www.OASUS.ca

Cody's Data Cleaning Techniques
Using SAS
Second Edition

by Ron Cody

Reviewed by Josée Blackburn

"Come out of the desert of ignorance to the OASUS of knowledge"

Ottawa Area SAS Users Society

www.OASUS.ca

About the author

- Private consultant/national instructor
- Presenter at many SAS conferences
- Author of 7 SAS books
- SAS courses

  Two courses are based on his books

  – Data Cleaning

  – SAS Functions by example

Why this book?

- Shows the importance of data cleaning
- Easy-to-follow examples
- Some new techniques with SAS 9 *
- DATA STEP vs MACROS Alternatives
- SQL approaches to data cleaning
- Integrity constraints and Audit trail *

* Second edition only

Ottawa Area SAS Users Society

www.OASUS.ca

Chapters

1. Checking Values of Numeric Variables
2. Checking for Missing Values
3. Working with Dates
4. Looking for Duplicates and ‘n’ Observations per Subject
5. Working with Multiples Files
6. Checking Values of Character Variables
7. Double Entry and Verification (PROC COMPARE)
8. Some PROC SQL Solutions to Data Cleaning
9. Correcting Errors
10. Creating Integrity Constraints and Audit Trails
11. DataFlux and dfPower Studio

Book example with the ?? Modifier

```
Title “Listing of missing and invalid dates”;
data _NULL_;
        file print;
        infile “c:\books\clean\patients.txt” truncover;
        input     @1 Patno $3.
                  @5 V_date $char10.;
        Visit = input (V_date, ?? Mmddyy10.);
        format Visit mmddyy10.;
        if missing(Visit) then put Patno= V_date=;
Run;
```

* You can also use the ?? Informat Modifier with the INPUT statement

?? Modifier

- Prevent long SAS log for known errors

NOTE: Invalid argument to function INPUT at line 111 column 13.
encPatWID=1449 labSingleResult=Noresult Result=. _ERROR_=1 _N_=7
NOTE: Invalid argument to function INPUT at line 111 column 13.
encPatWID=1449 labSingleResult=Noresult Result=. _ERROR_=1 _N_=8
NOTE: Invalid argument to function INPUT at line 111 column 13.
encPatWID=13644 labSingleResult=Not perf Result=. _ERROR_=1 _N_=345
NOTE: Invalid argument to function INPUT at line 111 column 13.
ERROR: Limit set by ERRORS= option reached. Further errors of this type will not be printed.
encPatWID=15315 labSingleResult=See Note Result=. _ERROR_=1 _N_=454
NOTE: Mathematical operations could not be performed at the following places. The results of the
operations have been set to missing values.
Each place is given by: (Number of times) at (Line):(Column).
1461 at 111:13

Lab test with ?? Modifier

```
/* Converting Character variable to numeric*/
/* with the ?? Modifier */

data labtest1;
    set dw.labtest;
    Result=INPUT(labSingleResult,?? 9.);

    if missing(result) then do;
        if labSingleResult in ('<10','< 10') then result=9;
        else if labsingleresult in ('<0.01','< 0.01') then result=0.009;
        else if labsingleresult='<0.05' then result=0.0499;
    end;
run;
```

Lab test with NOTDIGIT

NotDigit: looks for non numeric data

```
data labtest3;
set dw.labtest;
   if notdigit(trim(labSingleResult)) then do;
            if labSingleResult in ('<10','< 10') then result=9;
            if labsingleresult in ('<0.01','< 0.01') then
                     result=0.009;
            if labsingleresult='<0.05' then result=0.0499;
   end;
   else
   Result=INPUT(labSingleResult,9.);
run;
```

Good diet for your log

- Use ?? Modifier
  - only for errors that you are expecting or dealing with in other ways
  - May improve program efficiency
  - Sets the value of _ERROR_ to 0
  - Makes a cleaner log file
- Avoid SAS code that creates log notes or errors that you will ignore
- Look at your LOG

Ottawa Area SAS Users Society

www.OASUS.ca

Books-by-Users

- http://support.sas.com/publishing/

Buy directly from SAS or from Amazon

www.amazon.ca

(Free shipping available on orders over $39)

Ottawa Area SAS Users Society

www.OASUS.ca

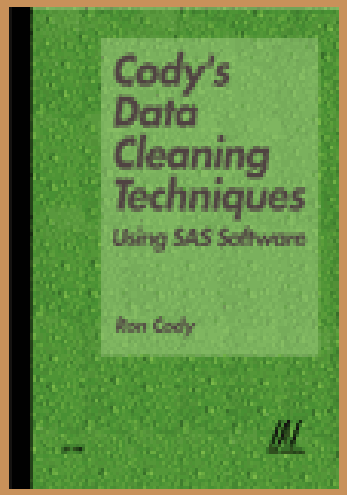
Books reviews you can do

- Interested in doing the next session book review?
  - Jot your name on the evaluation form
  - Come to see me or contact me
  - SAS will contact winner and work on next steps

Ottawa Area SAS Users Society

www.OASUS.ca

From SAS Press



ISBN: 978-1-59994-659-7

Ottawa Area SAS Users Society

www.OASUS.ca

My only frustration with this book…

- After getting all excited about the features of Dataflux and dfPower Studio in Chapter 11, you get this sentence:

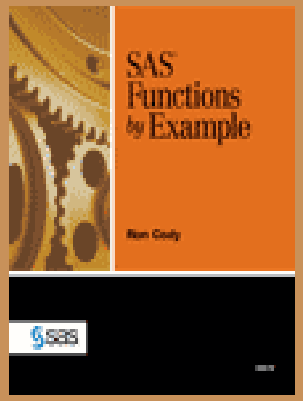
> *What is shown here is the tip of the iceberg. For more information on DataFlux and dfPower Studio, please contact SAS at (919) 677-8000 and request information or a demonstration of this powerful package. ”*

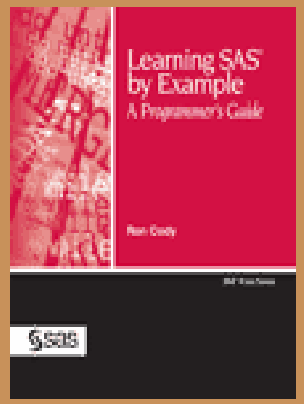
Another SAS product to add to my wish list.

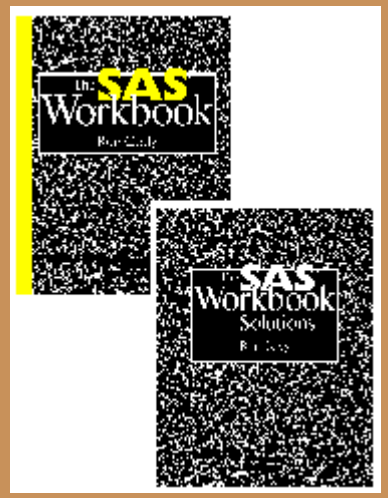
Ottawa Area SAS Users Society

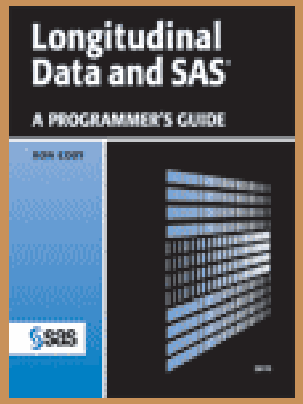
www.OASUS.ca

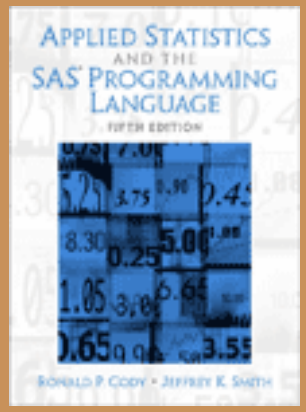
Other books by Ron Cody











Ottawa Area SAS Users Society

www.OASUS.ca

Questions?

- Josée Blackburn

  jblackburn@ohri.ca