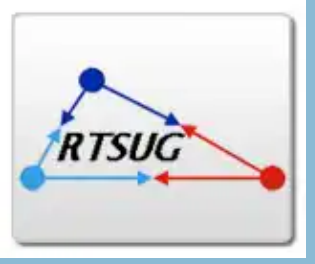
# Applications of PROC COMPARE to parallel programming and other projects

Jay Iyengar is director of Data Systems Consultants LLC.
He is a SAS consultant, trainer, and SAS Certified Advanced Programmer.
He was co-leader and organizer of the Chicago SAS Users Group (WCSUG) from 2015-19.

He's presented papers and training seminars at SAS Global Forum (SGF), and other regional and local SAS users groups. His main industry experience is in Health care, Public Health and Pharmaceutical. He obtained his Bachelor's in Public Policy and Economics from Syracuse University and his master's degree from the American University.




# Applications of PROC COMPARE to Parallel Programming and other projects

Jay Iyengar, Data Systems Consultants LLC

Research Triangle SAS Users Group
July 17, 2025

# INTRODUCTION

PROC COMPARE is an data validation tool with extensive capabilities.

PROC COMPARE compares  the data values and variable attributes of two SAS data sets; The BASE= and COMPARE= data sets.

It compares the data portion as well as the descriptor portion of SAS data sets.

By Default, PROC COMPARE will compare every matching variable and observation in both data sets.

# The Basics of PROC COMPARE

PROC COMPARE compares observations by position.

- 1st record in BASE data set with 1st record in COMPARE data set
- 2nd record in BASE data set with 2nd record in COMPARE data set

## Base Data Set

| Obs | COUNTRY | STATE | COUNTY | ACTUAL | PREDICT | PRODTYPE | PRODUCT | YEAR | QUARTER |
|---|---|---|---|---|---|---|---|---|---|
| 1 | U.S.A. | California | | $987.36 | $692.24 | FURNITURE | SOFA | 1995 | 1 |
| 2 | U.S.A. | California | | $1,782.96 | $568.48 | FURNITURE | SOFA | 1995 | 1 |
| 3 | U.S.A. | California | | $32.64 | $16.32 | FURNITURE | SOFA | 1995 | 1 |
| 4 | U.S.A. | California | | $1,825.12 | $756.16 | FURNITURE | SOFA | 1995 | 2 |
| 5 | U.S.A. | California | | $750.72 | $723.52 | FURNITURE | SOFA | 1995 | 2 |
| 6 | U.S.A. | California | | $2,426.24 | $2,428.96 | FURNITURE | SOFA | 1995 | 2 |
| 7 | U.S.A. | California | | $1,791.12 | $2,250.80 | FURNITURE | SOFA | 1995 | 3 |
| 8 | U.S.A. | California | | $2,282.08 | $350.88 | FURNITURE | SOFA | 1995 | 3 |
| 9 | U.S.A. | California | | $2,518.72 | $1,736.72 | FURNITURE | SOFA | 1995 | 3 |
| 10 | U.S.A. | California | | $1,436.16 | $2,167.84 | FURNITURE | SOFA | 1995 | 4 |
| 11 | U.S.A. | California | | $2,314.72 | $62.56 | FURNITURE | SOFA | 1995 | 4 |
| 12 | U.S.A. | California | | $1,410.32 | $1,670.08 | FURNITURE | SOFA | 1995 | 4 |
| 13 | U.S.A. | California | | $369.92 | $1,365.44 | FURNITURE | BED | 1995 | 1 |
| 14 | U.S.A. | California | | $2,014.16 | $2,358.24 | FURNITURE | BED | 1995 | 1 |
| 15 | U.S.A. | California | | $85.68 | $2,594.88 | FURNITURE | BED | 1995 | 1 |
| 16 | U.S.A. | California | | $2,694.16 | $1,403.52 | FURNITURE | BED | 1995 | 2 |
| 17 | U.S.A. | California | | $2,014.16 | $707.20 | FURNITURE | BED | 1995 | 2 |
| 18 | U.S.A. | California | | $1,500.08 | $2,461.60 | FURNITURE | BED | 1995 | 2 |
| 19 | U.S.A. | California | | $2,133.84 | $2,486.08 | FURNITURE | BED | 1995 | 3 |
| 20 | U.S.A. | California | | $1,834.64 | $2,884.56 | FURNITURE | BED | 1995 | 3 |

## Compare Data Set

| Obs | COUNTRY | STATE | COUNTY | ACTUAL | PREDICT | PRODTYPE | PRODUCT | YEAR | QUARTER |
|---|---|---|---|---|---|---|---|---|---|
| 1 | U.S.A. | California | | $726.00 | $509.00 | FURNITURE | SOFA | 1997 | 1 |
| 2 | U.S.A. | California | | $1,311.00 | $418.00 | FURNITURE | SOFA | 1997 | 1 |
| 3 | U.S.A. | California | | $24.00 | $12.00 | FURNITURE | SOFA | 1997 | 1 |
| 4 | U.S.A. | California | | $1,342.00 | $556.00 | FURNITURE | SOFA | 1997 | 2 |
| 5 | U.S.A. | California | | $552.00 | $532.00 | FURNITURE | SOFA | 1997 | 2 |
| 6 | U.S.A. | California | | $1,784.00 | $1,786.00 | FURNITURE | SOFA | 1997 | 2 |
| 7 | U.S.A. | California | | $1,317.00 | $1,655.00 | FURNITURE | SOFA | 1997 | 3 |
| 8 | U.S.A. | California | | $1,678.00 | $258.00 | FURNITURE | SOFA | 1997 | 3 |
| 9 | U.S.A. | California | | $1,852.00 | $1,277.00 | FURNITURE | SOFA | 1997 | 3 |
| 10 | U.S.A. | California | | $1,056.00 | $1,594.00 | FURNITURE | SOFA | 1997 | 4 |
| 11 | U.S.A. | California | | $1,702.00 | $46.00 | FURNITURE | SOFA | 1997 | 4 |
| 12 | U.S.A. | California | | $1,037.00 | $1,228.00 | FURNITURE | SOFA | 1997 | 4 |
| 13 | U.S.A. | California | | $1,923.90 | $222.20 | FURNITURE | SOFA | 1998 | 1 |
| 14 | U.S.A. | California | | $975.70 | $990.00 | FURNITURE | SOFA | 1998 | 1 |
| 15 | U.S.A. | California | | $973.50 | $949.30 | FURNITURE | SOFA | 1998 | 1 |
| 16 | U.S.A. | California | | $566.50 | $902.00 | FURNITURE | SOFA | 1998 | 2 |
| 17 | U.S.A. | California | | $1,456.40 | $1,710.50 | FURNITURE | SOFA | 1998 | 2 |
| 18 | U.S.A. | California | | $848.10 | $214.50 | FURNITURE | SOFA | 1998 | 2 |
| 19 | U.S.A. | California | | $1,203.40 | $69.30 | FURNITURE | SOFA | 1998 | 3 |
| 20 | U.S.A. | California | | $1,933.80 | $674.30 | FURNITURE | SOFA | 1998 | 3 |

# PROC COMPARE – The Basic Example

## Comparing All Variables In Different Data Sets

```
Proc Compare Base=SASHELP.PRDSAL2  Compare=SASHELP.PRDSAL3;
Run;
```

## Data Set Summary / Variable Summary

The COMPARE Procedure
Comparison of SASHELP.PRDSAL2 with SASHELP.PRDSAL3
(Method=EXACT)

Data Set Summary

| Dataset | Created | Modified | NVar | NObs | Label |
|---|---|---|---|---|---|
| SASHELP.PRDSAL2 | 05AUG20:21:05:49 | 05AUG20:21:05:49 | 11 | 23040 | Furniture sales data |
| SASHELP.PRDSAL3 | 05AUG20:21:05:48 | 05AUG20:21:05:48 | 11 | 11520 | Furniture sales data |

Variables Summary

Number of Variables in Common: 10.
Number of Variables in SASHELP.PRDSAL2 but not in SASHELP.PRDSAL3: 1.
Number of Variables in SASHELP.PRDSAL3 but not in SASHELP.PRDSAL2: 1.

## Observation Summary

Observation Summary

| Observation | Base | Compare |
|---|---|---|
| First Obs | 1 | 1 |
| First Unequal | 1 | 1 |
| Last Unequal | 11520 | 11520 |
| Last Match | 11520 | 11520 |
| Last Obs | 23040 | . |

Number of Observations in Common: 11520.
Number of Observations in SASHELP.PRDSAL2 but not in SASHELP.PRDSAL3: 11520.
Total Number of Observations Read from SASHELP.PRDSAL2: 23040.
Total Number of Observations Read from SASHELP.PRDSAL3: 11520.

Number of Observations with Some Compared Variables Unequal: 11520.
Number of Observations with All Compared Variables Equal: 0.

# PROC COMPARE OUTPUT

## Values Comparison Summary

Values Comparison Summary

Number of Variables Compared with All Observations Equal: 1.
Number of Variables Compared with Some Observations Unequal: 9.
Number of Variables with Missing Value Differences: 1.
Total Number of Values which Compare Unequal: 73727.
Maximum Difference: 3450.6.

The COMPARE Procedure
Comparison of SASHELP.PRDSAL2 with SASHELP.PRDSAL3
(Method=EXACT)

Variables with Unequal Values

| Variable | Type | Len | Label | Ndif | MaxDif | MissDif |
|---|---|---|---|---|---|---|
| COUNTRY | CHAR | 10 | Country | 4608 | | 0 |
| STATE | CHAR | 22 | State/Province | 9792 | | 0 |
| COUNTY | CHAR | 20 | County | 7488 | | 6336 |
| ACTUAL | NUM | 8 | Actual Sales | 11520 | 3417 | 0 |
| PREDICT | NUM | 8 | Predicted Sales | 11519 | 3451 | 0 |
| PRODTYPE | CHAR | 10 | Product Type | 5760 | | 0 |
| PRODUCT | CHAR | 10 | Product | 8640 | | 0 |
| YEAR | NUM | 8 | Year | 8640 | 3.000 | 0 |
| MONTH | NUM | 8 | Month | 5760 | 365 | 0 |

## Values Comparison Results For Variables

Value Comparison Results for Variables

| Obs | Clean STATENAME for geocoding<br>Base Value<br>StateName1 | <br>Compare Value<br>STATENAME2 |
|---|---|---|
| 1 | NEW YORK | NEWYORK |
| 2 | NEW YORK | NEWYORK |
| 3 | PUERTO RICO | PUERTORICO |
| 4 | PUERTO RICO | PUERTORICO |
| 5 | PUERTO RICO | PUERTORICO |
| 6 | PUERTO RICO | PUERTORICO |
| 7 | PUERTO RICO | PUERTORICO |
| 8 | PUERTO RICO | PUERTORICO |
| 9 | PUERTO RICO | PUERTORICO |
| 10 | PUERTO RICO | PUERTORICO |
| 11 | PUERTO RICO | PUERTORICO |
| 12 | PUERTO RICO | PUERTORICO |
| 13 | PUERTO RICO | PUERTORICO |
| 14 | PUERTO RICO | PUERTORICO |
| 15 | PUERTO RICO | PUERTORICO |
| 16 | PUERTO RICO | PUERTORICO |
| 17 | PUERTO RICO | PUERTORICO |
| 18 | PUERTO RICO | PUERTORICO |
| 19 | PUERTO RICO | PUERTORICO |
| 20 | PUERTO RICO | PUERTORICO |
| 21 | PUERTO RICO | PUERTORICO |
| 22 | PUERTO RICO | PUERTORICO |

# PROC COMPARE CAPABILITIES

**Comparing Specific Variables In Data Sets**

```
Proc Compare Base = HLTHDAT.ip2010claim Compare=HLTHDAT.ip2010line;
        Var CLM_ID;
        With CLM_ID;
Run;
```

**Different Variables Within The Same Data Set**

```
Proc Compare Base=ZIPCODE;
        Var StateName1;
        With StateName2;
 Run;
```

# PROC COMPARE Options

| Option | Description |
|---|---|
| | |
| ALLOBS | includes the values for all matching observations. |
| ALLSTATS | prints a table of summary statistics for all pairs of matching variables. |
| ALLVARS | includes in the report the values and differences for all matching variables. |
| BRIEFSUMMARY | prints only a short comparison summary. |
| MAXPRINT= | restricts the number of differences to be printed. |
| NOSUMMARY | suppresses the data set, variable, observation, and values comparison summary reports. |
| NOVALUES | suppresses the report of the value comparison results. |
| PRINTALL | invokes the following options: ALLVARS, ALLOBS, ALLSTATS, LISTALL, and WARNING. |
| STATS | prints a table of summary statistics for all pairs of matching numeric variables that are judged unequal. |
| TRANSPOSE | prints the reports of value differences by observation instead of by variable. |
| LISTALL | lists all variables and observations that are found in only one data set. |
| LISTBASE | lists all observations and variables that are found in the base data set but not in the comparison data set. |
| LISTBASEOBS | lists all observations that are found in the base data set but not in the comparison data set. |
| LISTBASEVAR | lists all variables that are found in the base data set but not in the comparison data set. |
| LISTCOMP | lists all observations and variables that are found in the comparison data set but not in the base data set. |
| LISTCOMPOBS | lists all observations that are found in the comparison data set but not in the base data set. |
| LISTCOMPVAR | lists all variables that are found in the comparison data set but not in the base data set. |
| LISTEQUALVAR | prints a list of variables whose values are judged equal at all observations in addition to the default list of variables whose values are judged unequal. |
| LISTOBS | lists all observations that are found in only one data set. |
| LISTVAR | lists all variables that are found in only one data set. |
| METHOD=ABSOLUTE \| EXACT \| PERCENT \| RELATIVE<(δ)> | specifies the method for judging the equality of numeric values. |
| OUT=*SAS-data-set* | names the output data set which contains differences between matching variables. |
| OUTALL | writes an observation for each observation in the BASE= and COMPARE= data sets. |
| OUTBASE | writes an observation to the output data set for each observation in the base data set, creating observations in which _TYPE_=BASE. |
| OUTCOMP | writes an observation to the output data set for each observation in the comparison data set, creating observations in which _TYPE_=COMP. |
| OUTDIF | writes an observation to the output data set for each pair of matching observations. |
| OUTNOEQUAL | suppresses the writing of an observation to the output data set when all values in the observation are judged equal. |
| OUTPERCENT | writes an observation to the output data set for each pair of matching observations. |
| OUTSTATS=*SAS-data-set* | writes summary statistics for all pairs of matching variables to the specified *SAS-data-set*. |

**OPTIONS WHICH LIMIT SECTIONS OF OUTPUT**

**The NOSUMMARY Option**

```
Proc Compare Base=PRDSAL2  Compare=PRDSAL3 NOSUMMARY;
Run;
```

**The BRIEFSUMMARY Option**

```
Proc Compare Base=PRDSAL2 Compare=PRDSAL3 BRIEFSUMMARY;
Run;
```

# PROC COMPARE Options

**The MAXPRINT Option**

- Sets the number of data value differences which appear in the values comparison results for variables section

```
Proc Compare Base=PRDSAL2 Compare=PRDSAL3 MAXPRINT=25;
Run;
```

**The NOVALUES Option**

- Suppresses the Values Comparison Report.

```
Proc Compare Base=PRDSAL2 Compare = PRDSAL3 NOVALUES;
Run;
```

# PROC COMPARE Options

**OPTIONS WHICH RESTRICT OBSERVATIONS AND VARIABLES DISPLAYED**

| OPTION | OUTPUT |
|---|---|
| **LISTBASE** | VARIABLES AND OBSERVATIONS FOUND ONLY IN THE BASE DATA SET |
| **LISTBASEVARS** | VARIABLES FOUND IN THE BASE DATA SET ONLY |
| **LISTBASEOBS** | OBSERVATIONS FOUND IN THE BASE DATA SET ONLY |
| **LISTCOMP** | VARIABLES AND OBSERVATIONS FOUND ONLY IN THE COMPARE DATA SET |
| **LISTCOMPVARS** | VARIABLES IN THE COMPARE DATA SET ONLY |
| **LISTCOMPOBS** | OBSERVATIONS IN THE COMPARE DATA SET ONLY |

# PARALLEL PROGRAMMING

Each programmer writes code independently from scratch using specifications.

Frequently used in Clinical Trials Programming / Observational Research

Data sets need to match in all aspects; Variable, Observations, Data Values, Attributes

Variable Attributes:  Variable Type, Length, Format, Label

```
104         Proc Compare Base=Prdsal2a Compare=Prdsal2b;
105         Run;

NOTE: There were 23040 observations read from the data set WORK.PRDSAL2A.
NOTE: There were 23040 observations read from the data set WORK.PRDSAL2B.
NOTE: PROCEDURE COMPARE used (Total process time):
      real time           0.10 seconds
      user cpu time       0.10 seconds
      system cpu time     0.01 seconds
      memory              5542.59k
      OS Memory           35760.00k
```

# PARALLEL PROGRAMMING

Observation discrepancies might point to data sets being incorrectly merged, or subset.

The COMPARE Procedure
Comparison of WORK.PRDSAL2A with WORK.PRDSAL2B
(Method=EXACT)

Value Comparison Results for Variables

| Obs | State/Province Base Value STATE | Compare Value STATE |
|---|---|---|
| 9217 | New York | NewYork |
| 9218 | New York | NewYork |
| 9219 | New York | NewYork |
| 9220 | New York | NewYork |
| 9221 | New York | NewYork |
| 9222 | New York | NewYork |
| 9223 | New York | NewYork |
| 9224 | New York | NewYork |
| 9225 | New York | NewYork |
| 9226 | New York | NewYork |
| 9227 | New York | NewYork |
| 9228 | New York | NewYork |
| 9229 | New York | NewYork |
| 9230 | New York | NewYork |
| 9231 | New York | NewYork |
| 9232 | New York | NewYork |
| 9233 | New York | NewYork |
| 9234 | New York | NewYork |
| 9235 | New York | NewYork |
| 9236 | New York | NewYork |
| 9237 | New York | NewYork |
| 9238 | New York | NewYork |
| 9239 | New York | NewYork |
| 9240 | New York | NewYork |
| 9241 | New York | NewYork |
| 9242 | New York | NewYork |
| 9243 | New York | NewYork |
| 9244 | New York | NewYork |
| 9245 | New York | NewYork |
| 9246 | New York | NewYork |
| 9247 | New York | NewYork |
| 9248 | New York | NewYork |
| 9249 | New York | NewYork |
| 9250 | New York | NewYork |
| 9251 | New York | NewYork |
| 9252 | New York | NewYork |
| 9253 | New York | NewYork |
| 9254 | New York | NewYork |
| 9255 | New York | NewYork |
| 9256 | New York | NewYork |
| 9257 | New York | NewYork |
| 9258 | New York | NewYork |
| 9259 | New York | NewYork |
| 9260 | New York | NewYork |

Value discrepancies might mean variable was derived or re-coded incorrectly.

The COMPARE Procedure
Comparison of WORK.PRDSAL2A with WORK.PRDSAL2B
(Method=EXACT)

Value Comparison Results for Variables

| Obs | Product Base Value PRODUCT | Compare Value PRODUCT |
|---|---|---|
| 3556 | SOFA | Coffee Tab |
| 3557 | SOFA | Coffee Tab |
| 3558 | SOFA | Coffee Tab |
| 3559 | SOFA | Coffee Tab |
| 3560 | SOFA | Coffee Tab |
| 3561 | SOFA | Coffee Tab |
| 3562 | SOFA | Coffee Tab |
| 3563 | SOFA | Coffee Tab |
| 3564 | SOFA | Coffee Tab |
| 3601 | SOFA | Coffee Tab |
| 3602 | SOFA | Coffee Tab |
| 3603 | SOFA | Coffee Tab |
| 3604 | SOFA | Coffee Tab |
| 3605 | SOFA | Coffee Tab |
| 3606 | SOFA | Coffee Tab |
| 3607 | SOFA | Coffee Tab |
| 3608 | SOFA | Coffee Tab |
| 3609 | SOFA | Coffee Tab |
| 3610 | SOFA | Coffee Tab |
| 3611 | SOFA | Coffee Tab |
| 3612 | SOFA | Coffee Tab |
| 3649 | SOFA | Coffee Tab |
| 3650 | SOFA | Coffee Tab |

NOTE: The MAXPRINT=50 printing limit has been reached for the variable PRODUCT. No more values will be printed for this comparison.

When there are a lot of value discrepancies, use the OUT= option to send the results to an output SAS data set.

```
Proc Compare Base=Prdsal2a
             Compare=Prdsal2b
             Out=PrdSal_Results Outnoequal OutDif Noprint;
Run;
```

# MERGING SAS DATA SETS

In a DATA STEP Match-Merge, the BY variables need to have the same name, and they should have the same attributes; Length, Type, Format, etc.

You can run PROC COMPARE to confirm that BY variables have the same attributes; Length, Type, etc.

Prior to the DATA STEP Merge, we'd like to know whether we have a one-to-one, one-to-many, or many-to-many relationship.

Using the ID statement, PROC COMPARE informs us if one or more data sets have duplicates.

# MERGING SAS DATA SETS

Using the ID statement, we can compare observations based on the values of BY variables, instead of position of observation or observation number.

```
81      Proc Sort Data = hlthcare.ip2010claim Out=ip2010claim;
82            By Bene_ID;
83      Run;

NOTE: There were 13916 observations read from the data set HLTHCARE.IP2010CLAIM.
NOTE: The data set WORK.IP2010CLAIM has 13916 observations and 36 variables.
NOTE: PROCEDURE SORT used (Total process time):
            real time       0.04 seconds
            user cpu time     0.05 seconds

85      Proc Sort Data = hlthcare.finder_attrib Out=finder_attrib;
86            By Bene_ID;
87      Run;

NOTE: There were 85872 observations read from the data set HLTHCARE.FINDER_ATTRIB.
NOTE: The data set WORK.FINDER_ATTRIB has 85872 observations and 2 variables.
NOTE: PROCEDURE SORT used (Total process time):
            real time       0.04 seconds
            user cpu time     0.04 seconds

89      Proc Compare Base=ip2010claim Compare=finder_attrib LISTBASEOBS LISTCOMPOBS;
90            ID Bene_ID;
91            Var AT_NPI;
92            With PRFNPI;
93      Run;

WARNING: The data set WORK.IP2010CLAIM contains a duplicate observation at observation number 4.
NOTE: At observation 4 the current and previous ID values are:  BENE_ID=0007F12A492FD25D.
NOTE: Further warnings for duplicate observations in this data set will not be printed.
NOTE: There were 13916 observations read from the data set WORK.IP2010CLAIM.
NOTE: There were 85872 observations read from the data set WORK.FINDER_ATTRIB.

NOTE: PROCEDURE COMPARE used (Total process time):
            real time       0.77 seconds
            user cpu time     0.75 seconds
```

```
                    The COMPARE Procedure
          Comparison of WORK.IP2010CLAIM with WORK.FINDER_ATTRIB
                          (Method=EXACT)

                        Observation Summary

                    Observation      Base  Compare  ID

          First Obs           1         1  BENE_ID=00013D2EFD8E45D1
          First Unequal       1         1  BENE_ID=00013D2EFD8E45D1
          Last  Unequal   13916     85865  BENE_ID=FFFA950301FCA748
          Last  Match     13916     85865  BENE_ID=FFFA950301FCA748
          Last  Obs           .     85872  BENE_ID=FFFF7C107A4E385A

                  Number of Observations in Common: 11347.
Number of Observations in WORK.IP2010CLAIM but not in WORK.FINDER_ATTRIB: 2569.
Number of Observations in WORK.FINDER_ATTRIB but not in WORK.IP2010CLAIM: 74525.
Number of Duplicate Observations found in WORK.IP2010CLAIM: 1947.
Total Number of Observations Read from WORK.IP2010CLAIM: 13916.
Total Number of Observations Read from WORK.FINDER_ATTRIB: 85872.

Number of Observations with Some Compared Variables Unequal: 11335.
Number of Observations with All Compared Variables Equal: 12.


                      Values Comparison Summary

Number of Variables Compared with All Observations Equal: 0.
Number of Variables Compared with Some Observations Unequal: 1.
Number of Variables with Missing Value Differences: 1.
Total Number of Values which Compare Unequal: 11335.


                  All Variables Compared have Unequal Values

Variable  Type  Len   Compare   Len   Label                                    Ndif Maxdif

AT_NPI    CHAR   10   PRFNPI     12   Attending Physician - National Provider Identifier No. 11335  9
```

# CONCATENATING SAS DATA SETS

Properly appending SAS data sets requires using data sets with the same variable attributes;
- Type, Length, Format, and Label.

Inconsistent variable attributes will impact data integrity and validity.

Use PROC COMPARE to confirm consistent variable attributes.

Best Practice, run PROC COMPARE before concatenating using PROC APPEND or DATA STEP

```
125     Proc Compare Base=US Compare=Mexico;
126     Run;

NOTE: There were 18432 observations read from the data set WORK.US.
NOTE: There were 4608 observations read from the data set WORK.MEXICO.
NOTE: PROCEDURE COMPARE used (Total process time):
   real time        0.15 seconds
   user cpu time    0.15 seconds
   system cpu time  0.01 seconds
   memory           4640.68k
   OS Memory        37696.00k
```

Appending Sales Records from The US to Sales Records in Mexico.

# CONCATENATING SAS DATA SETS

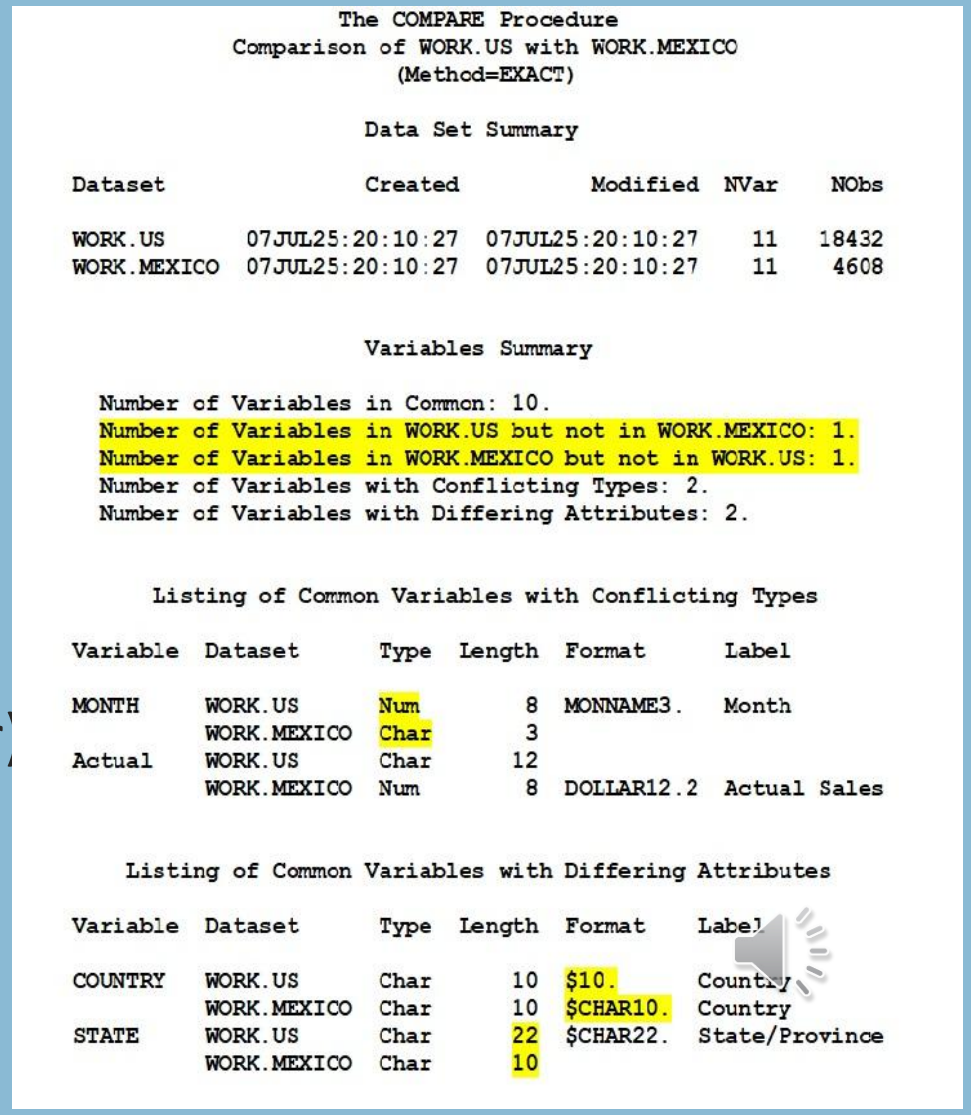
Discrepancies in observations should be expected
The data sets should have mutually exclusive records.

PROC COMPARE Output shows discrepancies to resolve.

1- Missing Variables.
- Variables present on one data set but missing from the other.

2. Variables with conflicting types (numeric vs. character)
- Might involve missing values generated

3. Variables with incongruent lengths.
- Could lead to truncation of data values.



The COMPARE Procedure
Comparison of WORK.US with WORK.MEXICO
(Method=EXACT)

Data Set Summary

| Dataset | Created | Modified | NVar | NObs |
|---|---|---|---|---|
| WORK.US | 07JUL25:20:10:27 | 07JUL25:20:10:27 | 11 | 18432 |
| WORK.MEXICO | 07JUL25:20:10:27 | 07JUL25:20:10:27 | 11 | 4608 |

Variables Summary

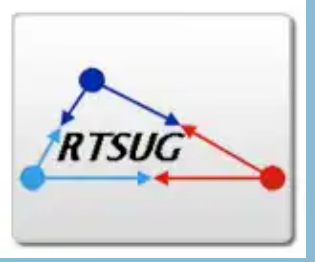
Number of Variables in Common: 10.
Number of Variables in WORK.US but not in WORK.MEXICO: 1.
Number of Variables in WORK.MEXICO but not in WORK.US: 1.
Number of Variables with Conflicting Types: 2.
Number of Variables with Differing Attributes: 2.

Listing of Common Variables with Conflicting Types

| Variable | Dataset | Type | Length | Format | Label |
|---|---|---|---|---|---|
| MONTH | WORK.US | Num | 8 | MONNAME3. | Month |
| | WORK.MEXICO | Char | 3 | | |
| Actual | WORK.US | Char | 12 | | |
| | WORK.MEXICO | Num | 8 | DOLLAR12.2 | Actual Sales |

Listing of Common Variables with Differing Attributes

| Variable | Dataset | Type | Length | Format | Label |
|---|---|---|---|---|---|
| COUNTRY | WORK.US | Char | 10 | $10. | Country |
| | WORK.MEXICO | Char | 10 | $CHAR10. | Country |
| STATE | WORK.US | Char | 22 | $CHAR22. | State/Province |
| | WORK.MEXICO | Char | 10 | | |




# RECONCILING VERSIONS OF SAS DATA SETS

A Master data set may be kept on a server and updated periodically with a transaction data set. A new version of the master data set is created.

Master data set updated on a monthly or quarterly basis.

SAS allows you to create generation data sets which are updated versions of master data sets.

With PROC COMPARE you can reconcile the versions of master data sets and document the changes and differences between them.

# RECONCILING VERSIONS OF SAS DATA SETS

Mass shootings database – Updated every Quarter

```
105         Proc Compare Base=Everytown_2023  Compare=Everytown_2025;
106         Run;

 NOTE: There were 284 observations read from the data set WORK.EVERYTOWN_2023.
 NOTE: There were 299 observations read from the data set WORK.EVERYTOWN_2025.
 NOTE: PROCEDURE COMPARE used (Total process time):
       real time           0.09 seconds
       user cpu time       0.09 seconds
       system cpu time     0.01 seconds
       memory              2063.40k
       OS Memory           32984.00k
```

The COMPARE Procedure
Comparison of WORK.EVERYTOWN_2023 with WORK.EVERYTOWN_2025
(Method=EXACT)

Data Set Summary

| Dataset | Created | Modified | NVar | NObs |
|---|---|---|---|---|
| WORK.EVERYTOWN_2023 | 08JUL25:17:17:20 | 08JUL25:17:17:20 | 14 | 284 |
| WORK.EVERYTOWN_2025 | 08JUL25:17:17:20 | 08JUL25:17:17:20 | 14 | 299 |

Variables Summary

Number of Variables in Common: 7.
Number of Variables in WORK.EVERYTOWN_2023 but not in WORK.EVERYTOWN_2025: 7.
Number of Variables in WORK.EVERYTOWN_2025 but not in WORK.EVERYTOWN_2023: 7.

Observation Summary

| Observation | Base | Compare |
|---|---|---|
| First Obs | 1 | 1 |
| First Unequal | 1 | 1 |
| Last Unequal | 229 | 229 |
| Last Match | 284 | 284 |
| Last Obs | . | 299 |

Number of Observations in Common: 284.
Number of Observations in WORK.EVERYTOWN_2025 but not in WORK.EVERYTOWN_2023: 15.
Total Number of Observations Read from WORK.EVERYTOWN_2023: 284.
Total Number of Observations Read from WORK.EVERYTOWN_2025: 299.

Number of Observations with Some Compared Variables Unequal: 16.
Number of Observations with All Compared Variables Equal: 268.

Values Comparison Summary

Number of Variables Compared with All Observations Equal: 1.
Number of Variables Compared with Some Observations Unequal: 6.
Total Number of Values which Compare Unequal: 68.
Maximum Difference: 34.429.

# RECONCILING VERSIONS OF SAS DATA SETS

Output shows differences

Understanding source of the  differences

The COMPARE Procedure
Comparison of WORK.EVERYTOWN_2023 with WORK.EVERYTOWN_2025
(Method=EXACT)

Variables with Unequal Values

| Variable | Type | Len | Label | Ndif | MaxDif |
|---|---|---|---|---|---|
| City | CHAR | 24 | City | 12 | |
| State | CHAR | 2 | State | 10 | |
| Latitude | NUM | 8 | Latitude | 12 | 13.419 |
| Longitude | NUM | 8 | Longitude | 12 | 34.429 |
| Narrative | CHAR | 764 | Narrative | 12 | |
| Last updated | CHAR | 24 | Last updated | 10 | |

Value Comparison Results for Variables

| Obs | City Base Value City | Compare Value City |
|---|---|---|
| 90 | Clarksburg | Hialeah |
| 91 | Hialeah | Clarksburg |
| 106 | Alturas | Indianapolis |
| 107 | Indianapolis | Alturas |
| 146 | Pike County | Appling |
| 147 | Appling | Pike County |
| 148 | Roswell | Orlando |
| 149 | Orlando | Roswell |
| 176 | Reading | Melcroft |
| 177 | Melcroft | Reading |
| 228 | Springfield | Moncure |
| 229 | Moncure | Springfield |

| Obs | Latitude Base Latitude | Compare Latitude | Diff. | % Diff |
|---|---|---|---|---|
| 90 | 39.2832 | 25.864 | -13.4192 | -34.1601 |
| 91 | 25.864 | 39.2832 | 13.4192 | 51.8837 |
| 106 | 41.4909 | 39.717 | -1.7739 | -4.2754 |
| 107 | 39.717 | 41.4909 | 1.7739 | 4.4663 |
| 146 | 39.0201 | 33.5989 | -5.4212 | -13.8934 |
| 147 | 33.5989 | 39.0201 | 5.4212 | 16.1351 |
| 148 | 33.3931 | 28.5195 | -4.8736 | -14.5946 |
| 149 | 28.5195 | 33.3931 | 4.8736 | 17.0887 |
| 176 | 40.3331 | 40.0521 | -0.2810 | -0.6967 |
| 177 | 40.0521 | 40.3331 | 0.2810 | 0.7016 |
| 228 | 37.210283 | 35.588419 | -1.6219 | -4.3586 |
| 229 | 35.588419 | 37.210283 | 1.6219 | 4.5573 |

# CONCLUSION

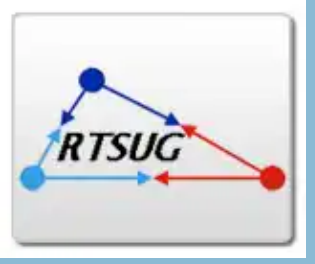
PROC COMPARE has broad applications in a variety of project tasks in SAS Programming.

Use PROC COMPARE options to control summary sections and volume of output.

Use PROC COMPARE with ID statement prior to running DATA STEP Merge

Use PROC COMPARE prior to appending data sets to check for discrepancies in attributes.

Carefully review all sections of PROC COMPARE output to properly validate your data sets.




# CONTACT INFORMATION

Thanks for Attending!

Questions?

Jay Iyengar

Data Systems Consultants LLC

Email: datasyscon@gmail.com

Linkedin: https://www.linkedin.com/in/datasysconsult/