13 Combining Data Sets Assignment

**Exercise 1**. Appending Like-Structured Data Sets

The data sets price\_current and price\_new are in the STA5066 sub directory.

1. Use two DATA steps to read price\_current and price\_new and create new data sets work.current and work.new.

2. Use two PROC CONTENTS steps to compare the variables in the two data sets and note the number of observations on each data set.

3. Use a PROC APPEND step to append work.new to work.current.

4. Use a PROC CONTENTS step to exam the contents of work.current and check that the number of observations is correct (the sum of the observations on work.current and work.new prior to the execution of PROC APPEND).

**Exercise 2**. Appending Unlike Structured Data Sets

1. Use two PROC CONTENTS steps to examine the data sets qtr1\_2007 and qtr2\_2007. Both of these SAS datasets are in the STA5066 subdirectory. Examine the variables on the two data sets and note which variable is not in both data sets.

2. Use a PROC APPEND step to append qtr1\_2007 to the (non-existing) data set called work.ytd.

3. Use a PROC CONTENTS step to confirm that the correct number of observations were copied to work.ytd.

4. Use a PROC APPEND step to append qtr2\_2007 to work.ytd. Note that the FORCE option is needed. Confirm that work.ytd has the correct number of observations.

**Exercise 3**. Concatenating Like Structured Data Sets

1. Use three PROC CONTENTS steps to examine how many observations are in the data sets mnth7\_2007, mnth8\_2007, and mnth9\_2007. These data sets are in the STA5066 sub directory.

2. Use a DATA step to concatenate the three data sets mnth7\_2007, mnth8\_2007, and mnth9\_2007 to create a new data set called work.thirdqtr.

3. Use a PROC CONTENTS step to verify that work.thirdqtr contains the correct number of observations.

**Exercise 4**. Concatenating Unlike-Structured Data Sets

1. Use two PROC CONTENTS steps to compare the variables in the two data sets sales and nonsales. Both of the data sets are in the STA5066 subdirectory. Note the number of observations on each data set and the names of the two variables that are different in the two data sets.

2. Use a DATA step to concatenate sales and nonsales to create a new data set called work.allemployees. Use a RENAME= data set option to change the names of the different variables in nonsales.

3. work.allemployees should include only the following five variables: Employee\_ID, First\_Name, Last\_Name, Job\_Title, and Salary.

4. Use a PROC PRINT step to print the first 100 observations in work.allemployees.

**Exercise 5**. Merging Two Data Sets One-to-One

The data sets employee\_payroll and employee\_addresses are in the STA5066 sub directory.

1. Use two PROC CONTENTS steps to verify that employee\_payroll and employee\_addresses have the same number of observations and note the number of variables on each file.

2. Use a PROC SORT step to sort employee\_payroll in ascending order by Employee\_ID to create a new data set, work.payroll, that contains the sorted data.

3. Use a PROC SORT step to sort employee\_addresses in ascending order by Employee\_ID to create a new data set called work.addresses.

4. Use a DATA step to merge the two sorted data sets by Employee\_ID to create a new data set called work.payadd.

5. Use a PROC CONTENTS step to verify that work.payadd has the correct number of observations and variables.

**Exercise 6**. Merging Three Data Sets One-to-One

The data sets employee\_addresses, employee\_payroll, and employee\_organization are in the STA5066 sub directory.

1. Use a PROC SORT step to sort employee addresses in ascending order by Employee\_ID to create a new data set called work.addresses.

2. Use a PROC SORT step to sort employee\_payroll in ascending order by Employee\_ID to create a new data set called work.payroll.

3. Use a PROC SORT step to sort employee organization in ascending order by Employee\_ID to create a new data set called work.organization.

4. Use three PROC CONTENT steps to examine the number of observations and the number of variables on each of the data sets work.addresses, work.payroll, and work.organization.

5. Use a DATA step to merge work.addresses, work.payroll, and work.organization to create a new data set work.payaddorg.

6. Use a PROC PRINT step to verify that work.payaddorg has the correct number of observations and variables.

**Exercise 7**. Merging Using the IN equals Option

The data sets product\_list and supplier are in the STA5066 sub directory.

1. Use a PROC SORT step to sort the data set product\_list by supplier\_id and create the new data set work.product

2. Use a PROC SORT step to sort the data set supplier by supplier\_id and create the new data set work.suppliersort

3. Use a DATA step to merge work.product and work.suppliersort by Supplier\_ID to create a new data work.prodsup. work.prodsup should contain only observations that are on the work.product data set but on the not work.suppliersort data set.

4. Use PROC PRINT to verify that work.prodsup has 75 observations, 10 variables, and that all values for supplier\_id are missing.

**Exercise 8**. Merging Using the IN= and RENAME= Options

The data sets customer and lookup\_country are in the STA5066 sub directory.

1. Use a PROC PRINT step to examine the data set lookup\_country.

(a) Note that the variable START is the two letter country abbreviation and the variable LABEL contains the full country name.

(b) Note that the data set is sorted in alphabetical order by the variable start.

2. Use a PROC PRINT step to examine the data set customer. Note that the variable country contains the two letter abbreviation for the customer’s country.

3. Use a PROC SORT step to sort the data set customer by Country to create a new data set called work.customer.

4. Use a DATA step to merge the data set customer with lookup\_country by Country to create a new data set called work.allcustomer. See the next page for further instructions.

(a) The variable Start needs to be renamed to Country and the variable Label needs to be renamed to Country\_Name in the lookup\_country data set.

(b) Include only the observations that contain both customer information and country information.

5. Use a PROC PRINT step to verify that allcustomer has 77 observations.

**Exercise 9**. Merge the Nhanes 3 exam and lab files

The data sets labsub2 and examsub2 are in the STA5066 sub directory.

1. Use a PROC SORT step to sort the data set labsub2 by seqn and create a new data set work.lab that contains the sorted data.

2. Use a PROC SORT step to sort the data set examsub2 by seqn and create a new data set work.exam that contains the sorted data.

3. Use a DATA step to merge the two datasets work.exam and work.lab by seqn and in doing this use the in= construct to create three new datasets:

(a) work.ExamOnly containing observations that are on the work.exam data set but are not on the work.lab data set.

(b) work.LabOnly containing observations that are on the work.lab data set but are not on the work.exam data set.

(c) work.LabAndExam containing observations that or on both the work.exam data set and the work.lab data set.

4. Use three PROC CONTENTS steps to determine the number of observations and the variables on each of the three data sets work.ExamOnly, work.LabOnly, and work.LabAndExam