How can I utilize the SAS SELECT DISTINCT statement in PROC SQL to remove duplicate records in my dataset?

Authored by

stats writer

June 23, 2024

The SAS SELECT DISTINCT statement in PROC SQL is a useful tool for removing duplicate records from a dataset. This statement allows the user to select unique values from a specific column or set of columns, eliminating any duplicate entries. By utilizing this statement, one can easily clean and organize their data by removing redundant information. This can be particularly helpful in cases where duplicate records may skew analytical results or create confusion in data interpretation. Overall, the SAS SELECT DISTINCT statement is an efficient way to streamline and improve the accuracy of datasets in SAS PROC SQL.

SAS: Use SELECT DISTINCT in PROC SQL

You can use the SELECT DISTINCT statement within PROC SQL in SAS to select only unique rows from a dataset.

The following example shows how to use this statement in practice.

Example: Using SELECT DISTINCT in SAS

Suppose we have the following dataset in SAS that contains information about various basketball players:

```
/*create dataset*/
data my_data;
input team $ position $ points;
datalines;
A Guard 14
A Guard 14
```

```
A Guard 24
A Forward 13
A Forward 13
B Guard 22
B Guard 22
B Forward 34
C Forward 15
C Forward 18
;
run;

/*view dataset*/
proc printdata=my_data;
```

Obs	team	position	points
1	A	Guard	14
2	A	Guard	14
3	A	Guard	24
4	A	Forward	13
5	A	Forward	13
6	B	Guard	22
7	B	Guard	22
8	B	Forward	34
9	C	Forward	15
10	C	Forward	18

We can use the SELECT DISTINCT statement within

PROC SQL to select all unique rows from the dataset:

```
/*select all unique rows*/
proc sql;
select distinct *
from my_data;
quit;
```

team	position	points
A	Forward	13
A	Guard	14
A	Guard	24
B	Forward	34
B	Guard	22
C	Forward	15
C	Forward	18

Note: The star (*) symbol after SELECT DISTINCT tells SAS to select *all* columns in the dataset.

Notice that all unique rows are shown in the output.

For example, there are multiple rows that have a team value of A, position value of Forward and points value of 13 but only one of these rows is shown.

Note that we can also specify which columns we'd like to select:

```
/*select all unique combinations of team and position*/
proc sql;
select distinct team, position
from my_data;
quit;
```

team	position
A	Forward
A	Guard
B	Forward
B	Guard
C	Forward

Notice that only the unique combinations of teams and positions are shown in the output.