The Missing(ness) Piece

Louise Hadden presented at her first SAS conference in 1996 and has never looked back, presenting at multiple conferences across the continent over the years. She supports file building and analytic programming for life sciences organizations, most frequently government agencies such as CMS and CDC, and specializes in reporting and data visualization.

THE MISSING(NESS) PIECE

Louise S. Hadden, Cormac Corporation

AGENDA

- The importance of missing data
- Preparing to analyze missing Values
  - Special Missing Values
  - Missing Functions
- PROC FREQ, ODS Output Objects, & NLEVELS
- Arrays, NOBS, and PROC UNIVARIATE OUTTABLE
- Assembling the Missing Piece(s)

MISSING DATA

What does missing data look like?

- In the context of this presentation, we are talking about missing values in columns / variables.
- These missing values can be identified programmatically and reported upon
- “Missing” is different for different types of variables
  - Missing character values are blank or null
  - Missing numeric values can take 28 different values: ., ._, .A through .Z

What does missing data look like?

```
%macro makerec(nn=.,cc=.);
    numvar=&nn;
    numvarf=&nn;
    charvar="&cc";
    charvarf="&cc";
    output;
%mend;
```

```
proc format;
    value ntos 95 = .O
               96 = .N
               97 = .R
               98 = .A
               99 = .U;
```

```
data test_missing (drop=i label="Test Missing Representation");
 length charvar charvarf $ 2 numvar numvarf 8;

    numvar=.;
    numvarf=.;
    charvar=' ';
    charvarf=' ';
    output;

    %makerec(nn=.,cc=.);
    %makerec(nn=._,cc=_);

    %makerec(nn=1,cc=1);
    %makerec(nn=2,cc=2);
```

What does missing data look like?

numvar	numvarf	charvar	charvarf	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	.			1	2.56	1	2.56
.	.	.	.	1	2.56	2	5.13
A	Not Answered	A	A	1	2.56	3	7.69
B	B	B	B	1	2.56	4	10.26
C	C	C	C	1	2.56	5	12.82
D	D	D	D	1	2.56	6	15.38
E	E	E	E	1	2.56	7	17.95
5	5	5	5	1	2.56	34	87.18
95	95	O	O	1	2.56	35	89.74
96	96	N	N	1	2.56	36	92.31
97	97	R	R	1	2.56	37	94.87
98	98	A	A	1	2.56	38	97.44
99	99	U	U	1	2.56	39	100.00

Reporting on Missing Values

- PROC UNIVARIATE and other statistical procedures report out on the number of missing values as a matter of course
- Depending on procedural options in PROC FREQ and PROC SURVEYFREQ, missing values can be included or not, in counts

Reporting on Missing Values

- Most statistical procedures simply drop records with missing values.
- Reporting on missing values can occur via “list” output, procedural output or ODS output objects.
- The focus for this presentation is on using PROC FREQ and PROC UNIVARIATE to report on missing values by variable in a data set

PREPARATION FOR REPORTING

- PROC FORMAT will report on the different missing values when used with reporting procedures and PROC FREQ/SURVEYFREQ.
- When the format shown on the slide below is applied to a variable, .O, .M, .V are all reported out separately

PREPARATION FOR REPORTING

```
proc format;
    value varx_f .O = '.O: Other Specify'
                 .M = '.M: Missing'
                 .V = '.V: Valid Skip'
              other = 'Non-Missing';
run;
```

Test special missing values

testvar	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.M: Missing	4	23.53	4	23.53
.O: Other Specify	3	17.65	7	41.18
.V: Valid Skip	4	23.53	11	64.71
Non-Missing	6	35.29	17	100.00

PROC FREQ NLEVELS

PROC FREQ NLEVELS

- SAS has an underutilized variant of PROC FREQ that allows you to produce a missingness report
- The procedural option to use is NLEVELS
- As we saw earlier, if your data has special missing NUMERIC values, these will be reported as “levels” of missing.
- CHARACTER variables have a single “level” of missing

PROC FREQ NLEVELS

```
Ods trace;
proc freq data=int.&infi. nlevels;
    ods output nlevels=nlevels0;
    tables _all_ / noprint;
run;
ods output close;
ods trace off;

proc print data=nlevels0 (obs=5) noobs;
title 'Test nlevels output';
run;

proc contents data=nlevels0 varnum;
run;
```

PROC FREQ NLEVELS

```
proc freq data=int.&infi. nlevels;
    tables testvar / noprint;
run;
```

NLEVELS

Number of Variable Levels			
Variable	Levels	Missing Levels	Nonmissing Levels
testvar	8	3	5

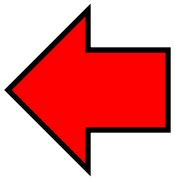
testvar	Frequency	Percent	Cumulative Frequency	Cumulative Percent
M	4	23.53	4	23.53
O	3	17.65	7	41.18
V	4	23.53	11	64.71
1	1	5.88	12	70.59
2	1	5.88	13	76.47
3	1	5.88	14	82.35
4	1	5.88	15	88.24
5	2	11.76	17	100.00

PROC FREQ NLEVELS

```
Ods trace;

proc freq data=testdata nlevels;
     ods output nlevels=nlevels0;
     tables _all_ / noprint;
run;
ods output close;

ods trace off;
```

PROC FREQ NLEVELS

```
data nlevels;
   set nlevels0;
   label TableVar = "Variable Name"
   TableVarLabel = "Variable Description"
   NLevels = "# of Variable values"
   NMissLevels = "# of Missing Value Levels"
   NNonMissLevels = "# of Non- Missing
      Value Levels";
   shadeit=(nnonmisslevels=0);
run;
```



PROC FREQ NLEVELS

```
proc report nowd data=nlevels split='|';
  columns TableVar TableVarLabel NLevels NmissLevels
                NNonMissLevels shadeit;
  define shadeit / display ' ' noprint;
  define TableVar / style(COLUMN)={just=l \
           font_face="Helvetica"
           font_size=8pt cellwidth=295 }
           style(HEADER)={just=l font_face="Helvetica"
           font_weight=bold font_size=8pt  };
        . . .
  compute shadeit;
          if (shadeit eq 1) then call
          define(_row_,"STYLE","STYLE=[BACKGROUND=PINK]");
  endcomp;
run;
```

PROC FREQ NLEVELS

We are able to produce a report from a data set with thousands of variables with a few lines of PROC FREQ and PROC REPORT code, instantly highlighting records for variables which may have a missingness problem. Using the SHADEIT variable to screen, we could produce a report with variables with only missing values to research.

Variable Name	Variable Description	# of Variable values	# of Missing Value Levels	# of Non-Missing Value Levels
IN_DATA_EXTRCT_DT	Mo 4: Date of data extraction	1	0	1
INF_IDENTIFIER1	Mo 4: Infant identifier-#1	1550	0	1550
IN_AMB_VISIT_DT	Mo 4: Date of any ambulatory care visit, including antenatal care, ED, telemedicine.	1	1	0
IN_CHLOROQ_END_DATE	Mo 4: First administration of treatment - End Date: Chloroquine Phosphate (Chloroquine)	1	0	1

BUT WAIT, THERE'S MORE!

```
%macro filecheck(inlib=out,
inmem=recover_surveillance_&delivdate.);
proc sql noprint;
     create table filelist0 as
     select libname, memname, nobs
     from dictionary.tables
     where libname = upcase("&inlib");
quit;
data filelist;
    set filelist0 (where=(memname = upcase("&inmem.")));
run;
%mend;

%filecheck(inlib=out,inmem=recover_surveillance_&delivdate.);
```

But wait, there's more!

```
data nlevels out.nlevels_surveillance_&delivdate.;
    length cr_type $ 12;
    set nlevels0;
    label TableVar = "Variable Name"
        TableVarLabel = "Variable Description"
        NLevels = "# of Variable values"
        NMissLevels = "# of Missing Value Levels"
        NNonMissLevels = "# of Non- Missing Value Levels";
      shadeit=(nnonmisslevels=0);
      /* shades variables with NO non missing values */
        nobs=&fileobs;
        cardinality=nobs/nlevels;
        label nobs = "# of Observations"
        cardinality = "Cardinality Ratio";
        select;
            when(nlevels eq 1 ) cr_type = '.unique';
            when(nlevels gt 10) cr_type = 'many';
            otherwise cr_type = 'few';
        end;
        label cr_type='Cardinality Type';
run;
```

BUT WAIT, THERE'S MORE!

Variable Name	Variable Description	# of Variable values	# of Missing Value Levels	# of Non-Missing Value Levels	# of Observations	Cardinality Ratio
target_4_target_1	Marshfield lab data: Lab-Swab 1: Target 4 Target: Name/type of target that the next 2 fields (Cq and call) are referencing.	3	1	2	304838	101612.67
target_4_cq_1	Marshfield lab data: Lab-Swab 1: Target 4 Cq: 'CQ' value is the numeric data point that comes from the PCR test for this target. There is a cut-off that determines positivity vs negativity which is assay specific.	1309	1	1308	304838	232.87853
target_4_call_1	Marshfield lab data: Lab-Swab 1: Target 4 Call: The 'call' is the written interpretation of the CQ value.	3	1	2	304838	101612.67
target_5_target_1	Marshfield lab data: Lab-Swab 1: Target 5 Target: Name/type of target that the next 2 fields (Cq and call) are referencing.	1	1	0	304838	304838
target_5_cq_1	Marshfield lab data: Lab-Swab 1: Target 5 Cq: 'CQ' value is the numeric data point that comes from the PCR test for this target. There is a cut-off that determines positivity vs negativity which is assay specific.	1	1	0	304838	304838
target_5_call_1	Marshfield lab data: Lab-Swab 1: Target 5 Call: The 'call' is the written interpretation of the CQ value.	1	1	0	304838	304838
assay_1	Marshfield lab data: Lab-Swab 1: This is the name of the assay that was being used at Marshfield when this test was conducted.	6	1	5	304838	50806.333

But wait, there's more!

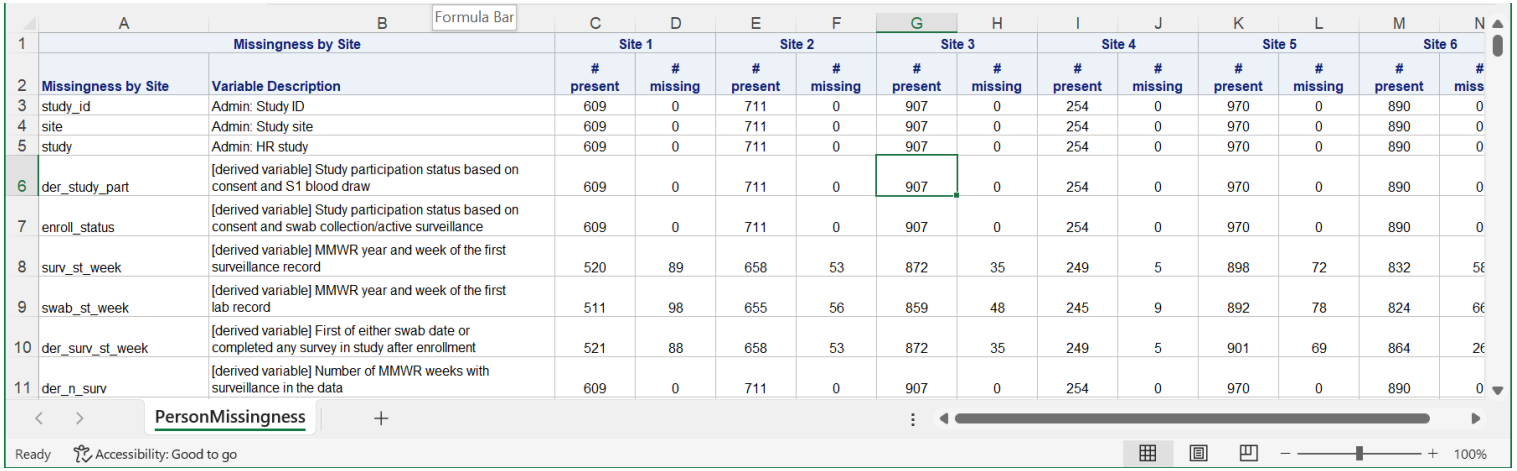
Frequency on CR_TYPE

Cardinality Type				
cr_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.unique	140	13.21	140	13.21
few	551	51.98	691	65.19
many	369	34.81	1060	100.00

PROC UNIVARIATE OUTTABLE

A USE CASE FOR A "SIMPLER" REPORT

- Large CDC "Illness" data base – thousands of records and variables
- Needed a broad stroke summary of presence and missingness by variable and site, with variable labels.
- Wanted to minimize the amount of user intervention



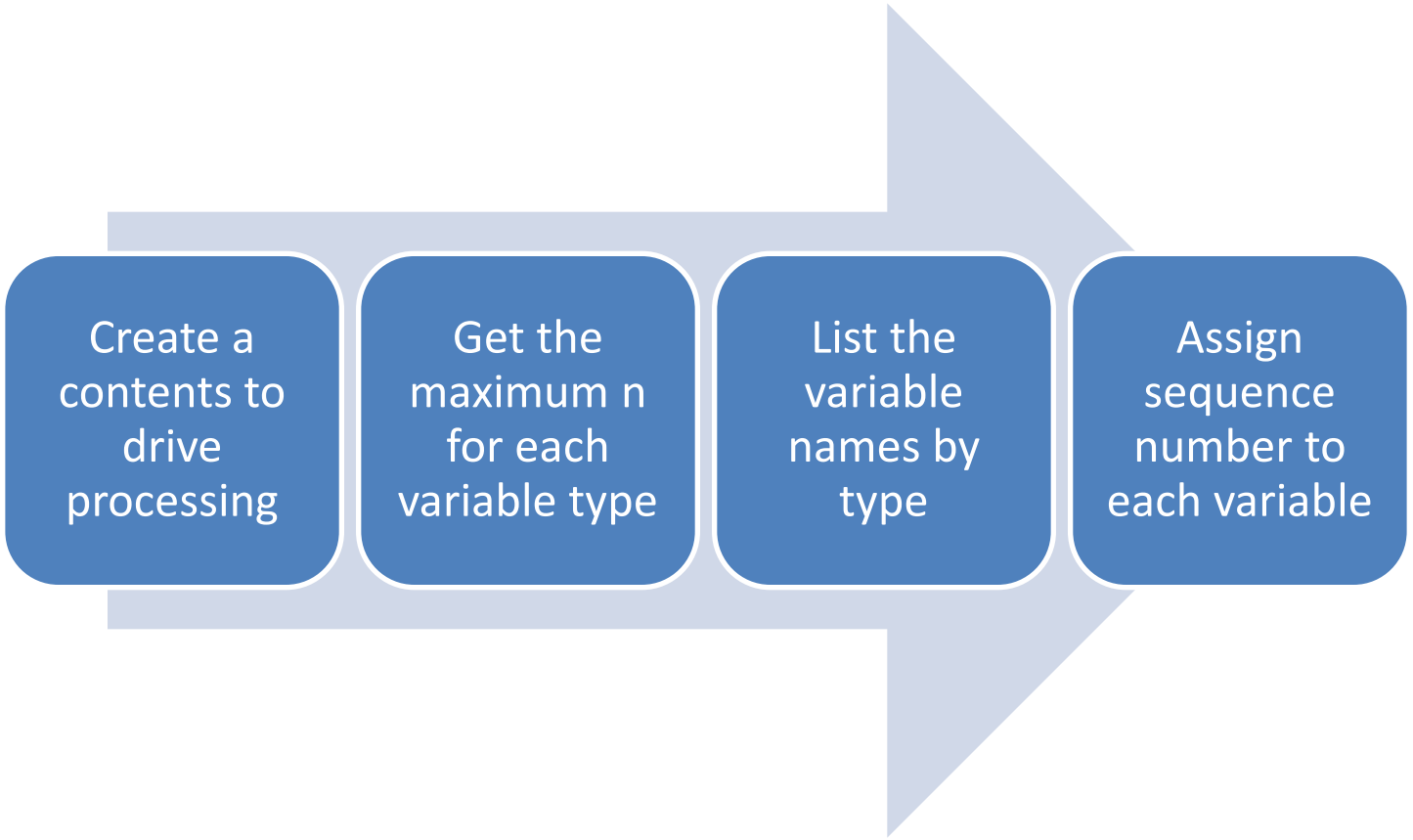
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Missingness by Site		Site 1		Site 2		Site 3		Site 4		Site 5		Site 6	
2	Missingness by Site	Variable Description	# present	# missing	# present	# missing	# present	# missing	# present	# missing	# present	# missing	# present	# miss
3	study_id	Admin: Study ID	609	0	711	0	907	0	254	0	970	0	890	0
4	site	Admin: Study site	609	0	711	0	907	0	254	0	970	0	890	0
5	study	Admin: HR study	609	0	711	0	907	0	254	0	970	0	890	0
6	der_study_part	[derived variable] Study participation status based on consent and S1 blood draw	609	0	711	0	907	0	254	0	970	0	890	0
7	enroll_status	[derived variable] Study participation status based on consent and swab collection/active surveillance	609	0	711	0	907	0	254	0	970	0	890	0
8	surv_st_week	[derived variable] MMWR year and week of the first surveillance record	520	89	658	53	872	35	249	5	898	72	832	58
9	swab_st_week	[derived variable] MMWR year and week of the first lab record	511	98	655	56	859	48	245	9	892	78	824	66
10	der_surv_st_week	[derived variable] First of either swab date or completed any survey in study after enrollment	521	88	658	53	872	35	249	5	901	69	864	26
11	der_n_surv	[derived variable] Number of MMWR weeks with surveillance in the data	609	0	711	0	907	0	254	0	970	0	890	0

Formula Bar

PersonMissingness

Ready Accessibility: Good to go 100%

Step 1: Collect data dictionary info



STEP 1: PROC SQL DATA DICTIONARY

```
PROC SQL NOPRINT;
CREATE TABLE CONTS_P AS
SELECT name format = $32. length = 32
, type format = $4. length = 4
, length format= 8. length = 8
, label format = $250. length = 250
, varnum format =8. length = 8
FROM DICTIONARY.COLUMNS
WHERE libname="INLIB" and
memname=upcase("&infi.");
QUIT;
```

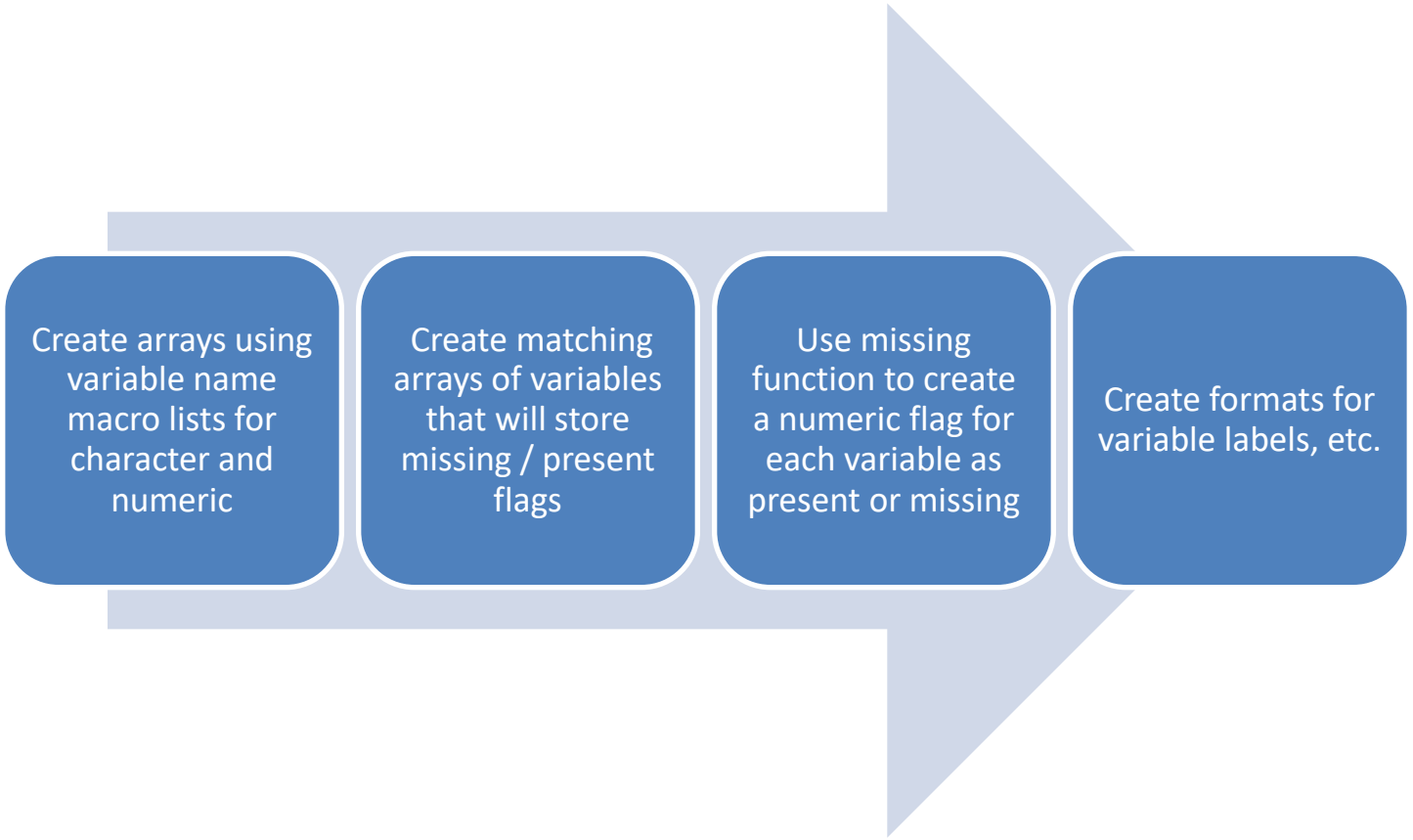
Step 1: Maximum Ns by type

```
proc freq data=conts_p;
    tables type / noprint out=max_type (keep=type count);
run;

data _null_;
    set max_type;

    if type='char' then do;
        ccount=put(count,z4.);
        CALL SYMPUTX("cmax",ccount);
    end;

    if type='num' then do;
        ncount=put(count,z4.);
        CALL SYMPUTX("nmax",ncount);
    end;

 run;
```

STEP 1: CREATE MACRO LISTS OF NAMES

```
proc sql noprint;
 select name format = $32. length = 32 into :clist
separated by ' '
 from conts_p
 where type='char';
quit;

proc sql noprint;
 select name format = $32. length = 32 into :nlist
separated by ' '
 from conts_p
 where type='num';
quit;
```

Step 2: Create and Manipulate Arrays



Step 2: Create arrays of names and ids

```
data temp;
    set inlib.&infi;

    array clist (*) &clist;
    array cton (*) cvar0001 - cvar&cmax. ;

    array nlist (*) &nlist;
    array nton (*) nvar0001 - nvar&nmax. ;

    tempc=.;
    tempn=.;
```

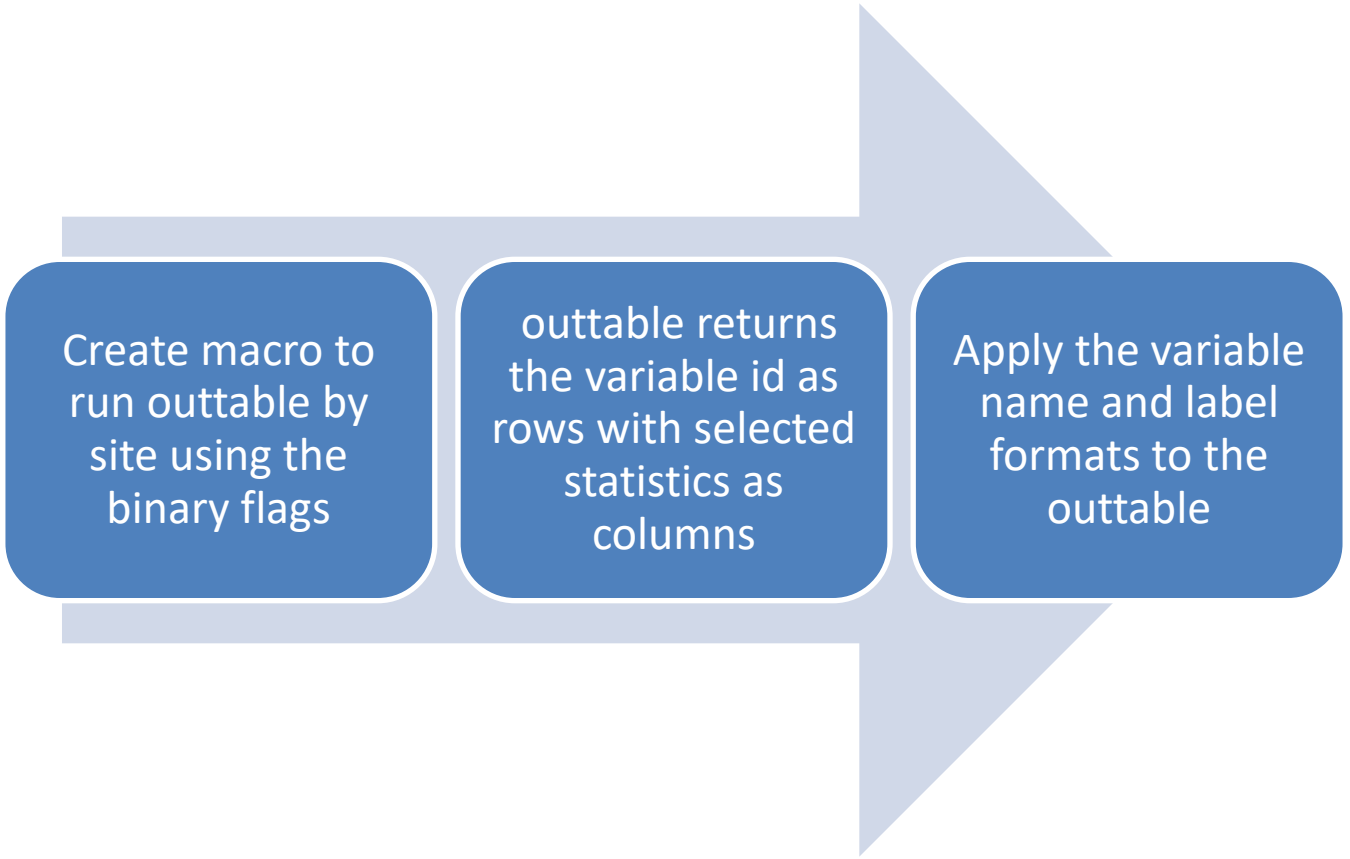
Step 2: Use missing function

```
    do i=1 to dim(clist);
       tempc=missing(clist(i));
       if tempc=0 then cton(i)=1;
       else if tempc=1 then cton(i)=.;
    end;

    do j=1 to dim(nlist);
       tempn=missing(nlist(j));
       if tempn=0 then nton(j)=1;
       else if tempn=1 then nton(j)=.;
    end;

    drop i j;
run;
```

STEP 2: CREATE FORMATS FROM CONTENTS

```
data labxwalk (keep=fmtname type start label);
    length start $ 8 vlabel label $ 250;
    set xwalk (keep=name type cnum label varnum
        rename=(type=vtype label=vlabel));
    fmtname = "labxwalk";
    type='c';
    start = cats(vtype,'var',put(cnum,z4.));
    label = vlabel;
run;

proc format library=work cntlin=labxwalk fmtlib;
run;
```

STEP 3: PROC UNIVARIATE OUTTABLE



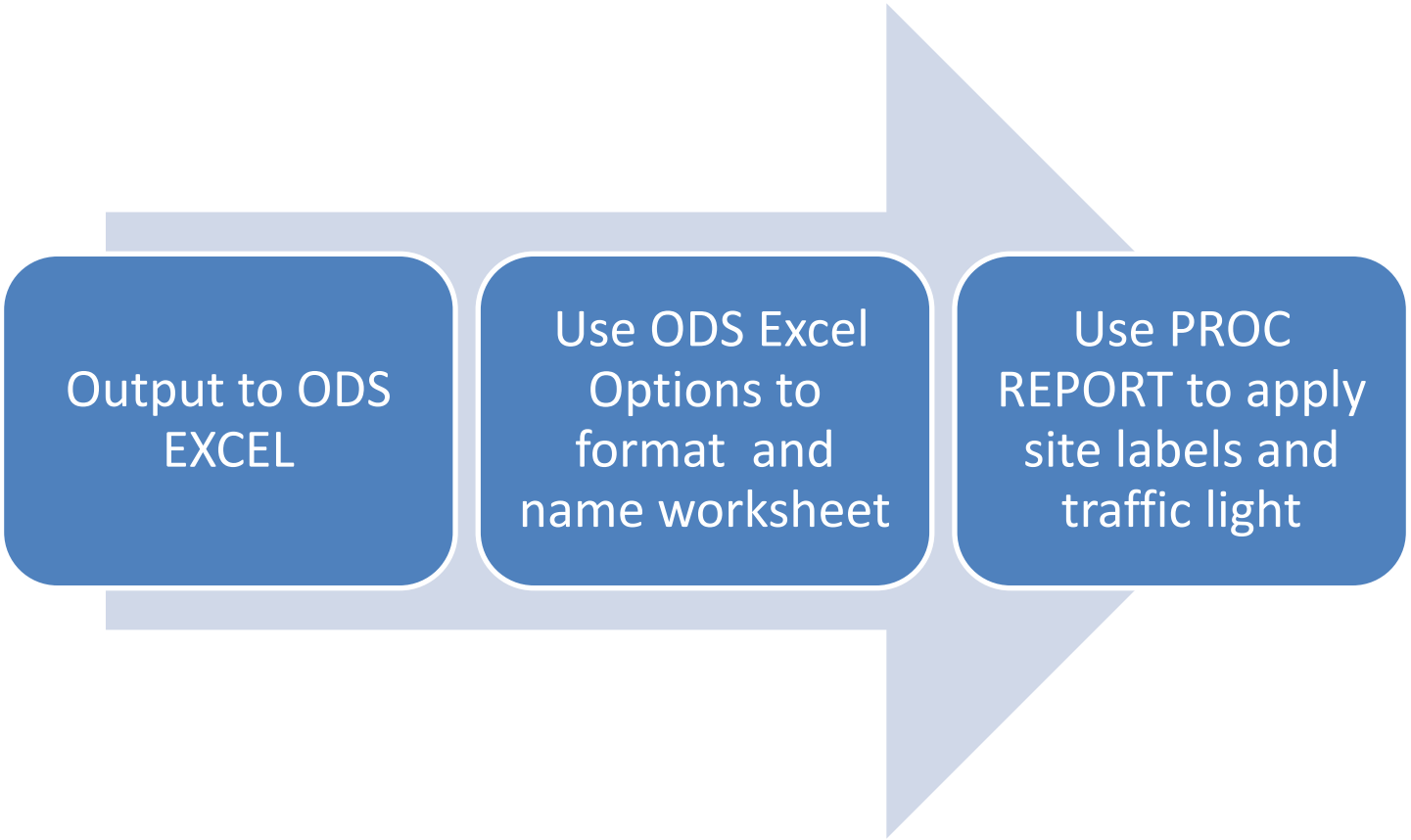
STEP 3: UNIVARIATE OUTTABLE

```
*** macro to process by site using univariate outtable;

%macro outtable(site=1);

proc univariate data=temp (where=(site=&site))
    outtable=TempTable_Site&site (keep=_var_ _nobs_ _nmiss_
    rename=(_var_=varname _nobs_=n&site _nmiss_=nmiss&site))
noprint;
    var cvar: nvar: ;
run;

proc sort data=temptable_site&site;
    by varname;
run;

%mend;

%outtable(site=1); …
```

STEP 3: COMBINE UNIVARIATE OUTPUTS

```
*** combine site level runs together with two columns per site
***;

data TempTable_AllSites;
     length name $ 32 varlabel $ 250;
     merge temptable_site: ;
     by varname;

     varname=trim(varname);
     varnum=input(put(varname,$ordxwalk.),8.);
     varlabel=put(varname,$labxwalk.);
     name=put(varname,$namexwalk.);
     format n1-n6 nmiss: comma7.0 name $32.;
     /* relabel */

     label name='Variable Name'
           varlabel='Variable Description'
           n1='# present'
           . . .
           nmiss1='# missing'
           . . . ;

run;
```

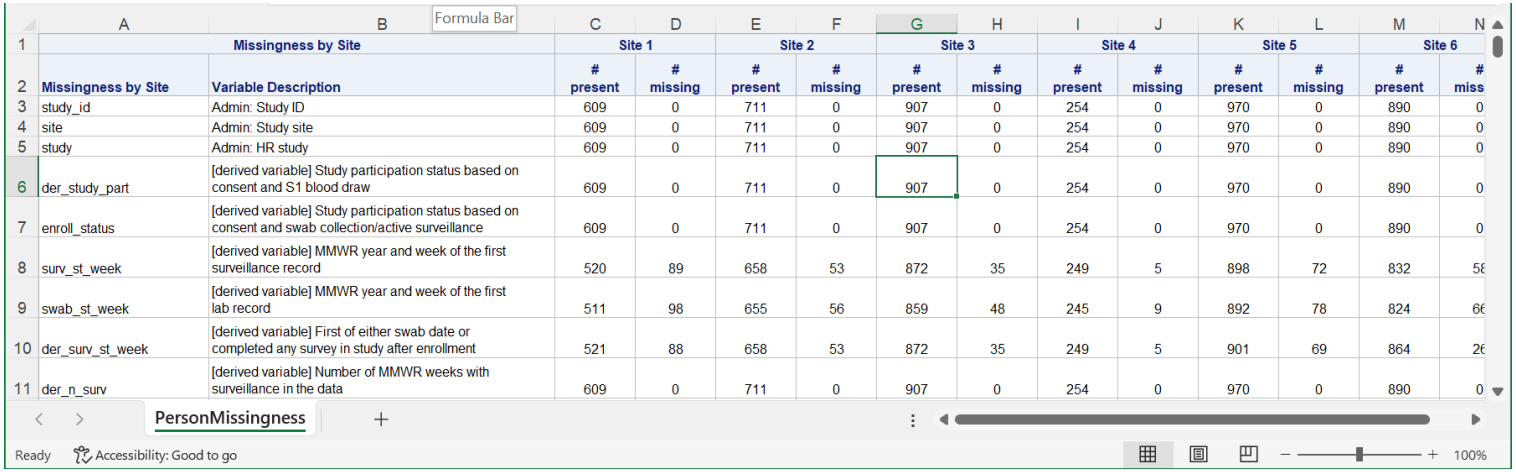
Step 4: Report on combined outtable



STEP 4: OUTPUT USING ODS EXCEL

```
ods excel file="&outfolder.\Person_Missing.xlsx" style=styles.excel
    options(sheet_interval="none" embedded_titles="yes"
            sheet_name="PersonMissingness");

proc report nowd data=temptable_allsites  . . .;
    columns ("^{style [just=l font_weight=bold font_size=8pt
              background=ligr]Missingness by Site}" name varlabel)
              . . .);
    define name / style(COLUMN)={just=l font_face="Helvetica"
              font_size=8pt cellwidth=250 } style(HEADER)={just=l
              font_face="Helvetica" font_size=8pt  };  . . .
run;

ods excel close;
```

WHEN LESS IS MORE



Missingness by Site		Site 1		Site 2		Site 3		Site 4		Site 5		Site 6	
Missingness by Site	Variable Description	# present	# missing	# present	# missing	# present	# missing	# present	# missing	# present	# missing	# present	# miss
study_id	Admin: Study ID	609	0	711	0	907	0	254	0	970	0	890	0
site	Admin: Study site	609	0	711	0	907	0	254	0	970	0	890	0
study	Admin: HR study	609	0	711	0	907	0	254	0	970	0	890	0
der_study_part	[derived variable] Study participation status based on consent and S1 blood draw	609	0	711	0	907	0	254	0	970	0	890	0
enroll_status	[derived variable] Study participation status based on consent and swab collection/active surveillance	609	0	711	0	907	0	254	0	970	0	890	0
surv_st_week	[derived variable] MMWR year and week of the first surveillance record	520	89	658	53	872	35	249	5	898	72	832	58
swab_st_week	[derived variable] MMWR year and week of the first lab record	511	98	655	56	859	48	245	9	892	78	824	66
der_surv_st_week	[derived variable] First of either swab date or completed any survey in study after enrollment	521	88	658	53	872	35	249	5	901	69	864	26
der_n_surv	[derived variable] Number of MMWR weeks with surveillance in the data	609	0	711	0	907	0	254	0	970	0	890	0

CONCLUSION

PROC FREQ with the NLEVELS option can provide an excellent broad stroke report on missingness in your data sets. PROC REPORT can generate a traffic-lighted report based on the number of total levels on each individual variable in the data set. Macro coding, the missing function, PROC UNIVARIATE outtable, PROC REPORT, and ODS EXCEL can produce a simple but effective table of presence and absence for an unlimited number of variables without user intervention. We have found the Missing(ness) Piece!

Contact Information

Louise S. Hadden
SASLouiseHadden@gmail.com