

Q:

How can I use PROC COMPARE to produce a report that shows the differences between two character values past the 20th character? Also, can PROC COMPARE show observations that occur only in one of the data sets?

A:

Creating an output data set with the OUT= option addresses both questions. The output data set shows the full character values and shows you what observations occur in only one of the data sets.

For this example we are comparing two data sets that contain employee information:

```
options ls=80 ps=60 nodate nonumber;

data emp1;
  input @1 idnum $4. @6 name $15. @22 address $42. @65 salary 9.;
  cards;
2388 James Schmidt 100 Apt. C Blount St. SW Raleigh NC 27693 92100
2457 Fred Williams 99 West Lane Garner NC 27509 33190
2776 Robert Jones 12988 Wellington Farms Ave. Cary NC 27512 29025
8699 Jerry Capalleti 222 West L St. Oxford NC 27587 39985
2100 Lanny Engles 293 Manning Pl. Raleigh NC 27606 30998
9857 Kathy Krupski 1000 Taft Ave. Morrisville NC 27508 38756
0987 Dolly Lunford 2344 Persimmons Branch Apex NC 27505 44010
3286 Hoa Nguyen 2818 Long St. Cary NC 27513 87734
6579 Bryan Samosky 3887 Charles Ave. Garner NC 27508 50234
3888 Kim Siu 5662 Magnolia Blvd Southeast Cary NC 27513 77558
;

data emp2;
  input @1 idnum $4. @6 name $15. @22 address $42. @65 salary 9.;
  cards;
2388 James Schmidt 100 Apt. C Blount St. SW Raleigh NC 27693 92100
2457 Fred Williams 99 West Lane Garner NC 27509 33190
2776 Robert Jones 12988 Wellington Farms Ave. Cary NC 27511 29025
8699 Jerry Capalleti 222 West L St. Oxford NC 27587 39985
3278 Mary Cravens 211 N. Cypress St. Cary NC 27512 35362
2100 Lanny Engles 293 Manning Pl. Raleigh NC 27606 30998
9857 Kathy Krupski 100 Taft Ave. Morrisville NC 27508 40456
0987 Dolly Lunford 2344 Persimmons Branch Trail Apex NC 27505 45110
3286 Hoa Nguyen 2818 Long St. Cary NC 27513 89834
6579 Bryan Samosky 3887 Charles Ave. Garner NC 27508 50234
3888 Kim Siu 5662 Magnolia Blvd Southwest Cary NC 27513 79958
;
```

Since the ID number is unique, we use IDNUM as an ID variable. Before we do a comparison, we must sort both data sets by the ID variable, IDNUM:

```
/* Sort both data sets by the ID variable. */
proc sort data=emp1;
  by idnum;
```

```
run;

proc sort data=emp2;
  by idnum;
run;
```

First, let's look at a PROC COMPARE step that uses the BRIEF option to show only a value comparison for matching variables. PROC COMPARE compares observations based on the value of IDNUM.

```
proc compare base=emp1 comp=emp2 brief;
  id idnum;
run;
```

Output 1

The SAS System					
COMPARE Procedure					
Comparison of WORK.EMP1 with WORK.EMP2					
(Method=EXACT)					
NOTE: Data set WORK.EMP2 contains 1 observations not in WORK.EMP1.					
NOTE: Values of the following 2 variables compare unequal: ADDRESS SALARY					
Value Comparison Results for Variables					
IDNUM	Base Value ADDRESS	Compare Value ADDRESS			
			+	+	
0987	2344 Persimmons Bran	2344 Persimmons Bran			
2776	12988 Wellington Far	12988 Wellington Far			
3888	5662 Magnolia Blvd S	5662 Magnolia Blvd S			
9857	1000 Taft Ave. Morri	100 Taft Ave. Morris			
IDNUM	Base SALARY	Compare SALARY	Diff.	% Diff	
0987	44010	45110	1100	2.4994	
3286	87734	89834	2100	2.3936	
3888	77558	79958	2400	3.0945	
9857	38756	40456	1700	4.3864	

Output 1 Notes

PROC COMPARE prints the nonmatching addresses and nonmatching salaries. For salaries, PROC COMPARE computes the numerical difference and the percent difference. Because ADDRESS is a character variable, PROC COMPARE displays only the first 20 characters. For addresses where the observation has an IDNUM of 0987, 2776, or 3888, the differences occur after the 20th character and we can not see what the differences are. The plus sign in the output indicates that the full value is not displaying **1**.

Now let's compare the same two data sets and create an output data set:

```
options ls=132;
```

```

proc compare base=empl comp=emp2 out=result outnoequal outbase
            outcomp outdif noprint;
    id idnum;
run;

proc print data=result label noobs;
    by idnum;
    id idnum;
run;

```

Program Notes

In the program above,

- OUT= specifies the name of the output data set.
- The following options determine the contents of the output data set:
 - OUTNOEQUAL includes information about unequal observations.
 - OUTBASE provides the observations from the base data set (EMP1).
 - OUTCOMP provides the observations from the comparison data set (EMP2).
 - OUTDIF gives the difference between the two observations. OUTDIF prints
 - an X or a period (.) for character variables. An X shows that the characters do not match. A period shows that the characters do match.
 - an E or the difference for numeric variables
- NOPRINT suppresses the printing of the PROC COMPARE output.
- PROC PRINT prints the output data set. See Output 2.

Output 1.1 Output 2

The SAS System						
IDNUM	Type of Observation	Observation Number	NAME	ADDRESS		SALARY
0987	BASE	1	Dolly Lunford	2344 Persimmons Branch Apex NC 27505		44010
	COMPARE	1	Dolly Lunford	2344 Persimmons Branch Trail Apex NC 27505		45110
	DIF	1XXXXXXXXXXXXX		1100
2776	BASE	5	Robert Jones	12988 Wellington Farms Ave. Cary NC 27512		29025
	COMPARE	5	Robert Jones	12988 Wellington Farms Ave. Cary NC 27511		29025
	DIF	5X.		E
3278	COMPARE	6	Mary Cravens	211 N. Cypress St. Cary NC 27512		35362
3286	BASE	6	Hoa Nguyen	2818 Long St. Cary NC 27513		87734
	COMPARE	7	Hoa Nguyen	2818 Long St. Cary NC 27513		89834
	DIF	6		2100
3888	BASE	7	Kim Siu	5662 Magnolia Blvd Southeast Cary NC 27513		77558
	COMPARE	8	Kim Siu	5662 Magnolia Blvd Southwest Cary NC 27513		79958
	DIF	7XX.....		2400
9857	BASE	10	Kathy Krupski	1000 Taft Ave. Morrisville NC 27508		38756
	COMPARE	11	Kathy Krupski	100 Taft Ave. Morrisville NC 27508		40456
	DIF	10XXXXXXXXXXXXX.XXXXX.XXXXXXXXXXXXX		1700

Output 2 Notes

The value **DIF** indicates the observations that show the differences. ❶ In these observations,

- the values of **NAME** are periods because all names match exactly.
- the values of **ADDRESS** are periods (.) for exact matches or **X**'s for unequal characters. For example, in the address of **Dolly Lunford**, the word **Trail** appears in the comparison data set (EMP2) but not in the base data set (EMP1) ❷.
- the value of **SALARY** is either an **E** for exact matches or the numeric difference.

In addition, the output shows us that **Mary Cravens** appears only in the comparison data set (EMP2) ❸. The output data set includes observations that occur only in one of the data sets by default. You do not have to use an option to make those observations appear in the output data set.