SAS Enterprise Guide® for Educational Researchers: Data Import to Publication without Programming

AnnMaria De Mars, University of Southern California, Los Angeles, CA

ABSTRACT

In this workshop, participants learn how to conduct a data analysis project from beginning to end with SAS Enterprise Guide. This will be done by completing practical exercises that are common data analysis tasks.

The workshop begins with creating a SAS dataset for analysis. Using these data, participants will produce output for a report on data quality, a table of sample characteristics and a comparison of results by group. Next, an Excel file is imported, new files created and merged, then variables computed to create a final analysis dataset to be used to evaluate whether the client's intervention was effective.

By the end of this workshop participants will be able to:

- Import Excel, Access and other files into Enterprise Guide,
- Create new SAS datasets,
- Merge files,
- Compute new variables,
- Recode variables,
- Conduct descriptive statistics,
- Conduct basic predictive and comparative analyses,
- Create publication-quality tables, and
- Create publication-quality graphics.

Can SAS Enterprise Guide do all of that, with no programming required? Yes, it can.

INTRODUCTION

SAS is extremely good at solving three types of problems:

1. Creating new data.
2. Finding information in data.
3. Reading and combining data from diverse sources.

This workshop addresses each type of problem in turn, all without any SAS programming required. The basic features and concepts in Enterprise Guide will be introduced through solving common problems.

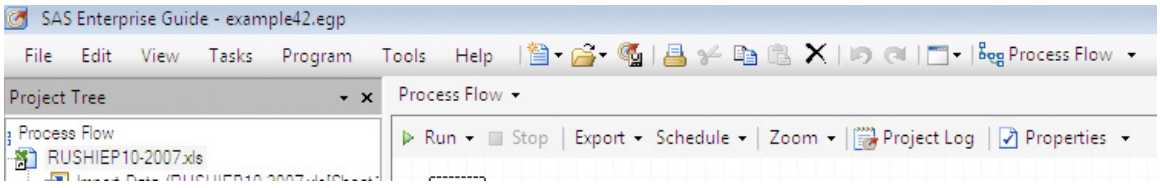
EXAMPLE 1: CREATING NEW DATA & DATA ANALYSIS

In the first example we are using a dataset of surveys on media usage collected in the Great Plains states. Our client, a federal agency, is interested in targeting information to Native Americans. The client wants to know how frequently various types of media are used and whether there are any demographic differences in the types of media used.

It has been said that 80% of any analysis project is getting the data in shape (Levesque, 2004). A common first step is to select a subset of records. In this example, the first step is to open the dataset and select a subset of Native American respondents, defined by our client specifications as those who are enrolled in a federally-recognized tribe.

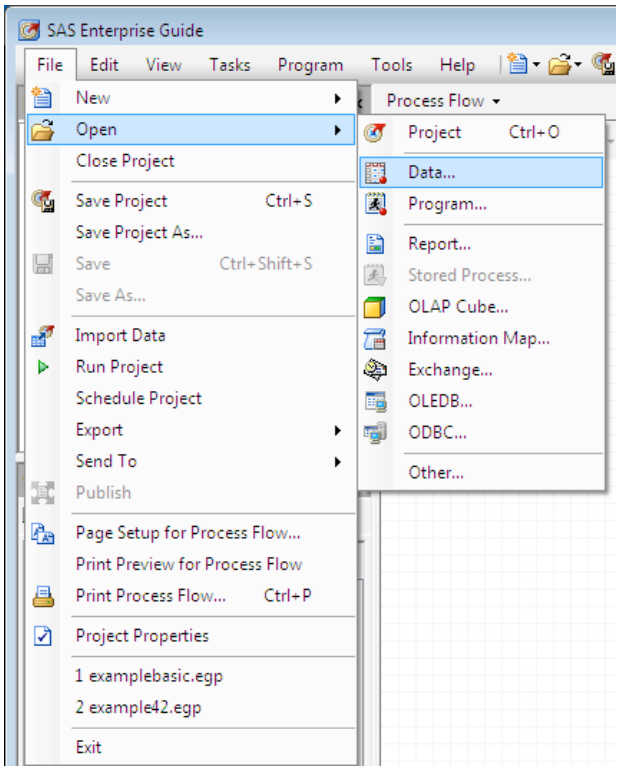
OPENING A SAS DATASET

Enterprise Guide is very menu-driven and quite intuitive. When you open Enterprise Guide for the first time, you'll see the view below:

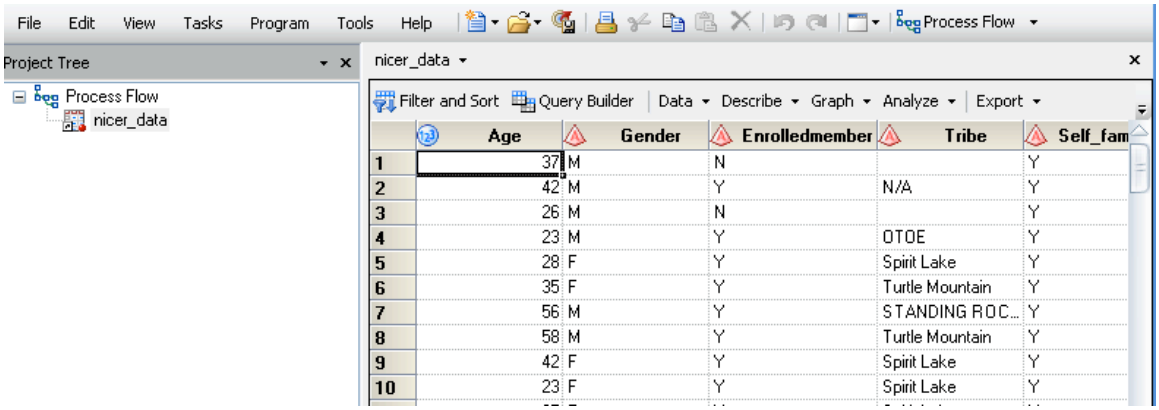


Notice the main menu at the top of your screen. Enterprise Guide has seven main menu options. Most of them are pretty self-explanatory. The **File** menu is used to open files, whether SAS files or imported from other programs, such as Excel.

Select "**File**" from the top menu. From the drop-down menu that appears select "**Open.**" Then, from the menu that appears next to "**Open**", select the "**Data"** option. A window will pop up that allows you to select your dataset. Open the dataset named *nicer_data.*



Once your dataset is opened, on the left side of your screen you see the **Process Flow**. This shows that so far you have only done one thing, opened the dataset. On your right, you see your SAS dataset. Columns that are numeric variables are identified with 123 inside a blue circle. Character variables are denoted by the letter A inside a red triangle.

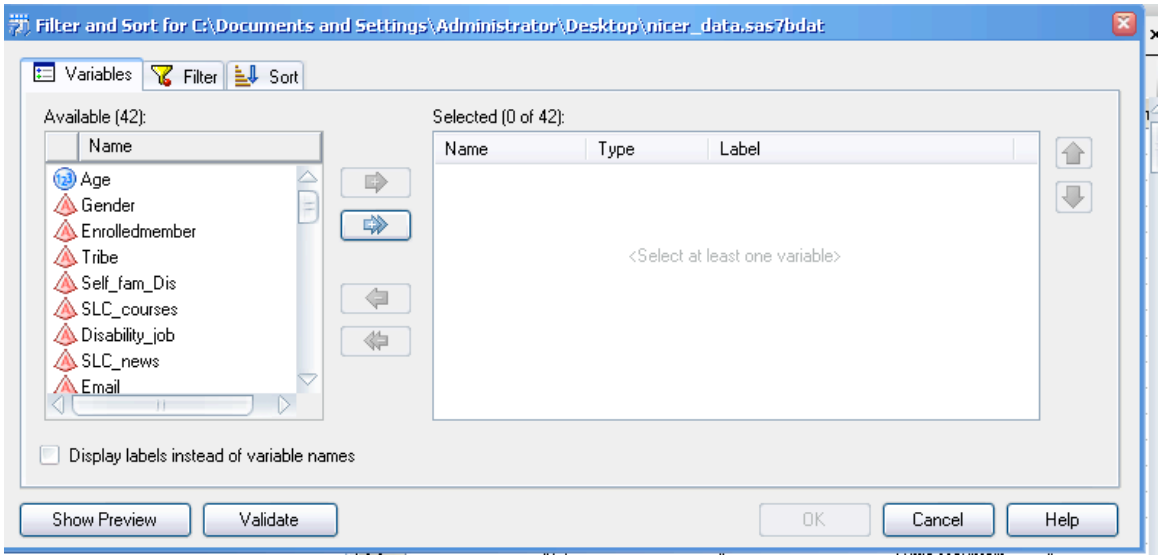


	Age	Gender	Enrolledmember	Tribe	Self_fam
1	37	M	N		Y
2	42	M	Y	N/A	Y
3	26	M	N		Y
4	23	M	Y	OTOE	Y
5	28	F	Y	Spirit Lake	Y
6	35	F	Y	Turtle Mountain	Y
7	56	M	Y	STANDING ROC...	Y
8	58	M	Y	Turtle Mountain	Y
9	42	F	Y	Spirit Lake	Y
10	23	F	Y	Spirit Lake	Y

The menus in the dataset window allow a lot of data manipulation without programming. An example of such manipulation follows.

Creating a subset of the original data

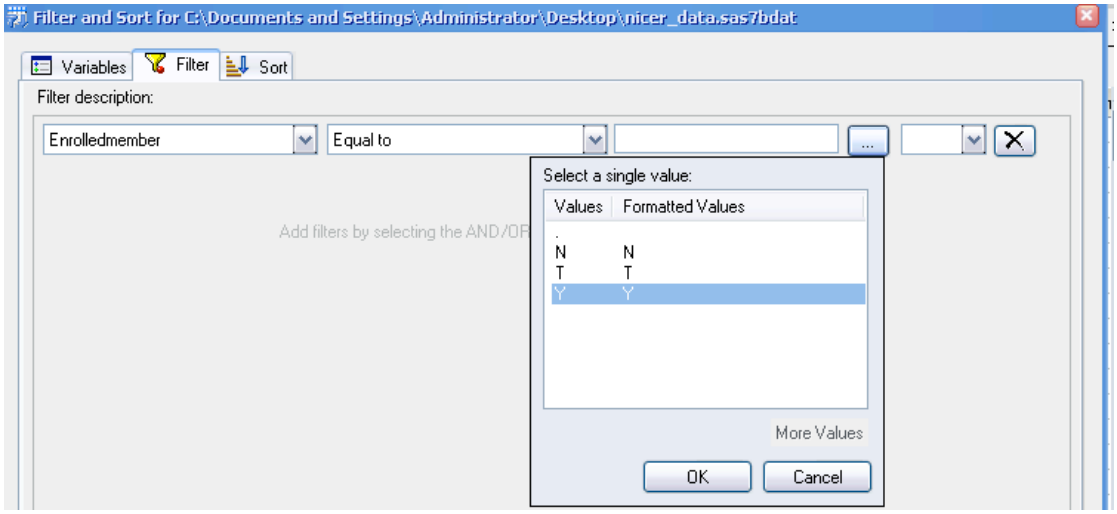
The first step in our example uses the first menu in the dataset window – **Filter and Sort.** Clicking on the **Filter and Sort** tab brings up the menu below. As the client is interested specifically in Native Americans, the first step is to include only those respondents who answered "Y" to the question, *"Are you an enrolled tribal member?"*



Clicking on the double arrows between the two windows will move all the variables over to the right, **Selected** window. This selects the variables to be included in the new dataset. (To select a single variable, click on the desired variable and click on the single arrow.)

Next, click the **Filter** tab. The window that appears has three boxes. Clicking the arrow to the right of the first box produces a drop down menu of all variables in the dataset. Select the variable named *enrolledmember.*

Clicking the arrow next to the second box produces a different drop down menu, of operations. For this example, select *Equal to*. Clicking the three dots next to the third box will bring up a list of all values of the variable .

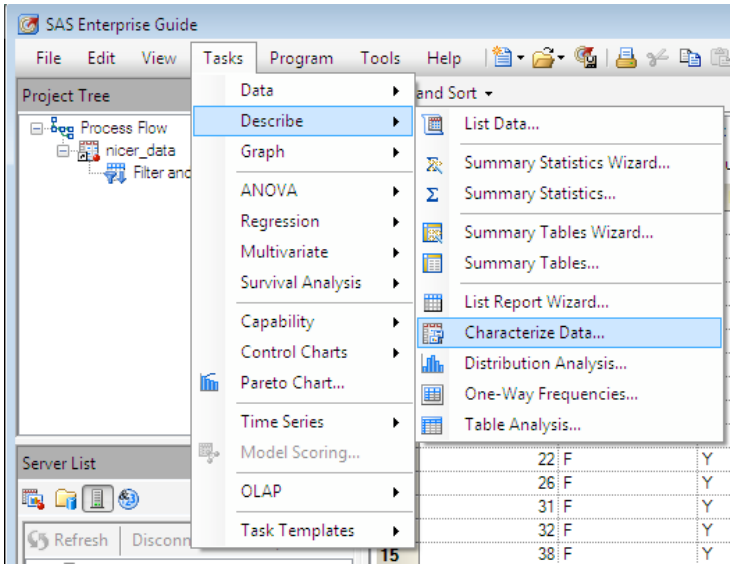


After selecting the desired value, 'Y', click on OK. Then, click on the RUN button. An experienced programmer may recognize that these steps have just written an IF statement, instructing SAS to create a new dataset that selects records if *enrolledmember* = 'Y'.

It may seem as if nothing has happened. There is no output and the dataset looks very similar. Look closely. Three differences are apparent. On the left, in the Process Flow, under the dataset, a blue Filter and Sort task is shown. In the dataset window pane, there are now four tabs above the dataset, which correspond to the original data (Input Data), the programming statements written by Enterprise Guide (Code), program notes and errors (Log) and the new dataset (Output Data). A particularly astute observer will notice that all the values for enrolled member in the output dataset are 'Y'. The filter worked.

Checking the data : The Characterize Data Task

Any reliable data analysis begins with checking the validity of the data. The **Characterize Data** task produces descriptive statistics for numeric variables, frequency distributions and plots of distributions.



The **Characterize Data** task will produce many pages of output. First, the frequencies are shown for the 30 most common categories of all categorical variables. If a variable has less than 30 categories, this amounts to a frequency distribution. Otherwise, it provides a useful first look at the data for verification purposes, e.g., that all the categories shown are names of actual tribes. The number and percent of records missing data are also shown for each variable.

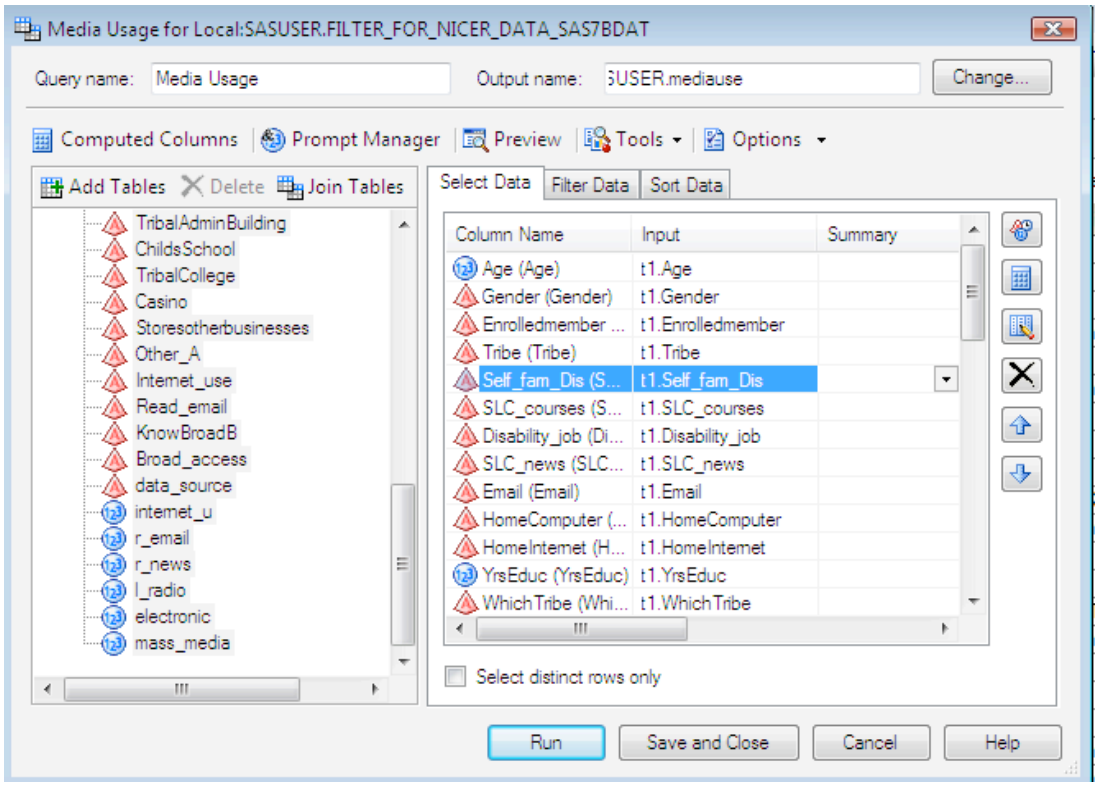
Next, the report shows the descriptive statistics for each numeric variable: mean, standard deviation, minimum and maximum. Again, the data provides a 'reality check'. Are the ages, years of education and other characteristics reasonable for what would be expected of a sample of this type?

Creating New Variables

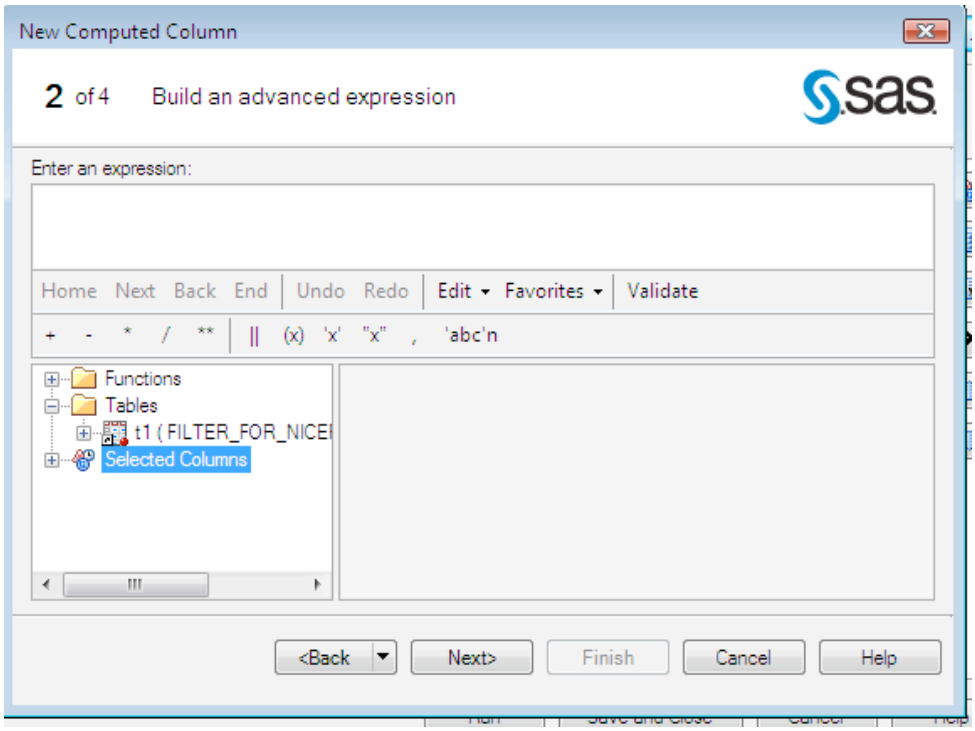
Now that you have verified that the data are good (a step you ignore at your peril), you can move on to creating new variables. The client wants to know about the relative frequency of electronic and mass media use. Click on the **Input Dataset** tab. Then click on **Query Builder**.

Good practices
Enterprise Guide gives datasets names like SASUSER.query_filter1_from_something.
In the pop-up window, give your query a name, in this case, Media Usage. Also give your output dataset a name.

Click on the first variable in the windowpane at left. Holding down the shift key, click on the last variable and drag all of the variables over to the **Select Data** pane at right.



Next, click on **Computed Columns**.
In the pop-up window, click on **NEW**.
Click the + sign next to selected columns to see all of the variables in your dataset.
Scroll down to *internet_u* and double-click on it.
Click on the +
Double-click on *r_email.*



Name the column “Electronic” and give it an Alias of “Electronic Media Usage”
Repeat the same steps to create a second column named MassMedia that is the sum of *l_radio* and *r_news.*
Click **CLOSE.**
Click **RUN.**
(SAS programmers will recognize the above as four statements, two assignment statements and two label statements.)

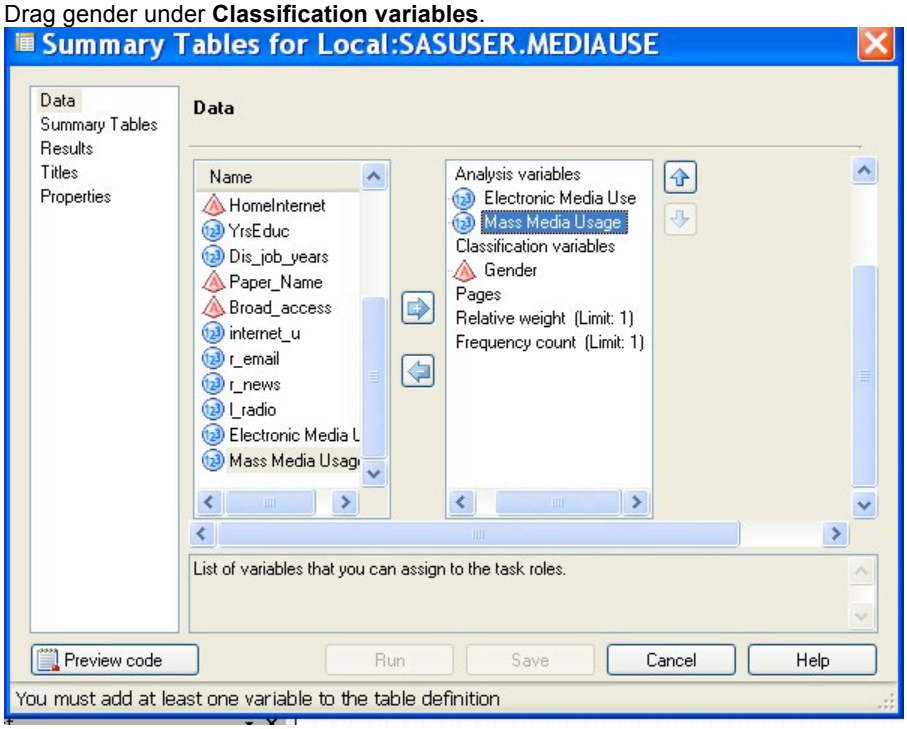
Summary Tables

The client is interested in distributing material on services for elementary school children and, believing that mothers are a better target market than fathers, would like to see the results broken down by gender. So, with the desired variables computed, usage of electronic media and mass media, it’s time to produce information on the average usage by men and women.
From the **Tasks** menu, select **Describe**.
Then, from the drop-down menu, select **Summary Tables**.

Selecting Task Roles

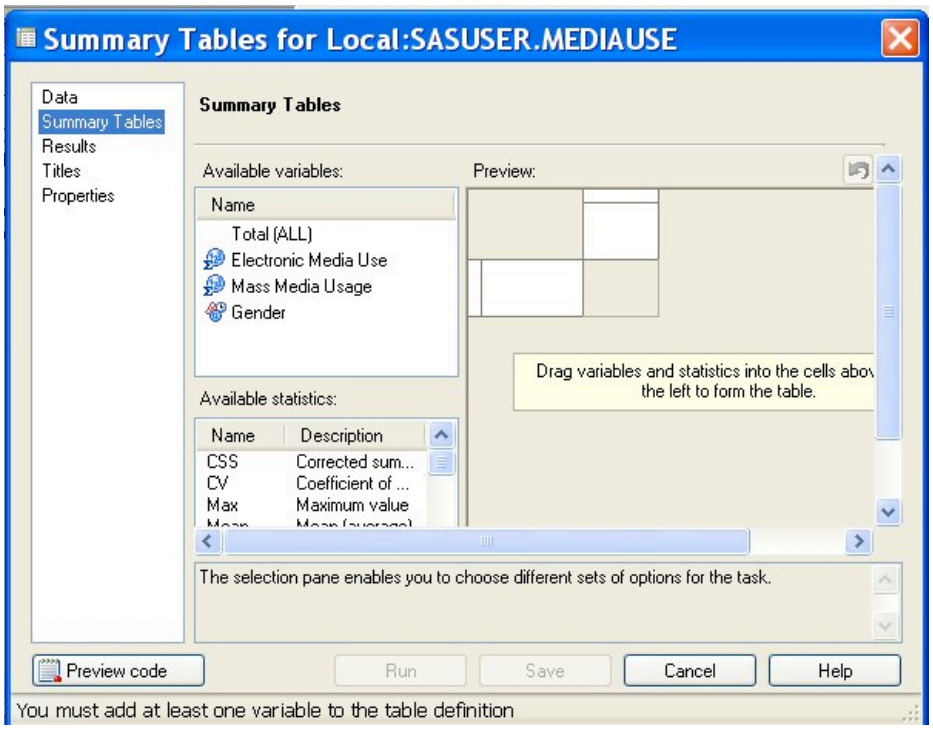
Task roles is a concept that comes up a lot in Enterprise Guide. There are a few major types of task roles. Analysis variables are those that are analyzed, i.e., on which statistics such as means, standard deviations and N (number of records) are produced. Classification variables are those by which statistics are classified. Variables can be assigned to a role by either right-clicking on the variable and assigning it to one of the roles in the drop-down menu or by dragging it under the appropriate role.

Drag Electronic Media Usage and Mass Media Usage under **Analysis variables.**
Drag gender under **Classification variables**.

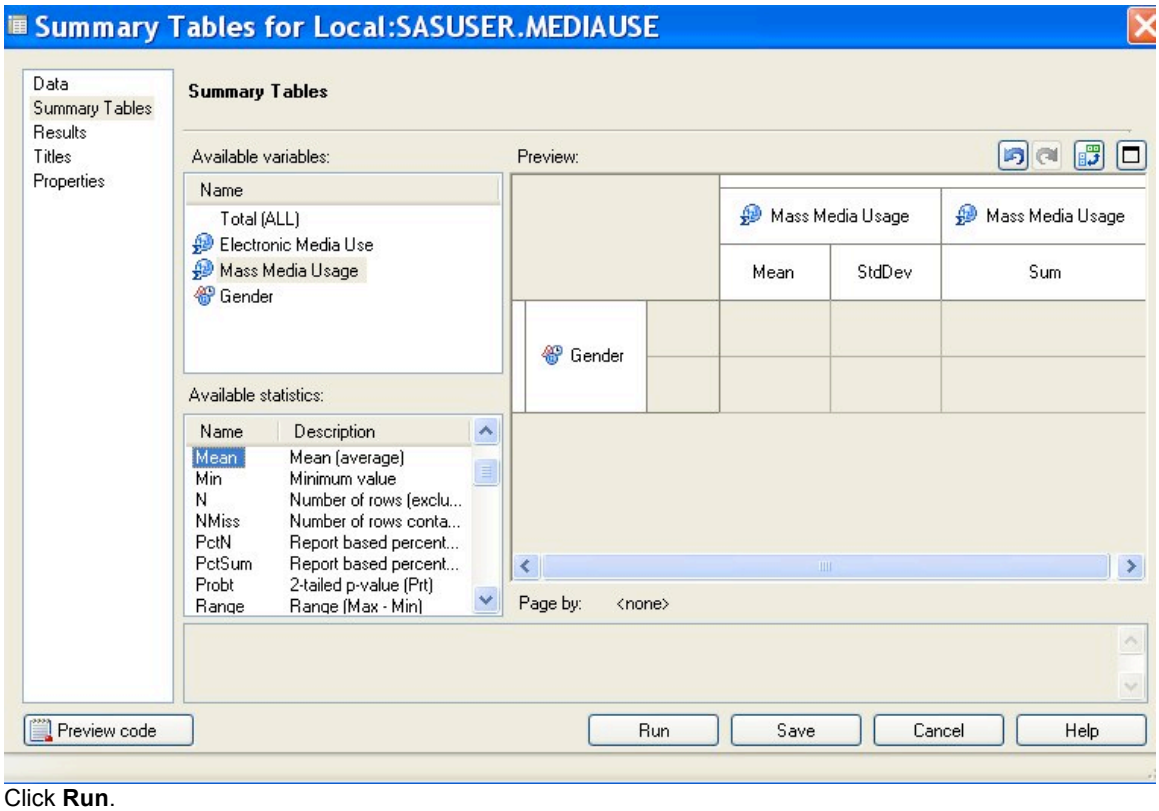


Next, click on Summary Tables in the left pane of the window.

The new pop-up window enables the user to design a table by dragging and dropping.



Drag Gender to the first box at the left of the table.
Drag Electronic and Media Usage to the top of the table window.
Initially, these two boxes will show **SUM** as the statistic to be reported. From the Available Statistics box, drag **N** on top of **SUM** to replace it. Drag **MEAN** and then **StdDev** next to **N**.



Click **Run**.

Our answer is shown below. On a 1(=never) to 8(= uses multiple sources daily) scale, Native American women score quite a bit higher on Mass Media Usage than Electronic Media. However, they do also use electronic media quite frequently. The average woman surveyed used at least one source (e-mail or Internet) on a daily basis.

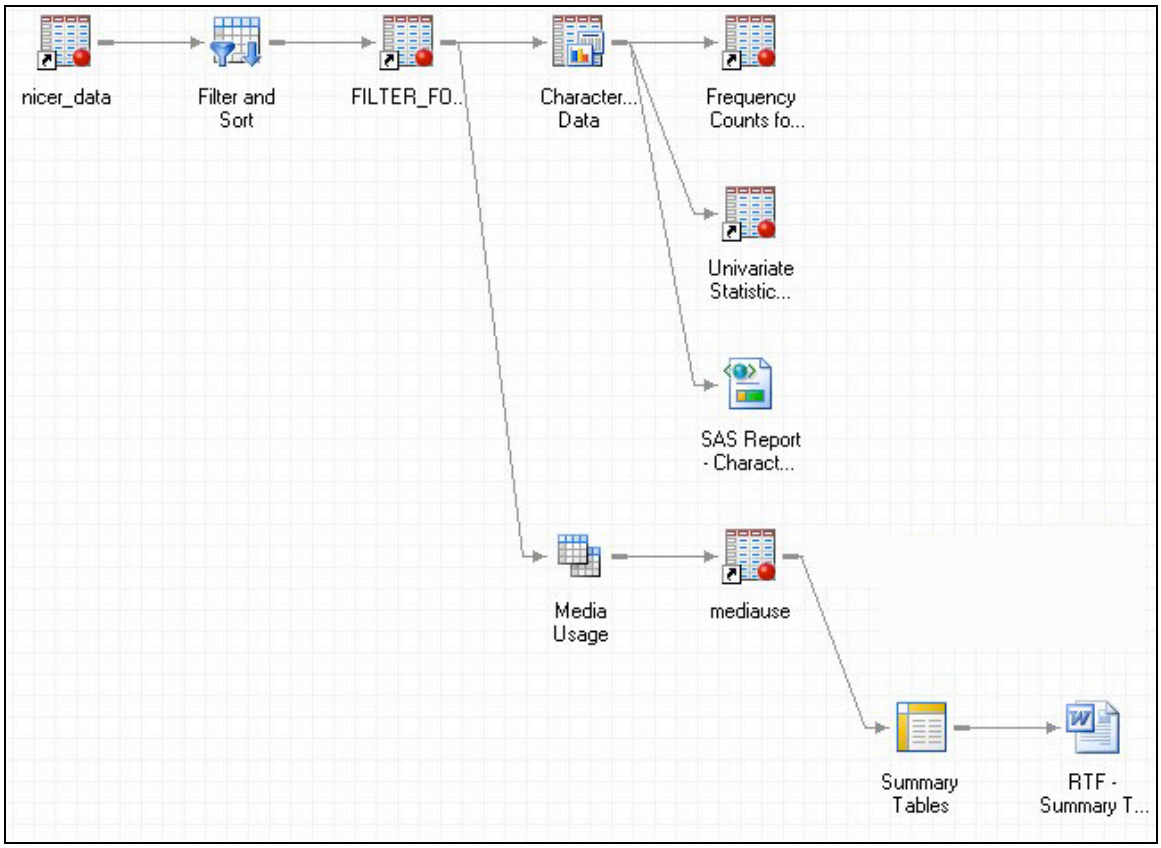
Electronic Media and Mass Media Usage by Gender

	Electronic Media Use			Mass Media Usage		
	N	Mean	StdDev	N	Mean	StdDev
Gender						
F	100	4.83	3.17	104	6.65	1.55
M	72	4.67	3.16	75	6.05	1.99

Research funded by the National Institute on Disability and Rehabilitation Research

Reviewing Your Project

Click on the Process flow tab in the left pane. Your process flow summarizes all of the tasks to this point, from reading in the dataset to filtering only the enrolled tribal members, through characterizing your data, creating new media usage variables and outputting your summary tables. It's a program without the programming.

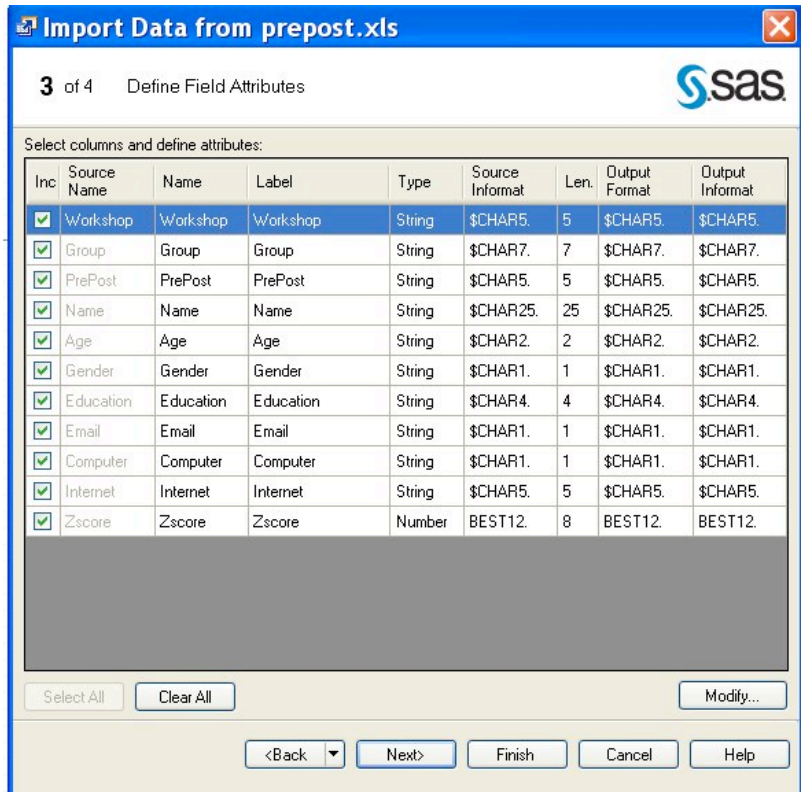


EXAMPLE 2: CREATING NEW DATA & DATA ANALYSIS

In this example, the data are from a study of the effectiveness of a training program. Participants were either in a control group that received no training or an experimental group that received online education. Each group was administered a pre-test and then took a post-test two months later. The client wants to know if there was a greater increase in test scores for the experimental group than the control group.

IMPORTING EXCEL DATA

From the **File** menu, select **Import Data**.
Select from the desktop the file named prepost.xls
Click on **NEXT** to accept all the defaults for input dataset and output dataset name in the first window.
Again, click **NEXT** to accept the defaults - variable names are given in the first row, you want to use the worksheet shown (this file only has one worksheet).

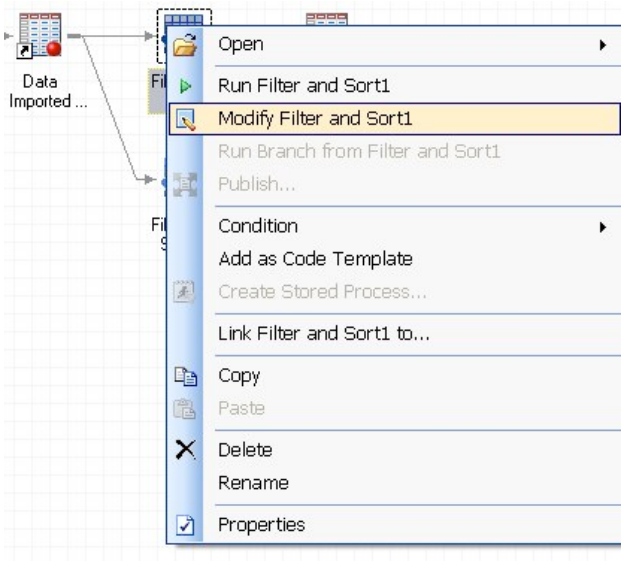


This is your chance to change any formats or names. If you are new to SAS, maybe you would not be interested in this. For now, click **NEXT**.
Click **FINISH**.

CREATING SUBSETS AND RE-RUNNING A TASK

The data are not as desired. Rather than one record for each person, the dataset has two, a pre-test record and a post-test record. You need to output the pre- and post-test records to separate files and then merge these two files by a unique identifier.

1. Click on **Filter and Sort.**
2. Select all variables.
3. Click on the **Filter** tab.
4. Select *Prepost* equal to PRE.
5. Click **Run**.



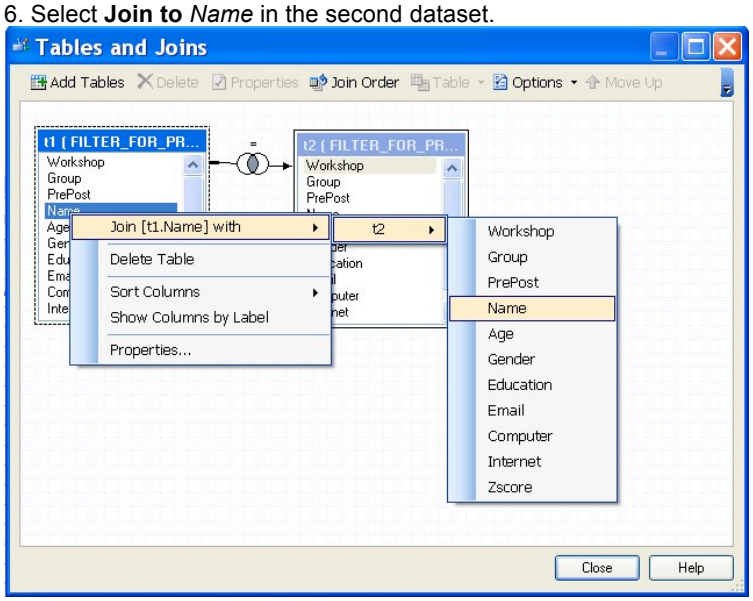
6. Right click on the Filter task.
7. Choose **Modify Filter and Sort.**
8. Click on the **Filter** tab.
9. Change the selection to *Prepos*t equal to POST.

The result of these two tasks is two datasets, one named PREPOST3, which contains the pre-test data and a second named PREPOST3_0000 that contains the post-test data. The next step is to merge these two files together.

MERGING (Joining) FILES

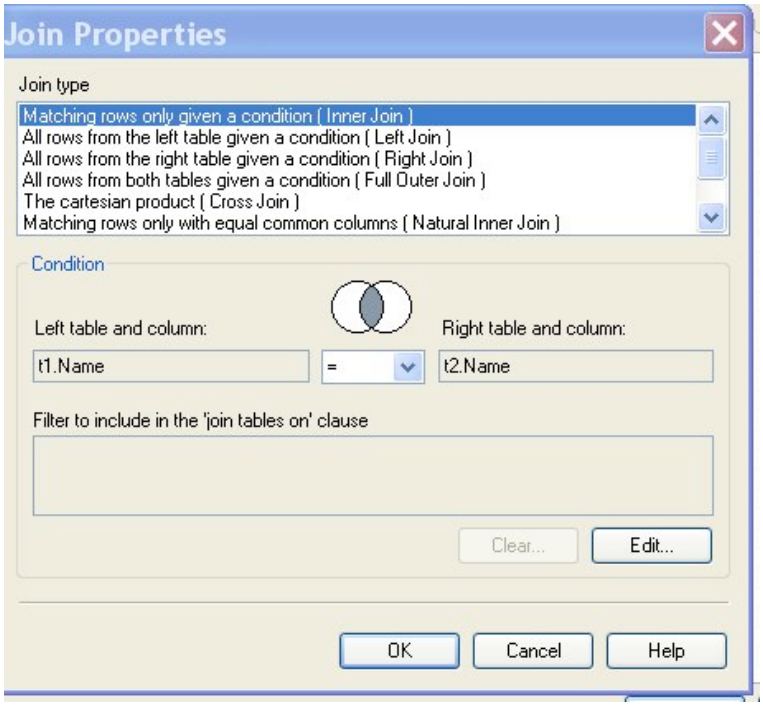
The previous task ends with a view of the newly-created dataset.

1. Click on the **Query Builder** button.
2. Click on the **Join Tables** tab.
4. Select pretest3 as the second table to join.
5. Right-click on *Name*.
6. Select **Join to** *Name* in the second dataset.

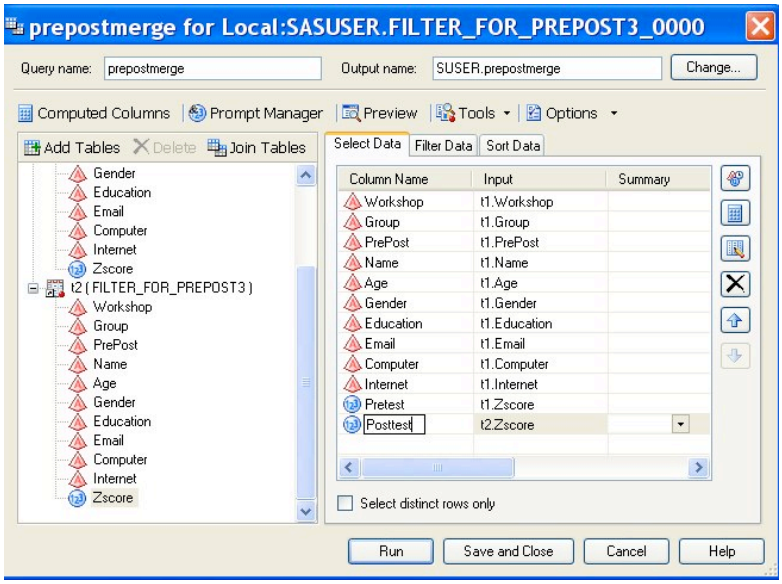


The Tables and Joins window now looks like the example above. A new window pops up with "Join Properties". Ignoring this may result in unexpected and very unwanted results.

The first Join type is selected here by default, (Inner Join) matching rows only on a given condition. This is the desired type.



7. Click **OK** .
8. Click **Close.**
9. Back at the Query Builder window again, drag over the variables desired. That would be all of the variables in the first dataset and just the last one in the second dataset.
10. Give the Query and Output descriptive names like prepostmerge.



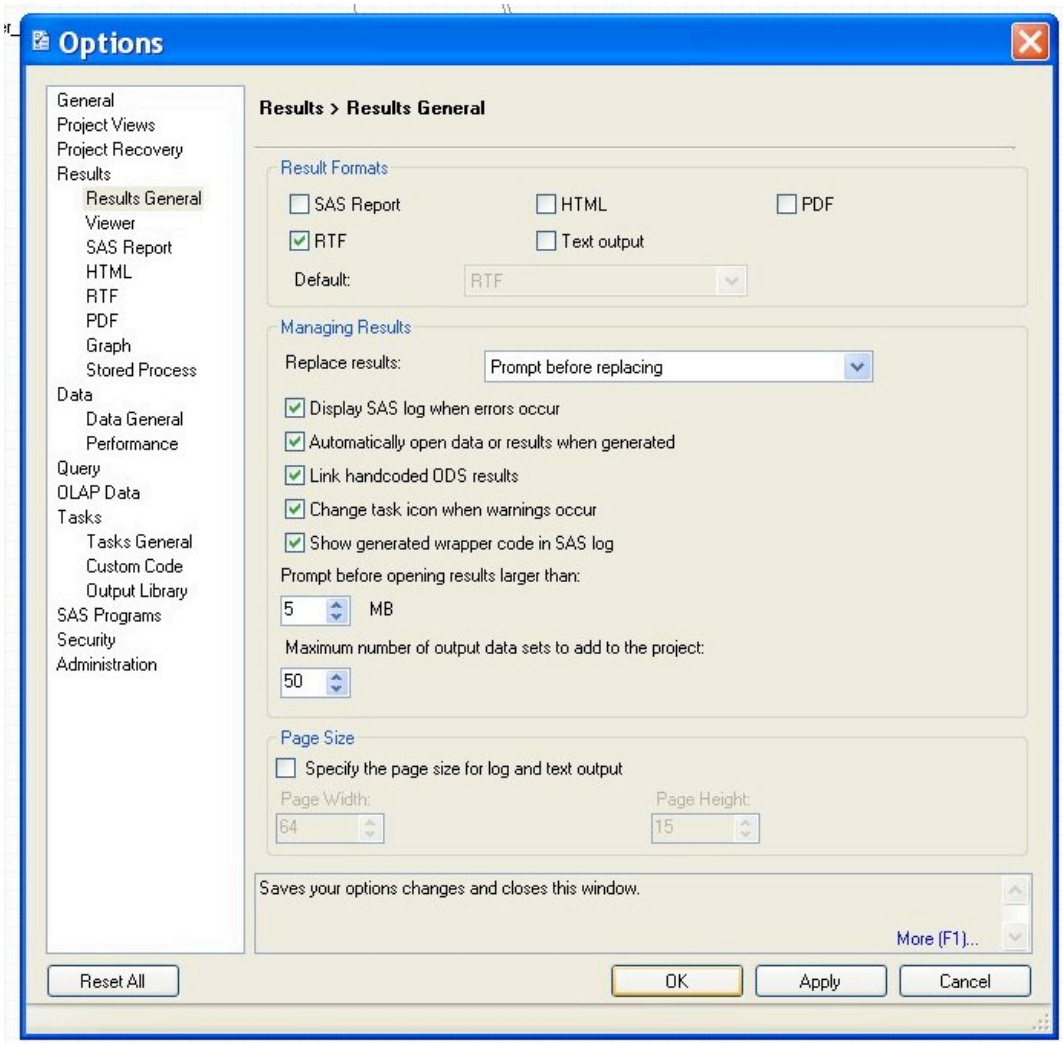
11. There are two variables named *Zscore*. Change these names to *pretest* and *posttest.*
12. Click **Run**.

Briefly, let's recap what you have done here. First, you have imported an Excel dataset, then output that data to two separate datasets of pre-test and post-test scores. Next, you merged those datasets back together by the respondents' names and renamed the variables.

Creating Output

As these results will be used in a report to the client, it is desirable to output to an RTF file that can be opened directly by Microsoft Word or other word processing program.

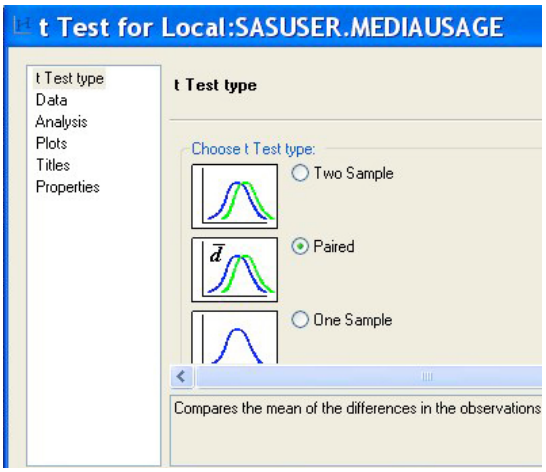
1.From the **Options** menu, choose **Tools**.
2. Click on **Results General**. (Although you might expect that clicking on RTF would be the way to get RTF output, that option actually specifies the type of RTF output desired.)
3. Under Results Format, click the box next to RTF.
4. Click **OK.**



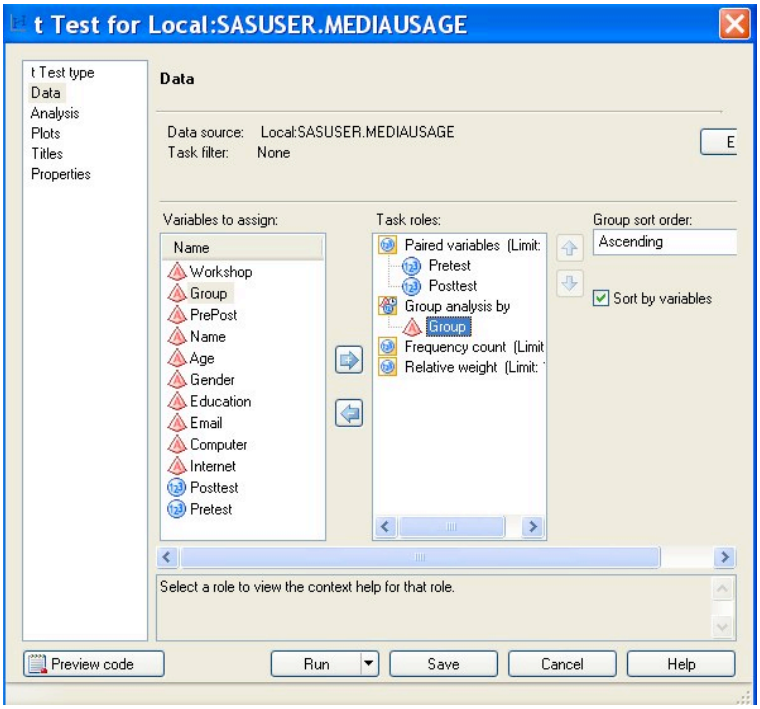
Comparing Groups

Now that you have the pretest and posttest scores together in one record, it is time to see if there is a significant difference between the two. Did the educational program actually work?

1. From the Tasks menu, select Analyze.
2. From the pull-down menu that appears, select ANOVA, select T-test
3. For t-test type, select Paired



4. Click on the Data tab in the left pane
5. In the Task Roles window, drag Group to Group Analysis by
6. In the Task Roles window drag Pretest and Posttest under Paired variables



7. Click **Run**

The t-test produces output that provides the pretest and posttest mean for each group, along with the number of subjects, standard deviation and tests of statistical significance.

t Test
The TTEST Procedure
Difference: Pretest - Posttest
Group=Control

N	Mean	Std Dev	Std Err	Minimum	Maximum
130	-2.8846	10.5144	0.9222	-50.0000	23.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-2.8846	-4.7092	-1.0601	10.5144	9.3730	11.9748

DF	t Value	Pr > |t|
129	-3.13	0.0022

t Test
The TTEST Procedure
Difference: Pretest - Posttest
Group=Exp

N	Mean	Std Dev	Std Err	Minimum	Maximum
111	-11.4414	14.2648	1.3540	-51.0000	27.0000

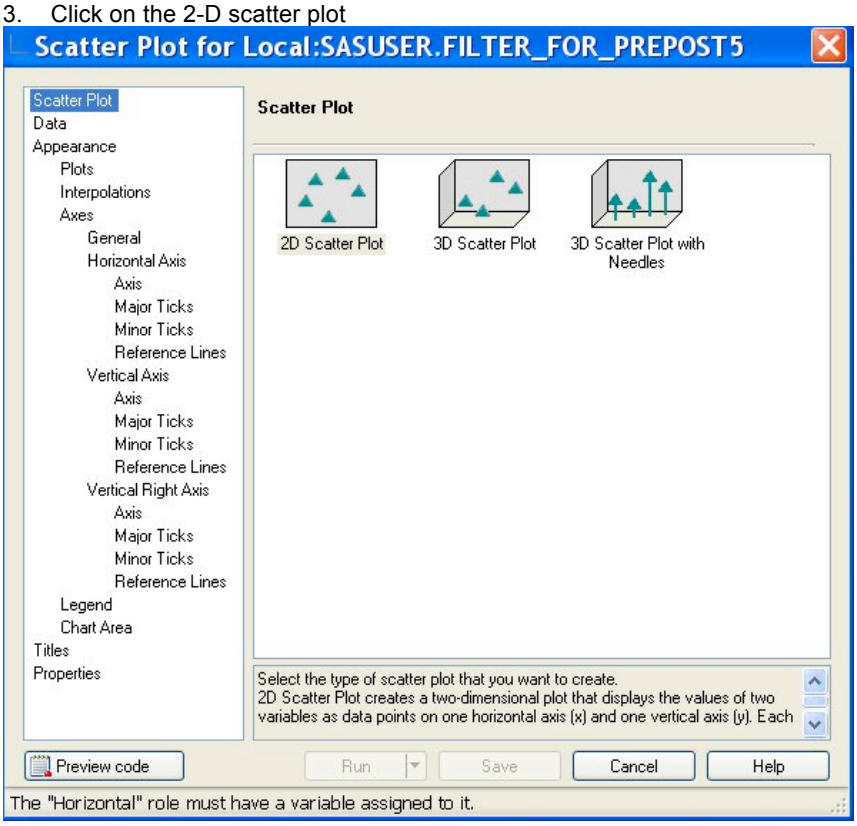
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-11.4414	-14.1247	-8.7582	14.2648	12.6032	16.4351

DF	t Value	Pr > |t|
110	-8.45	<.0001

Making a Graph

Perhaps the t-test results were not abundantly clear to you. For one last analysis, you may be interested to see the relationship between the pretest and posttest scores for those in the control group and the experimental group. An effective intervention with a reliable measure would show a strong relationship between pre-test and post-test for the control group and less of a relationship for the experimental group.

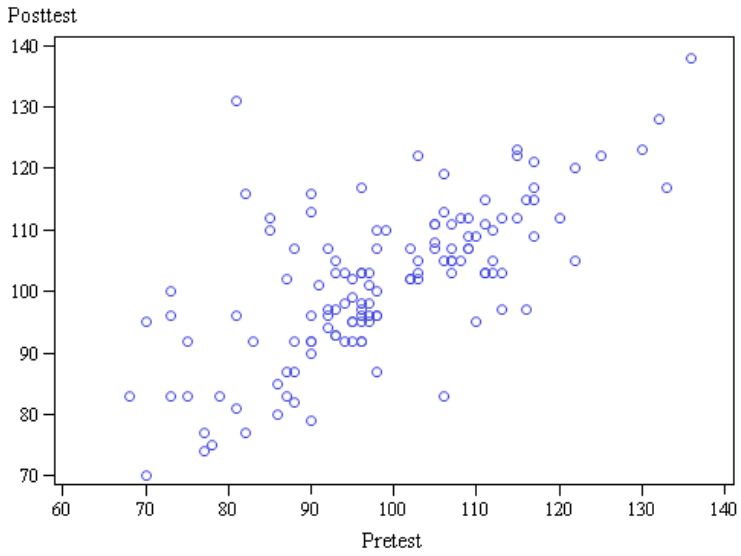
1. From the **Tasks** menu, select **Graph**
2. Select **Scatter**
3. Click on the 2-D scatter plot



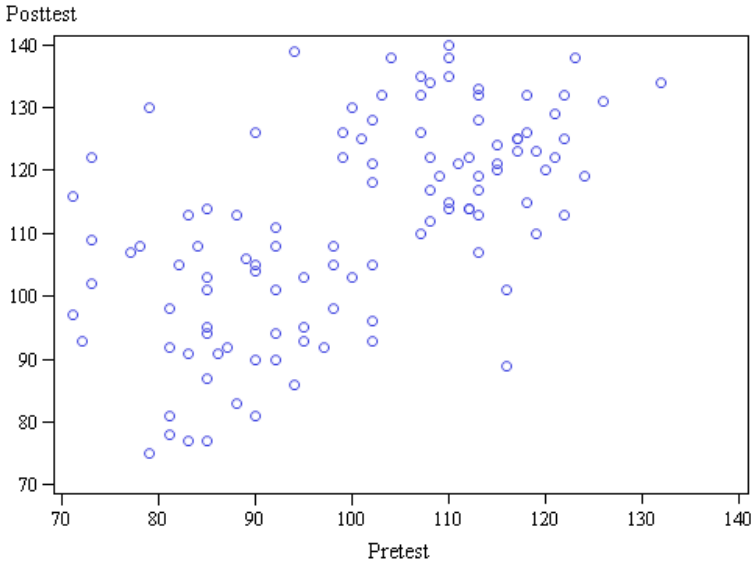
4. Click on **Data** in the left pane
5. Drag Pretest under horizontal variable
6. Drag Posttest under vertical variable
7. Click **Run**

The charts produced are shown below.

Group=Control



Group=Exp



CONCLUSION

Congratulations! In 90 minutes you have completed tasks that often take six months or more to learn as a SAS programmer. Enterprise Guide has a great many more capabilities that we have not even touched. With a few weeks of concerted effort, it will be an extremely powerful data analysis tool for you. Who knows, you may even become motivated to learn SAS programming.

SAS without programming is for programmers, too... SAS Enterprise Guide can be the solution of choice even for very experienced programmers. SAS code can be seamlessly integrated with Enterprise Guide. Code can be used within a project to create a dataset exactly the way you want it. Then, use the ease of the Enterprise Guide tables, graphics and statistical facilities for analysis.

In the current environment of "team science" and collaborative work groups, SAS Enterprise Guide is also a perfect solution to allow participation in design and analysis by non-programmers on your team who have specialized content area or statistical expertise.

References

Levesque, R. (2004). SPSS programming and data management: A guide for SPSS and SAS users. Chicago, IL : SPSS, Inc.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

AnnMaria De Mars
University of Southern California
3434 South Grand Ave
Los Angeles, CA
(213) 740-2840
ademars@usc.edu
Web: www.thejuliagroup.com/blog/