How do you sort data using SAS tutorials?

Authored by

stats writer

June 24, 2024

The How do you sort data using SAS tutorials? is a tutorial that provides step-by-step instructions on how to effectively sort data using the SAS software. SAS, or Statistical Analysis System, is a powerful tool used for data analysis and management. This tutorial aims to guide users, regardless of their level of experience with SAS, on how to properly sort data in an efficient and accurate manner. It covers the basics of sorting data, including understanding the sorting process, identifying the variables to be sorted, and using the appropriate commands and functions within SAS. By following this tutorial, users will be able to enhance their data organization skills and improve their data analysis capabilities.

Sorting Data

Sometimes you want to change the structure of your dataset so that observations are ordered according to one or more variables, called sorting. Data is easily sorted by one or more variables with a procedure called `PROC SORT`. You can sort data by both numeric and character variables. The general format of the sort procedure is:

```
PROC SORT <options>;
BY var;
RUN;
```

In the syntax above, `PROC` is the keyword that starts the proc step and `SORT` is the name of the procedure. Immediately following PROC SORT is where you put any options you want to include. Let's review some of the more common options:

`DATA = Dataset-Name`
Specifies the dataset to be sorted. If you don't include this statement then SAS will sort the most recently processed dataset by default.`NODUPLICATES`
This option will eliminate duplicate records in your dataset, as long as the duplicates are next to each other in the dataset. Note that if `NODUPLICATES` is used without `NODUPKEY`, then records are only considered duplicates if they have identical values for every variable.`NODUPKEY`
This option deletes duplicate observations from the dataset based on the variables specified in the BY statement. It considers an observation a duplicate if it has the same values for all of the variables specified in the BY statement.`OUT = New-Dataset-Name`
When SAS processes a sort procedure, it overwrites the unsorted dataset with the sorted dataset by default. If you would like your sorted dataset to be a new dataset, then use this option.

In the next line is the `BY` statement, where you tell SAS what variable(s) to sort the data on, and what order to do the sorting in.

If you list just one variable, then SAS will sort the observations in the dataset based on ascending

values of that variable.
Missing values are considered the smallest possible value or category. If sorting in ascending order, missing values will appear first.You can sort in descending order by placing the keyword `DESCENDING` before the variable name that you want the dataset to be sorted on.
When sorting in descending order, missing values will appear last (since they are considered the smallest possible value).You can sort by as many variables as are in the dataset. When more than one variable is listed in the BY statement, SAS will first sort the observations based on the values of the first variable, then sort observations by the values of the second variable *within each category of the first variable*, and so on. Variable names listed after the `BY` keyword should be separated by a space, and should be listed in the order you want SAS to order by.

The `RUN` statement is placed at the end of the block and tells SAS to execute the code.

SAS 9.2 Procedures Guide - PROC SORT

Example

Problem Statement

In the sample dataset, we have information about gender and birthday (but not necessarily age) for the subjects in the sample. How can we sort each gender from youngest (most recent birthdate) to oldest (least recent birthdate)?

The birthday variable (bday) in the sample dataset is a date variable. Recall that date variables are a special type of numeric variable; therefore, date variables will follow the same sorting rules as regular numeric variables. This means that when date variables are sorted in ascending order, missing values will come first, and then the dates will be sorted from oldest to newest. Conversely, if a date variable is sorted in descending order, the newest (most recent) dates will come first.

(Why is this? Recall that dates in SAS are internally measured as the amount of time that has elapsed since the reference date (January 1, 1960). This implies that more recent dates will technically be stored as larger numbers.)

Syntax

```
PROC SORT data=sample;
BY gender descending bday;
RUN;
```

Output

After sorting, your data will look similar to this:

	ids	bday	Gender	Athlete
1	23017	29NOV1995	.	0
2	33153	28JUL1995	.	0
3	42217	07APR1994	.	1
4	20248	01JAN1994	.	0
5	25714	27OCT1993	.	0
6	30418	31AUG1993	.	0
7	43678	24MAR1993	.	1
8	32415	07AUG1992	.	1
9	49165	16JAN1992	.	1
10	20230	02JAN1996	0	1
11	20826	25DEC1995	0	0
12	21939	12DEC1995	0	1

The data is sorted first by gender, in ascending order. Notice how missing gender values appear first, then 0 (coded for male). Within each gender, the data is then sorted in descending order by birth date. Among the cases with missing values for gender, we can see that SAS recognizes that November 29, 1995 > July 28, 1995 > April 7, 1994. In rows 10 through 12, we see the 3 "largest" (most recent) birthdates for males: January 2, 1996 > December 25, 1995 > December 12, 1995.