



Don't Be a SAS® Dinosaur: Modernize Your SAS Programs

by Warren Repole

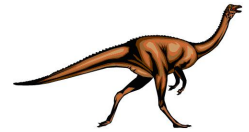
Duplicate Observations: Capture into a Data Set

Scenario:

You need to separate duplicate observations based on selected key variables. The first observation from each set of duplicates is placed into one data set while the remaining duplicates are output to a second data set.

The NODUPKEY option of PROC SORT can save the first observation of each BY group into the output data set, but, by default, all remaining duplicates are discarded.

The old way: Split the Sorted Data using a DATA Step



The original approach is to split the input observations into two output data sets using a DATA step with BY group processing. The first observation in each BY group is output to one data set. All other observations, those representing the duplicates, are output to the other data set.

```
proc sort data=sashelp.zipcode out=ByCity;
  where StateName="Delaware" and CountyNm="Kent";
  by City;
run;
data primary dups;
  set ByCity;
  by City;
  if first.City then output primary;
  else output dups;
run;
proc print data=primary noobs;
  var City ZIP;
  title1 "Primary ZIP Codes in Kent County, Delaware";
run;
proc print data=dups;
  by City; id City;
  var ZIP;
  title1 "Additional ZIP Codes for Kent County, Delaware Cities";
run;
```

Please direct all correspondence to:

Warren Repole, 1705 Palm Springs Dr, Vienna VA 22182-2331

sasdinosaur@repole.com

www.repole.com/dinosaur

The new way: Split the Data Set using the DUPOUT= Option (available in SAS 9)

An alternate approach is the DUPOUT= option of PROC SORT.

```
proc sort data=sashelp.zipcode out=primary
          nodupkey dupout=dups;
  where StateName="Delaware" and CountyNm="Kent";
  by City;
run;
proc print data=primary noobs;
  var City ZIP;
  title "Primary ZIP Codes in Kent County, Delaware";
run;
proc print data=dups;
  by City; id City;
  var ZIP;
  title "Additional ZIP Codes for Kent County, Delaware Cities";
run;
```

Advantages of the alternate approach:

- No additional passes of the data occur.
- Only a single PROC SORT step is required.

Disadvantages of the alternate approach:

- This approach does not permit control over which observations are output into which data set. Applying BY group processing in the DATA step would allow unique observations (FIRST.variable=1 and LAST.variable=1) to be placed into one data set and all other observations into another data set.

Additional documentation for this technique can be found in *Base SAS® 9.2 Procedures Guide*. Cary, NC: SAS Institute Inc.

Visit <http://support.sas.com/documentation/onlinedoc/sas9doc.html> for SAS 9 documentation.

Go to <http://www.repole.com/dinosaur/dupout.html> for the sample code and output for this topic.

These techniques are mentioned in other SAS references and publications:

SAS Sample 24626(<http://support.sas.com/kb/24/626.html>)

Identifying Duplicate Values; Christopher J. Bost

<http://www.nesug.info/Proceedings/nesug06/cc/cc24.pdf>.