Match-Merging

So far in this section, we've combined data sets based on the order of the observations in the input data sets. But sometimes you need to combine observations from two or more data sets into a single observation in a new data set according to the values of a common variable. This is called match-merging.

When you match-merge, you use a MERGE statement rather than a SET statement to combine data sets.

General form, basic DATA step for match-merging:

```
DATA output-SAS-data-set;
    MERGE SAS-data-set-1 SAS-data-set-2;
    BY <DESCENDING> variable(s);
RUN;
```

where

¤ output-SAS-data-set names the data set to be created.

¤ SAS-data-set-1 and SAS-data-set-2 specify the data sets to be read.

¤ variable(s) in the BY statement specifies one or more variables whose values are used to match observations.

¤ DESCENDING indicates that the input data sets are sorted in descending order (largest to smallest numerically, or reverse alphabetical for character variables) by the variable that is specified. If you have more than one variable in the BY statement, DESCENDING applies only to the variable that immediately follows it.

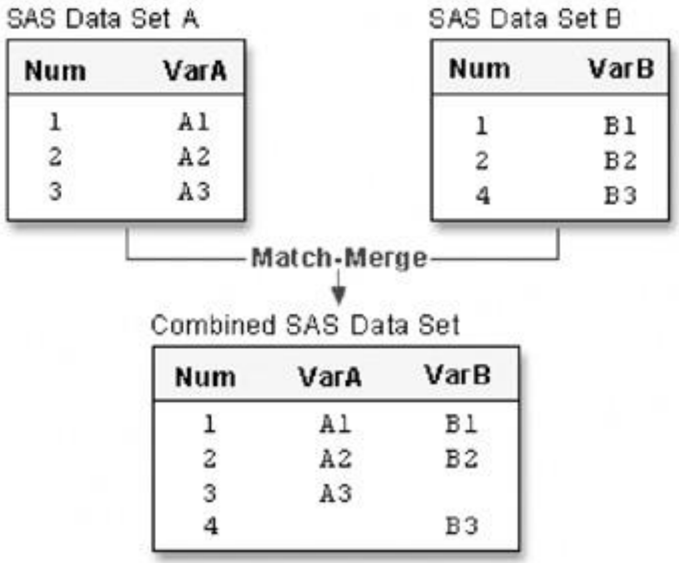
Additional Note: Each input data set in the MERGE statement must be sorted in order of the values of the BY variable(s), or it must have an appropriate index. Each BY variable must have the same type in all data sets to be merged.

Additional Note: You cannot use the DESCENDING option with indexed data sets because indexes are always stored in ascending order.

How Match-Merging Selects Data

Generally speaking, during match-merging, SAS sequentially checks each observation of each data set to see whether the BY values match, and then writes the combined observation to the new data set.

```
data merged;
   merge a b;
   by num;
run;
```



SAS Data Set A

Num	VarA
1	A1
2	A2
3	A3

SAS Data Set B

Num	VarB
1	B1
2	B2
4	B3

Combined SAS Data Set

Num	VarA	VarB
1	A1	B1
2	A2	B2
3	A3	
4		B3

Basic DATA step match-merging produces an output data set that contains values from all observations in all input data sets. You can add statements and options to select only matching observations.

If an input data set doesn't have any observations for a particular value of the by-variable, then the observation in the output data set contains missing values for the variables that are unique to that input data set.

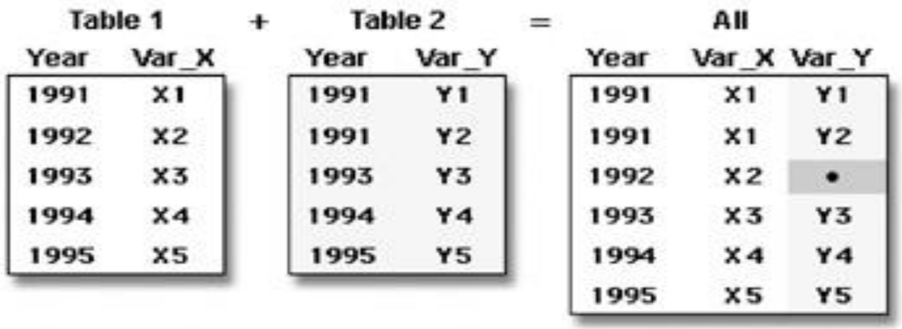


Table 1

Year	Var_X
1991	X1
1992	X2
1993	X3
1994	X4
1995	X5

+

Table 2

Year	Var_Y
1991	Y1
1991	Y2
1993	Y3
1994	Y4
1995	Y5

=

All

Year	Var_X	Var_Y
1991	X1	Y1
1991	X1	Y2
1992	X2	•
1993	X3	Y3
1994	X4	Y4
1995	X5	Y5

Additional Note: In match-merging, often one data set contains unique values for the BY-variable and other data sets contain multiple values for the BY-variable.