESPONDENT_ID = 'Respondent ID'
60                      WAVE          = 'Study Wave'
61                   ;
62 /*****************************************************************/
63 /* RUN MACRO TO RENAME VARIABLES FOR A PARTICULAR DATASET.       */
64 /*****************************************************************/
65             rename
66                %rename&z
67             ;
68             drop
69                RESPID
70             ;
71          run;
72
73          proc sort data=&dsnm
74             out=lbnm.&dsnm.(
75                compress=binary
76                reuse=yes
77                label="Data 2003 &file."
78             );
79             by RESPONDENT_ID;
80          run;
81
82          proc contents data=lbnm.&dsnm;
83          title3 "Contents from lbnm.&dsnm";
84          run;
85
86          proc print data=lbnm.&dsnm(obs=6) heading=h;
87          title3 "Proc print procedure for first 6 obs from lbnm.&dsnm";
88          run;
89
90          proc means data=lbnm.&dsnm noprint;
91             output
92                out=var_means(drop=_freq_ _type_)
93             ;
94          run;
95
96          proc transpose data=var_means
97             out=transp(
```

Paper 016-30

```
 98              drop=_LABEL_
 99              std
100              rename=(_NAME_=NAME)
101           );
102           id _STAT_;
103        run;
104
105        proc sort data=transp;
106           by NAME;
107        run;
108
109        data cols;
110           set sashelp.vcolumn;
111           if LIBNAME = "lbnm";
112           if MEMNAME = upcase("&dsnm");
113           keep
114              NAME
115              LABEL
116              VARNUM
117              TYPE
118           ;
119        run;
120
121        proc sort data=cols;
122           by TYPE VARNUM;
123        run;
124
125        data _null_;
126           set cols(where=(TYPE='num'));
127           by type;
128           if first.TYPE then call symput('fst',NAME);
129           if last.TYPE  then call symput('lst',NAME);
130        run;
131
132        proc summary data=lbnm.&dsnm missing;
133           var &fst -- &lst ;
134           output
135              sum=
136              out=sum_means
137           ;
138        run;
139
140        proc transpose data=sum_means
141           out=transps(
142              drop=_LABEL_
143              rename=(_NAME_=NAME)
144           );
145        run;
146
147        proc sort data=cols;
148           by NAME;
149        run;
150
151        proc sort data=transp;
152           by NAME;
153        run;
154
```

Paper 016-30

```
155         proc sort data=transps;
156            by NAME;
157         run;
158
159         data transpout;
160            merge
161               cols(in=INA)
162               transp
163               transps(rename=(COL1=SUM))
164            ;
165            by NAME;
166            if INA;
167         run;
168
169         proc sort data=transpout;
170            by VARNUM;
171         run;
172
173         proc print data=transpout noobs heading=h;
174         title3 "Means for lbnm.&dsnm";
175         run;
176
177         proc append base=lbnm.DATA_2003_means data=transpout force;
178         run;
179
180         proc datasets lib=work nolist;
181            delete
182               &dsnm
183               &dsnm._sp
184               cols
185               sum_means
186               transp
187               transps
188               transpout
189               var_means
190            ;
191         run; quit;
192
193         %put Loop &z.: Dataset &dsnm created from file &file with
&vars variables.;
194
195      %end;
196
197   %mend;
198
199   %multiple_read;
```