Merging data sets to create paired data

In Class Exercise 1, you input the Fall 2008 first-year cohort of students at University of South Carolina in both SAS and R. Using your SAS code, import the Fall 2008 data and the Fall 2009 data as well; name them `FALL08` and `FALL09` respectively. We would like to create some paired data for Homework 1, so we will combine the 3364 records in each file by student ID, and save students' GPA for each semester as a large paired data set.

The code below demonstrates two different methods for merging files by ID and retaining only ID, Fall 2008 GPA and Fall 2009 GPA. The first method uses the SQL database query language, which SAS has implemented as PROC SQL; the second method uses a more traditional DATA step match merge. Note that both data sets have the same variable name–GPA–that needs to be assigned two separate names in the combined data set (which we have named `PAIRED`).

In PROC SQL, a merge is referred to as a "join". The `COALESCE` statement ensures that all IDs from both data sets are included in the output data set. This is an unnecessary step for these two data sets, since we know they have an identical set of IDs, but is generally good programming practice for a join, otherwise.

In the DATA step match merge, both data sets need to be sorted by ID beforehand, even though they are pre-sorted in Excel.

```
*PROC SQL;
proc sql;
create table paired as select coalesce(fall08.ID, fall09.ID) as ID,
       fall08.GPA as GPA2008,
       fall09.GPA as GPA2009
       from fall08 full join fall09 on fall08.ID=fall09.ID;
quit;


*DATA step;
proc sort data=fall08; by id; proc sort data=fall09;
data paired (keep=ID GPA2008 GPA2009);
merge fall08 (rename=(GPA=GPA2008)) fall09 (rename=(GPA=GPA2009));
keep id gpa2008 gpa2009;
by ID;
run;
```

Carry out both merges and check the data set `PAIRED` in your WORK directory each time to make sure it was constructed correctly. Comment on any features of the GPA variables that concern you. Which method is more familiar to you? Do you see advantages/disadvantages in the two approaches?