Filtering Data in SAS Enterprise Guide

Working with SAS Enterprise Guide (EG) can seem a bit overwhelming, at first. This is the first in a series of articles that will make the transition to using SAS Enterprise Guide a little bit easier. Please note that this article was written using Enterprise Guide 4.3.

To begin an EG project, there are a few simple initial steps:

1. Organize the project – use multiple process flows as needed and use good naming conventions for columns, datasets and tasks.

   **Sample Process Flows within a Project:**
   System Documentation, Startup/Initialization, Data Extraction, Data Clean Up/Transformation, Conversion, Updates, Reports-Custom SAS Reports/Custom HTML Reports

2. Identify the SAS and non-SAS data that will be needed for the project. Determine if the data should be part of the project or simply referenced. Explore the variable attributes and values. Determine any links/keys that associate the various data sets.

3. Establish the type of analysis required by your project
   a. Combining multiple data sets
   b. Sub-setting
   c. Sorting
   d. Summarizing
   e. Reporting – style of output

Once these initial planning steps are completed, it is time to start working with the data!

1. **Identify the Location of Data**

   The first step is to identify the location (directory) of the data sets that will be used in the EG project. It will be given a nickname, or Library. To do this, on the 'Tool' menu, select the 'Assign Project Library' wizard to designate the location of our SAS data sets. In this example, we will use the nickname 'PIELIB'. After completing this step, the 'Assign Project Library' icon appears in the Project Tree as well as the Process Flow Diagram.

2. **Determine the Variable Attributes and Values:**

   In the 'Basic Analysis-File Attributes' Process Flow that was created, the EG user can run the 'Data Set Attributes' and 'List Data' tasks to determine the variable attributes and values. For these examples, we will be using sample Pie Production and Nutrition Data.
3. **FILTERING AND SUB-SETTING:**
One of the basic processes that projects may require is the ability to filter or sub-set data sets. There are two ways to accomplish this process. One way is by using the 'Filter and Sort' task. The second is by using 'Query Builder'.

**Sub-Selection Based on a Character Variable:**

For the Pie Project, one may want to select all the rows from the Production data set that are in the 'Chocolate Pies' category. This can be accomplished using the 'Filter and Sort' task.

Highlight the 'Basic Analysis – Subsetting' process flow. Double click the Production data set in the Server list. This will link this data set to this flow without physically including the data in the project.

Highlight the Production data set by clicking it in either the Project Tree or Process Flow. Then select the 'Filter and Sort' task by either:

1. Right clicking the data set and selecting 'Filter and Sort' or
2. Clicking the 'Tasks' dropdown, click 'Data' and then select 'Filter and Sort'

On the 'Variables' tab, select all of the variables and drag them from 'Available' to 'Selected'.

Click the 'Filter' tab and select the variable 'Category'. Select 'Equal To' for the operator. Select 'Chocolate Pies' as the comparison value. Click 'OK'.
The task runs and the filtered data set is created.

To create the filtered data set with a meaningful name, click the ‘Results’ tab and change the name.

Using Query Builder

As with the ‘Filter and Sort’ task, highlight the Production data set by clicking it in either the Project Tree or Process Flow. Select the ‘Query Builder’ task by either:

1. Right clicking the data set and selecting ‘Query Builder’
2. Clicking the ‘Tasks’ dropdown, click ‘Data’ and then select ‘Query Builder’

Select all the variables from the Production data set and drag them to the ‘Select Data’ tab. To create the filtered data set with a meaningful name, change the output name.
Click the 'Filter Data' tab and click the 'New filter' icon.

Select the 'Category' variable

Use the default 'Equal to' and click the 'Value' dropdown arrow. Click 'Get Values'.
Click ‘Fruit Pies’. Because ‘Category’ is a character variable, the ‘Enclose values in quotes’ box should be checked. Click ‘Next’. Click ‘Finish’ to create the filtered file.

**Sub-Selection Based on a Numeric Variable:**

This next example will select all the rows from the Production data for pie makers who made 10 or more pies for a particular pie type and size.

One could use either of the techniques that was used to create the Chocolate Pie or Fruit Pie subsets. However, this example will select the technique in a different manner. First, open the Production data set and select the tab named ‘Query Builder’ task. Using the ‘Filter and sort’ tab works as well.
The first step is to change the Query name to make it more meaningful. The next step is to select all the variables from the original data set. Then, click the 'New filter' icon to subset the file and use the basic filter. Use the numeric variable 'Quantity' to perform the sub-setting.

Select the rows that have a quantity of 10 or more. Notice that the 'Enclose values in quotes' option is not available since Quantity is a numeric field.
Click 'Finish'. The filter appears in the Query Builder. Run the Query Builder to complete the subset process.

**Sub-Selection Based on a Date Variable:**

Lastly, select all the rows from the Production data set for pie makers who made pies between June 20, 2006 and June 21, 2006.

Highlight the Production data set and select the ‘Query Builder’ task. As in the previous example, change the task and output data set names to be more meaningful. Select all the variables on the 'Select Data' tab. On the 'Filter Data' tab, click the ‘New filter’ icon. Using 'Basic Filter', select the ‘Date’ variable to use in the selection criteria.

From the 'Operator' drop-down menu select ‘Between’. The 'Start value' and 'End value’ options appear. Use the down arrow and 'Get Values' options to display the date values in the data set.

Select 06/20/2006 as the ‘Start value’ and 06/21/2006 as the 'End value'. Note the 'Use formatted dates' box is checked so that the formatted dates as well as the SAS dates appear. Click 'Next' to see the filter 'Where' statement or click 'Finish'. Click 'Run' to create the new file.
The new task and output file appear in the Project Tree and Process Flow.

Now that the data has been subsetted in one of the previous ways, it is ready for further analysis. Future articles in *The Missing Semicolon* will explore what to do next in Enterprise Guide.