2.sumrData

March 3, 2025

0.0.1 Summarize Data

In Summarizing Data you learn new syntax that enables you to alter the default behavior of the DATA step to solve a problem. First you learn to create an accumulating column, or in other words generate a running total. Then you learn to process data in groups, so you can perform an action when each group begins or ends.

Creating an Accumulating Column

- At the beginning of the first iteration of the DATA step, all column values are set to missing.
- By default, all computed columns are reset to missing at the beginning of each subsequent iteration of the DATA step. This is called reinitializing the PDV. Columns read from the SET statement automatically retain their value in the PDV.
- To create an accumulating column, this default behavior must be modified.

Directing DATA Step Output

RETAIN *column <initial-value>*;
column+expression;

```
TotalRain+Rain_mm;
```

PDV

...other columns...	Rain_mm	TotalRain
	.	3

- creates accumulating column and sets initial value to 0
- retains value of the accumulating column
- adds right column value to accumulating column for each row
- ignores missing values

Processing Data in Groups

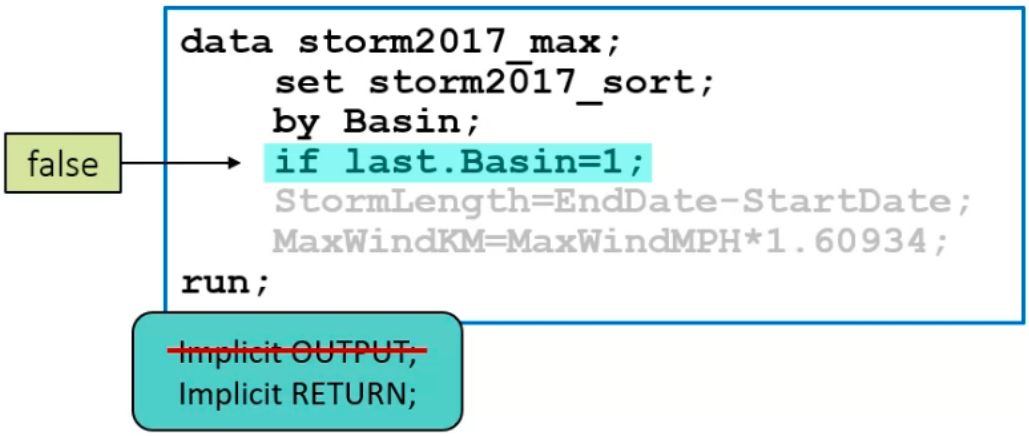
- To process data in groups, the data first must be sorted by the grouping column or columns. This can be accomplished with PROC SORT.
- The BY statement in the DATA step indicates how the data has been grouped. Each unique value of the BY column will be identified as a separate group.
- The BY statement creates two temporary variables in the PDV for each column listed as a BY column: **First.*bycol*** and **Last.*bycol***.
- **First.*bycol*** is 1 for the first row within a group and 0 otherwise. **Last.*bycol*** is 1 for the last row within a group and 0 otherwise.
- Conditional IF-THEN logic can be used based on the values of the **First./Last.** variable to execute statements in the DATA step.

> **BY** <**DESCENDING**> *col-name(s)*;
> **FIRST.***bycol*
> **LAST***bycol*

- **First./Last.** variables can be used in combination with IF-THEN logic to execute one or more statements at the beginning or end of a group.
- The subsetting IF statement affects which rows are written from the PDV to the output table. The expression can be based on values in the PDV.
- When the subsetting IF expression is true, the remaining statements are executed for that iteration, including any explicit OUTPUT statements or the implicit OUTPUT that occurs with the RUN statement.

- If the subsetting IF expression is not true, the DATA step immediately stops processing statements for that particular iteration, likely skipping the output trigger, and the row is not written to the output table.

> **IF** *expression*;



[ ]: