SAS ENTERPRISE GUIDE: AN OVERVIEW

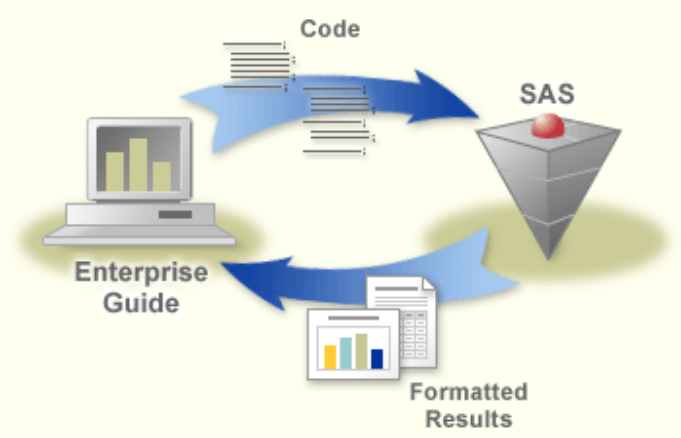
R.S. Tomar, Rajender Parsad, Seema Jaggi, Sanju and Sachin Kumar
I.A.S.R.I., Library Avenue, New Delhi – 110 012
tomar@iasri.res.in; rajender@iasri.res.in; seema@iasri.res.in; san.iss26@gmail.com; sachinhere@gmail.com

1. Introduction

SAS Enterprise Guide is an easy-to-use module on local computer as well as Windows client application that provides the following features:

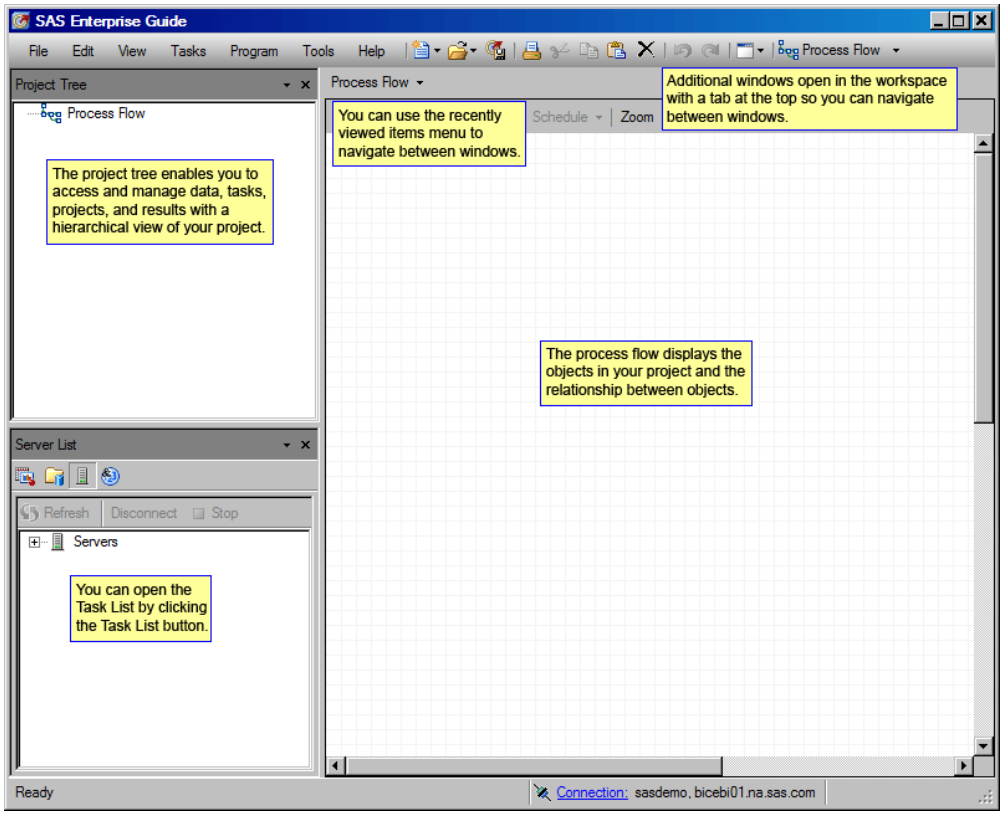
- Access the functionality of SAS
- An intuitive, visual, customizable interface
- Transparent access to data
- Ready-to-use tasks for analysis and reporting
- Easy to export data and results to other applications
- Scripting and automation
- A code editing facility

SAS Enterprise Guide can connect to SAS on local computer or SAS server. SAS Enterprise Guide generates SAS code as soon as we access data and build tasks,. When we run a task, the generated code is passed on to SAS for processing and the results are returned to SAS Enterprise Guide.



SAS Enterprise Guide also connects to a SAS Metadata Repository where information about objects is stored.

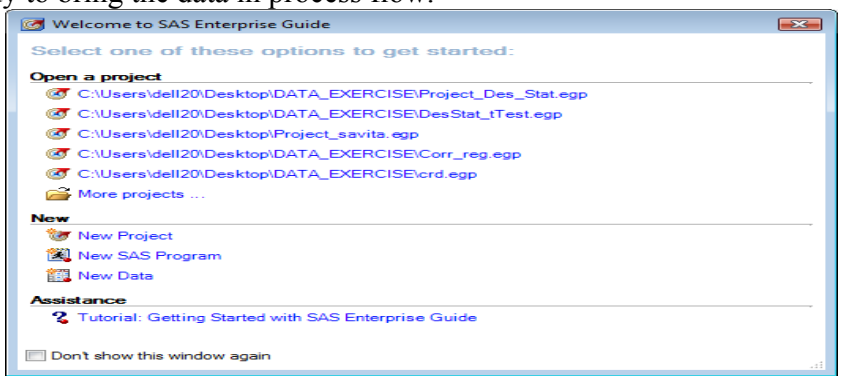
When we start SAS Enterprise Guide first time, the windows are arranged in the default application layout. This layout consists of the project tree, the Server List window, and the workspace area. The workspace area displays data, code, logs, task results, and process flows. First of all, the process flow window gets opened in the workspace area. When we open data or generate reports, other windows open in the workspace in tabbed interface fashion. We can use the recently viewed items menu in the upper-left corner of the workspace to navigate between the windows.



If one wants to customize the layout by closing, opening, or changing the position of windows, gets automatically saved on exiting from SAS Enterprise Guide. If we want to restore the default layout, we can select **Tools→ Options**, and then click **Restore Window Layout**. If we close one of the application windows and want to restore it, we can select the window name from the **View** menu.

2. Start Enterprise Guide

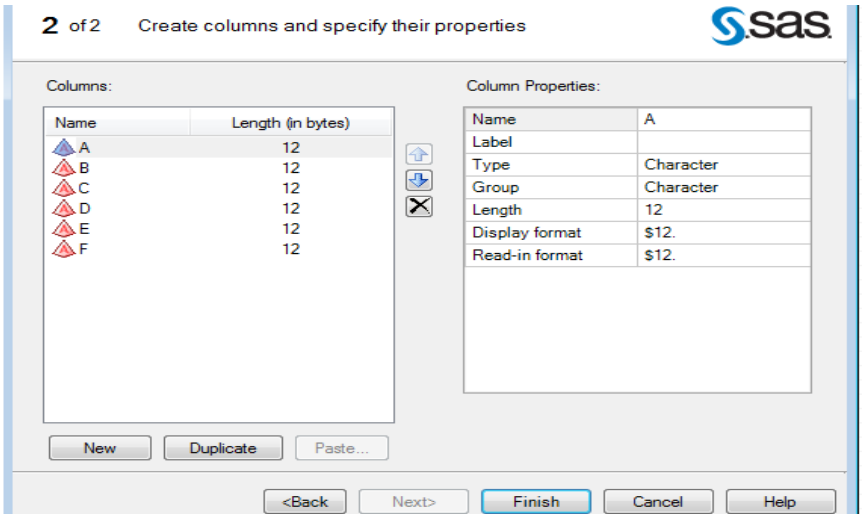
To open SAS Enterprise Guide, click the **Start Menu→ SAS→ Enterprise Guide 4.2** (or the version of Enterprise guide available on the system) or just click the icon of SAS EG on desk top. At the time of opening the Enterprise Guide, a Welcome window appears on the screen containing few options, select **New Project,** it will not ask the project or file name but Enterprise guide gets ready to bring the data in process flow.



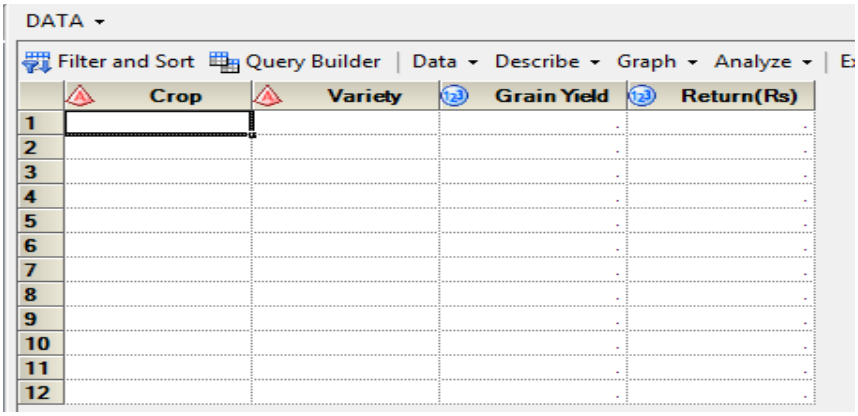
Welcome window can be closed by clicking on the cross button without opening New Project. In case SAS Enterprise Guide is already open then the new project can be opened by selecting, **File→ New→ Project.**

Data Entry in SAS EG

One can enter the data directly in the data sheet of enterprise guide. At the time of opening of the enterprise guide, an option "NEW DATA" appears in welcome window . On clicking **New Data → Next→ Type the data name** and select the required library and then click on **Next**. Enterprise Guide default data sheet appears on the screen with default variables ie A, B,C...etc. with default variable length 12 bytes.

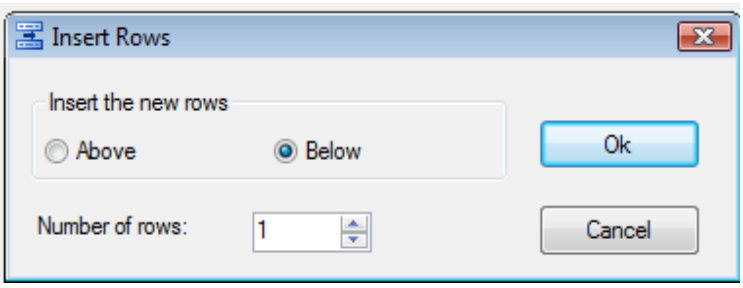


Now we can type the our own variable name. First select the default variable ie A, B,C...etc from left side "Columns' window than click on right side 'Columns properties' window and then type the name variable in place of default variable name ie A, B...etc, by clicking on type we can change the type of variable i.e. Character or Numeric similarly we can change the length and format of the variable according to our requirement. After making the necessary entries, delete rest of the default variables by clicking the cross (X) button and then click **finish.** A data sheet with our own variables look like as following snap shot.

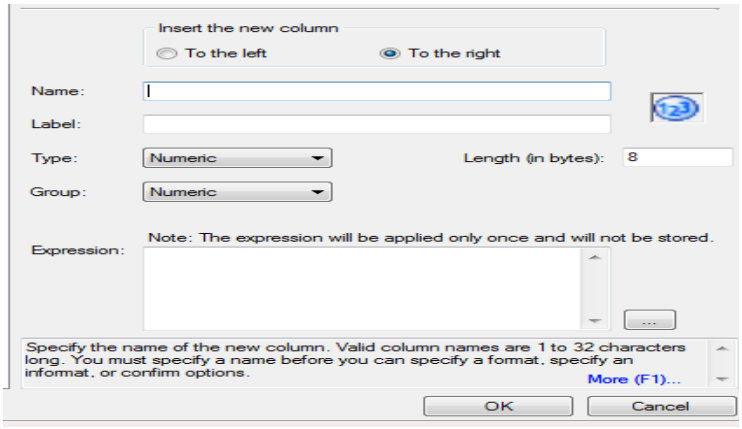


In case one forget these step for data entry then Click **File →New →Data ,** the above mentioned windows will appears. As soon as data entries are completed one has to protect the data sheet by following the steps **Edit→ Protect Data**. If one forget to protect the data sheet and start performing the task in enterprise guide, it will ask automatically to protect the data sheet.

One can add or any number of row and columns to the data sheet after unprotecting the data sheet. To add row, select the row where one wants to add the new rows, click right mouse button, select insert row option from drop down menu, following dialogue box will appear on the screen with options. If one wants delete the row select the delete row option from the drop down menu.



To add the numbers of columns, select the column where one wants to add a column or numbers of columns and then click right mouse button a dialogue box will appear where one has to make desired entry to add the columns.



If data file is already created in any ASCII format it can be imported in enterprise guide. The steps are:

File → Import Data → Local Computer → Desktop (Location of data file) → Folder name (in which the data file is available) → File name (containing data) → Open → Finish or

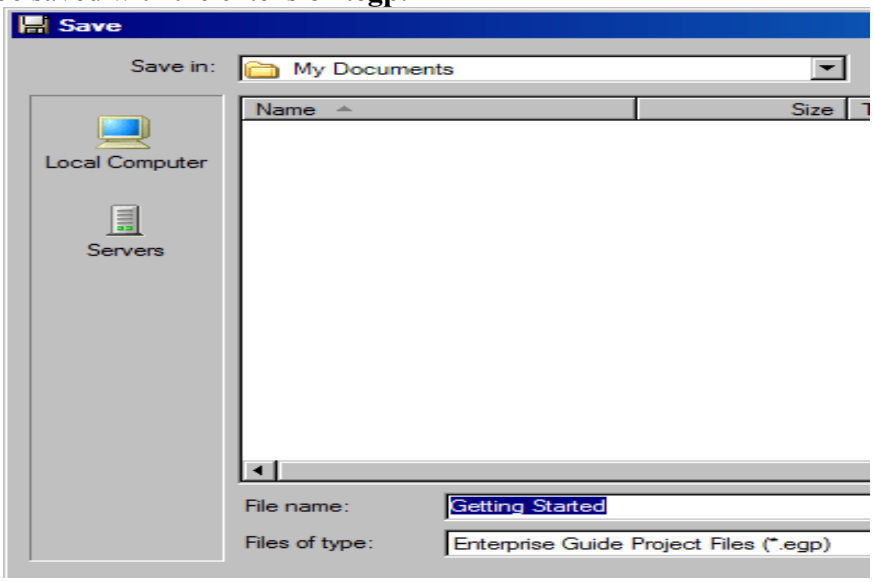
File → Open → Data → Local Computer → Location of data file (Folder name) → Double click the file → Finish

3. Save the Project

One can save a project and its contents to any location, including a location of the server. Projects are saved as a single file.

1. Select ***File→ Save Project As***.
2. The Save window opens and prompts to choose whether to save the project on local computer or on a server. (default location is local computer, one cannot save on server until the connectivity with the server is established)

In the Save window, select a location for the project. In the **File name** box, type Project file name, it will be saved with the extension **.egp**.



Now click the save.

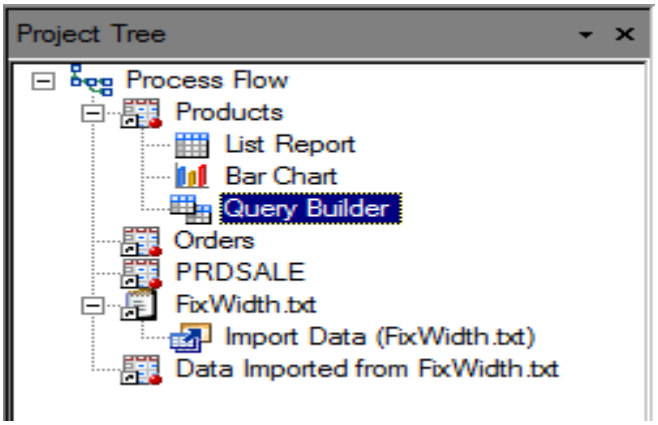
Please note that the any component of the project data file of the result file can be saved as SAS Report, HTML, RTF, PDF, Graph, Stored Process, etc. by clicking on **Tools → Options** and then selecting sub-options.

4. To Open Saved Project

The Project already saved, can be opened by clicking **File→ Open →Project → Local Computer or Servers** (depending upon the location of the project), and then select project from the location where it is saved. All task performed earlier will be shown in the process flow but to get it activated we have click the Run.

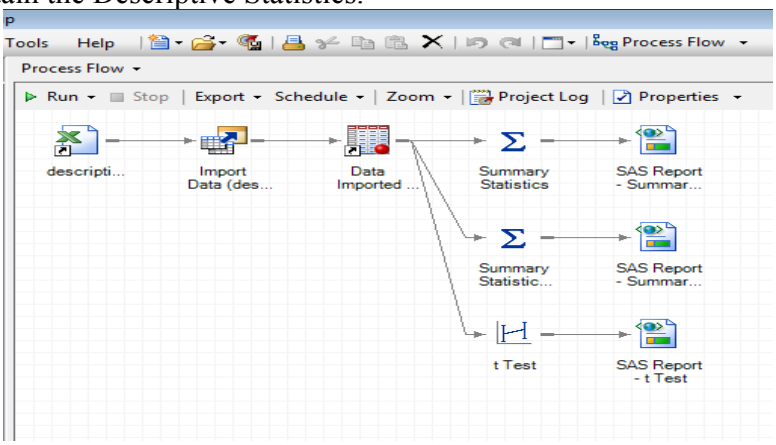
5. The Project Tree

A project is a collection of related data, tasks, programs, and results. The project tree displays a hierarchical view of the active project and its associated data, programs, notes, and results. One can use the Project Tree window to manage the objects in the project. Items in the project tree can be deleted, renamed and reordered.



6. The Workspace and Process Flow Windows

On creating a new project, an empty Process Flow window opens. As soon as we add data, run tasks, and generate output, an icon for each object is added automatically to the process flow. When we give the run command, the objects in the process flow runs in the same order as they are displayed in process flow. In the following Process Flow window, the SAS data of MS Excel is imported to obtain the Descriptive Statistics.

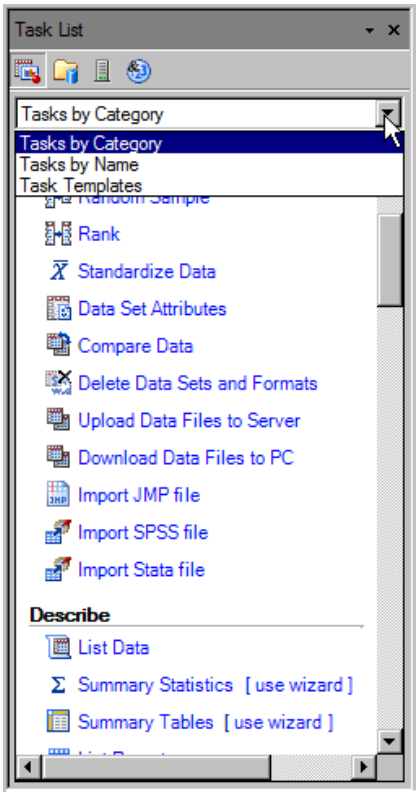


7. The Task List

The Task List is hidden by default, but it can be displayed by clicking the **Task List** button in the resources pane in the lower-left corner of the workspace. We can use tasks to do everything from manipulating data, running specific analytical procedures, creating reports. Many tasks are also available as wizards for this choose tasks and wizards by using the Task List or by using menus it can provide a quick and easy way to use some of the tasks. The views of Task List are as under:

(i) **Tasks by Category**: View lists individual tasks, grouped by type.
(ii) **Tasks by Name**: View lists individual tasks alphabetically.
(iii) **Tasks by Name:** View also lists the SAS procedures that are related to the task.
(iv) **Task Templates:** View to save our settings for a specific task to a template. We can then run that template with any input data source.

In window shown below, upper portion show that we can choose the type of task where the lower portion which is blue in colour, shows the different parts of task which we have selected.



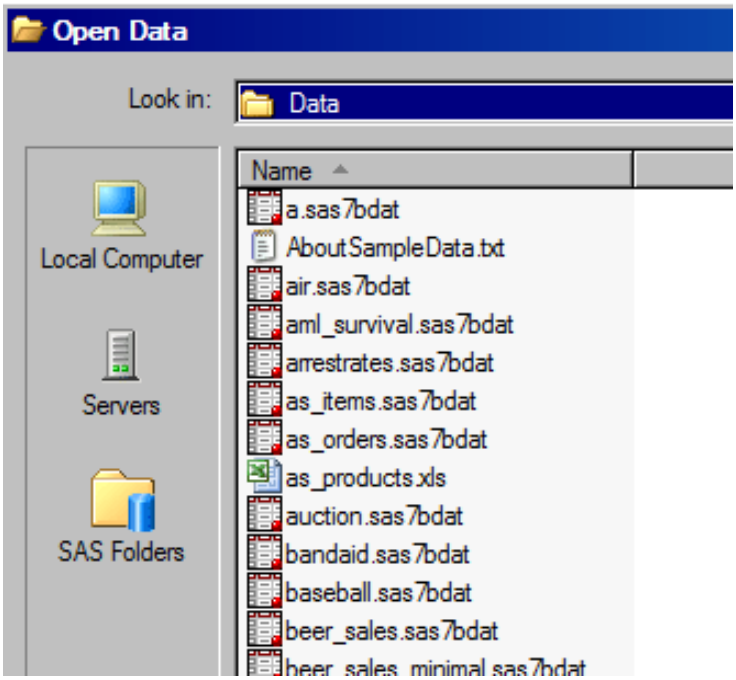
8. Data Access in SAS Enterprise Guide

Enterprise Guide provides the facility to access the data kept in the following formats:

- SAS data sets (files containing data that is logically arranged in a form that SAS can understand)
- Data tables from databases (DB2 and Oracle that use licensed SAS database engines)
- Local data in other formats (such as Excel, Access, Lotus, Text, HTML, ODBC, and OLE/DB)
- OLAP cubes (with a connection to an OLAP server)
- SAS Enterprise Guide can open and run tasks on various types of data but if we want to edit the data, it must open as a SAS data set. SAS Enterprise Guide enables us to import many data files to create SAS data sets.

Local and Remote Data

When we open data in SAS Enterprise Guide, we must select whether we want to look for the data on local computer, a SAS server, or in a SAS folder.

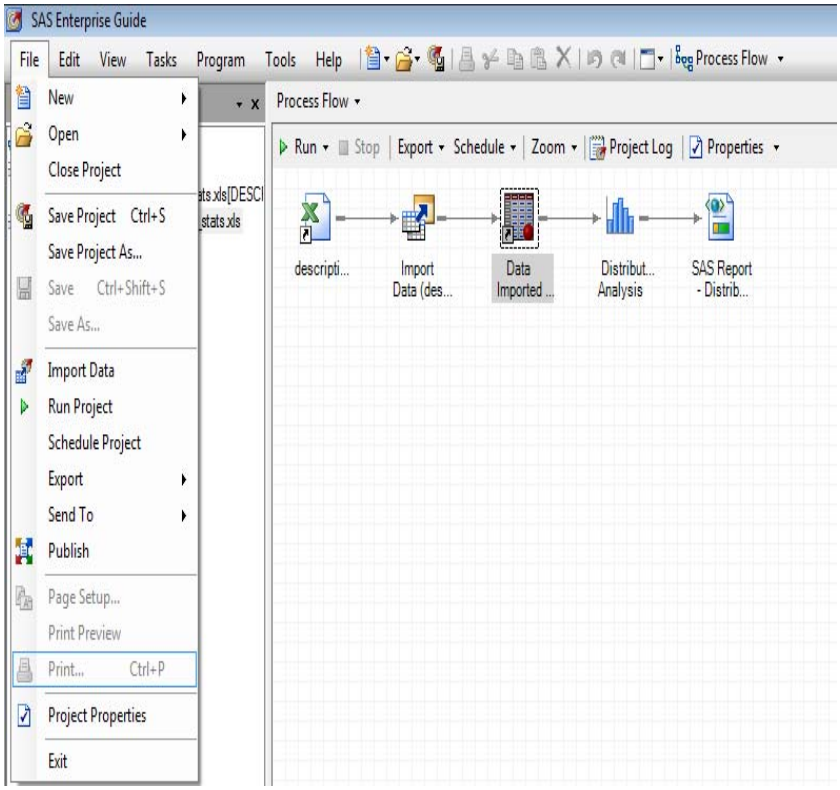


On clicking **Local Computer**, by browsing the directory structure of local computer, one can open any type of data file that SAS Enterprise Guide can read. On clicking **Servers** as shown in the above snap shot, it looks for our data on a server. In the server there are icons that we can select for **Libraries** and **Files**. Libraries are shortcut names for directory locations that SAS knows about. Some libraries are defined by SAS, and some are defined by SAS Enterprise Guide. Libraries contain only SAS data sets.

9. Menu and Sub Menu

On opening SAS Enterprise Guide, Window shows following list of menu at the top of the screen. On clicking the main menu item list of tasks appears in drop down sunmenu.

On clicking **File** it shows submenu as shown below



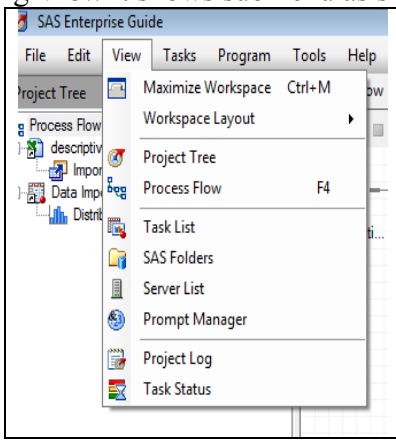
New: To open the new project or data or programme etc.
Open: To open the already existing project or data or programme Report OLAP Cube, Information Map etc. etc.

Close Project: To close the project presently opened.

Save Project: It is used to save the already created project and available presently in the process flow

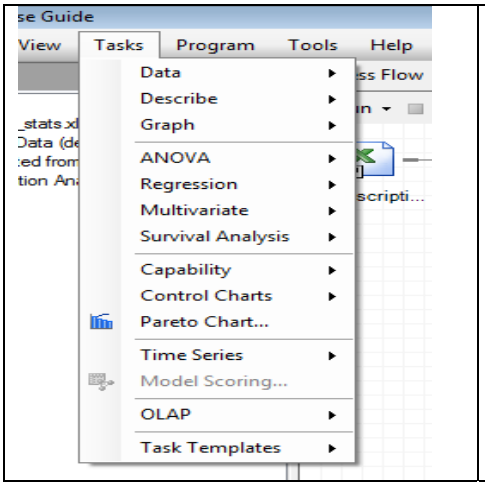
Save Project As: To save the project which has not been created already.

On clicking **View** it shows submenu as shown below



View: To view the Project tree, Process flow, Task list, SAS Folder, Prompt Manager, Task Status etc.

On clicking **Tasks** it shows submenu as shown below



Task: This part of menu contains the submenu *Data, Describe, Graph ANOVA, Regression, Multivariate, Survival Analysis, Time Series, OLAP* etc. Which are required for data management and statistical analysis? These are described in brief in the sequel:

Data: This part of sub menu contain the sub menu (i) Filter & Sort (ii) Query Builder (iii) Append Table (iv) Sort data (v) Creating Format (vi) Transpose (vii) Random Sample (viii) Rank.

Describe: All the components of descriptive statistics like mean, median, mode, skewness, kurtosis etc. can be obtained by using the submenu's of Describe. The submenu are : (i) List data (ii) Summary Statistics Wizard (iii) Summary Statistics (iv) Summary table Wizard (v) Summary table (vi) Distribution Analysis (vii) One Way Frequencies (viii) Table Analysis.

Graph: Different types of graph like Bar Chart, Pie Chart, Line Plot, Scatter Plot can be prepared.

ANOVA: t Test, One way ANOVA, Non Parametric, Linear Model and Mixed model statistics can be obtained by using this part of submenu.

Regression: Different regression analysis linear, non linear, logistic and general linear model can be obtained by regression submenu.

Multivariate: The correlation, Canonical Correlation, Principal Component Analysis, Cluster Analysis and Discriminant Analysis statistics are available here.

Time Series: It includes Prepare Time Series Data, Basic Forecasting…, ARIMA Modelling and Forecasting…, Regression analysis with Autoregressive errors…, Regression analysis of Panel Data…, Create Time Series Data etc.

10. Assign Library

By default SAS files are temporarily saved in work library but we can create our own library to keep our file for further use.

- To create the own permanent library for saving the project, we click the **Tools → Assign Library → Name,** here type the name of library. The name must be all uppercase and can

contain a maximum of eight characters. The name must be unique for each server on which the library is created. The name cannot contain any of these characters / : * ? " < > .

- Select the server from the list of available servers. By default it is Local Computer.
- Click **Next** to select the engine for the library. The types of engine are:
  a. **file system:** If one selects **File System** as the engine type then select the **Let SAS choose the engine based on the contents of the specified path** check box. Do not select this check box if the library path contains mixed file types. If one selects **BASE**, **V9**, **V8**, **V7**, **V604**, or **V6** as the engine, then no need to specify a path. For all other engines, the path is required then Click **Browse** to select location.
  b. **database system:** Select the engine to use when accessing files in the library. Select the external database server for the library. Specify the database schema used to access the data on the server. The schema name that we provide must match the name of a schema that has already been defined on the database server. Enter a user ID and password for the database server
  c. **WebDAV:** Select the WebDAV server to use as the library source. Enter the physical path where the library resides on the server in the **Path** box. Enter a user ID and password for the WebDAV server.

11. Graphs in Enterprise Guide

Enterprise Guide provides the facility to draw the various types of graphs. To get the graph options one has to click **Tasks→Graph**, It displays list of options from where one can choose the desired types of graph.

Bar Chart

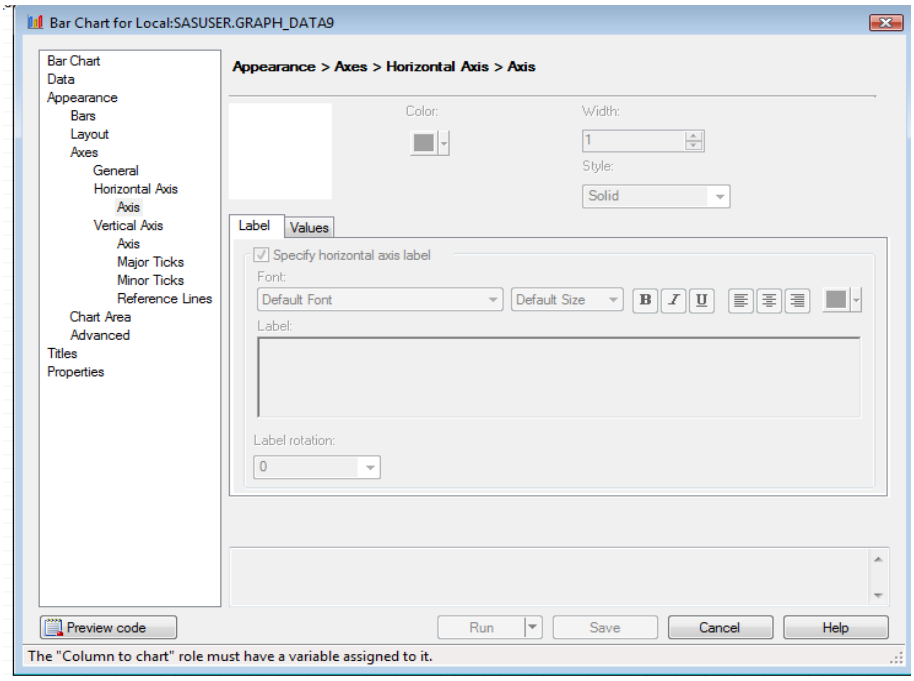
(i) Bar Chart Wizard

To create the bar chart one has to click **Tasks →Graph →Bar Chart wizard→Next→Bars**(from drop down list select the variable for x axis) → **Bar height** (select the variable of y axis from drop down list) → **Next→ Color** (For same colour for all bars select '**All bar same'** for different colours of bars select '**Bar Category'**) → **Data label** (select from the drop down list as per requirement) → **Click Axis label** type x axis title in front of 'Bars' and y axis title in front of 'Bar height'→ **Next** type Chart title in front of ' Graphs' → **Finish.** There are options for 2D and 3D bar charts which can be selected as per requirements.

(ii) Bar Chart

To create the bar chart one has to click **Tasks →Graph →Bar Chart,** types of bar charts (i) Simple (ii) Grouped (iii) Stacked (iv) Grouped/Stacked (v) 3D Grouped etc appears on the screen. Select the desired type of bar chart and then from the selection pane click **Data,** by default, the data source is that which we have selected before opening the task. Under the **Appearance** heading, click **Bars** to access these options. There are option to have custom colours and number of bars. One can specify the different colours for each bar and can **Specify number of bars** by check box to specify how many bars appear in the chart. By default, the number of bars is determined automatically. Click **Layout** to access these options. By default **2D** check box is selected, the bar size is set automatically and outline colours of bar is black. To create a three-dimensional chart, clear the **2D** check box and select a bar shape from the **Shape** drop-down list. One can specify the bar width or spacing between the bars by selecting the

option from drop down list. Under Axes option one can suppress axes and tick marks by clicking **General** in the selection pane and then select the **Turn off Axes and Ticks** check box. Click on **Horizontal** axis in selection pane , select label and type the x-axis title, one can change the font and position of label by selecting required button. Now click **Vertical** axis in the selection pane, select label and type the y axis title in the label window and then select the label rotation 90 so that y-axis title will appear along the axis, by default it will appear at top of y-axis



Pie Chart

The Pie Chart task creates simple, group, or stacked charts that represent the relative contribution of the parts to the whole by displaying data as slices of a pie. Each slice represents a category of data. The size of a slice represents the contribution of the data to the total chart statistic. One can draw the pie chart by using the following steps.

1. Select **Tasks →Graph→ Pie Chart**.
2. In the **Pie Chart** gallery, We have to select **Type of Pie chart**
3. In the selection pane, click **Data**. We have to assign the **Variable Name** column to the **Column to chart** role so that the pie chart has a slice for each category. Assign the **Variables Name** column to the **Sum of** role so that the size of each slice represents the relative contribution of each of the variable we have assigned in column to chart.
4. In the selection pane, click **Layout**. Click the **Percentage** drop-down list we have to select position **Inside, Outside** etc.
5. In the selection pane, click **Titles**. In the **Section** box, select **Graph**. Clear the **Use default text** check box, delete the default title, and type our own title.
6. In the **Section** box, select **Footnote**. Clear the **Use default text** check box and delete the default footnote text and our own footnote if any, and Click **Run**.

Now let us try to draw the different graphs by considering the following example

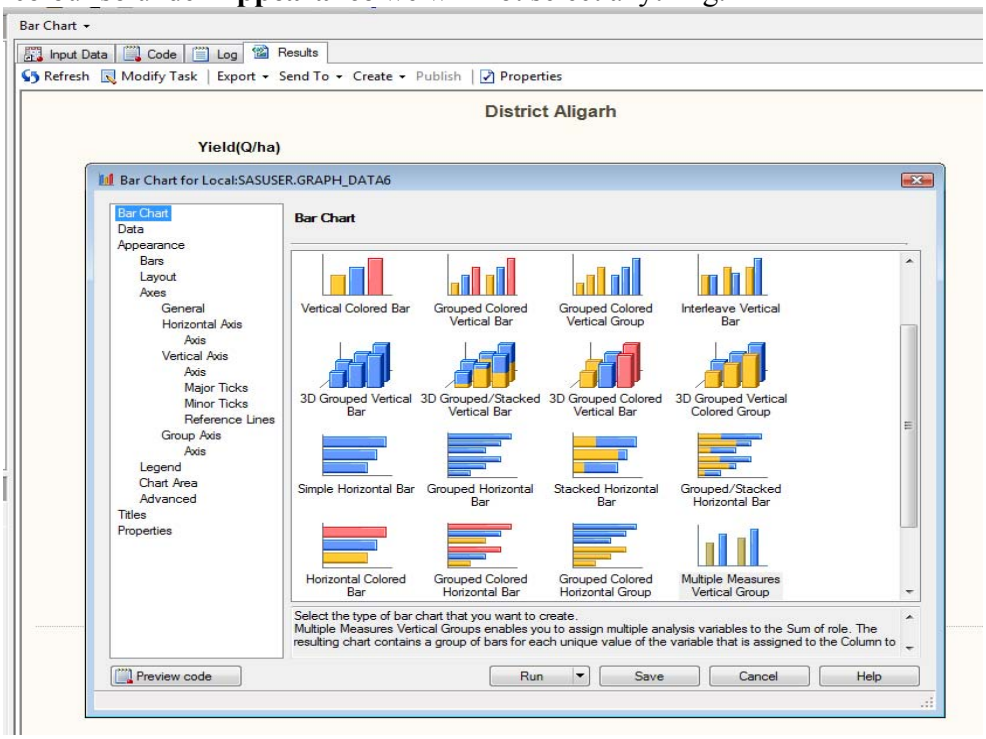
Example 11.1: The average yield of barely crop in five district of Uttar Pradesh as under

	Aligarh	Bulandshahar	Ghaziabad	Meerut	Bijnor
2000-01	18.02	19.05	20.09	30.33	23.01
2001-02	22.03	25.06	20.05	29.78	30.08
2002-03	17.09	27.33	20.1	32.55	35.08
2003-04	20	23	21.08	32.45	29.34
2004-05	26.03	24.98	20.45	33.8	30.55

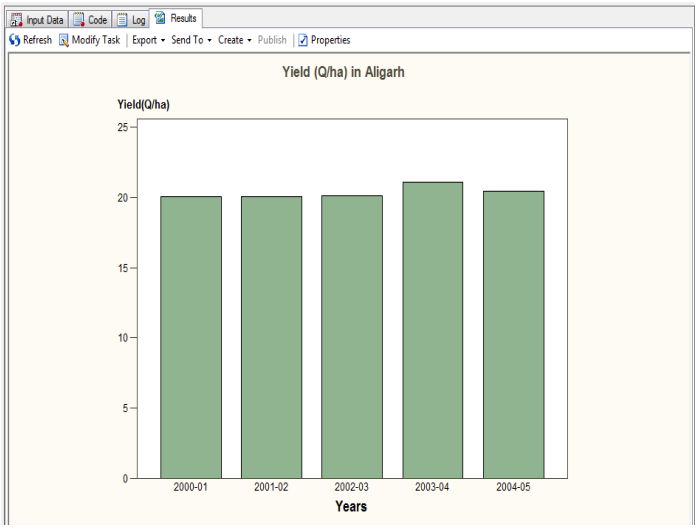
i) Draw the bar chart for each district separately; ii) Draw the bar chart all district for all year on single axis iii) draw Pie Chart iv) Draw line graph for all five district on single axis.

Solution:

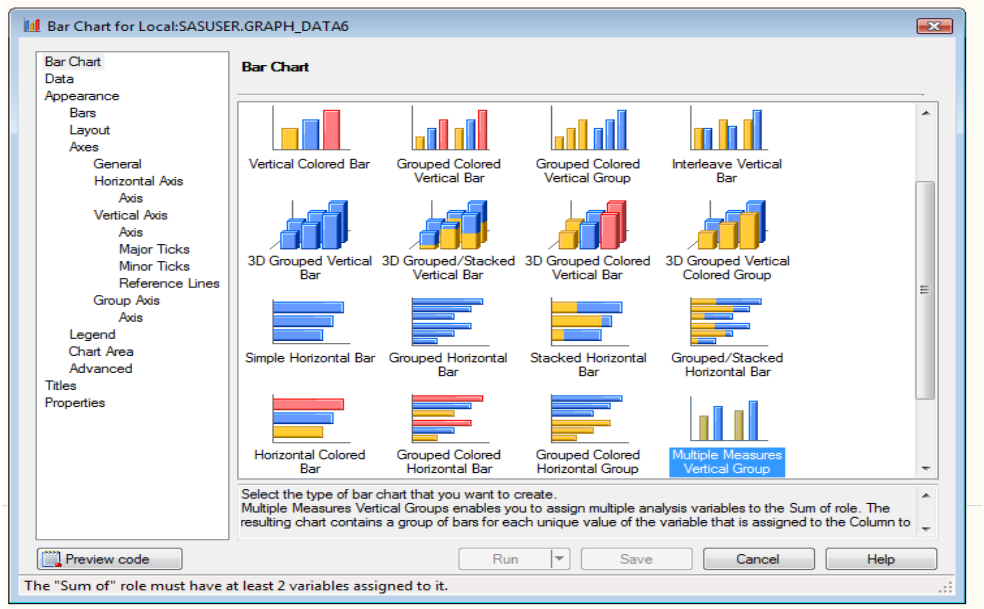
(i) : a) **Tasks →Graph →Bar Chart→ Select Simple Vertical Bar** b) From the selection pane click **Data,** From Column to Assign, the select **Year,** drag it to task role column and drop under **Column to chart** c) select Aligarh and drop it under **Sum of** d) Since we are not selecting any particular colour so under **Appearance** we will not select anything.



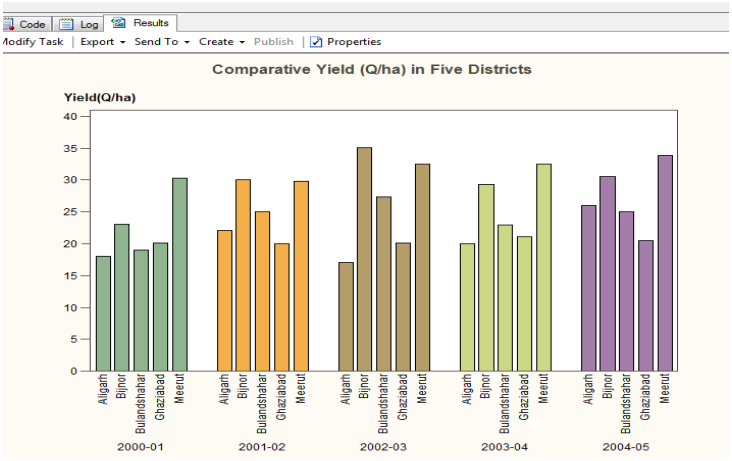
e) Select **Vertical Axis** from selection pane**,** Click **Label and** check the box **specify vertical axis label** and type **Yield(q/ha)** under the Label **,** select **90** from the drop down list of label rotation f) In the selection pane, click **Titles**. In the **Section** box, select **Graph**. Clear the **Use default text** check box, delete the default title, and type the chart title"Yield (Q/ha) in Aligarh". And clear the check box for footnote as we are not interested in footnote and then Click **Run.** Following the steps of bar chart we obtain the Bar Chart as under



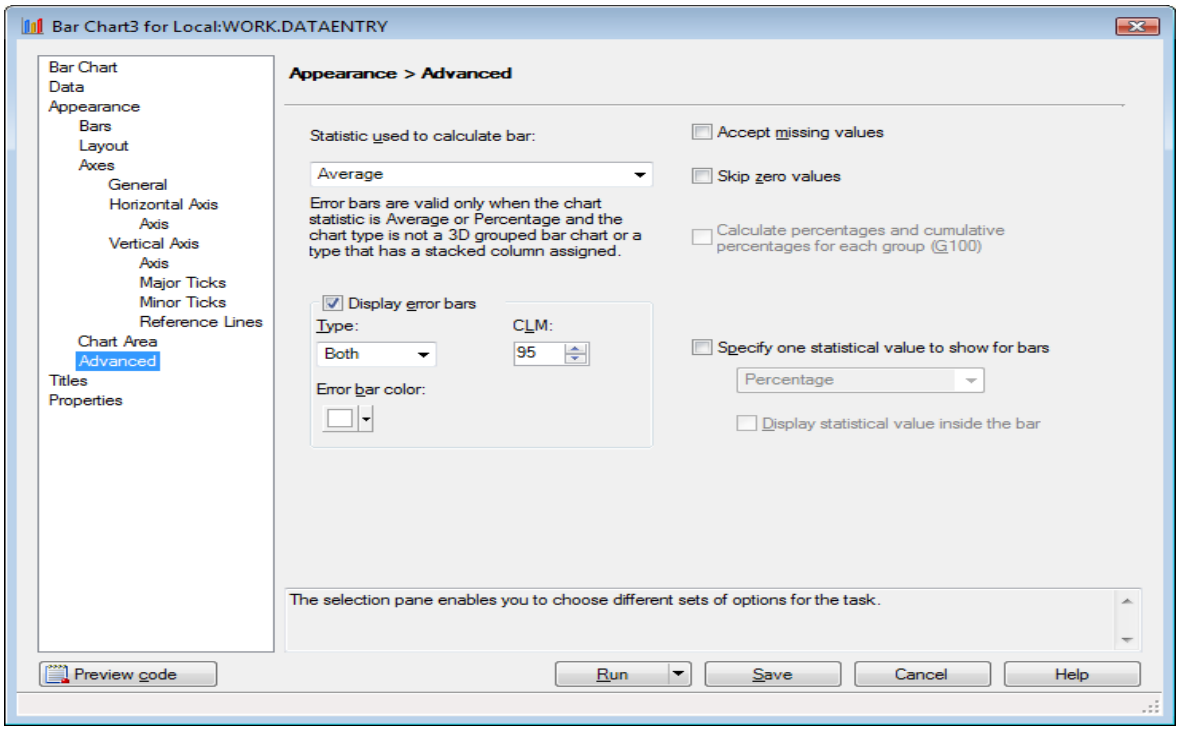
Solution (ii) : a) **Tasks →Graph →Bar Chart→ Select Multiple Measure Vertical Group** and double click it b) From the selection pane click **Data,** From Column to Assign, the select **Year,** drag it to task role column and drop under **Column to chart** c) select **Aligarh Bulandshhar, Meerut Bijnor Ghaziabad** and drop it under **Sum of**



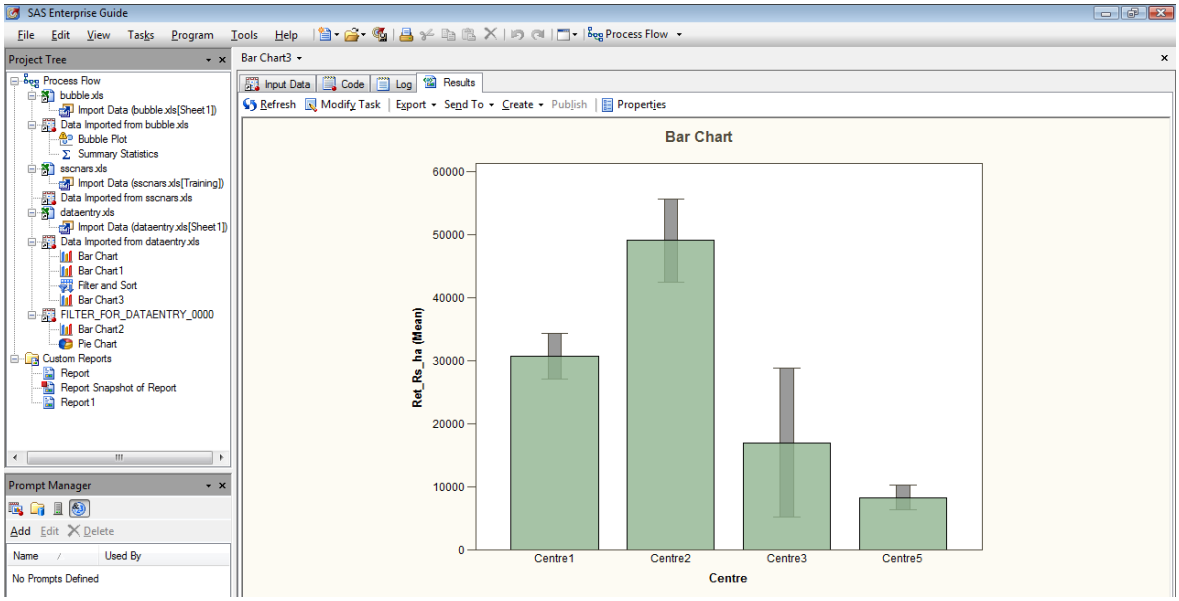
d) Select **Vertical Axis** from selection pane**,** Click **Label** and check the box **specify vertical axis label** and type **Yield (q/ha)** under the Label e) In the selection pane, click **Titles**. In the **Section** box, select **Graph**. Clear the **Use default text** check box, delete the default title, and type the chart title ”Comparative Yield (q/ha) in Five Districts”. And clear the check box for footnote as we are not interested in footnote and then Click **Run.** We obtain the Bar Chart as under



Sometime it is required to obtain the bar graph with Standard Error Bars on Bars. To obtain standard error bars our data must be replicated otherwise Standard Error bar cannot be drawn. Now select **Advanced** tab → **Statics used to calculate bars** (Select Average from drop down) → Check the box **Display Error Bars**.



Finally click on RUN. We will obtain the Standard Error bar chart as under:-



Solution (iii) Pie Chart

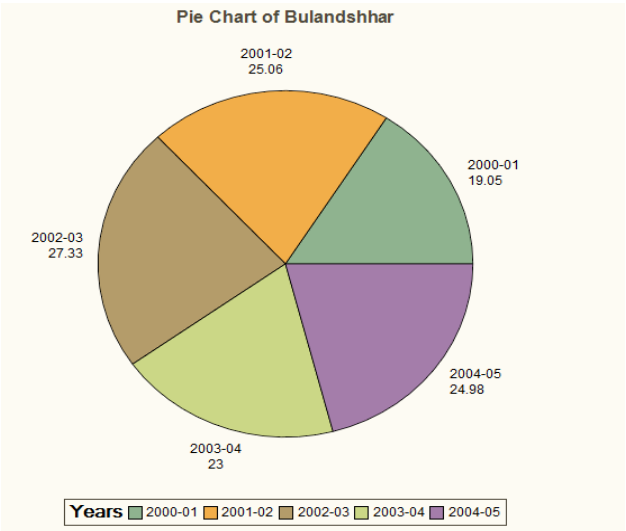
a) Select **Tasks→ Graph→ Pie Chart**.

b) In the **Pie Chart** gallery, select type of pie chart Simple Pie, group or stacked pie. Here we are selecting **Simple Pie.**

c) In the selection pane, click **Data**. Assign the **Year** column to the **Column to chart** role so that the pie chart has a slice for each category. Assign the **Bulandshhar** column to the **Sum of** role so that the size of each slice represents the relative contribution of each category to overall profit.

d) In the selection pane, click **Layout**. In dimension select the radio button 2D or 3D. We are selecting 2D. Now click the **Percentage** drop-down list and select **Outside**

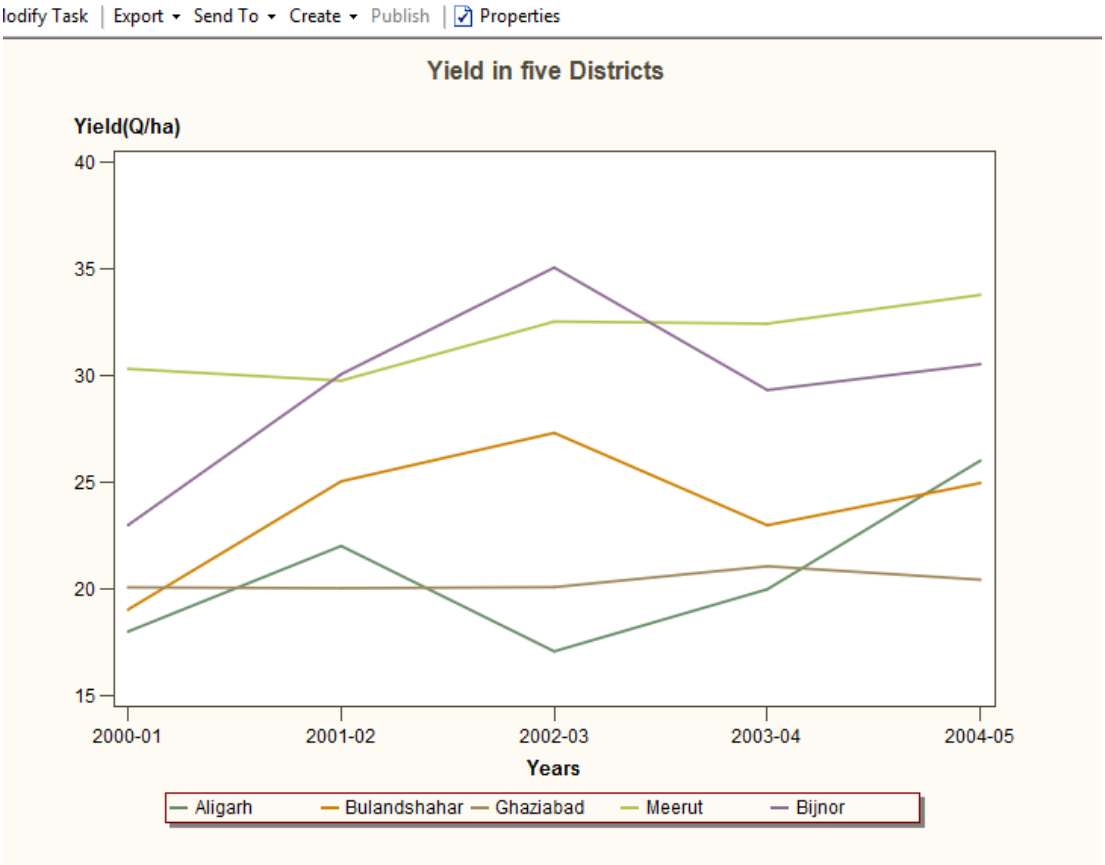
e) In the selection pane, click **Titles**. In the **Section** box, select **Graph**. Clear the **Use default text** check box, delete the default title, and type the chart title" Pie Chart of Bulandshhar". Similarly we can select Footnote and can give the footnote we desire and then click **Run**.



Solution (iv): Line Graph

a) **Tasks →Graph → Line Plot → Select Multiple Vertical Line Plot** and double click it b) From the selection pane click **Data,** From Column to Assign, the select **Year,** drag it to task role column and drop under **Column to chart** c) select **Aligarh Bulandshhar, Meerut Bijnor Ghaziabad** and drop it under **Verticle** d) Select **Vertical Axis** from selection pane**,** Click **Label and** check the box **specify vertical axis label** and type **Yield(q/ha)** under the Label e) In the selection pane, click **Titles**. In the **Section** box, select **Graph**. Clear the **Use default text** check box, delete the default title, and type the chart title" Yield in Five Districts". And clear the check box for footnote as we are not interested in footnote and then Click **Run.**

We obtained the following output of the line graph



Scatter Plot and Bubble Chart

Scatter plots are used to plot data points on a horizontal and a vertical axis to see how much one variable is affected by another. Each row in the data table is represented by a marker whose position depends on its values in the columns set on the X- and Y-axes. Multiple scales can be used on the Y-axis to when one wants to compare several markers with significantly different value ranges.

A bubble chart is a variation of a scatter plot in which the data points are replaced with bubbles, and an additional dimension of the data is represented in the size of the bubbles. Just like a

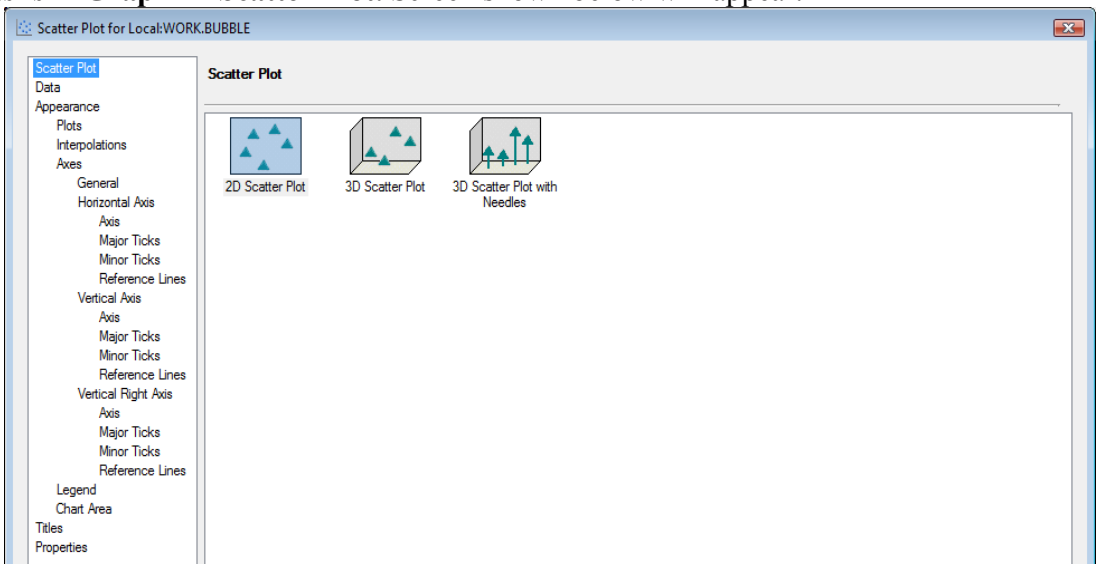
scatter plot, a bubble chart does not use a category axis — both horizontal and vertical axes are value axes. In addition to the x values and y values that are plotted in a scatter plot, a bubble chart plots x values, y values, and z (size) values. For example, if one use a scatter plot to illustrate the relationship between gas mileage (y) and weight (x) for various automobiles, the size of the scatter points might be determined by the cost (z) of the automobiles. Bubble plots use circles of varying sizes to summarize data in which the radius (r) of each circle is proportional to the size of the data value (z).

Example 11.2: Considering the population projection data five state of India. In 2012, population (Millions), distance travelled by people of the states by train and the gross domestic product of the state is

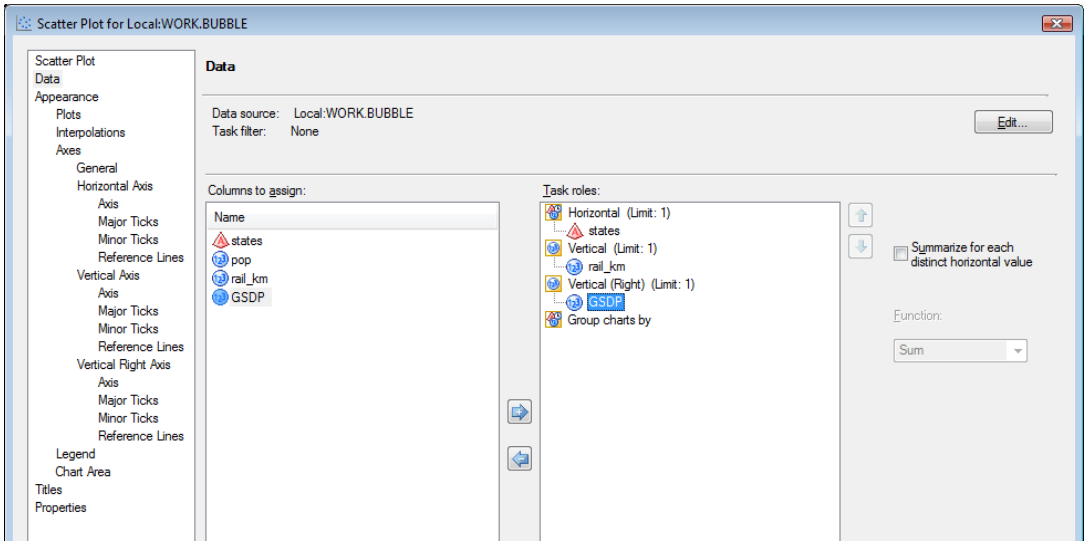
States	Population (million)	Distance Travelled by Train (km) (rail_km)	Gross Domestic Product of State (GSDP)
Punjab	28.16	2,156	1565
Tamil Nadu	68.11	3,943	4165
Uttar Pradesh	206.31	8,800	4200
Uttarakhand	10.17	345	609
West Bengal	90.8	4,000	3336

Solution: Scatter Plot: To draw scatter chart we have consider the two variable GSDP and rail_km from the given data. Now select

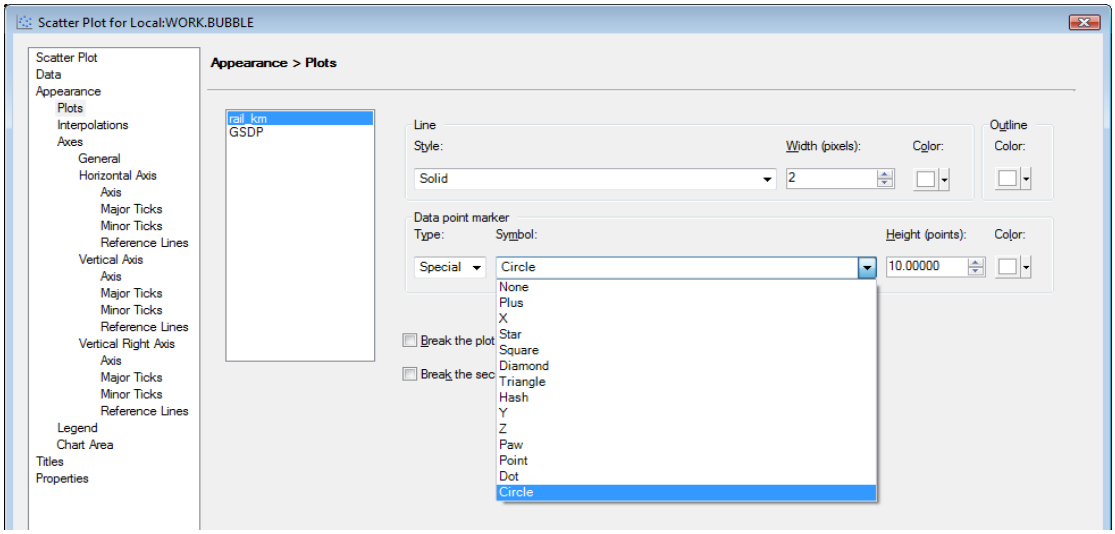
a) **Tasks →Graph → Scatter Plot.** Screen shown below will appear.



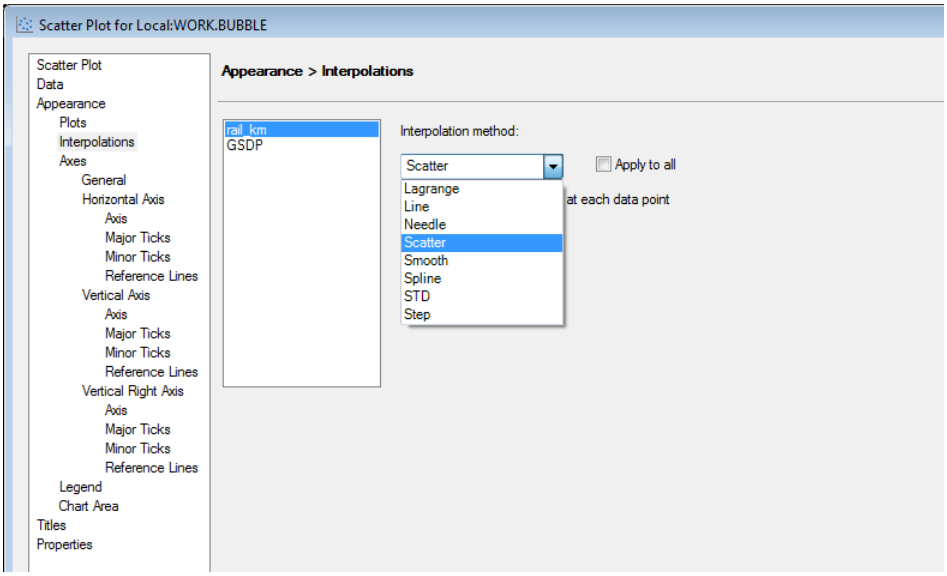
b) Select the desired 2D or 3D scatter plot and then select the Data tab from selection pane window. Now select the variables from **Column to assign** window to **Task role** window. For example, select variable **State** to give the task role **Horizontal** ie X-axis variable. Similarly selecting the variables **rail_km** and **GSDP** and send to task role window under **Vertical** and **Vertical (Right)** tabs respectively.



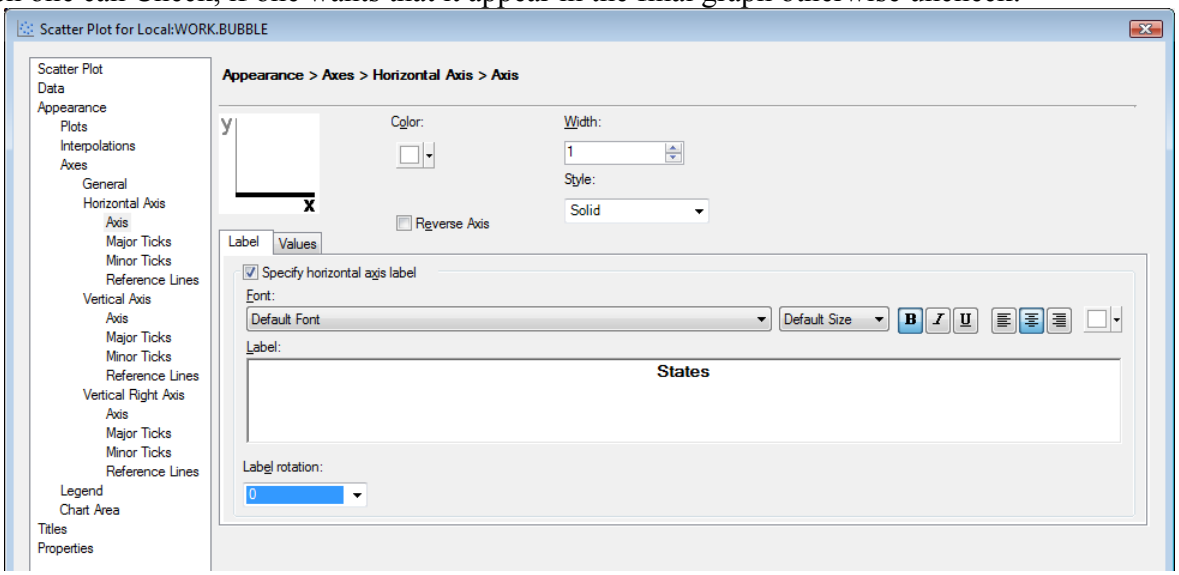
c) Under **Appearance** tab selecting **Plot *options,*** following window will appear on the screen. From drop down menu under data point maker select the circle (or any other style).



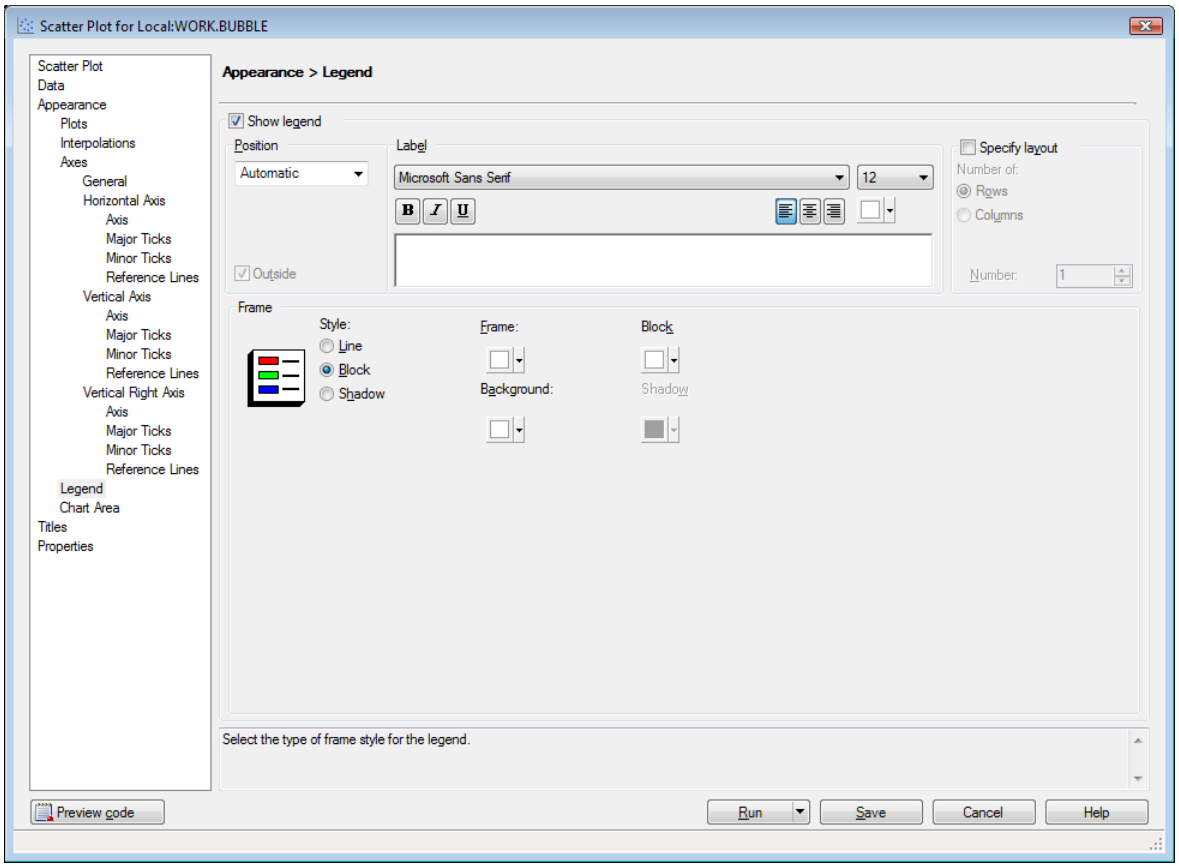
d) Select **Interpolations** tab and then select the method **Scatter** as shown in the screen shot



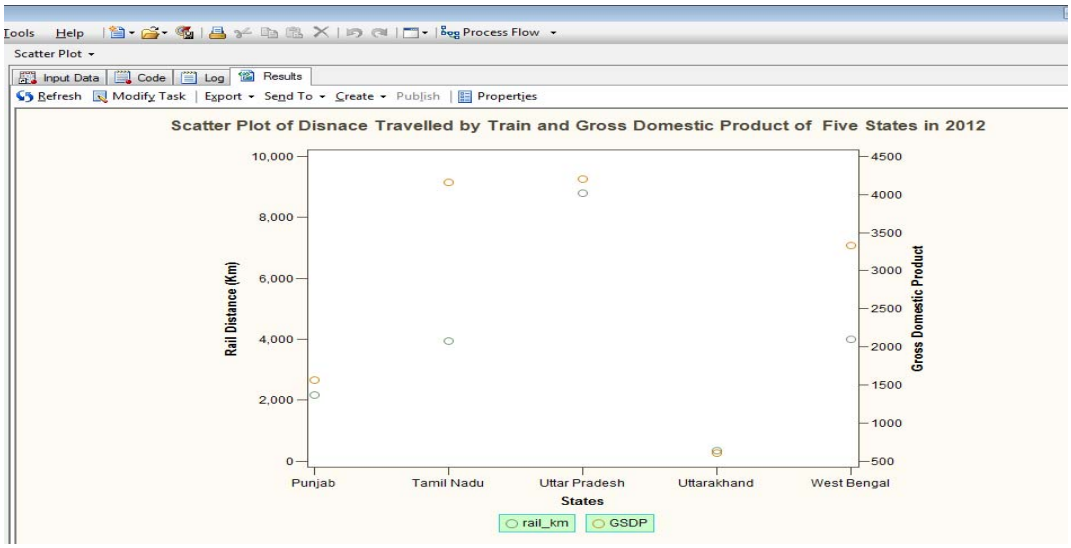
e) Select the **Axes** tab. In Axes tab, **Horizontal Axis** (X-axis) under **Label** tab type the X-axis title and keep label of rotation default (0 degree). Similarly select the **Vertical** and **Vertical (Right)** and type the Y-axis (Left and Right) titles under the label tab and select the label of rotation 90 for both the axes. There are options for major and minor ticks and reference lines which one can Check, if one wants that it appear in the final graph otherwise uncheck.



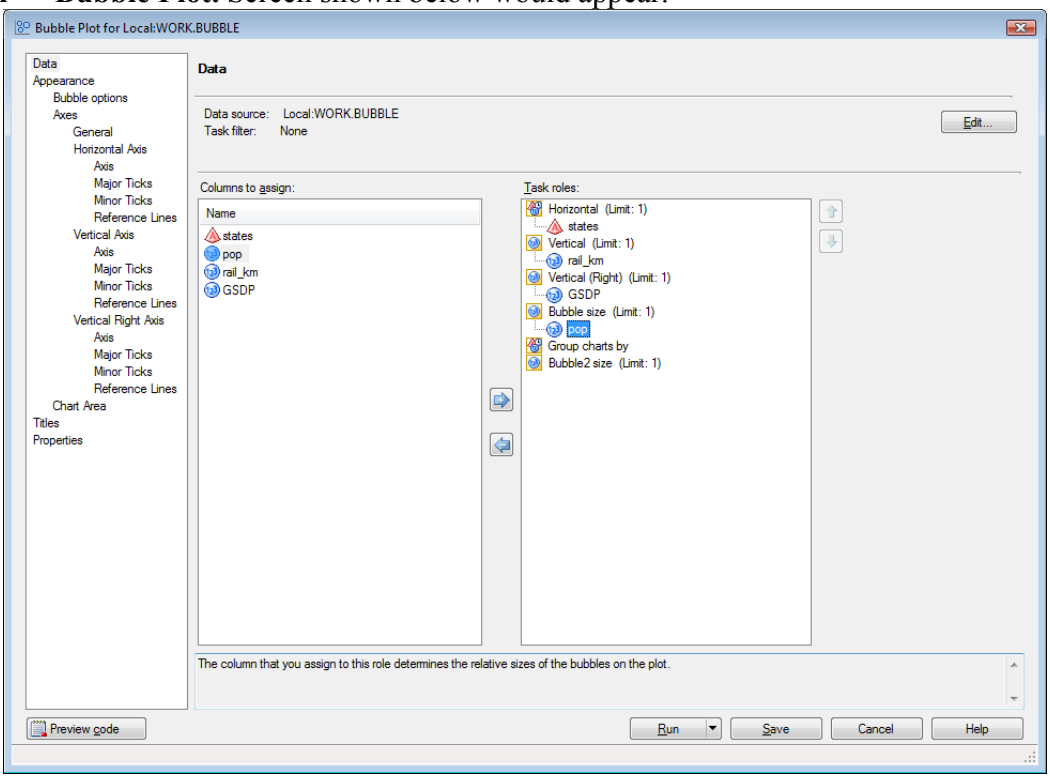
f) One can select the legend and can fix the position of legend in the graph by selecting the desired or required available options. In case one does not want legend in the graphs then unselect the **Show legend tab**. Leave the Chart Area option default. To write the title and footnote of Chart select **Titles** option and then select the **Graph** tab, uncheck the **Use default text** box and type the chart title in space provided. Similarly select the **Footnote** option to write the footnote. Leave the Properties option default.



g) Select the **RUN**. Output is as under:

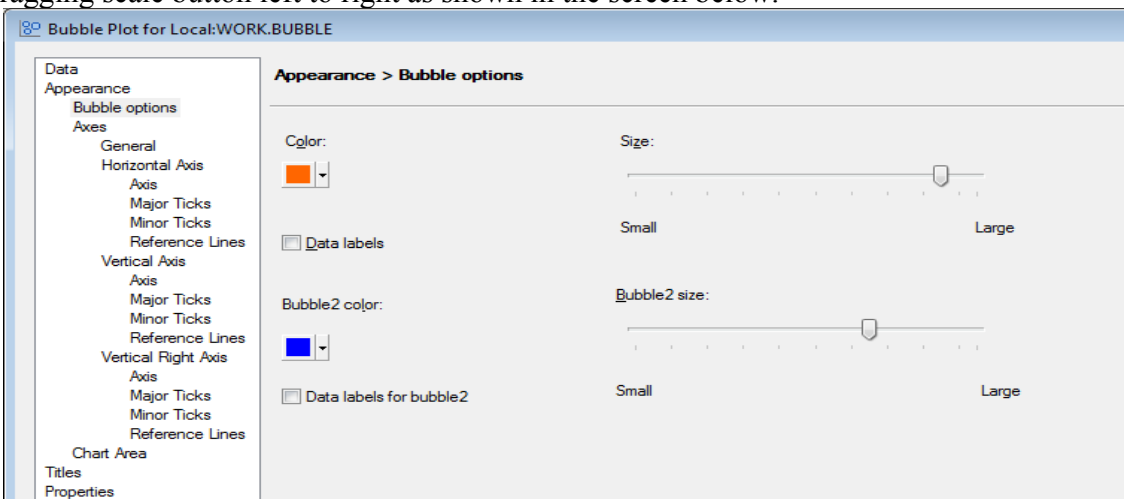


Bubble Chart: To draw bubble chart, one has to consider all the three variable. GSDP and rail_km will be X-Y axis variables and Population as Z-axis variable. Now select a) **Tasks →Graph → Bubble Plot.** Screen shown below would appear.



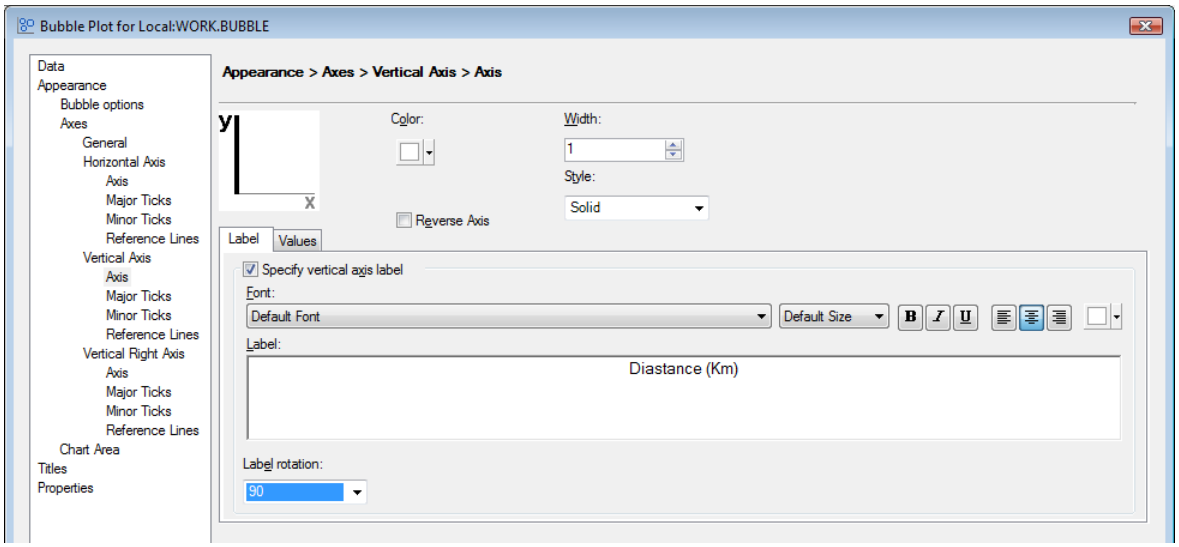
b) Select **Data** from task pane (left most column of the window). Move variable **State** from **column to assign** window to **task roles** window under **Horizontal;** rail_km and GSDP variables under **Vertical** and **Vertical (Right)** and population variable under **Bubble size.**

c) Under appearance select Bubble Option and select the color and size of bubbles small to large by dragging scale button left to right as shown in the screen below.



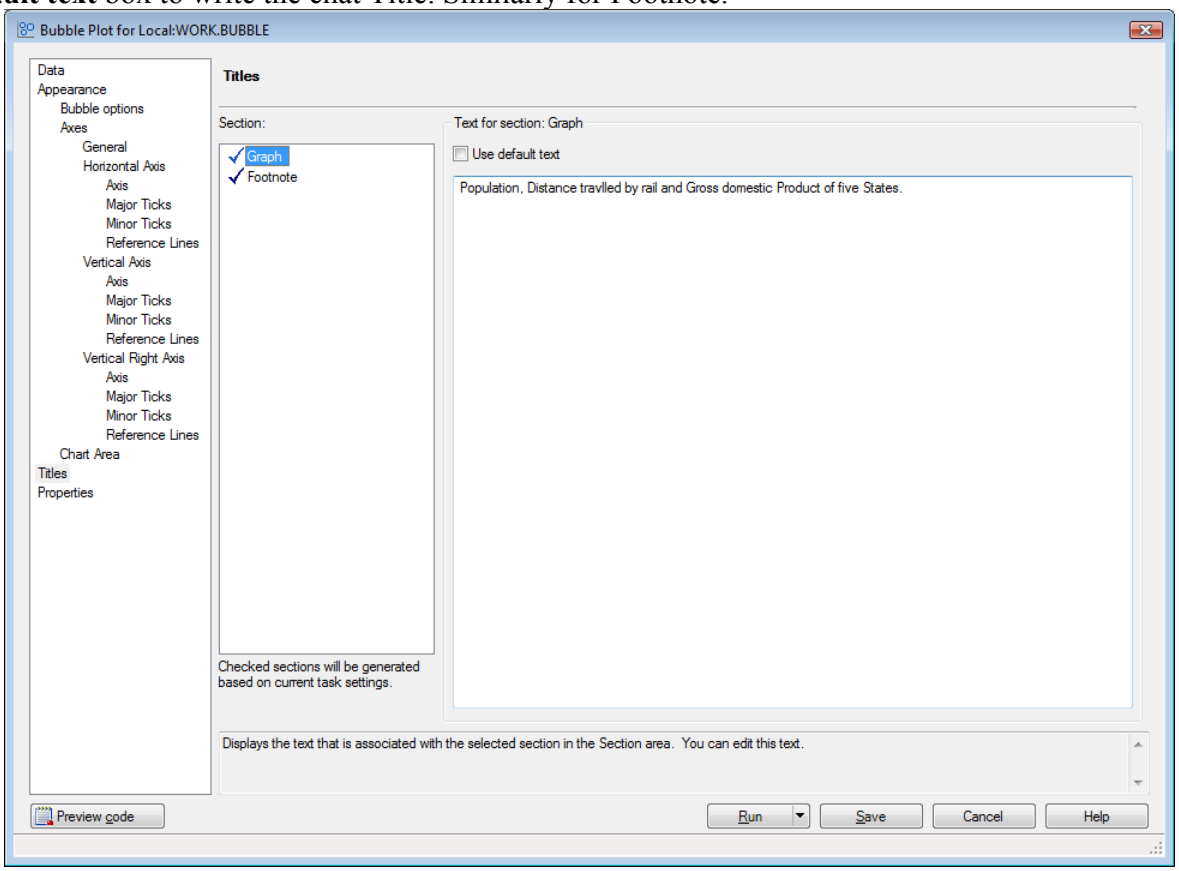
d) Select **Axes**, keep General Uncheck so that one can see the axis and the ticks on axis.

e) Select **Horizontal, Vertical** and **Vertical (right)** axis one by one and type the axis title under the tab **label** and align centre by select the option given under this tab and shown in screen shot given below.

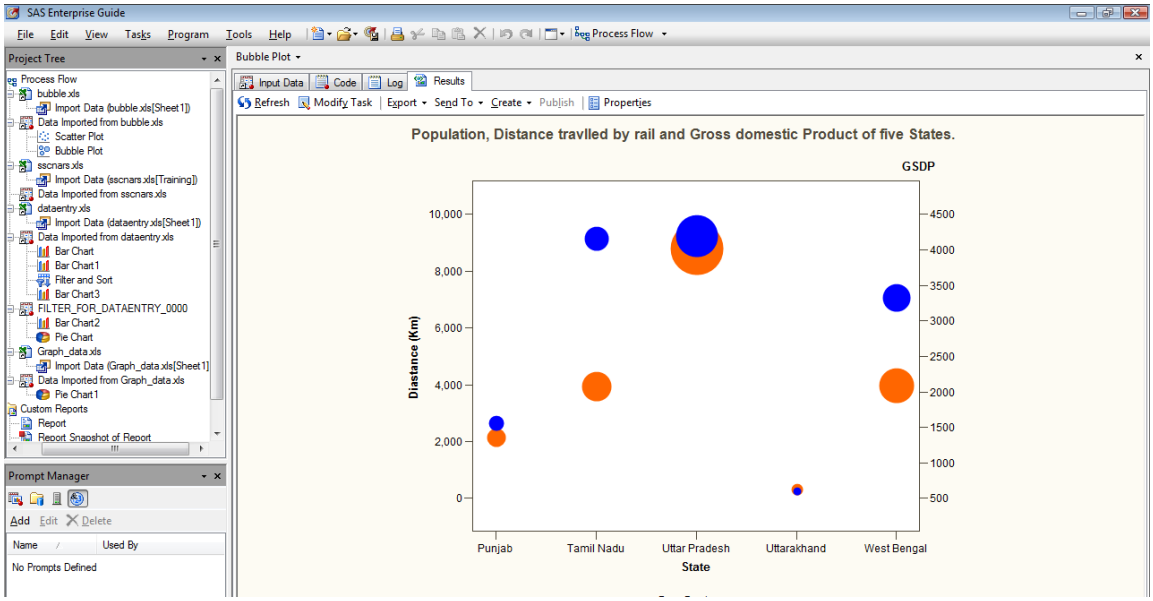


f) Leave Chart Area and Properties tabs as default.

g) Select **Title** tab to type the chart title and footnote. Select **Graphs** and then uncheck the **Use default text** box to write the chat Title. Similarly for Footnote.

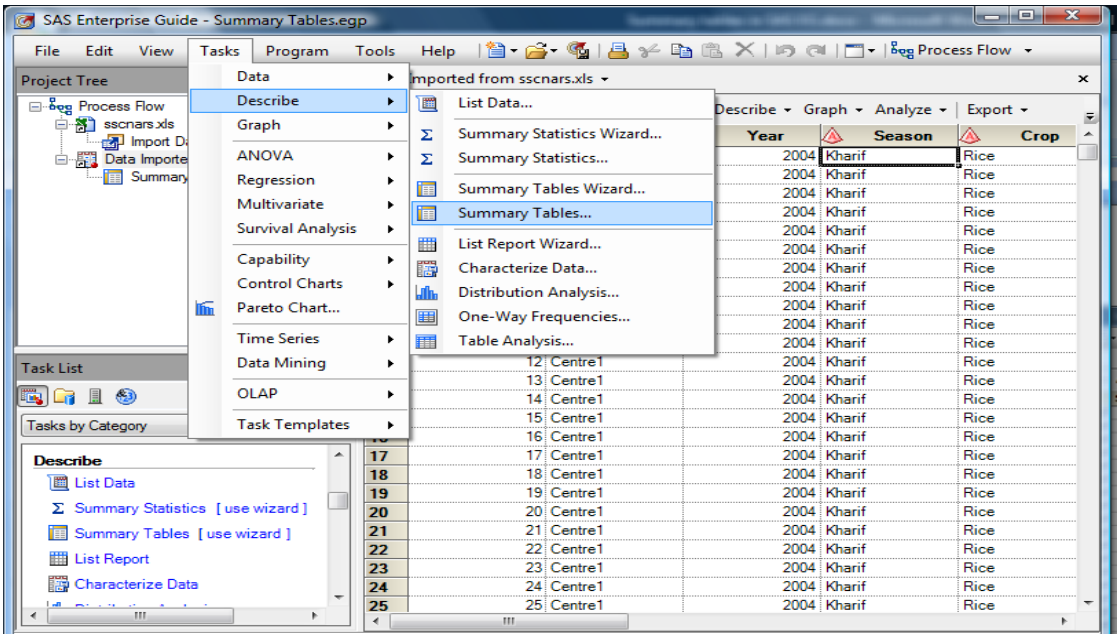


h) Select RUN, one obtains a bubble chart as shown below:

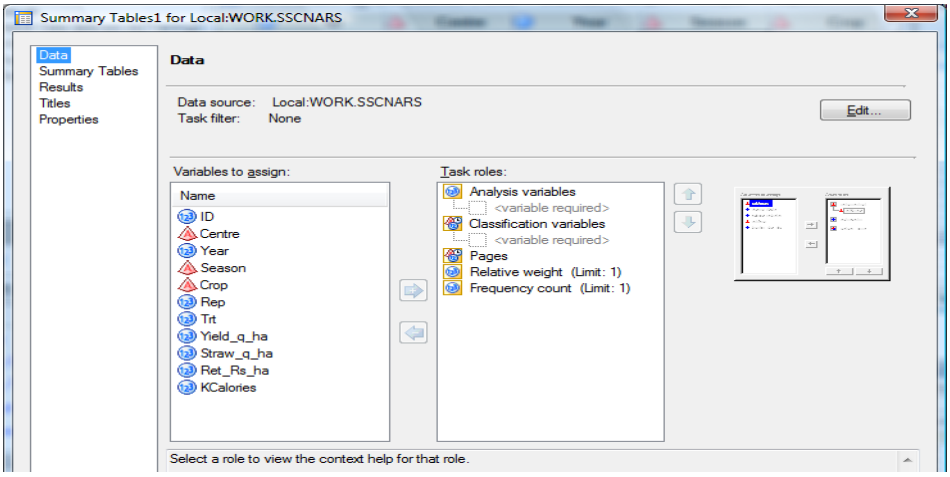


12. Creating Summary Report

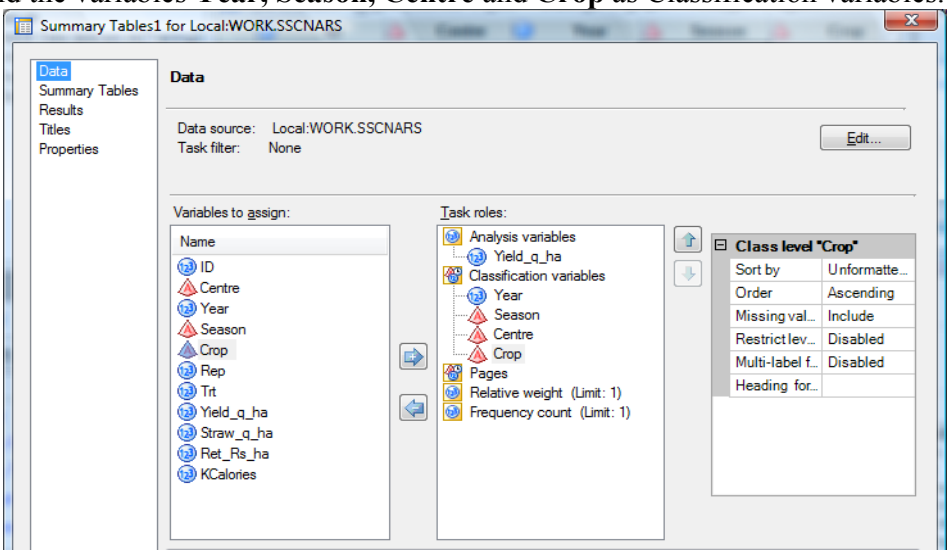
To create a summary report using the data on yield in q/ha of various crops for different centres, years, seasons, crops, go to the Menu bar and select **Tasks option → Describe → Summary Tables.**



Selecting Summary Tables option would open a variable selection wizard where one can assign task roles to variables by simply dragging and dropping them at appropriate roles. Here one will see the variables on the left side and task roles on the right side of the window. To move the variables from left window to right window one can either drag and drop the variables or using the arrows shown in the screenshot.

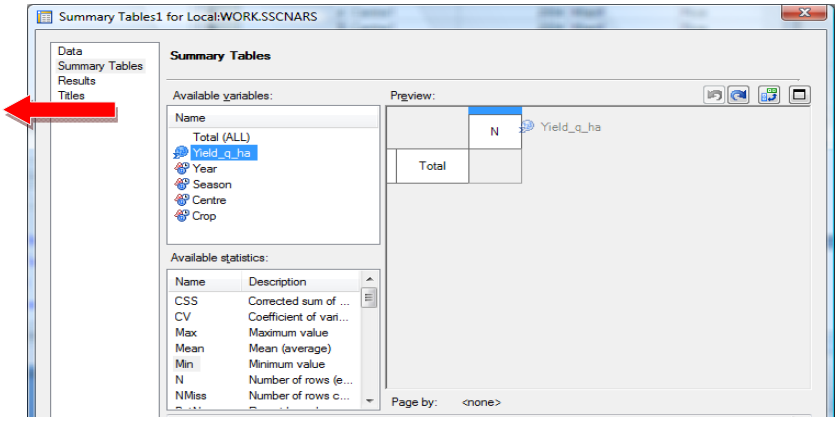


Now one has to select the variables from variables to assign window by moving them to the right side in the task role window. In this case one is taking the variable **Yield_q_ha** as analysis variable and the variables **Year, Season, Centre** and **Crop** as Classification variables.

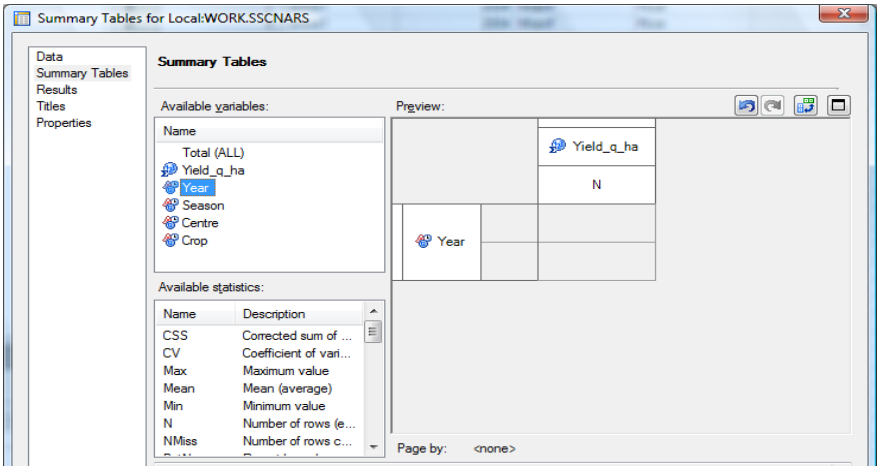


After assigning the roles to the variables, select **Summary Tables** option form the left selection pane (highlighted by red arrow) to define the table layout and output statistics options one wants in summary table for the variables.

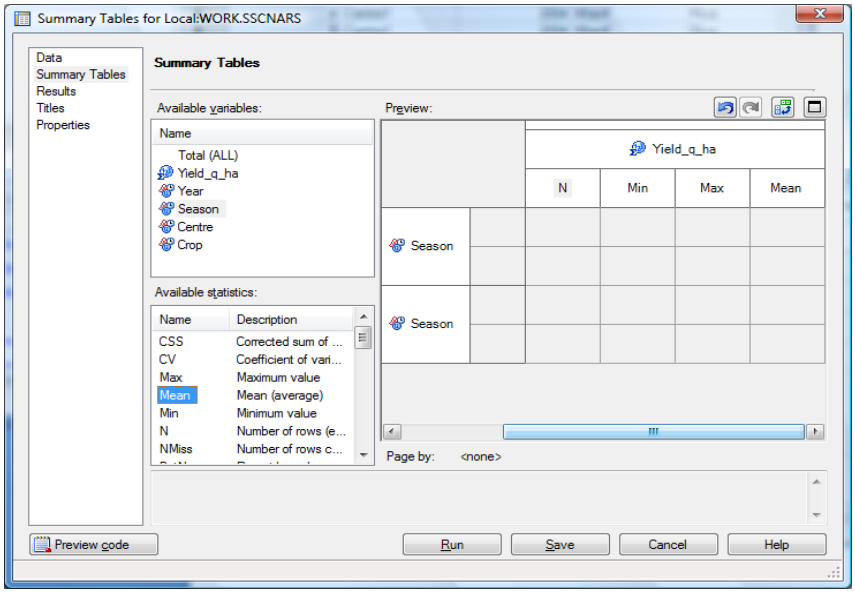
Here one can simply drag and drop desired variables to the table preview window to define their positions in the table. Select variable **yield_q_ha** and drag it to the preview window and drop above the space highlighted in blue.



Similarly select variables **year, Season, Centre** and **Crop**, drag and drop them inside the preview window.



Now select **Min, Max** and **Mean** from the Available Statistics and drag and drop them as shown in the screenshot below.



Now Click run button to execute the task which will generate the summary table as shown in above.

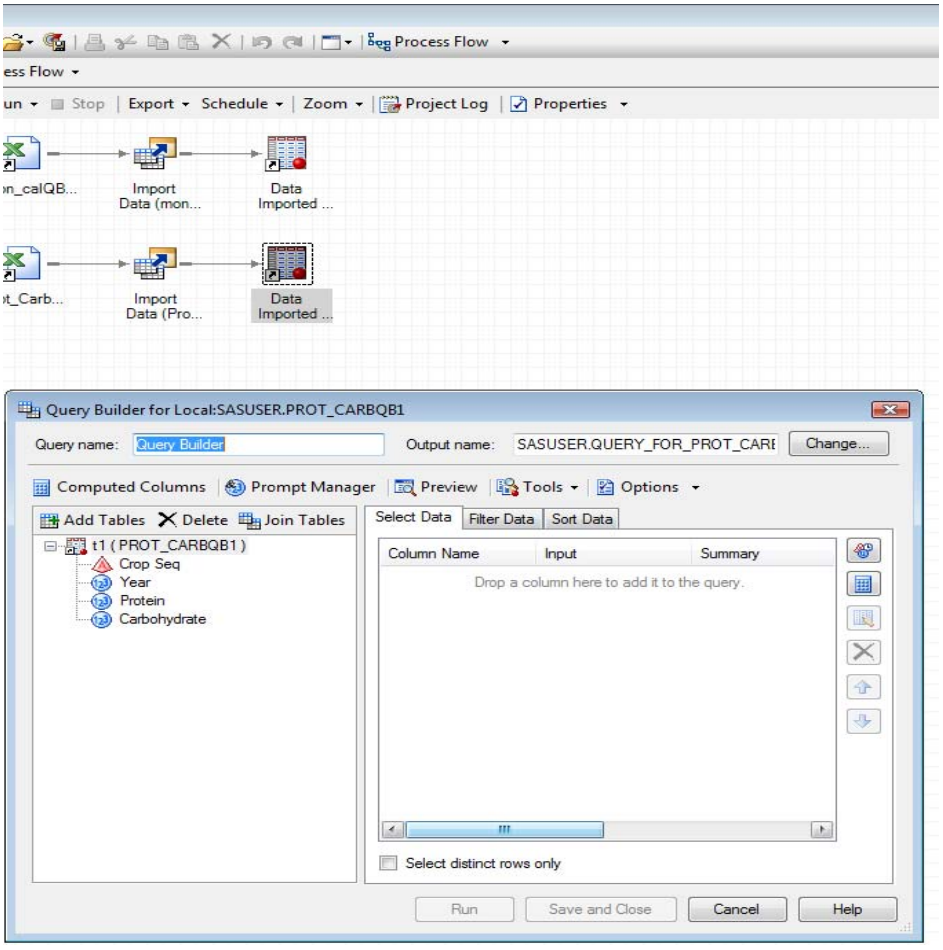
13. Query Builder

The Query Builder is the tool that we use to query the data and request to retrieve data from one or more data sources. It help us to specify the columns that we want to include in our results, and we can specify the order in which the columns appear. We can also compute new columns and replace values in existing columns.

Creating Query Builder

First of all we have to open the data that we want to query and click **Query Builder** in the workspace.

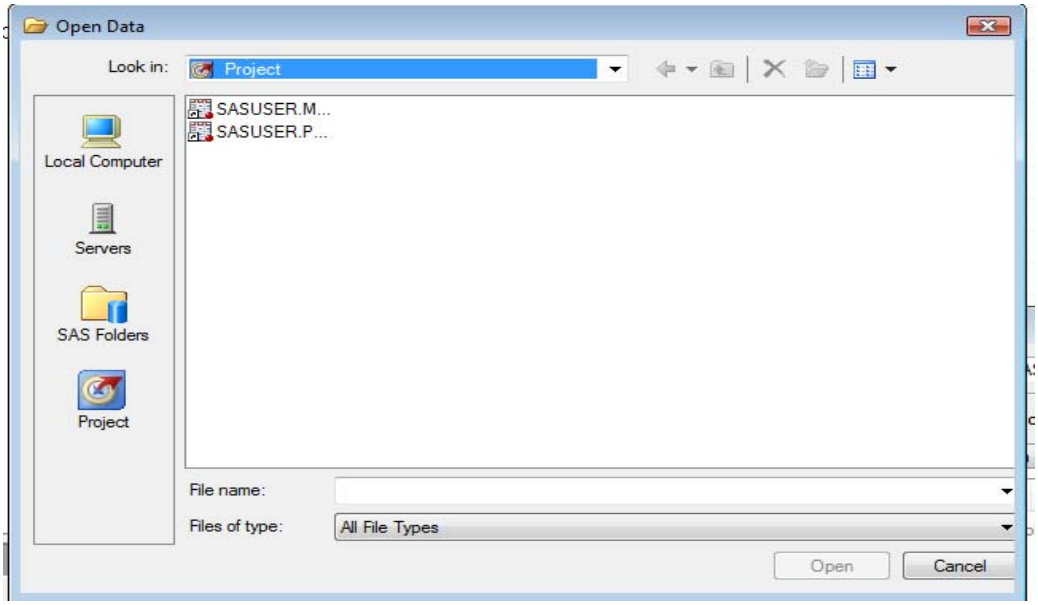
1. Select **Tasks→ Data →Query Builder**. This will Open dialog box opens so that we can select the table that we want to use.



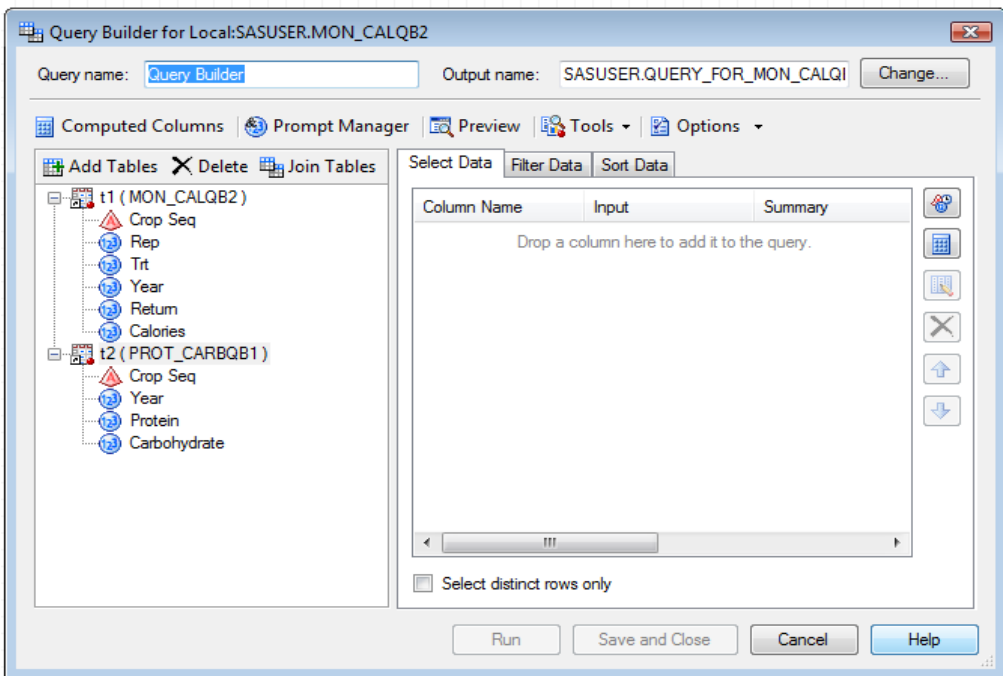
Click **Add Tables**, data tables will be displayed under the current project as shown in the following snap shot 1. Now select the table and click **open,** the selected tables will appear one below other (See snap Shot 2). Click **Join Table,** it will show the suitable join of the table (See Snap Shot 3) click close. Select the items from the from the tables 1 and table on which query is required. If other than auto join is proffered than click right mouse button of first table it will

display the option for the join and then select required option. After the selection of required field click **Run**. The query table will the show the data of all the field we have selected. Now we can sort or filter the data from this table as per our requirement.

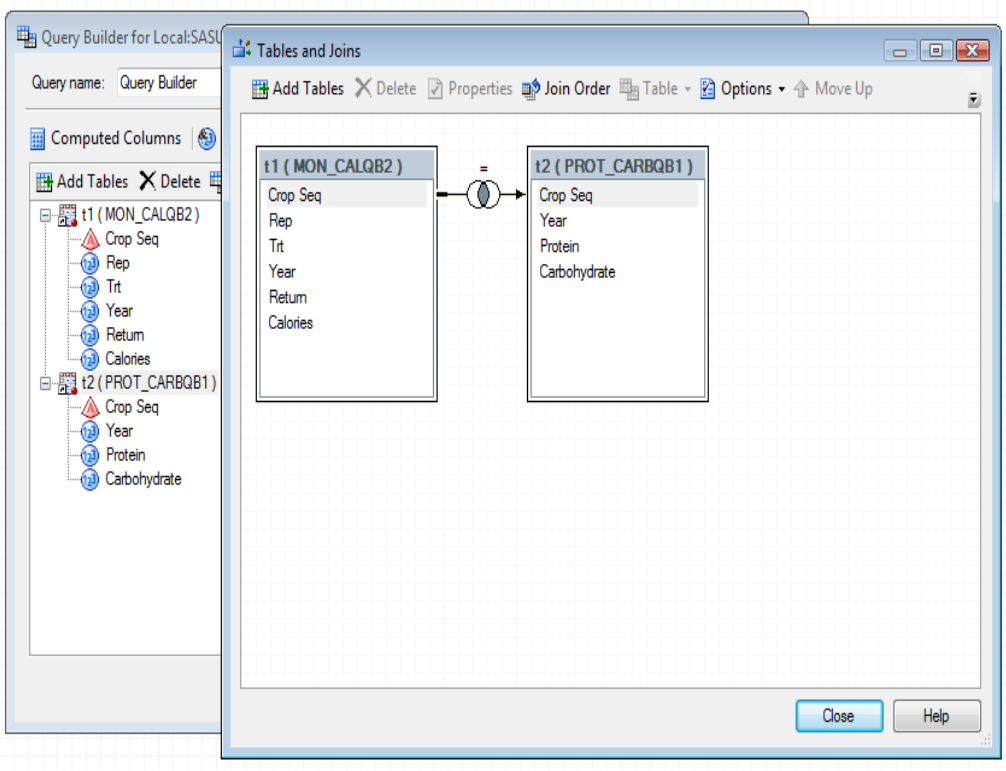
Snap Shot 1



Snap Shot 2



Snap Shot 3



In case we are interested in the tables other than current project then we have to choose the path accordingly. For example we want to join the table laying on the desktop of the computer than we have to follow the path **Local Computer→Desktop→ Folder Name → File Name,** if file is not a SAS data even than it will be imported automatically but it must have suitable otherwise error will occur.

14. Transformation of Data

Most commonly used transformation in the analysis of experimental data are Arcsine, Logarithmic and Square root.

Arcsine Transformation

Arcsine (ARSIN) is the angle whose sine is number and this number must be from -1 to 1. The return angle is given in radians in the range $-\pi/2$ to $\pi/2$. To obtain Arcsine transformation, first of all one has replace zeroes by 1/(4n) and 100 by 100-1/(4n), where n is number of observations on which percentage is based then apply the transformation formula by using the following command lines.

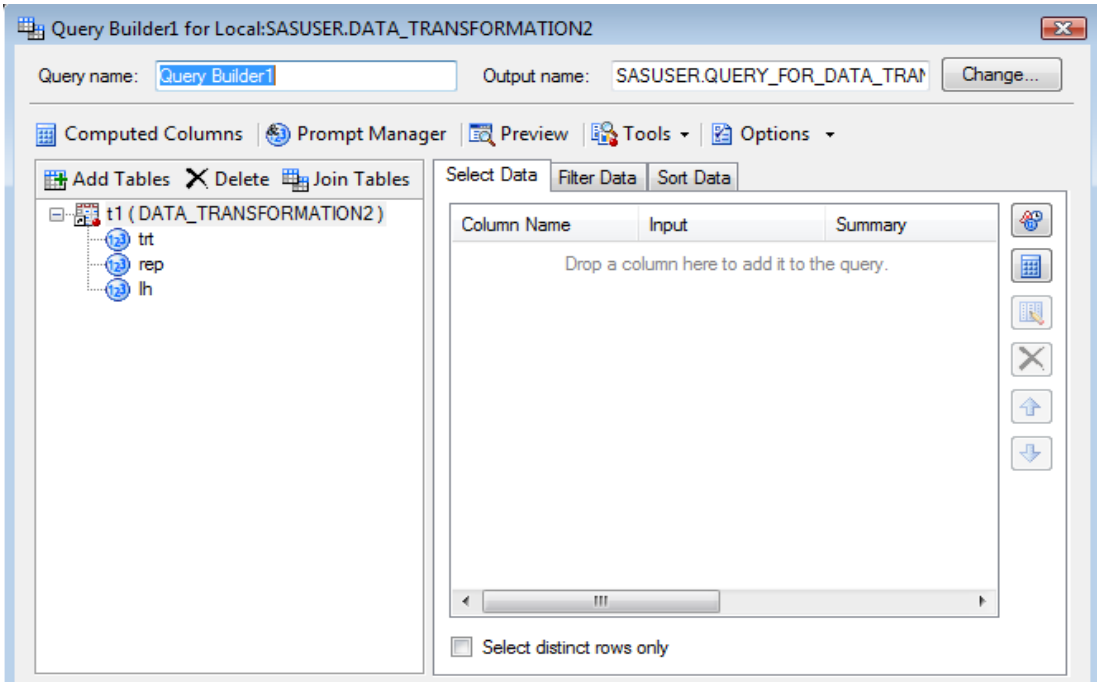
Task→ Data→ Query Builder, now select the whole table or selected columns of the data table which is to be transform and place it in the right side window under 'select data' tab. Select, **Compute columns→ New→ Select radio button Advance Expression→ Next→** Expand the **functions** folder, list of function displays, select the required function from the list and Double click it . The function will appears in ' Enter an Expression' window. Type name of the variable within brackets alongwith the required transformation arguments. One can type **arsin(sqrt(CASE WHEN t1. *Variable name*=0 THEN (1/4n) WHEN t1. *Variable name*=100**

THEN 100-(1/4n) ELSE t1.*Variable name* END)/100)*(180*7)/22 directly in space given under the "Enter an Expression". Here n is number of observations.

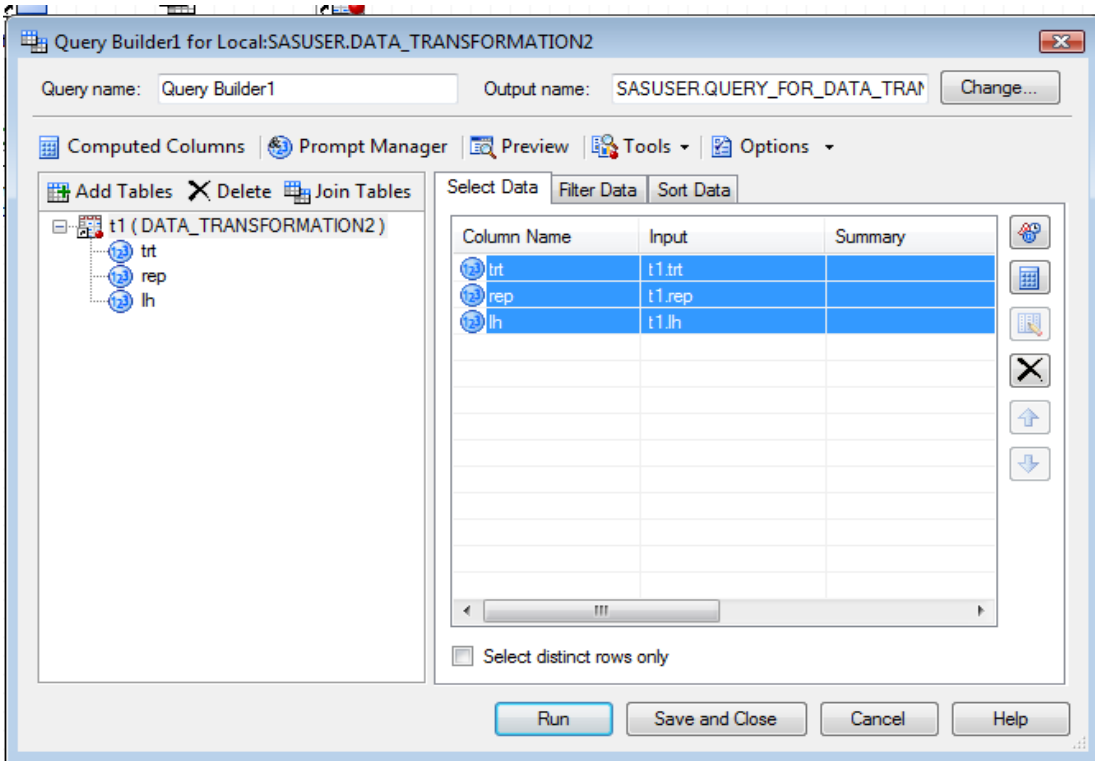
Exercise 14.1: Consider the following data and obtain Arcsine Transformation

TRT	REP	Leafhoppers
1	1	44.00
1	2	21.33
1	3	0.00
1	4	25.33
1	5	24.00
1	6	0.00
1	7	32.00
1	8	0.00
1	9	17.33
1	10	93.33
1	11	13.33
1	12	46.66
2	1	25.30
2	2	49.30
2	3	0.00
2	4	26.66
2	5	26.66
2	6	0.00
2	7	29.33
2	8	0.00
2	9	33.33
2	10	80.00
2	11	36.00
2	12	46.66
3	1	48.00
3	2	80.00
3	3	0.00
3	4	49.33
3	5	54.66
3	6	20.00
3	7	28.00
3	8	0.00
3	9	10.66
3	10	78.00
3	11	33.33
3	12	16.00

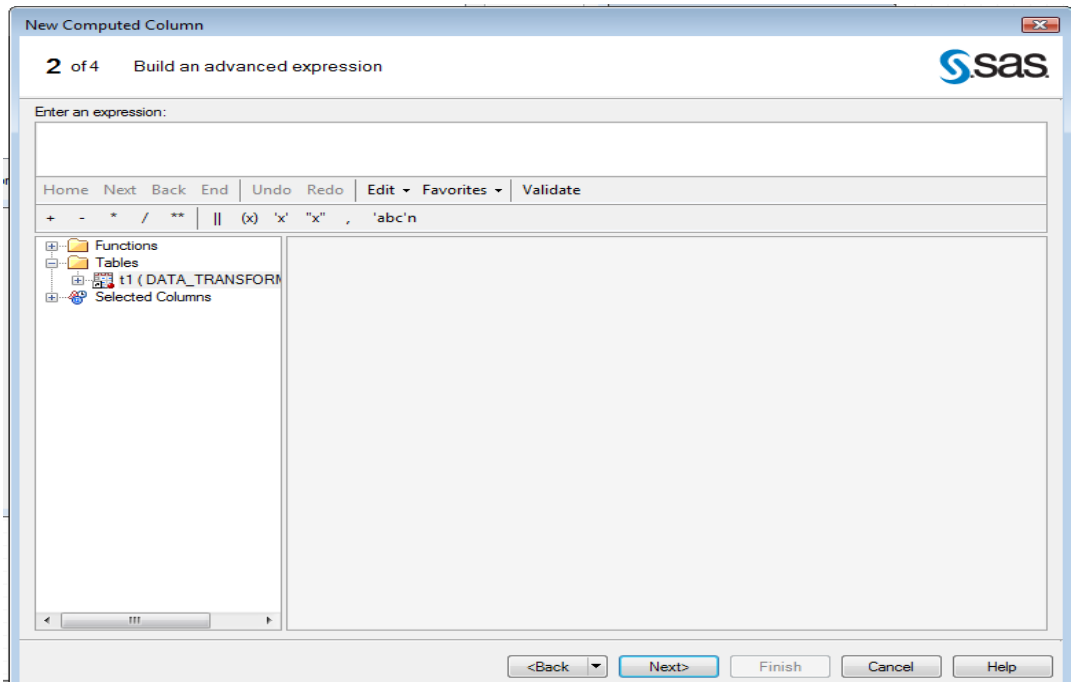
Solution: To obtain the transformed data use the steps: **Task→ Data→ Query Builder**



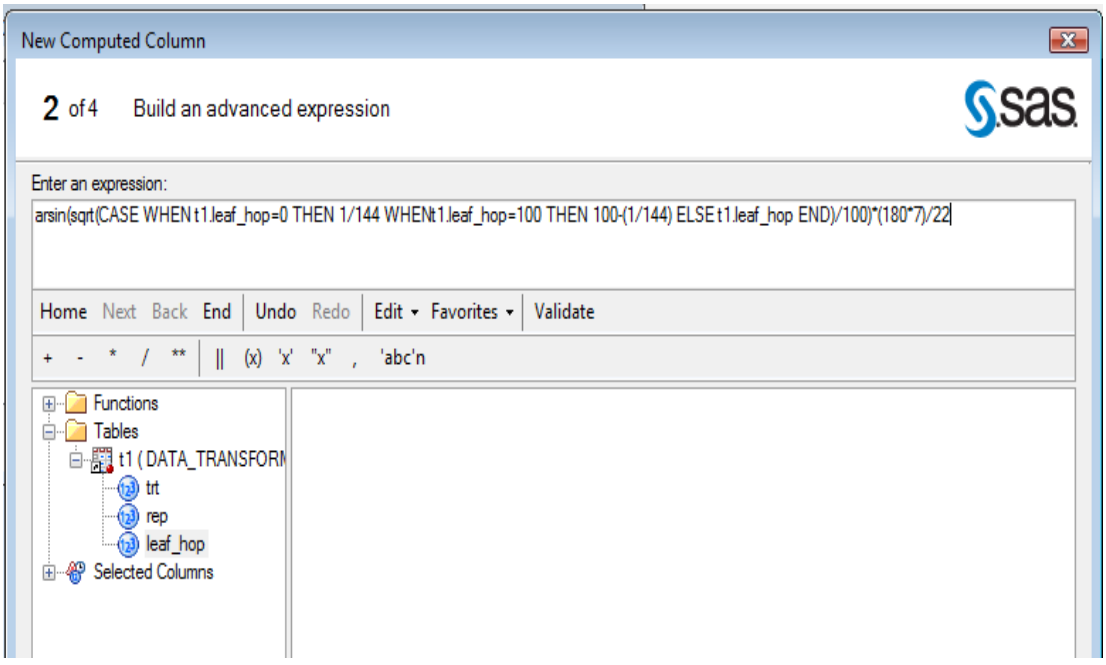
select the whole table t1 and drag and drop under select Data



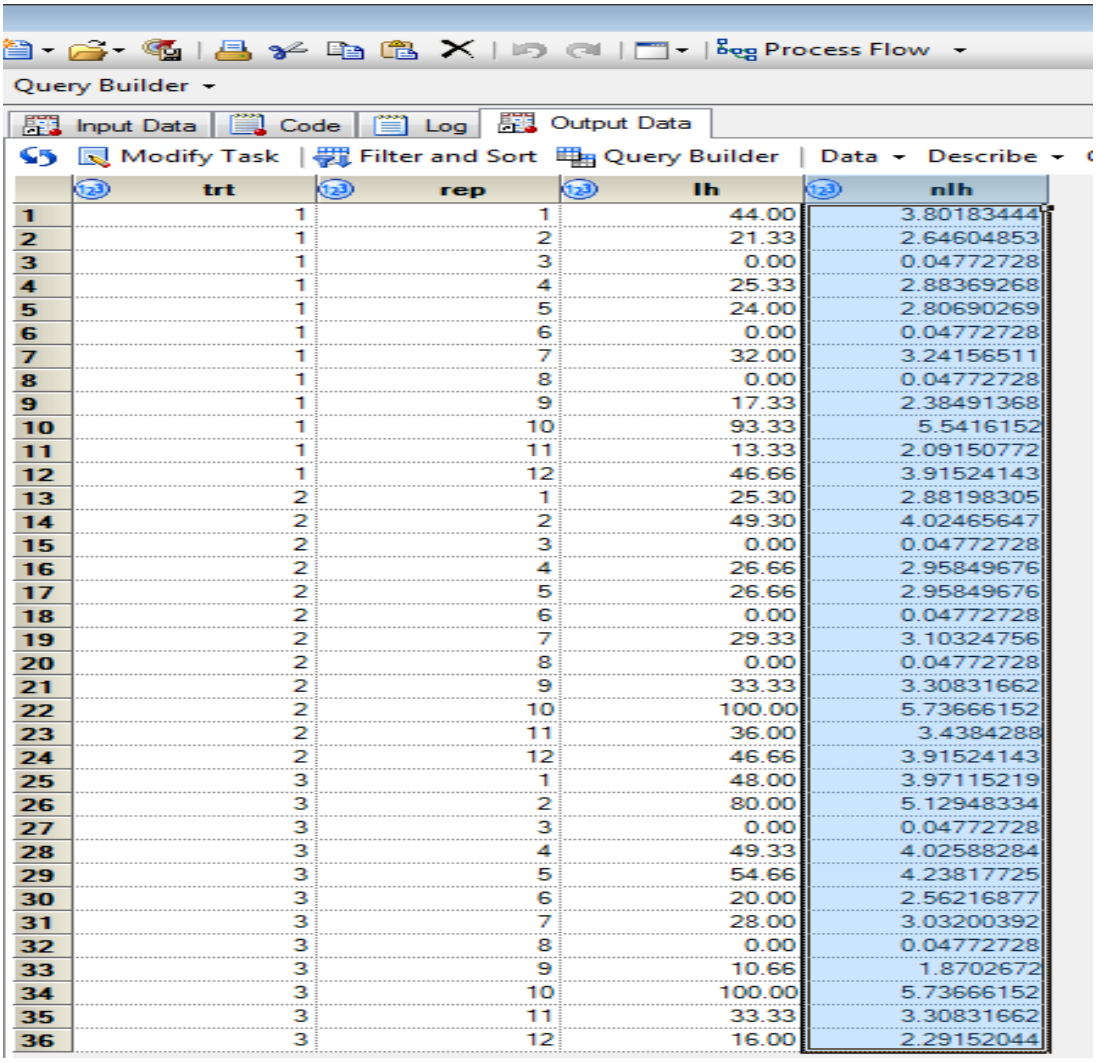
Compute columns→ New→ Select radio button Advance Expression→ Next



Now type **arsin(sqrt(CASE WHEN t1. *LH*=0 THEN 1/144 WHEN t1. *LH*=100 THEN 100-(1/144) ELSE t1.*LH* END)/100)*(180*7)/22.** (Note :One can use upper or lower case letters).



Click **Next** we obtain the **Modify Additional Option** window. Type the New Column name in front of Column heading and Alias then **Next→ Finish.** Close all the windows except Main table window and Click **Run** we obtain the transform data as under



	trt	rep	lh	nlh
1	1	1	44.00	3.80183444
2	1	2	21.33	2.64604853
3	1	3	0.00	0.04772728
4	1	4	25.33	2.88369268
5	1	5	24.00	2.80690269
6	1	6	0.00	0.04772728
7	1	7	32.00	3.24156511
8	1	8	0.00	0.04772728
9	1	9	17.33	2.38491368
10	1	10	93.33	5.5416152
11	1	11	13.33	2.09150772
12	1	12	46.66	3.91524143
13	2	1	25.30	2.88198305
14	2	2	49.30	4.02465647
15	2	3	0.00	0.04772728
16	2	4	26.66	2.95849676
17	2	5	26.66	2.95849676
18	2	6	0.00	0.04772728
19	2	7	29.33	3.10324756
20	2	8	0.00	0.04772728
21	2	9	33.33	3.30831662
22	2	10	100.00	5.73666152
23	2	11	36.00	3.4384288
24	2	12	46.66	3.91524143
25	3	1	48.00	3.97115219
26	3	2	80.00	5.12948334
27	3	3	0.00	0.04772728
28	3	4	49.33	4.02588284
29	3	5	54.66	4.23817725
30	3	6	20.00	2.56216877
31	3	7	28.00	3.03200392
32	3	8	0.00	0.04772728
33	3	9	10.66	1.8702672
34	3	10	100.00	5.73666152
35	3	11	33.33	3.30831662
36	3	12	16.00	2.29152044

Logarithmic and Square Root Transformation

Logarithmic (LGAMMA) gives the natural logarithm with base e of a positive number, LOG10 gives logarithm with base 10 of a positive numbers. The other logarithmic functions in the Enterprise Guide are LOG2, LOGBETA, LOGCDF, LOGPDF and LOGSDF. Square Root (SQRT) transformation gives the square root of a positive number.

To obtain the required transformation, first of all, we select the data and then follow the steps, **Task→ Data→ Query Builder ,** now select the whole table or selected columns of the data table which is to be transform and place it in the right side window under 'select data' tab. Select, **Compute columns→ New→ Select radio button Advance Expression→ Next→** Expand the **functions** folder, list of function displays, select the required function from the list and Double click it. If we do not select the whole table or columns, Enterprise guide will transform the desired data as separate query. Generally we keep the transformed data column along with the original data so we have to place the table in the right side window before performing the transformation task. Click **Next**, we obtain the new window as "Modify Additional Option". Now we shall give the name to the column, obtained by the transformation.

Type the name of new column or variable in front of "column" and then click **Next** and **Finish.** Close the window where new variable is created and then click **Run.**

Exercise 14.2: Obtain Logarithmic transformation of the following data.

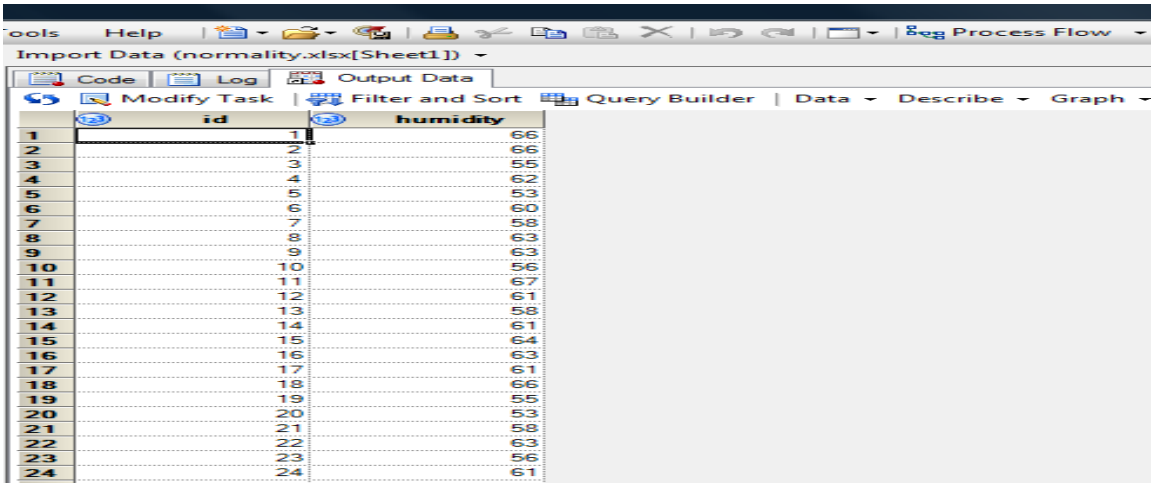
TRT	REP	White heads
1	1	1.39
1	2	8.43
1	3	7.58
1	4	8.95
1	5	4.16
1	6	4.68
1	7	2.37
1	8	0.95
1	9	26.09
1	10	26.39
1	11	21.99
1	12	3.58
1	13	0.19
1	14	0
1	1	0.92
1	2	4.38
1	3	3.79
1	4	12.81
1	5	17.39
1	6	1.32
1	7	5.32
1	8	0.7
1	9	25.36
1	10	22.29
1	11	12.88
1	12	2.62
1	13	0
1	14	3.64
1	1	2.63
1	2	6.94
1	3	1.91
1	4	3.22
1	5	8.06
1	6	2.09
1	7	4.86
1	8	0.98
1	9	15.69
1	10	1.98
1	11	5.15
1	12	2.91
1	13	0.61
1	14	4.44

Exercise 14.3: Consider the following data and obtain the Square Root.

TRT	REP	Larval
1	1	9
1	2	4
1	3	6
1	4	9
1	5	27
1	6	35
1	7	1
1	8	10
1	9	4
2	1	12
2	2	8
2	3	15
2	4	6
2	5	17
2	6	28
2	7	0
2	8	0
2	9	10
3	1	0
3	2	5
3	3	6
3	4	4
3	5	10
3	6	2
3	7	0
3	8	2
3	9	15
4	1	1
4	2	1
4	3	2
4	4	5
4	5	10
4	6	15
4	7	0
4	8	1
4	9	5

15. Testing for Normality

Exercise 15.1: Consider a data set consisting of a single measured variable, **humidity**, based on relative humidity at 14 hrs. Pertaining to Raipur district from 1991 to 1995 for kharif season per month. A portion of the data set is shown below. Months were assigned identification numbers (**id**) in the data set in addition to the **humidity**.

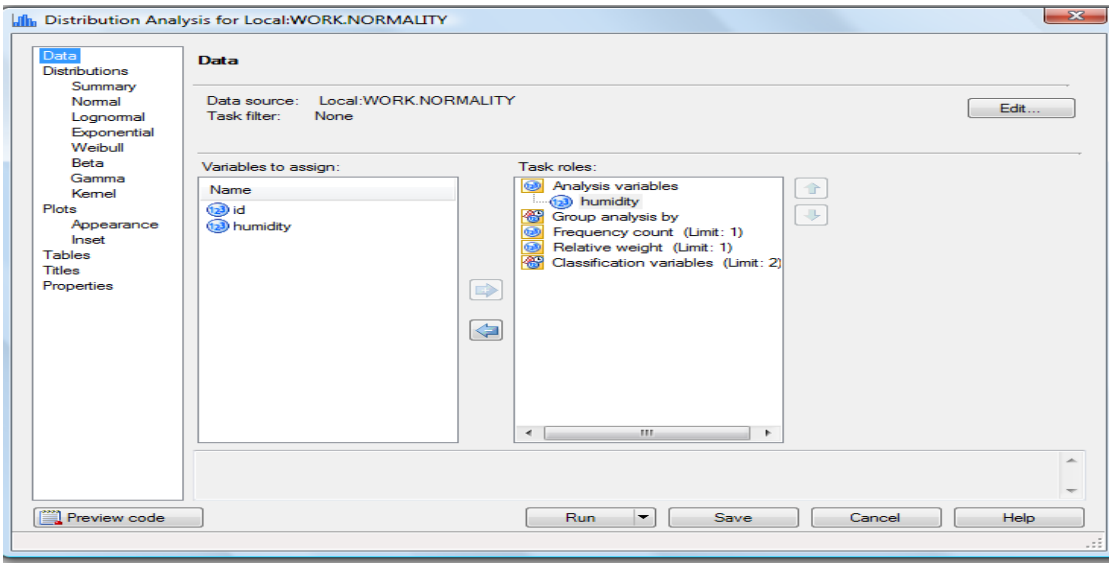


	id	humidity
1	1	66
2	2	66
3	3	55
4	4	62
5	5	53
6	6	60
7	7	58
8	8	63
9	9	63
10	10	56
11	11	67
12	12	61
13	13	58
14	14	61
15	15	64
16	16	63
17	17	61
18	18	66
19	19	55
20	20	53
21	21	58
22	22	63
23	23	56
24	24	61

Setting up the analysis

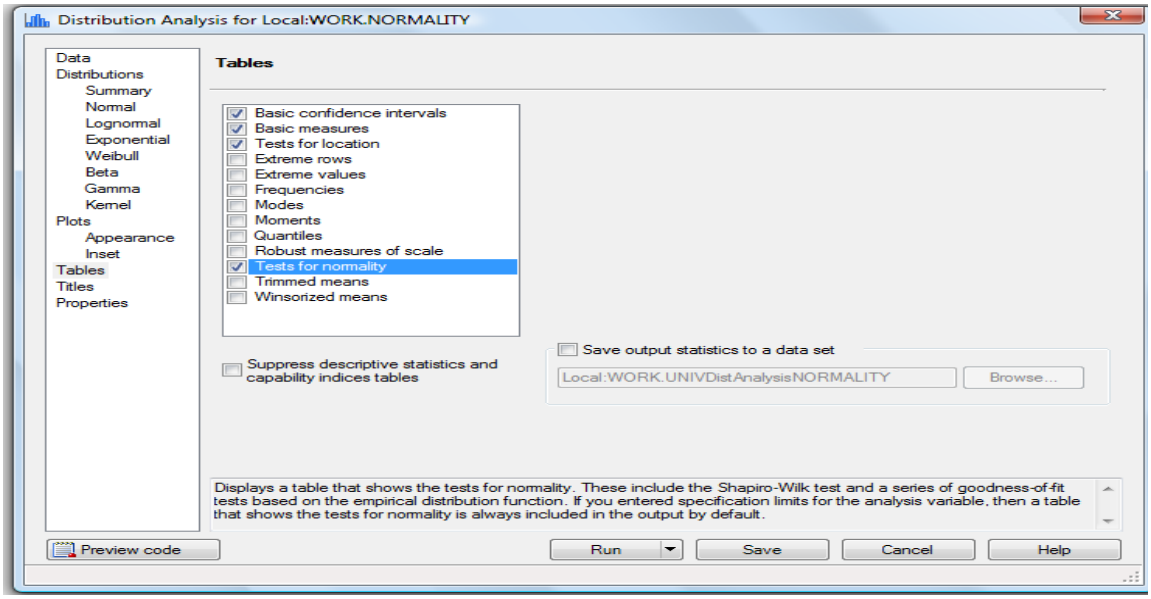
Step 1: Obtaining the normality assessments

From the main menu, select **Tasks→Describe→ Distribution Analysis**. This brings one to the **Task Roles** window. Drag **humidity** to the slot under **Analysis variables** in the rightmost panel. This is shown below.



The **Task Roles** screen for the **Distribution Analysis** procedure

Step 2:Click **Tables** from the navigation panel on the far left. Select **Tests for normality** and click the checkbox to place a check mark there as shown in screen below. Click the **Run** push button to perform the analysis.



The **Tables** screen for the **Distribution Analysis** procedure

The results of the analysis are shown in the following table

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.965873	**Pr < W**	0.0914
Kolmogorov-Smirnov	**D**	0.111814	**Pr > D**	0.0617
Cramer-von Mises	**W-Sq**	0.117093	**Pr > W-Sq**	0.0680
Anderson-Darling	**A-Sq**	0.690944	**Pr > A-Sq**	0.0715

Each test occupies a row in the output table. The last column in the table is the test of significance against the null hypothesis that the values of the measured variable are distributed in a normal manner. All four tests gives a statistically non-significant result, indicating that one cannot reject the null hypothesis; that is, it appears that the distribution does not significantly depart from normality.

16. How to Analyse the Data?

The data set after importing in the process flow of Enterprise Guide can be analysed by using the respective menu and sub menu and the results and reports, can be saved for further use. All data sets used for exercises and the questions to be answered are taken from the link Analysis of Data (http://iasri.res.in/design/Analysis%20of%20data/Analysis%20of%20Data.html) available on Design Resources Server (www.iasri.res.in).

16.1 Descriptive Statistics

Exercise 1: An experiment was conducted to study the hybrid seed production of bottle gourd (*Lagenaria siceraria (Mol) Standl)* Cv. Pusa hybrid-3 under open field conditions during Kharif-2005 at Indian Agricultural Research Institute, New Delhi. The main aim of the investigation was to compare natural pollination and hand pollination. The data were collected on 10 randomly selected plants from each of natural pollination and hand pollination on number of fruit set for the period of 45 days, fruit weight (kg), seed yield per plant (g) and seedling length (cm). The data obtained is as given below:

Group	No. of fruit	Fruit weight (kg)	Seed yield/plant (g)	Seedling length (cm)
1	7.0	1.85	147.70	16.86
1	7.0	1.86	136.86	16.77
1	6.0	1.83	149.97	16.35
1	7.0	1.89	172.33	18.26
1	7.0	1.80	144.46	17.90
1	6.0	1.88	138.30	16.95
1	7.0	1.89	150.58	18.15
1	7.0	1.79	140.99	18.86
1	6.0	1.85	140.57	18.39
1	7.0	1.84	138.33	18.58
2	6.3	2.58	224.26	18.18
2	6.7	2.74	197.50	18.07
2	7.3	2.58	230.34	19.07
2	8.0	2.62	217.05	19.00
2	8.0	2.68	233.84	18.00
2	8.0	2.56	216.52	18.49
2	7.7	2.34	211.93	17.45
2	7.7	2.67	210.37	18.97
2	7.0	2.45	199.87	19.31
2	7.3	2.44	214.30	19.36

Consider the data set of Example 1 and obtain the **Descriptive Statistics**

1. Obtain mean, standard deviation, minimum and maximum values of all the characters.
2. Obtain mean, standard deviation, minimum and maximum values of all the characters for each group separately.
3. Obtain mean, median, coefficient of skewness, coefficient of kurtosis of all the characters.
4. Obtain mean, median, coefficient of skewness, coefficient of kurtosis of all the characters for each of the group separately.
5. Obtain the partition values.

6. Test whether the data follows a normal distribution or not for all the characters? Do it separately for each of the two groups.
7. Create a box plot for all the characters.
8. Create a box plot for all the characters for each group separately.

Solution: To obtain descriptive statistics perform the following commands
For preparation of Data: First of all Import the data file, click **File→Import Data** select the file and click **Finish.**

Analysis:

1. Select **Tasks → Describe → Summary Statistics**. On the Summary Statistics Dialog Box: Select Number of Fruit set (**fs45),** Fruit weight **(fw)**, Seed yield/plant **(syp),** Seedling length (**sl).** Click the **right arrow** and select **Analysis variables**. Under **Statistics** in the selection pane, select **Basic** and select the desired basic statistics: Mean, Standard Deviation, minimum, maximum, Standard Error, variance, Mode, Range, Sum, Number of Observations, and Number of Missing values. Default selection is Mean, Standard deviation, Minimum, Maximum and Number of observations. In this case, uncheck the **Number of observations** box and retain **Mean, Standard deviation**, **Minimum**, and **Maximum** check boxes selected) and then **Run**.
2. Follow the same steps as in 1, except the change in the selection of variable list as: From Variables list, select Group. Click the right arrow and select Classification variables. Again Select Number of Fruit set (fs45), Fruit weight (fw), Seed yield/plant(syp), Seedling length(sl). Click the right arrow and select Analysis variables.
3. In addition to the Steps in 1, before run, Under Statistics selection pane, select additionally Percentile and then Check on Median. For obtaining skewness and kurtosis, Right-click on the task in the project tree or in the process flow and select Add As Code Template. This would generate SAS code, in this SAS code add *skewness and kurtosis.* A new program item named Code For *task-Skew Kurt* is added to the project. Save changes and click RUN.
4. In addition to the Steps in 1, except the change in the selection of variable list as: From Variables list, select Group. Click the right arrow and select Classification variables. Again Select Number of Fruit set (fs45), Fruit weight (fw), Seed yield/plant(syp), Seedling length(sl). Click the right arrow and select Analysis variables.
5. Following the similar steps, the partition values can be obtained after selecting Percentile from Statistics Selection Pane.
6. Task → Describe → Distribution Analysis → From Variables list, select Group. Click the right arrow and select Classification variables. Again Select Number of Fruit set (fs45), Fruit weight (fw), Seed yield/plant(syp), Seedling length(sl). Click the right arrow and select Analysis variables. From the Plots selection pane, select appearance, one gets display option for Histogram Plot, Probability Plot, Quartile Plot, Box Plot and Text-based Plots. From this display check on Histogram Plot and then Select Inset from Plots Selection Pane, Chaeck the Include Inset Box and then among the options in the Combo Box select Test Statistic for Normality and click on Run.
7. In addition to the Steps in 1, form the plots Selection Pane, select Box and Whisker Plots and then run.
8. In addition to the Steps in 2, form the plots Selection Pane, select Box and Whisker Plots and then run.

Please note that the variables on which analysis is to be performed are Analysis Variables and Classification Variables are group variables.

16.2 Performing t-test :

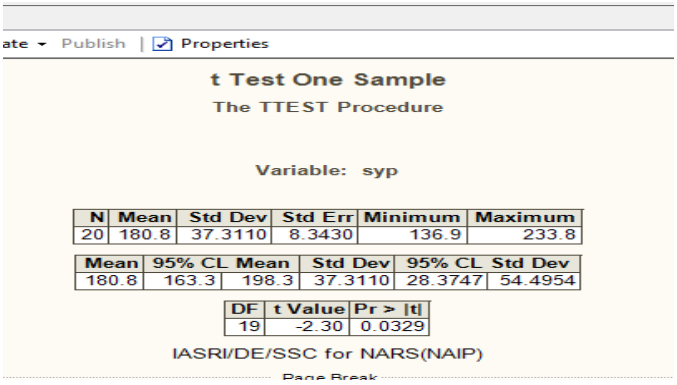
Exercise 2: Using the data from exercise 1, test the following:

1. Whether the mean of the population of Seed yield/plant (g) is 200 or not?
2. Whether the natural pollination and hand pollination under open field conditions are equally effective or are significantly different?
3. Whether hand pollination is better alternative in comparison to natural pollination?

Solution: To answer the question 1, use the following steps

t Test (one Sample)

a) Select Tasks→ANOVA→t Test.
b) Under t Test Type of Selection pane, select One Sample.
c) In the selection pane, select Data.
d) From Variables list, select syp Click the right arrow and select Analysis variables.
e) Select Analysis from Selection pane and define H0=200.
f) Click Run to run the one-sample t-test.


ate ▾ Publish | Properties

t Test One Sample

The TTEST Procedure

Variable: syp

N	Mean	Std Dev	Std Err	Minimum	Maximum
20	180.8	37.3110	8.3430	136.9	233.8

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
180.8	163.3	198.3	37.3110	28.3747	54.4954

DF	t Value	Pr > |t|
19	-2.30	0.0329

IASRI/DE/SSC for NARS(NAIP)

Page Break

Two Sample

a) Select Tasks→ANOVA→t Test.
b) Under t Test Type of Selection pane, select radio button Two Sample (The default option is also Two Sample).
c) In the selection pane, select Data.
d) From Variables list, select the variable to be analysed. Click the right arrow and select Analysis variables.
e) From Variables list, identify classificatory variable. Click the right arrow and select Classification Variable and define it as Group Variable.
f) Select Analysis from Selection pane and define H0=0, if interested in equality of two population means, otherwise define the desired value. The default option is zero.
g) Click Run to run the two-sample t-test.

A Snap Shot of the results are as under:

The TTEST Procedure

Variable: fs45

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	10	6.7000	0.4830	0.1528	6.0000	7.0000
2	10	7.4000	0.5907	0.1868	6.3000	8.0000
Diff (1-2)		-0.7000	0.5395	0.2413		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		6.7000	6.3544	7.0456	0.4830	0.3323	0.8819
2		7.4000	6.9775	7.8225	0.5907	0.4063	1.0783
Diff (1-2)	Pooled	-0.7000	-1.2069	-0.1931	0.5395	0.4077	0.7979
Diff (1-2)	Satterthwaite	-0.7000	-1.2084	-0.1916			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	-2.90	0.0095
Satterthwaite	Unequal	17.318	-2.90	0.0098

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	1.50	0.5585

16.3 Correlation and Regression

Exercise 3: The following data was collected through a pilot sample survey on Hybrid Jowar crop on yield and biometrical characters. The biometrical characters were average PlantPopulation (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot).

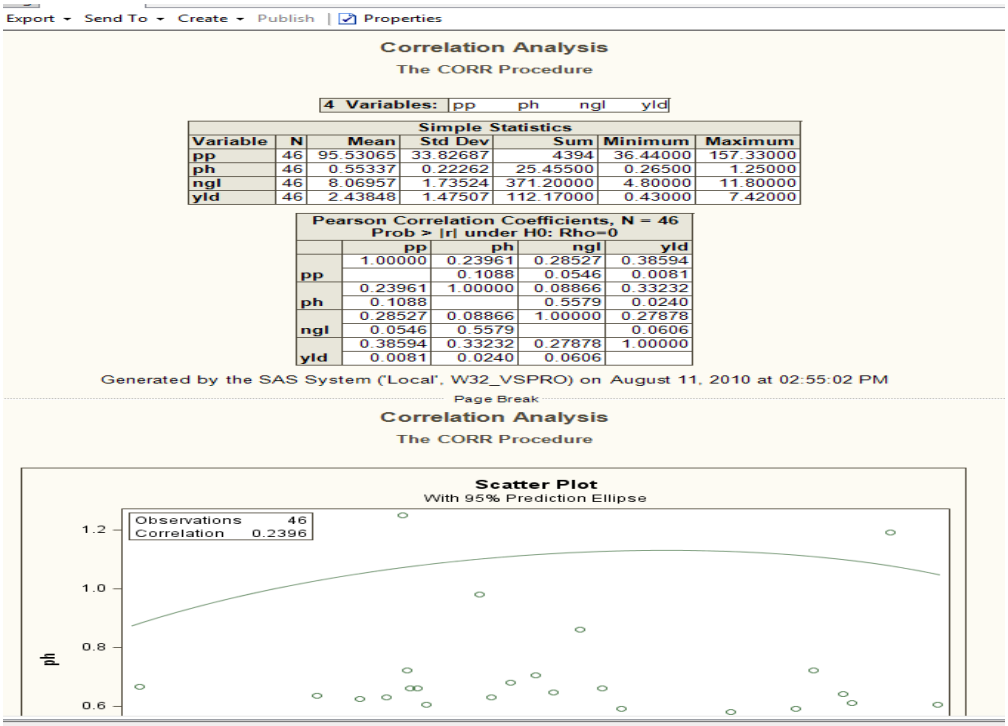
S.No.	PP	PH	NGL	Yield	S.No.	PP	PH	NGL	Yield
1	142.00	0.525	8.2	2.470	24	55.55	0.265	5.0	0.430
2	143.00	0.640	9.5	4.760	25	88.44	0.980	5.0	4.080
3	107.00	0.660	9.3	3.310	26	99.55	0.645	9.6	2.830
4	78.00	0.660	7.5	1.970	27	63.99	0.635	5.6	2.570
5	100.00	0.460	5.9	1.340	28	101.77	0.290	8.2	7.420
6	86.50	0.345	6.4	1.140	29	138.66	0.720	9.9	2.620
7	103.50	0.860	6.4	1.500	30	90.22	0.630	8.4	2.000
8	155.99	0.330	7.5	2.030	31	76.92	1.250	7.3	1.990
9	80.88	0.285	8.4	2.540	32	126.22	0.580	6.9	1.360
10	109.77	0.590	10.6	4.900	33	80.36	0.605	6.8	0.680
11	61.77	0.265	8.3	2.910	34	150.23	1.190	8.8	5.360
12	79.11	0.660	11.6	2.760	35	56.50	0.355	9.7	2.120
13	155.99	0.420	8.1	0.590	36	136.00	0.590	10.2	4.160
14	61.81	0.340	9.4	0.840	37	144.50	0.610	9.8	3.120
15	74.50	0.630	8.4	3.870	38	157.33	0.605	8.8	2.070
16	97.00	0.705	7.2	4.470	39	91.99	0.380	7.7	1.170

17	93.14	0.680	6.4	3.310	40	121.50	0.550	7.7	3.620
18	37.43	0.665	8.4	1.570	41	64.50	0.320	5.7	0.670
19	36.44	0.275	7.4	0.530	42	116.00	0.455	6.8	3.050
20	51.00	0.280	7.4	1.150	43	77.50	0.720	11.8	1.700
21	104.00	0.280	9.8	1.080	44	70.43	0.625	10.0	1.550
22	49.00	0.490	4.8	1.830	45	133.77	0.535	9.3	3.280
23	54.66	0.385	5.5	0.760	46	89.99	0.490	9.8	2.690

1. Obtain correlation coefficient between each pair of the variables PP, PH, NGL and yield.
2. Give a scatter plot of the variable PP.
3. Obtain partial correlation between NGL and yield after removing the linear effect of PP and PH.
4. Fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables.
5. Obtain the predicted values corresponding to each observation in the data set.
6. Check for the linear relationship among the biometrical characters, i.e., multi-colinearity in the data.
7. Fit the multiple linear regression model without intercept.

Solution:
First create a data file in MS- Excel or in any ASCII mode like CSV(Comma Separated Values) or .txt having fixed length. Now we Import this file in SAS Enterprise guide to perform the following steps to perform the analysis:.

1. **Correlation:** To obtain the correlation between each pair of variables
   - Select Tasks → Multivariate → Correlations.
   - From Variables list, select the desired variable PP, PH, NGL and yield. Click the right arrow and select Analysis variables. Select Option (from the selection pane) and the check box for types of Correlation as Pearson. By Check in the Box: Fisher Options, we can select the level of significance for testing the significance of correlation coefficients, we can also select type of alternative hypothesis as lower, upper or two sided from the combo box, Type of confidence limits.
   - From the Results Pane uncheck the Show statistics for each variable check box. Check the Show significance probabilities associated with correlations check box. Click Run to generate the Pearson correlation coefficient along with their test if significance.
2. **Scatter Plot:** In addition to the above steps, In the selection pane, select **Results, select plots,** if scatter plots is required as in question 3.



Export ▾ Send To ▾ Create ▾ Publish | Properties

Correlation Analysis

The CORR Procedure

4 Variables:	pp	ph	ngl	yld

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
pp	46	95.53065	33.82687	4394	36.44000	157.33000
ph	46	0.55337	0.22262	25.45500	0.26500	1.25000
ngl	46	8.06957	1.73524	371.20000	4.80000	11.80000
yld	46	2.43848	1.47507	112.17000	0.43000	7.42000

Pearson Correlation Coefficients, N = 46
Prob > |r| under H0: Rho=0

	pp	ph	ngl	yld
pp	1.00000	0.23961	0.28527	0.38594
		0.1088	0.0546	0.0081
ph	0.23961	1.00000	0.08866	0.33232
	0.1088		0.5579	0.0240
ngl	0.28527	0.08866	1.00000	0.27878
	0.0546	0.5579		0.0606
yld	0.38594	0.33232	0.27878	1.00000
	0.0081	0.0240	0.0606	

Generated by the SAS System ('Local', W32_VSPRO) on August 11, 2010 at 02:55:02 PM

Page Break

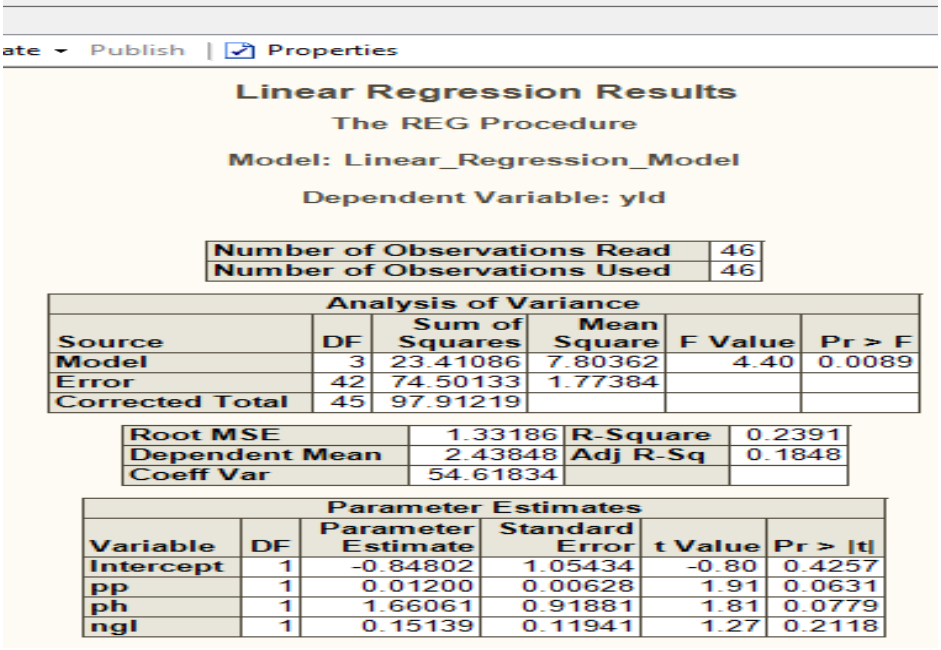
Correlation Analysis

The CORR Procedure

3. **Partial Correlations:** To obtain the partial correlations between a pair of variables
   - Select Tasks → Multivariate → Correlations.
   - From Variables list, select the desired variables NGL and yield (whose partial Correlation is to be obtained) and Click the right arrow and select Analysis variables. Select Option (from the selection pane) and the check box for types of Correlation as Pearson. By Check in the Box: Fisher Options, we can select the level of significance for testing the significance of correlation coefficients, we can also select type of alternative hypothesis as lower, upper or two sided from the combo box, Type of confidence limits.
   - From **Variables** list, select PP, PH(**Partial variable)**. Click the **right arrow** and select **Partial variables**.
   - From the Results Pane uncheck the Show statistics for each variable check box. Check the Show significance probabilities associated with correlations check box. Click Run to generate the Pearson partial correlation coefficient along with their test if significance.

Please note: For title and footnote we have to select the **Title** from selection pane and uncheck the box and type the name of the title and footnote.

4. **Multiple Regression**
   - Select **Tasks → Regression →Linear Regression**.
   - From **Variables** list, select the variables which are **explanatory** (PP, PH, NGL). Click the **right arrow** and select **Explanatory variables**.
   - From **Variables** list, select **Dependent** variable (yield). Click the **right arrow** and select **Dependent variable**.
   - Click **Run** to run the linear regression.



ate ▾ Publish | Properties

Linear Regression Results

The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: yld

Number of Observations Read	46
Number of Observations Used	46

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	23.41086	7.80362	4.40	0.0089
Error	42	74.50133	1.77384		
Corrected Total	45	97.91219			

Root MSE	1.33186	R-Square	0.2391
Dependent Mean	2.43848	Adj R-Sq	0.1848
Coeff Var	54.61834		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.84802	1.05434	-0.80	0.4257
pp	1	0.01200	0.00628	1.91	0.0631
ph	1	1.66061	0.91881	1.81	0.0779
ngl	1	0.15139	0.11941	1.27	0.2118

16.4 Analysis of Data Generated from Completely Randomised Design

Exercise 4: A feeding trial with 3 feeds namely (i) Pasture(control), (ii) Pasture and Concentrates and (iii) Pasture, Concentrates and Minerals was conducted at the Yellachihalli Sheep Farm, Mysore, to study their effect on wool yield of Sheep. For this purpose twenty-five ewe lambs were allotted at random to each of the three treatments and the three treatments and the weight records of the total wool yield (in gms) of first two clipping were obtained. The data for two lambs for feed 1 {Pasture (control)}, three for feed 2 { Pasture and Concentrates}and one for feed 3 {Pasture, Concentrates and Minerals} are missing. The details of the experiment are given below: Yield (in gms)

FEED 1	FEED 2	FEED 3
850.50	510.30	992.25
453.60	963.90	850.50
878.85	652.05	1474.20
623.70	1020.60	510.30
510.30	878.85	850.50
765.45	567.00	793.80
680.40	680.40	453.60
595.35	538.65	935.55
538.65	567.00	1190.70
850.50	510.30	481.95
850.50	425.25	623.70
793.80	567.00	878.85
1020.60	623.70	1077.30
708.75	538.65	850.50

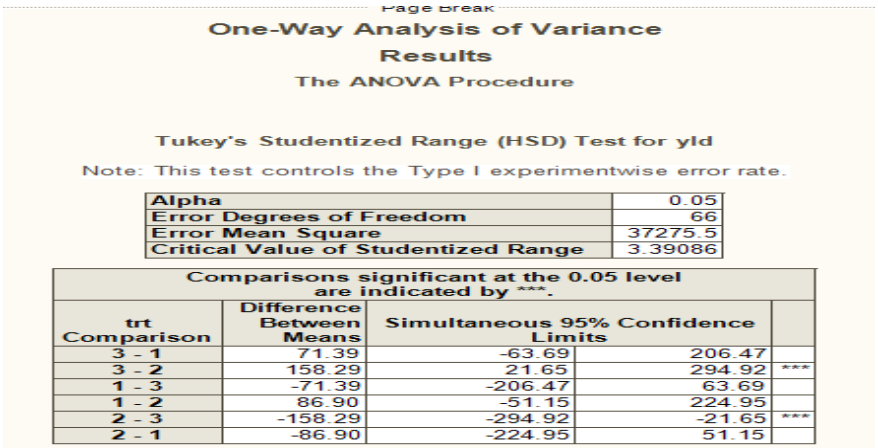
652.05	737.10	680.40
623.70	453.60	737.10
396.90	481.95	737.10
822.15	368.55	708.75
680.40	567.00	708.75
652.05	595.35	652.05
538.65	567.00	567.00
850.50	595.35	453.60
680.40	.	652.05
.	.	567.00
.	.	.

1. Perform the analysis of variance of the data to test whether there is any difference between treatment effects.
2. Perform all possible pair wise treatment comparisons and identify the best treatment i.e. the treatment giving highest yield.

Solution: Above data can be analysed by following steps

- Prepare the data for analysis by taking FEED1, FEED2 and FEED3 as treatments and providing the code 1,2 and 3 respectively (or any other desired codes), the corresponding numerical figure written under the FEED1, FEED2 and FEED3 are yield. Use Ms-Excel Worksheet to create a data file by giving the codes and typing the corresponding yield. In the first column write the treatment number and in the second column write the corresponding yield.
- Import the file in SAS Enterprise guide and then performs the following steps:
- Select **Tasks → ANOVA → One-Way ANOVA.**
- From **Variables to assign** list, select **the independent Variable** (trt). Click the **right arrow** and select **Independent variable**.
- In the **Variables to assign** list, select **Dependent Variable** (yield). Click the **right arrow** and select **Dependent variables**.
- Under **Means** in the selection pane, select **Comparison**.
- If you reject the null hypothesis of equality of treatment effects, then use multiple comparison procedure for all possible pair wise treatment comparisons to determine which of the means are different. A desired Multiple Comparison Procedure from the available options, say **Tukey's studentized range test (HSD)**.
- For this example, you could choose any of the procedures shown. In this case, select **Tukey's studentized range test (HSD)**. Tukey's method examines the difference between all possible combinations of two treatment means.
- Click **Run** to run the one-way ANOVA.

Results are shown in the following snap shot



Page Break

One-Way Analysis of Variance

Results

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for yld

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	66
Error Mean Square	37275.5
Critical Value of Studentized Range	3.39086

Comparisons significant at the 0.05 level are indicated by ***.

trt Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 1	71.39	-63.69	206.47	
3 - 2	158.29	21.65	294.92	***
1 - 3	-71.39	-206.47	63.69	
1 - 2	86.90	-51.15	224.95	
2 - 3	-158.29	-294.92	-21.65	***
2 - 1	-86.90	-224.95	51.15	

16.5 Two Way ANOVA

Exercise 5: An initial varietal trial (Late Sown, irrigated) was conducted to study the performance of 20 new strains of mustard vis-a-vis four checks (Swarna Jyoti: ZC; Vardan: NC; Varuna: NC; and Kranti: NC) using a Randomized complete Block Design (RCB) design at Bhatinda with 3 replications. The seed yield in kg/ha was recorded. The details of the experiment are given below:

Yield in kg/ha

Strain	Code	Replications		
		1	2	3
RK-04-3	MCN-04-110	1539.69	1412.35	1319.73
RK-04-4	MCN-04-111	1261.85	1065.05	1111.36
RGN-124	MCN-04-112	1389.19	1516.54	1203.97
HYT-27	MCN-04-113	1192.39	1215.55	1157.66
PBR-275	MCN-04-114	1250.27	1203.97	1366.04
HUJM-03-03	MCN-04-115	1296.58	1273.43	1308.16
RGN-123	MCN-04-116	1227.12	1018.74	937.71
BIO-13-01	MCN-04-117	1273.43	1157.66	1088.20
RH-0115	MCN-04-118	1180.82	1203.97	1041.90
RH-0213	MCN-04-119	1296.58	1458.65	1250.27
NRCDR-05	MCN-04-120	1122.93	1065.05	1018.74
NRC-323-1	MCN-04-121	1250.27	926.13	1030.32
RRN-596	MCN-04-122	1180.82	1053.47	717.75
RRN-597	MCN-04-123	1146.09	1180.82	856.67
CS-234-2	MCN-04-124	1574.42	1412.35	1597.57
RM-109	MCN-04-125	914.55	972.44	659.87
BAUSM-2000	MCN-04-126	891.40	937.71	798.79
NPJ-99	MCN-04-127	1227.12	1203.97	1389.19
SWARNA JYOTI(ZC)	**MCN-04-128**	1389.19	1180.82	1273.43

VARDAN(NC)	**MCN-04-129**	1331.31	1157.66	1180.82
PR-2003-27	MCN-04-130	1250.27	1250.27	1296.58
VARUNA(NC)	**MCN-04-131**	717.75	740.90	578.83
PR-2003-30	MCN-04-132	1169.24	1157.66	1111.36
KRANTI-(NC)	**MCN-04-133**	1203.97	1296.58	1250.27

1. Perform the analysis of variance of the data to test whether there is any difference between treatment effects.
2. Perform all possible pair wise treatment comparisons and identify the best treatment i.e. the treatment giving highest yield. Also identify the other treatments which are non-significantly different from this treatment.
3. The varieties Swarna Jyoti (MCN-04-128), Vardan (MCN-04-129), Varuna (MCN-04-131) and Kranti (MCN-04-133) were check varieties and rest of them were strains. Test whether the performance of check varieties is significantly different from strains.

Solution: Above data can be analysed by following steps

- Prepare the data for analysis by taking FEED1, FEED2 and FEED3 as treatments and providing the code 1,2 and 3 respectively (or any other desired codes), the corresponding numerical figure written under the FEED1, FEED2 and FEED3 are yield. Use Ms-Excel Worksheet to create a data file by giving the codes and typing the corresponding yield. In the first column write the treatment number and in the second column write the corresponding yield.
- Import the file in SAS Enterprise guide and then performs the following steps:
- Select **Tasks → ANOVA → Linear Models.**
- From **Variables to assign** list, select press **CTRL** and select Replication (**rep)** and Treatment (**trt)** Click the **right arrow** and select **Classification variables**.
- In the **Variables to assign** list, select **Dependent Variable** (yield). Click the **right arrow** and select **Dependent variables**.
- In the selection pane, select **Model**.
- In the **Class and quantitative variables** list, press **CTRL** and select **Rep** and **Trt** Click **Main**.
- For desired type of Sum of Squares select from Model Options
- For multiple comparison procedure, select from Post Hoc Tests, Least squares and then select the trt under the options for means tests, select the adjustment method from Comparisons, select all pair wise differences from Show p-values of differences.
- If the data is balanced as in case of RCB design, then for multiple comparison procedure, select from Post Hoc Tests, Arithmetic and then select the trt under the options for means tests, select the comparison method from Comparisons, select show means for all model variables from Means Option along with Join non-significant subsets and sort means in ascending/descending order and then select yes for show for all pair wise differences in the option Confidence Intervals.
  (If analysis of covariance is to be performed, then select the covariate in the Covariates option)
- Click **Run**
  A part of the result is shown in the following snap shot

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	287872.433	143936.216	3.86	0.0259
Error	66	2460182.924	37275.499		
Corrected Total	68	2748055.357			

R-Square	Coeff Var	Root MSE	yld Mean
0.104755	27.83783	193.0686	693.5478

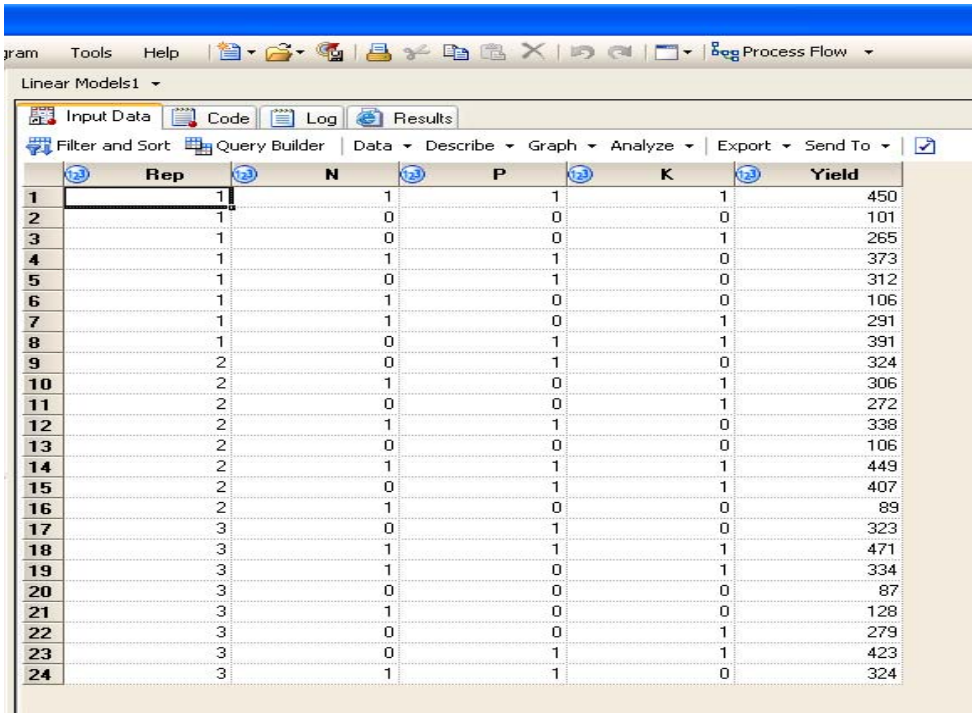
Source	DF	Anova SS	Mean Square	F Value	Pr > F
trt	2	287872.4328	143936.2164	3.86	0.0259

16.6 Factorial Randomised Complete Block Design:

Exercise 6: Carry out the analysis of following data

Rep	N	P	K	Yield
1	1	1	1	450
1	0	0	0	101
1	0	0	1	265
1	1	1	0	373
1	0	1	0	312
1	1	0	0	106
1	1	0	1	291
1	0	1	1	391
2	0	1	0	324
2	1	0	1	306
2	0	0	1	272
2	1	1	0	338
2	0	0	0	106
2	1	1	1	449
2	0	1	1	407
2	1	0	0	89
3	0	1	0	323
3	1	1	1	471
3	1	0	1	334
3	0	0	0	87
3	1	0	0	128
3	0	0	1	279
3	0	1	1	423
3	1	1	0	324

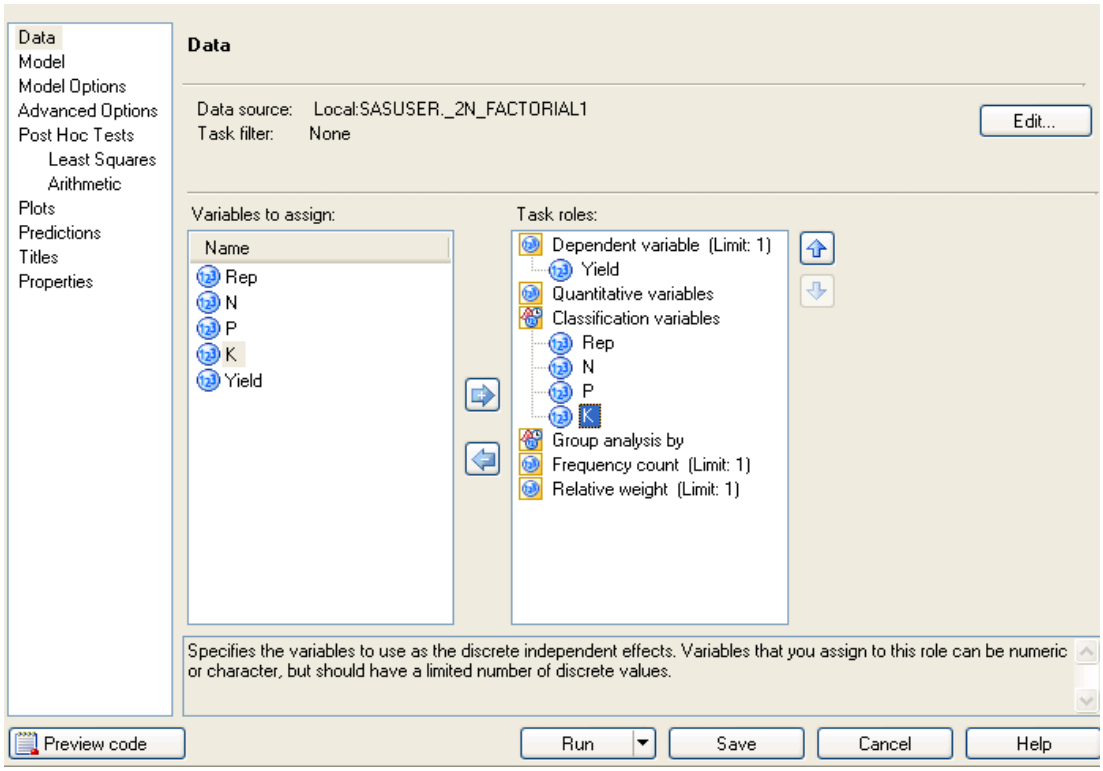
Solution: First enter the above data in Enterprise Guide or enter the data in MS-Excel and then import that file in the Enterprise Guide. The snap shot of data in enterprise guide is as under.



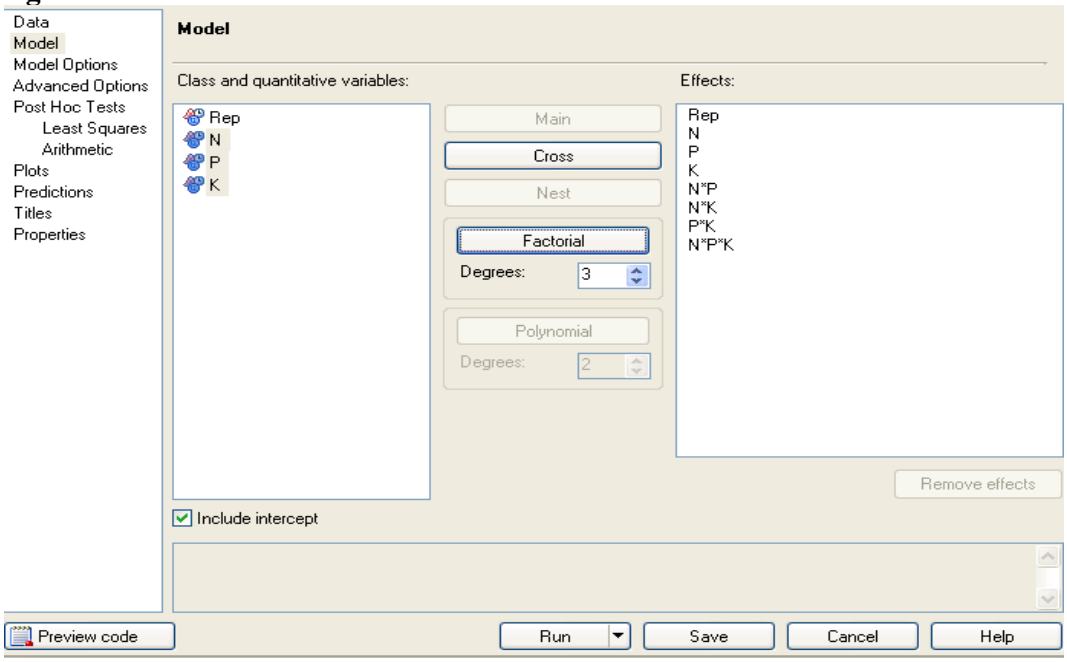
	Rep	N	P	K	Yield
1	1	1	1	1	450
2	1	0	0	0	101
3	1	0	0	1	265
4	1	1	1	0	373
5	1	0	1	0	312
6	1	1	0	0	106
7	1	1	0	1	291
8	1	0	1	1	391
9	2	0	1	0	324
10	2	1	0	1	306
11	2	0	0	1	272
12	2	1	1	0	338
13	2	0	0	0	106
14	2	1	1	1	449
15	2	0	1	1	407
16	2	1	0	0	89
17	3	0	1	0	323
18	3	1	1	1	471
19	3	1	0	1	334
20	3	0	0	0	87
21	3	1	0	0	128
22	3	0	0	1	279
23	3	0	1	1	423
24	3	1	1	0	324

- Select **Task → ANOVA →Linear Model**
- From Variable **to assign** list select **Yield** and place in **Task roles** window under **Dependent Variable** (see snapshot Selecting Variables for Analysis). Again from **Variable to assign list,** select **Rep, N, P** and **K** and place all these variable under **Classification Variables.**
- Click the **Model** select **Rep,** click **Main, Variable** Rep will get placed in **Effects** window, similarly select **N, K** one by one from the **Class and quantitative variables** window and place in **Effects** window**. Now** select **N** press **CTRL** key then select P(Interaction of N and P) click **Cross** or **Factorial** it will get placed in the **Effects** window**,** similarly N* K, K* P and N*P*K. (see snapshot Defining Model)**.**
- Select **Post Hoc Test→ Arithmetic → Add** option means for test window get activated, select variable **Rep** and then from drop down option **True** and again click **Add** we find that **Rep** get placed under the **Effect to Estimate** window**.** Similarly we can place the other variable N, P, K and interactions N*P, P*K, N*K and N*P*K (See Snap shot Post-hoc Test).
- For comparision Select Rep from **Effect to Estimate** window and then click **Comparision Methods,** select the desired method from drop down, we are selecting **Pairwise T test .** Similarly we repeat this process for **N**, P, K etc. (See Snap shot Post-hoc Test).
- Select the **Plots** from selection pane, Uncheck the box "Show plots for linear model analysis" as we do not want the graphs.
- Click **Run**

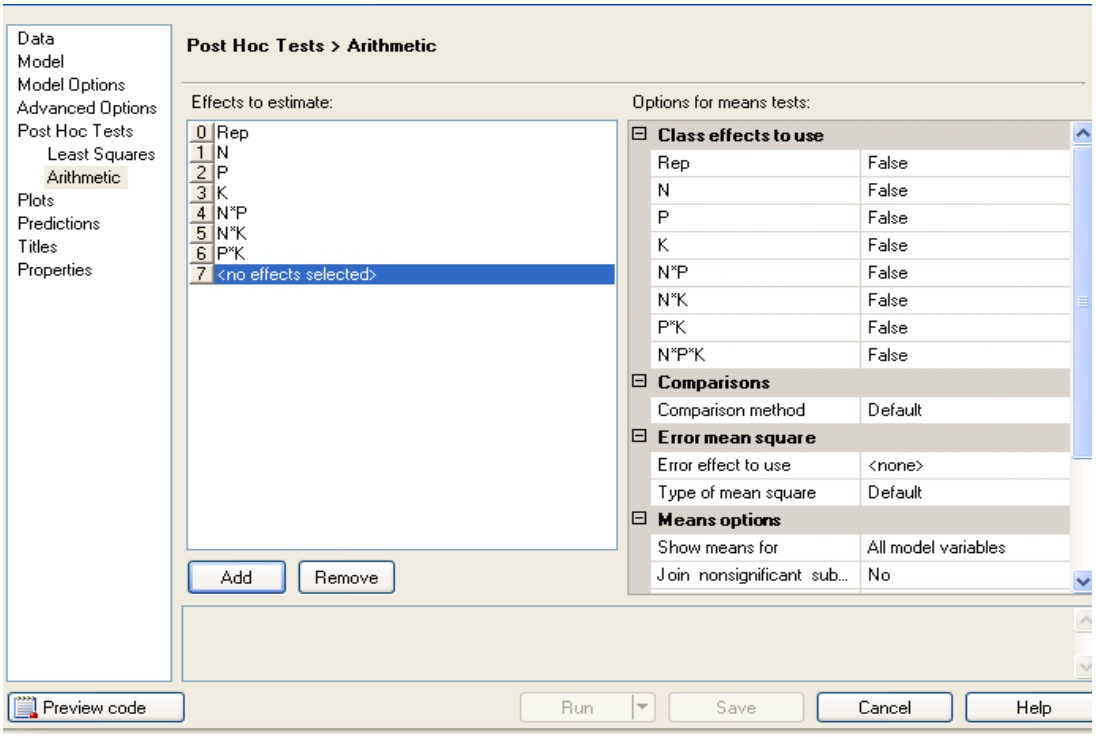
Selecting Variables for Analysis



Defining Model



Post-hoc Test



Output

Linear Models

The GLM Procedure

Dependent Variable: Yield Yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	349171.6250	38796.8472	147.90	<.0001
Error	14	3672.3333	262.3095		
Corrected Total	23	352843.9583			

R-Square	Coeff Var	Root MSE	Yield Mean
0.989592	5.593659	16.19597	289.5417

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Rep	2	520.3333	260.1667	0.99	0.3955
N	1	5673.3750	5673.3750	21.63	0.0004
P	1	205535.0417	205535.0417	783.56	<.0001
K	1	124272.0417	124272.0417	473.76	<.0001
N*P	1	273.3750	273.3750	1.04	0.3246
N*K	1	1053.3750	1053.3750	4.02	0.0648
P*K	1	11837.0417	11837.0417	45.13	<.0001
N*P*K	1	7.0417	7.0417	0.03	0.8722

16.7 Split Plot Analysis

Exercise 7:. The experiment was conducted using a split plot design with method of sowing in paddy in main plots and six sub-plot treatments consisting of organic, inorganic fertilizers and micronutrients (B1: 150 kg/ha of N as Urea+60 kg/ha of P_2O_5 as Super+40 kg/ha of K_2O as Murate of Potash. as recommended inorganic fertilizer, B2: B1 + 150 q/ha of FYM, B3=B1+residual effect of Green manure(Sesbania), B4=B1+$MnSO_4$ @ 0.5% as foliar spray, B5= B1 + 150 q/ha of FYM+ $MnSO_4$ @ 0.5% as foliar spray and B6= B1+residual effect of Green manure(Sesbania)+ $MnSO_4$ @ 0.5% as foliar spray. There were 3 replications, and the data of wheat yield in kg/ha is:

	Rep I		**Rep II**		**Rep III**	
	M1	M2	M1	M2	M1	M2
S1	4940	4900	4830	5020	5080	5090
S2	4810	4920	5110	5110	5160	5130
S3	5150	5070	4920	5230	5180	4980
S4	5090	4890	4900	5120	5190	5200
S5	5130	5150	4880	5160	5160	4920
S6	5140	5070	4930	5200	5280	5250

Perform the analysis of the data to test the significance of effects

Solution: First of all we have to prepare the data for split plot analysis by entering the data in following format:

Data Entry

Import Data (split.xls[SPLITPLOT])

Code | Log | Output Data

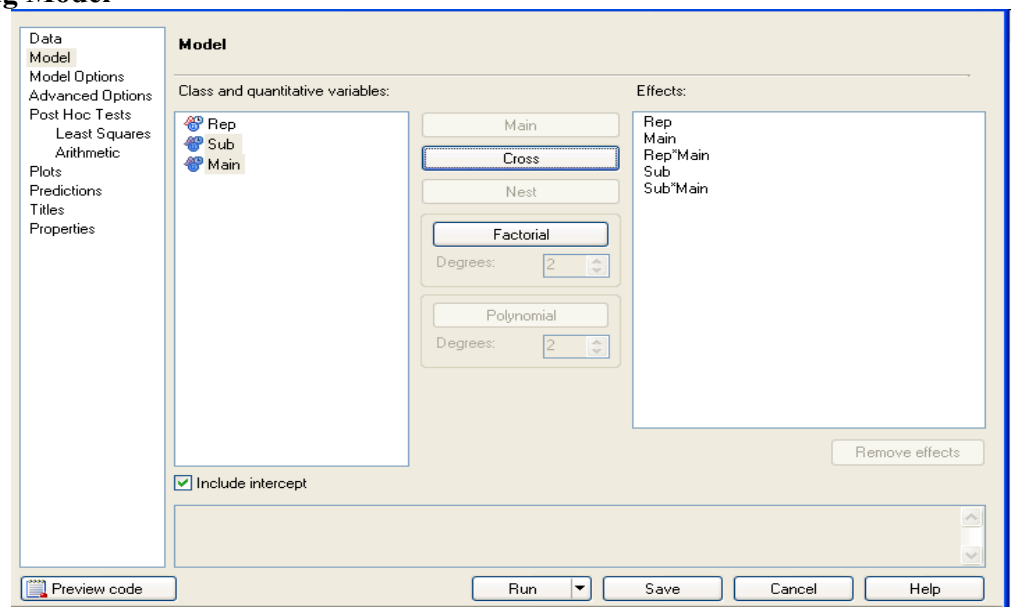
Modify Task | Filter and Sort | Query Builder | Data | Describe | Grap

	Rep	Main	Sub	Yield
1	1	1	1	4940
2	1	1	2	4810
3	1	1	3	5150
4	1	1	4	5090
5	1	1	5	5130
6	1	1	6	5140
7	1	2	1	4900
8	1	2	2	4920
9	1	2	3	5070
10	1	2	4	4890
11	1	2	5	5150
12	1	2	6	5070
13	2	1	1	4830
14	2	1	2	5110
15	2	1	3	4920
16	2	1	4	4900
17	2	1	5	4880
18	2	1	6	4930
19	2	2	1	5020
20	2	2	2	5110
21	2	2	3	5230
22	2	2	4	5120
23	2	2	5	5160
24	2	2	6	5200
25	3	1	1	5080
26	3	1	2	5160
27	3	1	3	5180
28	3	1	4	5190
29	3	1	5	5160
30	3	1	6	5280
31	3	2	1	5090
32	3	2	2	5130
33	3	2	3	4980

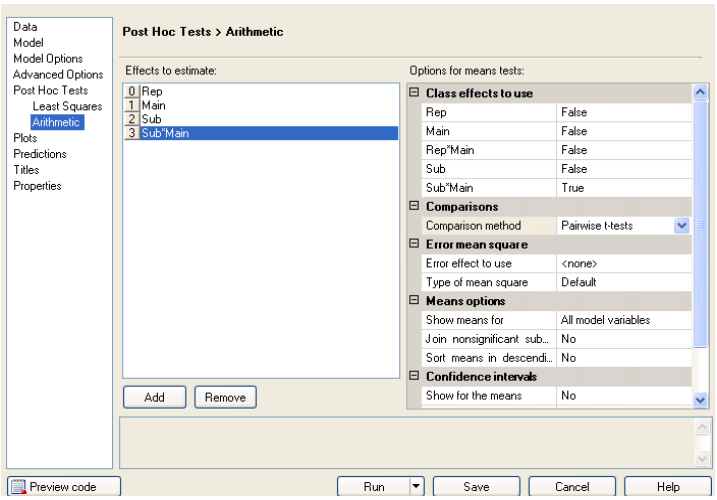
Steps of Analysis

- Select **Task → ANOVA →Linear Model**
- From **Variable to assign** list select **Yield** and place in the **Task roles** window as **Dependent Variable.** Select **Rep, Main,** and **Sub** and place this entire variable as **Classification Variables.**
- Select the **Model** from selection pane, the list of variable appear in "Class and quantitative variable" window, select **Rep,** click **Main** it will get placed in **Effects** window, similarly select **Main, Rep*Main (**select Rep press CTRL key select Main)**, Sub and Sub*Main** (select Sub press CTRL key select Main) **,** click **Cross** or **Factorial** it will get placed in the window **Effects,** (See snapshot Defining Model)**.**
- Select **Model Option** from selection pane, Model option window appears on the screen, Check the box **Type III** and Uncheck all other boxes.
- Select **Post Hoc Test→ Arithmetic → Add ,** the "option means for test" window get activated, select variable **Rep** and then select **True** from drop down options and again click **Add** we find that **Rep** get placed in the **Effect to Estimate** window**.** Similarly we can place the **Main**, **Rep*Main**, **Sub** and **Sub*Main** (See Snap shot Post-hoc Test).
- For comparison Select **Rep** from **Effect to Estimate** window and then click **Comparison Methods,** select the desired method from drop down, we are selecting **Pairwise T test .** Similarly we repeat this process for Main and Sub (See Snap shot Post-hoc Test).
- Select the **Plots** from selection pane, Uncheck the box "Show plots for linear model analysis" as we do not want the graphs.
- Click **Run**

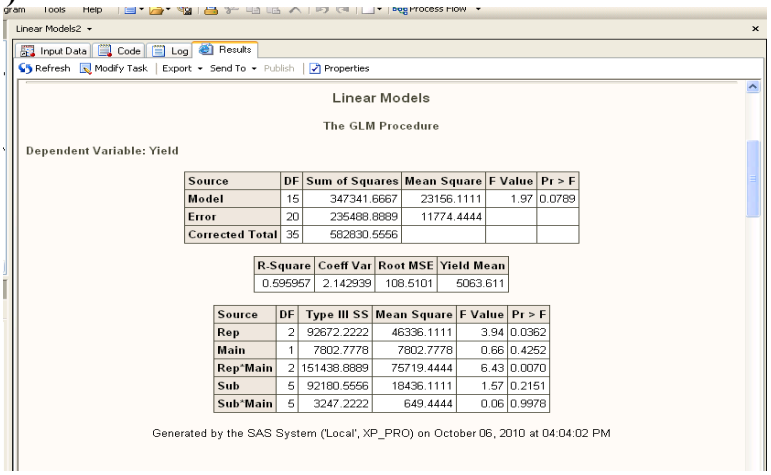
Defining Model



Post-Hoc test



Output (ANOVA)



Linear Models2

Input Data | Code | Log | Results

Refresh | Modify Task | Export | Send To | Publish | Properties

Linear Models

The GLM Procedure

Dependent Variable: Yield

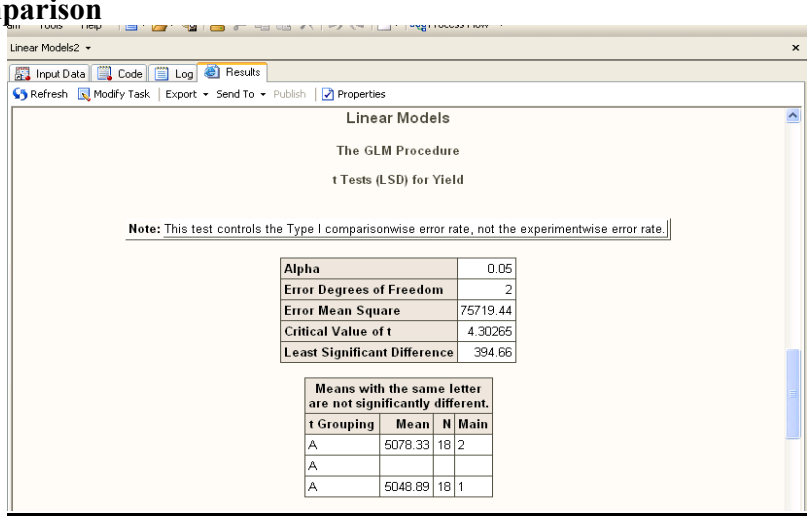
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	347341.6667	23156.1111	1.97	0.0789
Error	20	235488.8889	11774.4444		
Corrected Total	35	582830.5556			

R-Square	Coeff Var	Root MSE	Yield Mean
0.595957	2.142939	108.5101	5063.611

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Rep	2	92672.2222	46336.1111	3.94	0.0362
Main	1	7802.7778	7802.7778	0.66	0.4252
Rep*Main	2	151438.8889	75719.4444	6.43	0.0070
Sub	5	92180.5556	18436.1111	1.57	0.2151
Sub*Main	5	3247.2222	649.4444	0.06	0.9978

Generated by the SAS System ('Local', XP_PRO) on October 06, 2010 at 04:04:02 PM

Main plot Comparison



Linear Models2

Input Data | Code | Log | Results

Refresh | Modify Task | Export | Send To | Publish | Properties

Linear Models

The GLM Procedure

t Tests (LSD) for Yield

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	2
Error Mean Square	75719.44
Critical Value of t	4.30265
Least Significant Difference	394.66

Means with the same letter are not significantly different.

t Grouping	Mean	N	Main
A	5078.33	18	2
A			
A	5048.89	18	1

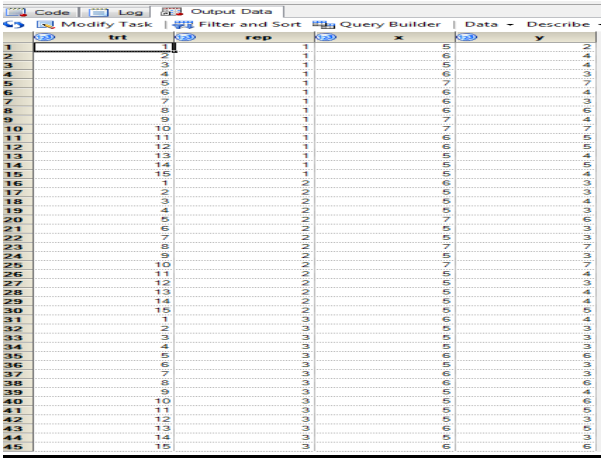
Please note that for split plot design, the minimum significant differences can only be obtained through Syntax approach.

16.8 Analysis of Covariance

Exercise 8: A trial was designed to evaluate *15* rice varieties grown in soil with a toxic level of iron. The experiment was in a RCB design with three replications. Guard rows of a susceptible check variety were planted on two sides of each experimental plot. Scores for tolerance for iron toxicity were collected from each experimental plot as well as from guard rows. For each experimental plot, the score of susceptible check (averaged over two guard rows) constitutes the value of the covariate for that plot. Data on the tolerance score of each variety (Y variable) and on the score of the corresponding susceptible check (X variable) are shown below:

Variety Number	Replication I		Replication II		Replication III	
	X	Y	X	Y	X	Y
1	5	2	6	3	6	4
2	6	4	5	3	5	3
3	5	4	5	4	5	3
4	6	3	5	3	5	3
5	7	7	7	6	6	6
6	6	4	5	3	5	3
7	6	3	5	3	6	3
8	6	6	7	7	6	6
9	7	4	5	3	5	4
10	7	7	7	7	5	6
11	6	5	5	4	5	5
12	6	5	5	3	5	3
13	5	4	5	4	6	5
14	5	5	5	4	5	3
15	5	4	5	5	6	6

Data Entry

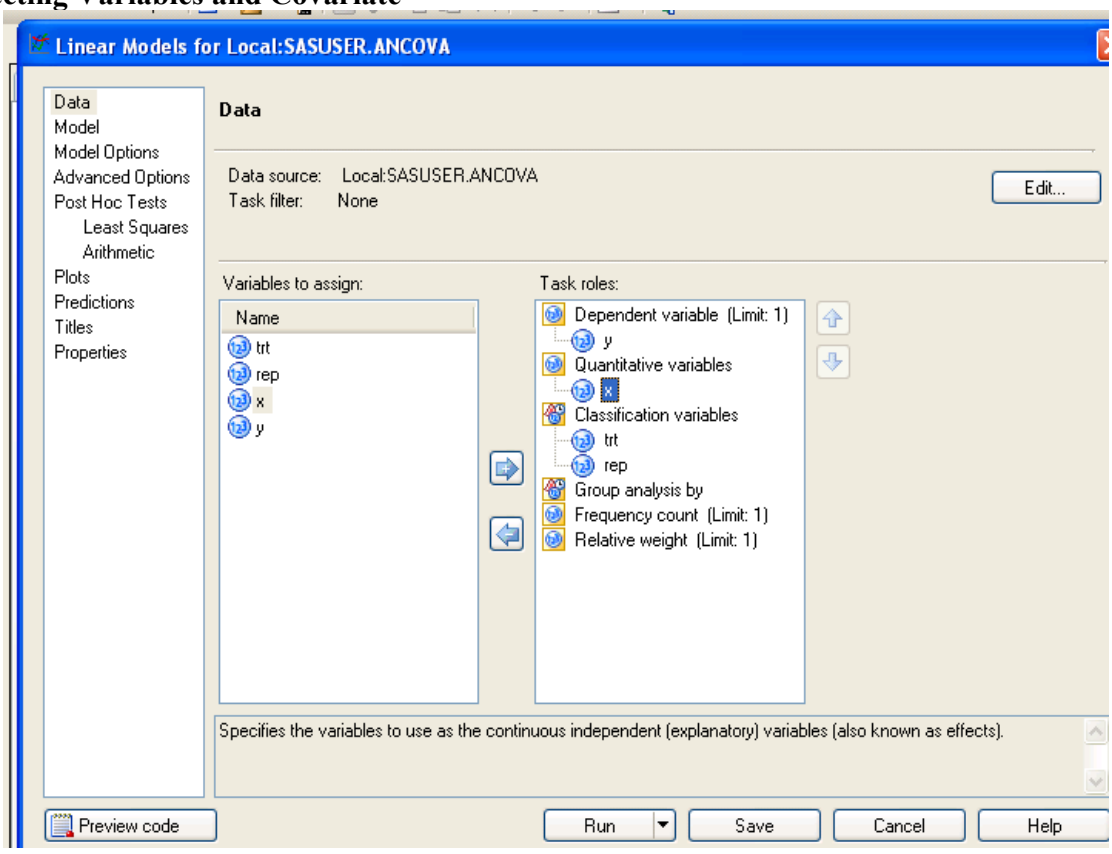


Step for Analysis

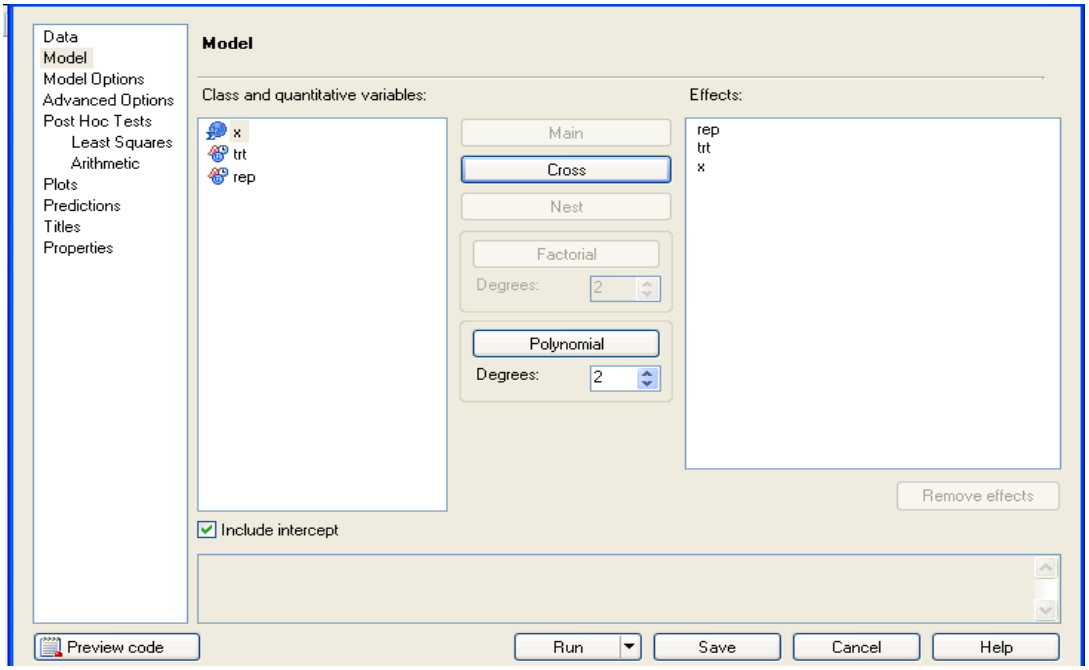
- Select **Task → ANOVA →Linear Model**

- From **Variable to assign** list select "**y"** and place in the **Task roles** window as **Dependent Variable.** Select "**x"** from Variable to assign and place it as **Quantitative variable,** select **trt** and **rep** and place in as **Classification variable.**
- Select the **Model** from selection pane, the list of variable appear in "Class and quantitative variable" window, select **x ,** click **Main** it will get placed in **Effects** window**.** Similarly select **trt** and **rep** one by one and click **Main** it will get placed in **Effects** window**. from '**Class and quantitative variables' window**,** click **Main** it will get placed in **Effects** window, similarly select **Main, Rep*Main (**select Rep press CTRL key select Main)**, Sub and Sub*Main** (select Sub press CTRL key select Main) **,** click **Cross** or **Factorial** it will get placed in the window **Effects,** (See snapshot Defining Model)**.**
- **S**elect **Model Option** from selection pane, Model option window appear on the screen, Check the box **Type III** and Uncheck all other boxes.
- Select the **Plots** from selection pane, Uncheck the box "Show plots for linear model analysis" as we do not want the graphs.
- Click **Run.**

Selecting Variables and Covariate



Defining Model



Output

Linear Models

The GLM Procedure

Class Level Information		
Class	Levels	Values
rep	3	1 2 3
trt	15	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Number of Observations Read	45
Number of Observations Used	45

Generated by the SAS System ('Local', W32_VSPRO) on October 07, 2010 at 05:08:09 PM

Page Break

Linear Models

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	819.2000000	819.2000000	435.32	<.0001
Error	44	82.8000000	1.8818182		
Uncorrected Total	45	902.0000000			

R-Square	Coeff Var	Root MSE	y Mean
0.000000	32.15142	1.371794	4.266667

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Intercept	1	819.2000000	819.2000000	435.32	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	4.266666667	0.20449494	20.86	<.0001

Generated by the SAS System ('Local', W32_VSPRO) on October 07, 2010 at 05:08:09 PM

16.9 Analysis of BIB design

Exercise 9: Considering the following data to carry out BIB design analysis

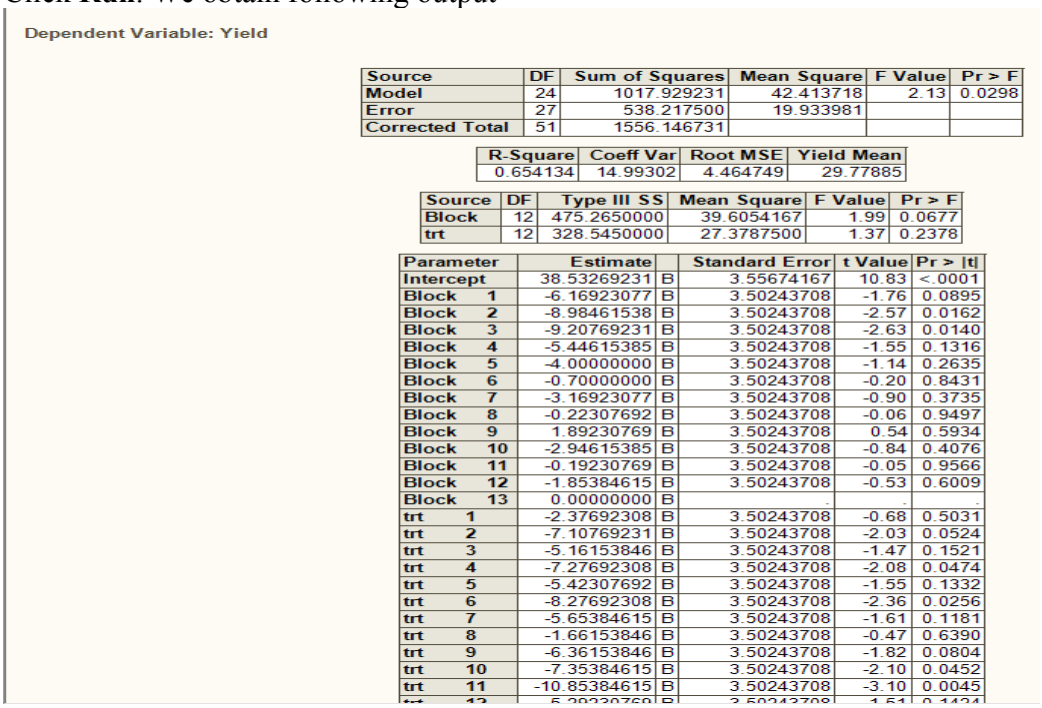
Block	trt	Yield

1	11	24.6
1	6	19.9
1	9	29
1	3	25.3
2	4	19.8
2	8	33.3
2	3	23
2	12	22.7
3	12	31.7
3	13	26.6
3	11	19.3
3	10	16.2
4	2	27.3
4	5	27
4	8	35.6
4	11	17.4
5	7	23.4
5	8	30.5
5	9	30.8
5	10	32.4
6	4	30.6
6	5	32.4
6	6	27.2
6	10	32.8
7	1	34.7
7	5	31.1
7	9	25.7
7	12	30.5
8	3	34.4
8	5	32.4
8	7	33.3
8	13	36.9
9	1	38.2
9	2	32.9
9	3	37.3
9	10	31.3
10	2	28.7
10	4	30.7
10	9	26.9
10	13	35.3
11	1	36.6
11	4	31.1
11	7	31.1
11	11	28.4
12	1	31.8

12	6	33.7
12	8	27.8
12	13	41.1
13	2	30.3
13	6	31.5
13	7	39.3
13	12	26.7

Step for Analysis

- Select **Task → ANOVA →Linear Model**
- From **Variable to assign** list select "**yield"** and place in the **Task roles** window as **Dependent Variable.** Select **Block** and **trt** and place in as **Classification variable.**
- Select the **Model** from selection pane, select **Block**, click **Main** it will get placed in **Effects** window**.** Similarly select **trt** and click **Main** it will get placed in **Effects** window.
- **S**elect **Model Option** from selection pane, Model option window appears on the screen, Check the box **Type III** and uncheck all other boxes.
- Select the **Plots** from selection pane, Uncheck the box "Show plots for linear model analysis" as we do not want the graphs.
- Click **Run**. We obtain following output



Dependent Variable: Yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	24	1017.929231	42.413718	2.13	0.0298
Error	27	538.217500	19.933981		
Corrected Total	51	1556.146731			

R-Square	Coeff Var	Root MSE	Yield Mean
0.654134	14.99302	4.464749	29.77885

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Block	12	475.2650000	39.6054167	1.99	0.0677
trt	12	328.5450000	27.3787500	1.37	0.2378

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		38.53269231	B	3.55674167	10.83	<.0001
Block	1	-6.16923077	B	3.50243708	-1.76	0.0895
Block	2	-8.98461538	B	3.50243708	-2.57	0.0162
Block	3	-9.20769231	B	3.50243708	-2.63	0.0140
Block	4	-5.44615385	B	3.50243708	-1.55	0.1316
Block	5	-4.00000000	B	3.50243708	-1.14	0.2635
Block	6	-0.70000000	B	3.50243708	-0.20	0.8431
Block	7	-3.16923077	B	3.50243708	-0.90	0.3735
Block	8	-0.22307692	B	3.50243708	-0.06	0.9497
Block	9	1.89230769	B	3.50243708	0.54	0.5934
Block	10	-2.94615385	B	3.50243708	-0.84	0.4076
Block	11	-0.19230769	B	3.50243708	-0.05	0.9566
Block	12	-1.85384615	B	3.50243708	-0.53	0.6009
Block	13	0.00000000	B	.	.	.
trt	1	-2.37692308	B	3.50243708	-0.68	0.5031
trt	2	-7.10769231	B	3.50243708	-2.03	0.0524
trt	3	-5.16153846	B	3.50243708	-1.47	0.1521
trt	4	-7.27692308	B	3.50243708	-2.08	0.0474
trt	5	-5.42307692	B	3.50243708	-1.55	0.1332
trt	6	-8.27692308	B	3.50243708	-2.36	0.0256
trt	7	-5.65384615	B	3.50243708	-1.61	0.1181
trt	8	-1.66153846	B	3.50243708	-0.47	0.6390
trt	9	-6.36153846	B	3.50243708	-1.82	0.0804
trt	10	-7.35384615	B	3.50243708	-2.10	0.0452
trt	11	-10.85384615	B	3.50243708	-3.10	0.0045

- Once again repeating the all above steps with minor change in the selection of model. Select the **Model** from selection pane, select **trt,** click **Main** it will get placed in **Effects** window. Now select **Block** and click **Main** it will get placed in **Effects** window**.** This time we have selected **trt** first so it adjust for **Block**. Clicking the **Run** we obtain following output.

16.10 Non Parametric Tests

One Sample location test

Exercise 10: Following data is related to the length(in cm) of the ear-head of a wheat variety 9.3, 18.8, 10.7, 11.5, 8.2, 9.7, 10.3, 8.6, 11.3, 10.7, 11.2, 9.0, 9.8, 9.3, 10.3, 10, 10.1 9.6, 10.4. Test the data that the median length of ear-head is 9.9 cm.

Solution: Given data is one sample, so we apply the Sign test or Wilcoxon signed rank test. The step of analysis is as under:-

- Select **Tasks → Describe →Distribution**
- In the **Variables to assign** list, select **Analysis variables.**
- In the selection pane, select **Tables.** Clear the **Basic confidence intervals** and **Basic measures** check boxes. Keep selecting **Test for Location** and type $\mathbf{H_0=9.9}$
- Click **Run**.

k-Independent Samples

Exercise 11: An experiment was conducted with 21 animals to determine if the four different feeds have the same distribution of Weight gains on experimental animals. The feeds 1, 3 and 4 were given to 5 randomly selected animals and feed 2 was given to 6 randomly selected animals. The data obtained is presented in the following table.

Feeds	Weight gains (kg)					
1	3.35	3.8	3.55	3.36	3.81	
2	3.79	4.1	4.11	3.95	4.25	4.4
3	4	4.5	4.51	4.75	5	
4	3.57	3.82	4.09	3.96	3.82	

Solution: First enter the data in SAS EG data sheet or in other ASCII format in the following manner

feed	wt
1	3.35
1	3.8
1	3.55
1	3.36
1	3.81
2	3.79
2	4.1
2	4.11
2	3.95
2	4.25
2	4.4
3	4
3	4.5
3	4.51

3	4.75
3	5
4	3.57
4	3.82
4	4.09
4	3.96
4	3.82

For k-independent samples, we are using Kruskal-Wallis test to analyze the above data by using the following steps.

- Select **Tasks → ANOVA → Nonparametric One-Way ANOVA**
- In the **Variables to assign** list, select **wt** as **Dependent variables.**
- In the **Variables to assign** list, select **feed** as **Independent variable.**
- In the selection pane, select **Analysis.** Under **Test scores**, clear the **Median**, **Savage**, and **Van der Waerden** check boxes. In this case it will perform the analysis of Kruskal-Wallis test along with Median, Savage, and Van der Waerden. If we want some selected test then we have to check that particular box.
- Clear the **Calculate empirical distribution function statistics (EDF)** check box.
- Click **Run.**

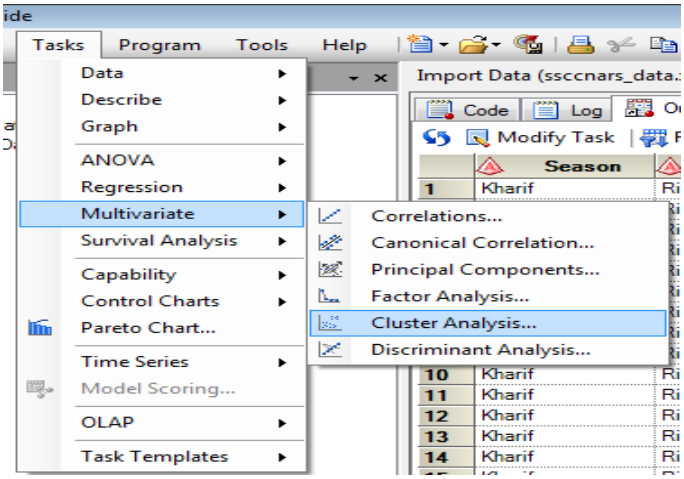
16.11 Cluster Analysis

Cluster Analysis task creates hierarchical clusters of the observations that contains either co-ordinate data or distance data. If the data set contains coordinate data, the task computes Euclidean distances before applying the clustering methods. Result can be obtained in the graphical form of the hierarchical clustering to produce a tree diagram. This tree diagram is called a dendrogram. By using the K-means method of the cluster analysis, it creates non-hierarchical clusters of coordinate data. Cluster analysis can be used to analyse the population data. We can use this task to analyze population data. For example, suppose that we are interested to determine whether national figures for birth rates, death rates, and infant death rates can be used to determine certain types or categories of countries. One can perform a cluster analysis to determine whether the observations can be formed into groups that are suggested by the data.

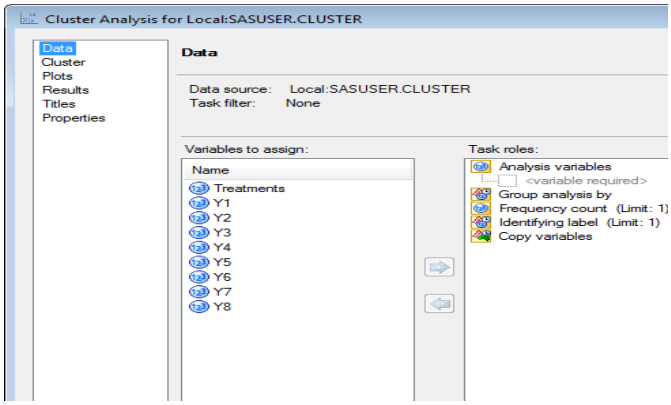
Exercise 12: For illustration of cluster analysis let us download example of the cluster analysis from the link analysis of Data on Design Resources Sever

(http://iasri.res.in/design/Analysis%20of%20data/cluster_analysis.html), in which 8 characters observed in an experiment to evaluate 110 genotypes of Lentil conducted using an alpha-design in 3 replications with block size 10.

To analyse this data, open or import the data file downloaded from design resource sever then from top menu select **Task → Multivariate → Cluster Analysis**

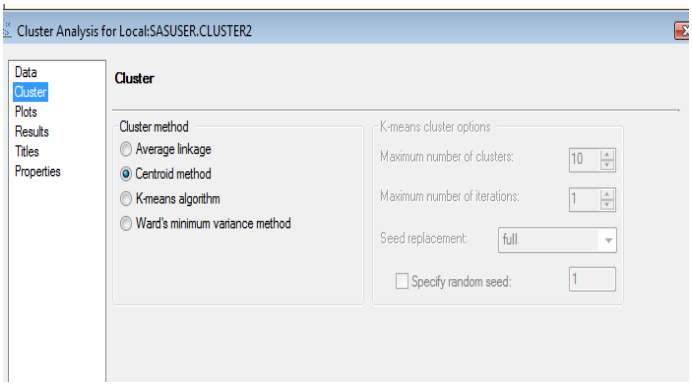


In the selection pane, click **Data** to access these options



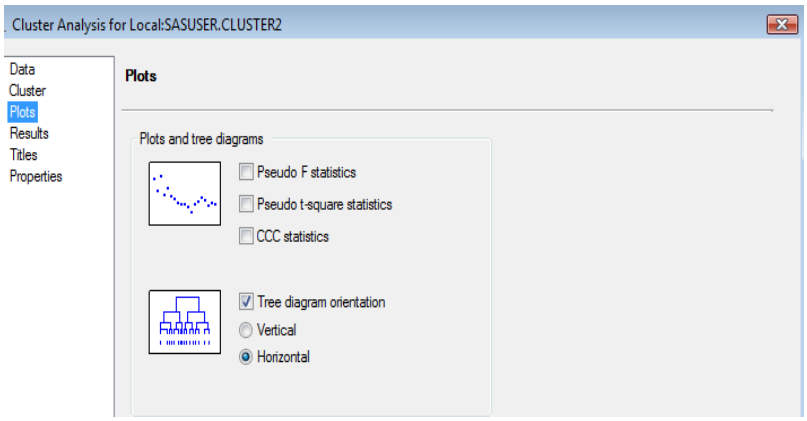
Select the **Analysis variable** from the **Variable to assign** then drag and drop to **task roles** window .In our case analysis variable is **Treatment** .There must be at least one variable to this role, analyses will be performed on each variables that we assign to this role.

Click on **Cluster** to choose the method of Cluster analysis.



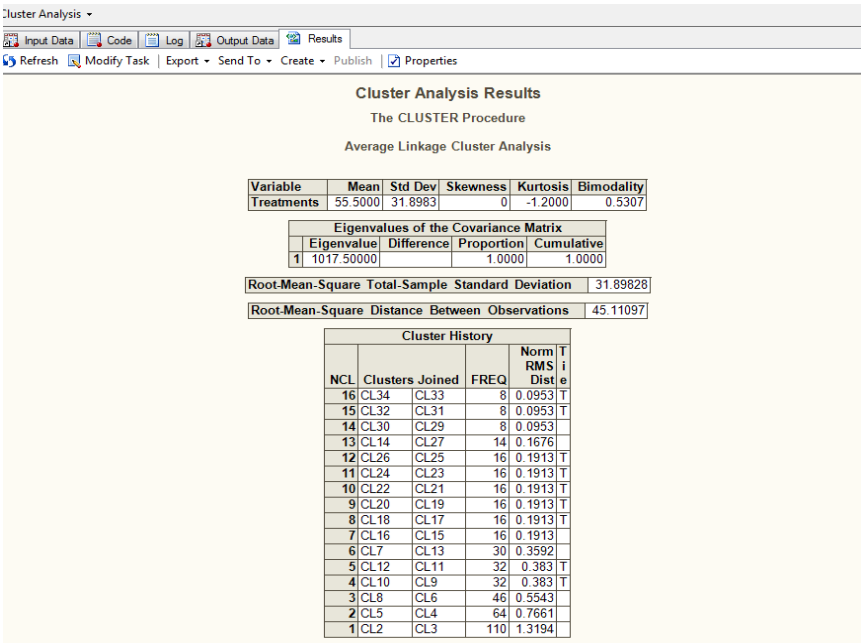
Click on **Plot** to select the **plot** and **tree diagram.** Here we are selecting **Tree diagram orientation.** A graphical view, help us in interpreting the clusters. Plots are generated using a

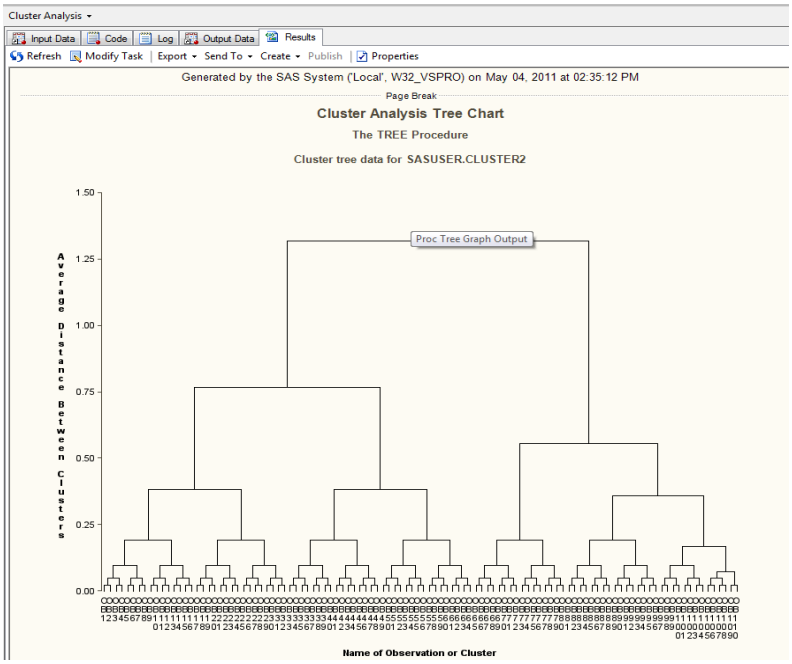
subset of the input data source. Tree diagram orientation box is the **dedogram** which is most useful in the cluster analysis**. Pseudo F statistics** displays a scatter plot for the pseudo *F*-



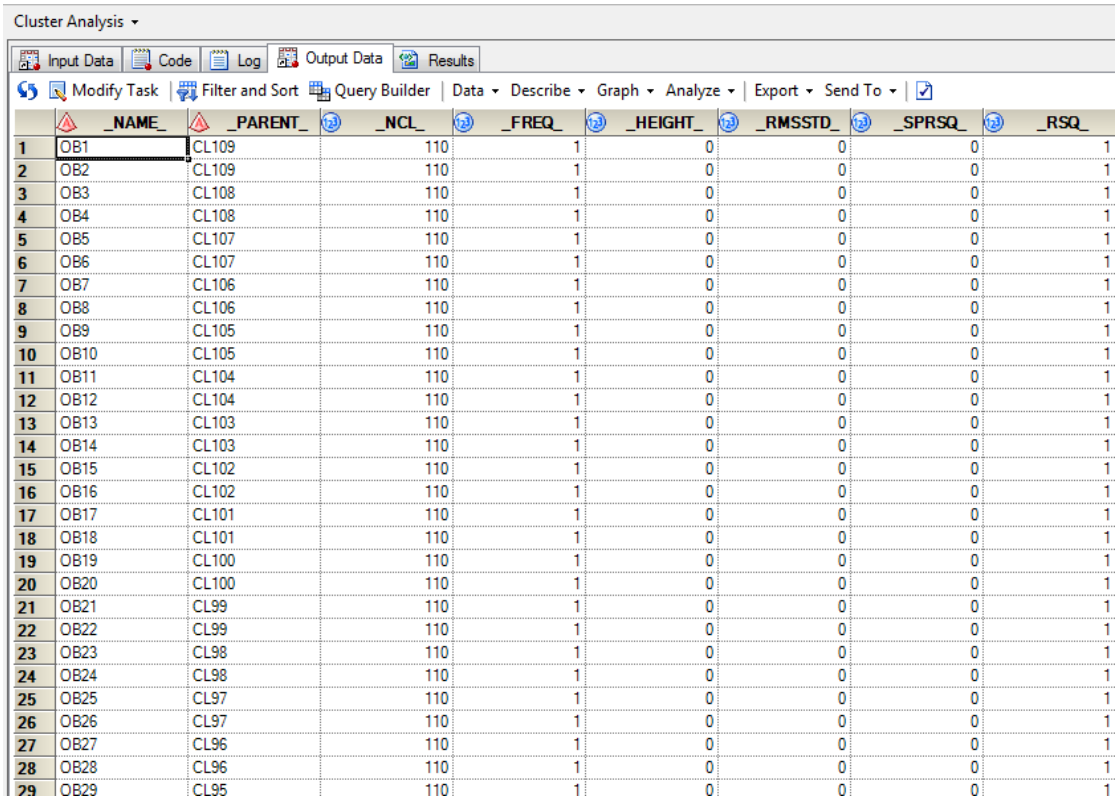
statistic against the number of clusters. **Pseudo t-square statistics** displays a scatter plot for the pseudo t^2 statistic against the number of clusters. **CCC statistics** displays a scatter plot for the cubic clustering criterion (CCC) statistic against the number of clusters.

Click **Result** and check the boxes **Display output** and **Simple Summary Statistics. Cluster generation** box can be used to set the desired number of cluster**, if any.**





The output data generated is as under



	NAME	_PARENT_	_NCL_	_FREQ_	_HEIGHT_	_RMSSTD_	_SPRSQ_	_RSQ_
1	OB1	CL109	110	1	0	0	0	1
2	OB2	CL109	110	1	0	0	0	1
3	OB3	CL108	110	1	0	0	0	1
4	OB4	CL108	110	1	0	0	0	1
5	OB5	CL107	110	1	0	0	0	1
6	OB6	CL107	110	1	0	0	0	1
7	OB7	CL106	110	1	0	0	0	1
8	OB8	CL106	110	1	0	0	0	1
9	OB9	CL105	110	1	0	0	0	1
10	OB10	CL105	110	1	0	0	0	1
11	OB11	CL104	110	1	0	0	0	1
12	OB12	CL104	110	1	0	0	0	1
13	OB13	CL103	110	1	0	0	0	1
14	OB14	CL103	110	1	0	0	0	1
15	OB15	CL102	110	1	0	0	0	1
16	OB16	CL102	110	1	0	0	0	1
17	OB17	CL101	110	1	0	0	0	1
18	OB18	CL101	110	1	0	0	0	1
19	OB19	CL100	110	1	0	0	0	1
20	OB20	CL100	110	1	0	0	0	1
21	OB21	CL99	110	1	0	0	0	1
22	OB22	CL99	110	1	0	0	0	1
23	OB23	CL98	110	1	0	0	0	1
24	OB24	CL98	110	1	0	0	0	1
25	OB25	CL97	110	1	0	0	0	1
26	OB26	CL97	110	1	0	0	0	1
27	OB27	CL96	110	1	0	0	0	1
28	OB28	CL96	110	1	0	0	0	1
29	OB29	CL95	110	1	0	0	0	1

16.12.Principal Component Analysis

Principal component analysis is a variable reduction procedure. It is useful when we have to obtained data on a number of variables (possibly a large number of variables), and believe that there is some redundancy in those variables. Principal component analysis is appropriate when we have to obtained measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables. The principal components may then be used as predictor or criterion variables in subsequent analyses. In this case, redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same construct. Because of this redundancy, we believe that it should be possible to reduce the observed variables into a smaller number of principal components (artificial variables) that will account for most of the variance in the observed variables. Since it is a variable reduction procedure, principal component analysis is similar in many respects to exploratory factor analysis.

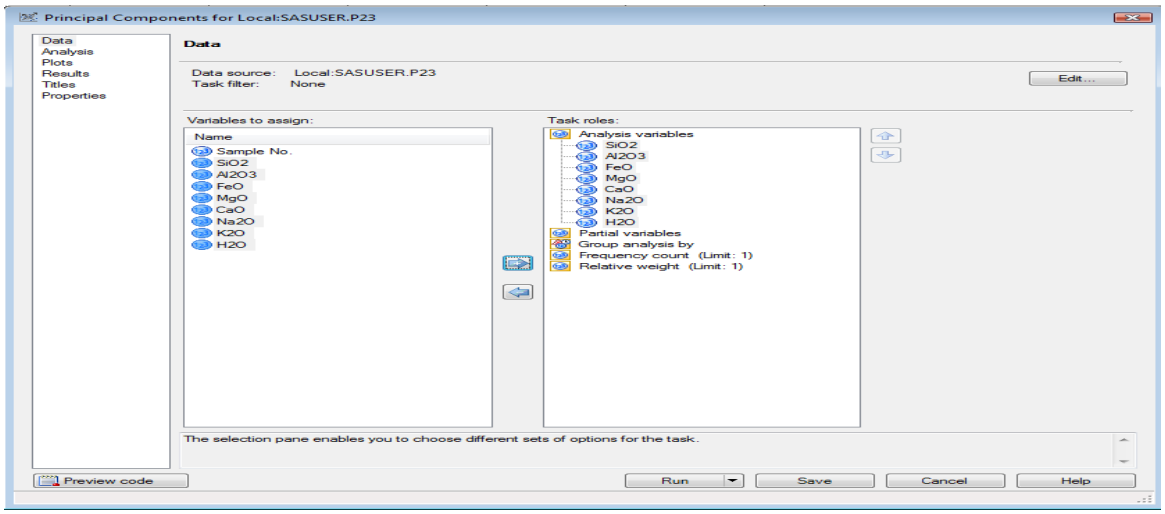
Exercise 13: Let us consider a data having fifteen samples of a rhyolite–basailt complex from the Gardiner River,Yellowstone National Park,(Miesch,1976) were collected and percentages of SiO_2, Al_2O_3,FeO, MgO,CaO,Na_2O, K_2O, and H_2O were recorded.

Sample No.	**SiO_2**	**Al_2O_3**	**FeO**	**MgO**	**CaO**	**Na_2O**	**K_2O**	**H_2O**
1	51.64	16.25	10.41	7.44	10.53	2.77	0.25	0.44
2	54.33	16.06	9.49	6.70	8.98	2.87	1.04	0.53
3	54.49	15.74	9.49	6.75	9.30	2.76	0.98	0.49
4	55.07	15.72	9.40	6.27	9.25	2.77	1.13	0.40
5	55.33	15.74	9.40	6.34	8.94	2.61	1.13	0.52
6	58.66	15.31	7.96	5.35	7.28	3.13	1.58	0.72
7	59.81	14.97	7.76	5.09	7.02	2.94	1.97	0.45
8	62.24	14.82	6.79	4.27	6.09	3.27	2.02	0.51
9	64.94	14.11	5.78	3.45	5.15	3.36	2.66	0.56
10	65.92	14.00	5.38	3.19	4.78	3.13	2.98	0.61
11	67.30	13.94	4.99	2.55	4.22	3.22	3.26	0.53
12	68.06	14.20	4.30	1.95	4.16	3.58	3.22	0.53
13	72.23	13.13	3.26	1.02	2.22	3.37	4.16	0.61
14	75.48	12.71	1.85	0.37	1.10	3.58	4.59	0.31
15	75.75	12.70	1.72	0.40	0.83	3.44	4.80	0.37

Perform Principal Component Analysis on the data.

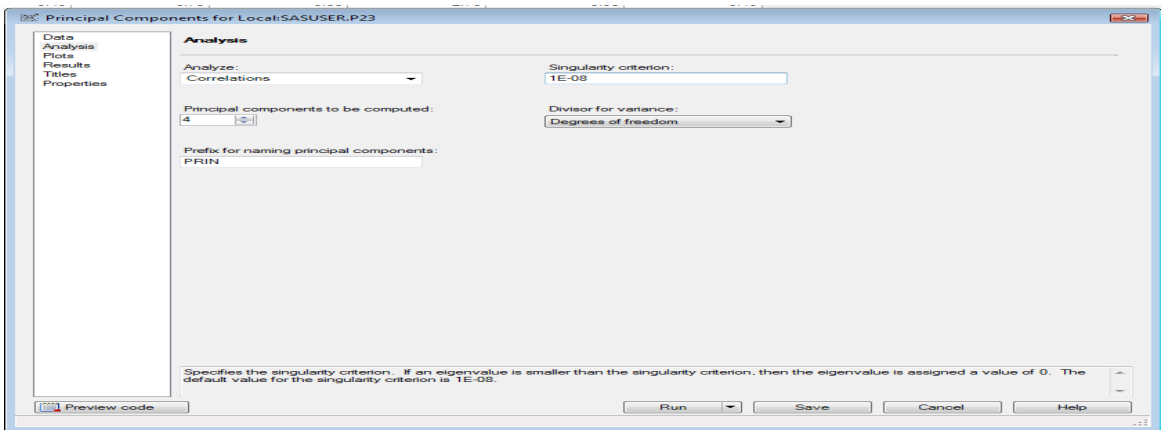
Solution:

Step1: From the main menu select **Tasks→Analyze→Multivariate→ Principal Component Analysis**. As shown in the **Task Roles** as shown below, drag all of the variables except for **sample No.** to the icon for **Analysis variables**.



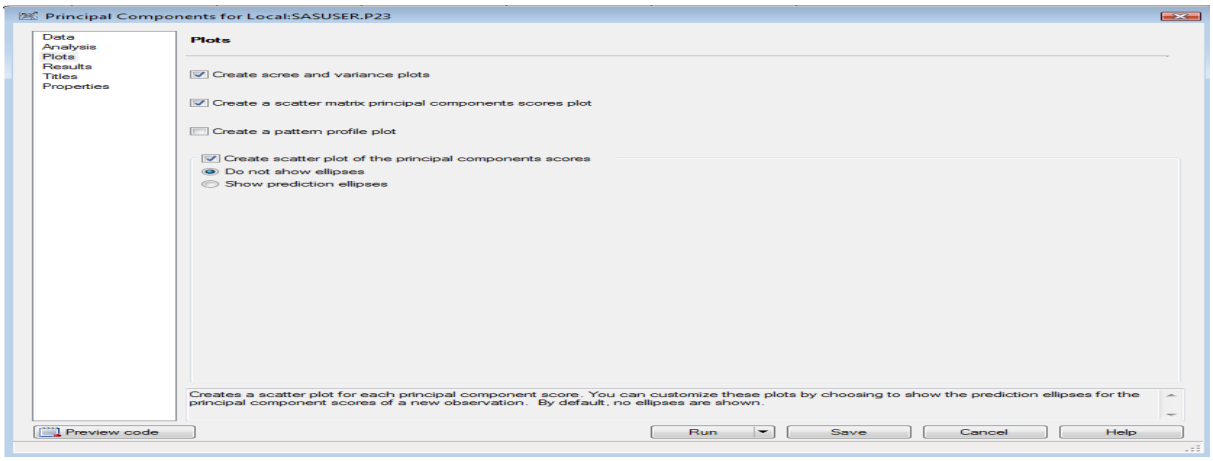
The **Task Roles** screen of the **Principal Component Analysis** procedure

Step 2: Select **Analysis** in the navigation panel. This brings us to the screen shown below. In the **Analysis** panel, there are several analyze methods available in the dropdown menu, including **Correlations, Covariances, Uncorrelated correlation and Uncorrelated covariance's**. For illustration purposes, we select **Correlations,**, the simplest of the factoring methods. In the panel for **Principal Component to be computed** from that drop-down menu, select the value of **4**. This instructs SAS to reduce the sample numbers from 15 to 4.



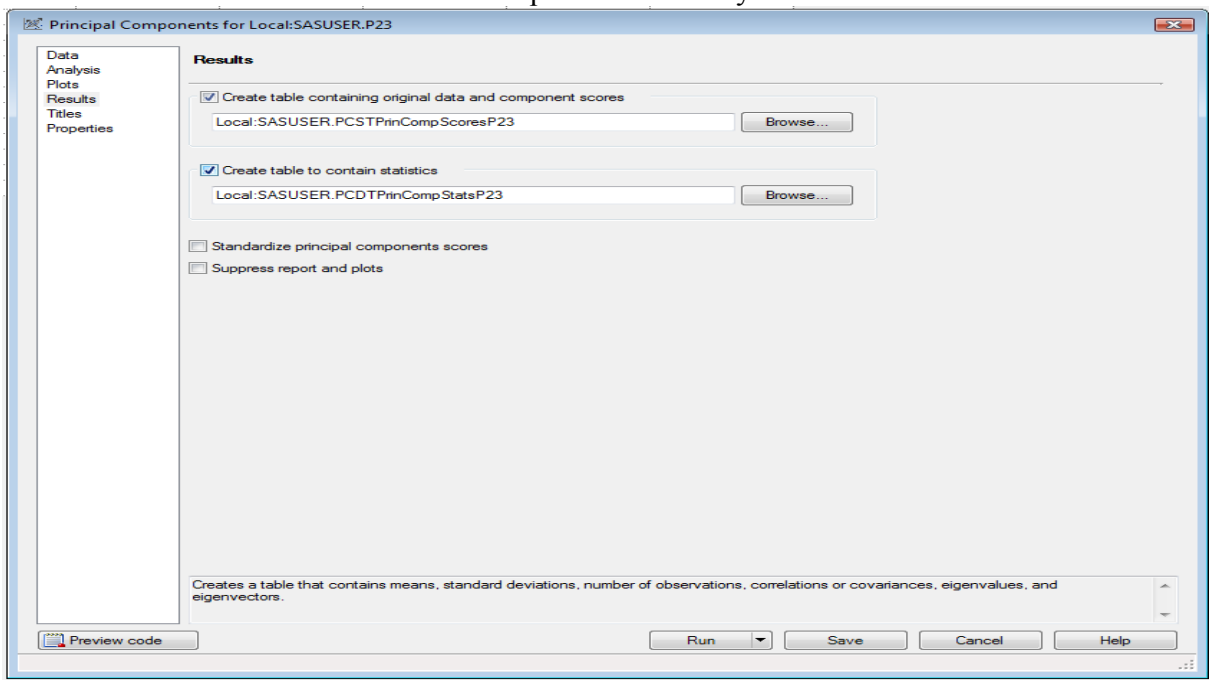
The **Analysis** screen of the **Principal Component Analysis** procedure

Step 3: Select **Plots** in the navigation panel. This brings us to the screen shown below. In the **Plots** panel, there are several plots available. Click the check boxes of **Create score and variance Plot, Create a scatter matrix principal component scores plot and Create scatter plot of the principal component scores.**



The **Plot** screen of the **Principal Component Analysis** procedure

Step 4 : Select **Result** in the navigation panel. This brings us to the screen shown below. Click the check boxes of **Create table containing original data and component scores, Create tables to contain statistics.** Click **Run** to perform the analysis.



The **Result** screen of the **Principal Component Analysis** procedure

The principal analysis output

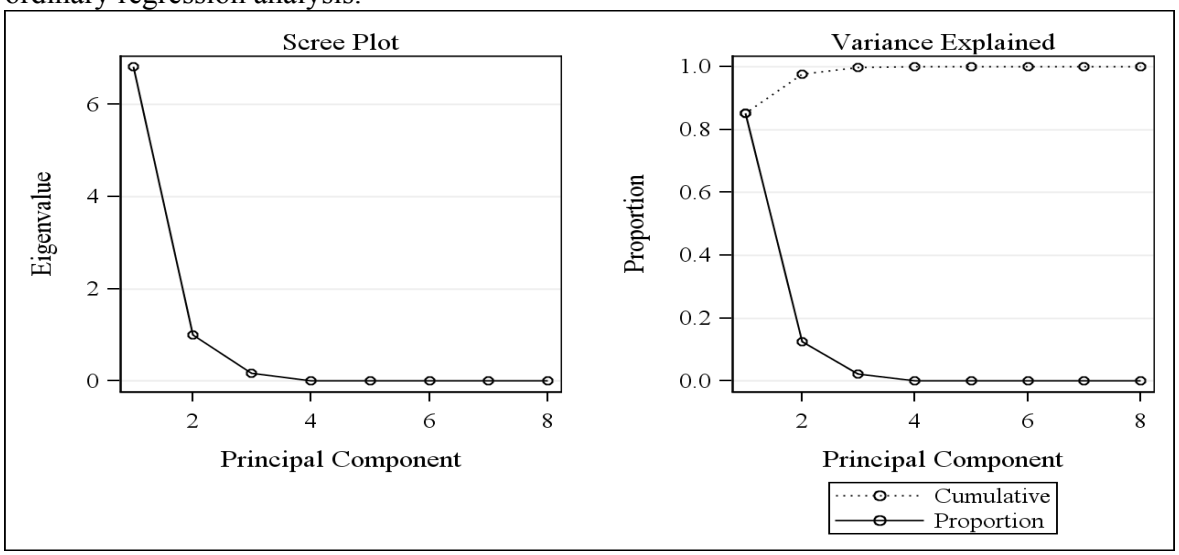
Observations	15
Variables	8

Simple Statistics								
	SiO_2	Al_2O_3	FeO	MgO	CaO	Na_2O	K_2O	H_2O
Mean	62.75000000	14.62666667	6.532000000	4.076000000	5.990000000	3.120000000	2.384666667	0.5053333333
StD	7.95924261	1.19298944	2.898739135	2.428879224	3.116488408	0.317129987	1.418806675	0.1028776435

The standard deviation of the variables are highly variable,and therefor it would be more appropriate to use correlation PCA than to use covariances, Our analysis used the default (correlation) option.

Correlation Matrix								
	SiO_2	Al_2O_3	FeO	MgO	CaO	Na_2O	K_2O	H_2O
SiO_2	1.0000	-.9930	-.9989	-.9970	-.9985	0.8968	0.9962	-.1685
Al_2O_3	-.9930	1.0000	0.9887	0.9852	0.9911	-.8631	-.9916	0.1741
FeO	-.9989	0.9887	1.0000	0.9967	0.9969	-.9076	-.9938	0.1731
MgO	-.9970	0.9852	0.9967	1.0000	0.9937	-.9082	-.9921	0.1454
CaO	-.9985	0.9911	0.9969	0.9937	1.0000	-.8967	-.9955	0.1435
Na_2O	0.8968	-.8631	-.9076	-.9082	-.8967	1.0000	0.8660	-.0320
K_2O	0.9962	-.9916	-.9938	-.9921	-.9955	0.8660	1.0000	-.1772
H_2O	-.1685	0.1741	0.1731	0.1454	0.1435	-.0320	-.1772	1.0000

From the above, one can see some variables are highly correlated with one another, with correlation coefficients as high as 0.99. It would not be a good idea to include all of them in an ordinary regression analysis.



From Scree Plot, one can see that there is only one large eigenvalue, and the remaining eigenvalues are small. It may be helpful to look at the proportion of variance accounted for by the component. the variance explained plot suggests only one large component on the Proportion curve, although this interpretation is subjective.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	6.81576651	5.82476385	0.8520	0.8520
2	0.99100265	0.81791630	0.1239	0.9758
3	0.17308635	0.16151712	0.0216	0.9975
4	0.01156923	0.00625374	0.0014	0.9989
5	0.00531549	0.00336659	0.0007	0.9996
6	0.00194890	0.00066463	0.0002	0.9998
7	0.00128427	0.00125768	0.0002	1.0000
8	0.00002659		0.0000	1.0000

The eigenvalues are shown below in the table labelled **Eigenvalues of the Correlation Matrix**. The columns in the table, from left to right, represent the following:

- The first column represents the **variable number**. The column is not labelled but each row represents a variable The numbers down the column thus start at 1 and end at the number of variables in the analysis, in this case 8.
- The second column represents the **eigenvalue**. This is the amount of variance accounted for by the variable.
- The third column represents the **difference**. This is the difference between successive eigenvalues. It gives us a sense of how much more variance is accounted for by the next variable.
- The fourth column represents **proportion**. This is the proportion of the total variance accounted for by the varaible. In principal components analysis, the total variance is equal to the number of variables in the analysis; here, there is a total of 7 units of variance.
- The fifth column represents cumulative proportion. This is the cumulative proportion of the variance accounted for by the first k variable.

One can see in the **Eigenvalues of the Correlation Matrix** table, the first two cumulatively accounted for 97.58% or approximately 97% of the variance. All else being equal, two variable solution accounting for this much variance would be considered reasonably good.

The total variation explained by the principal components is $\sum_{i=1}^{p} \lambda_i = 7.98$

The proportion of total variation accounted for by the first principal component is

$$\frac{\lambda_1}{\sum_{i=1}^{p} \lambda_i} = \frac{6.81}{7.98} = 0.86$$

Continuing, the first two component account for a proportion of the total variance.

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{p} \lambda_i} = \frac{7.80}{7.98} = 0.97$$

The first two principal components account for more than 97% of the total variation in these data and so must suffice for all practical purposes. The first principal component that explains about 86% of the variability seems to represent a contrast between the compounds.

Eigenvectors				
	PRIN1	**PRIN2**	**PRIN3**	**PRIN4**
SiO_2	-.382587	0.001793	-.110383	0.036284
Al_2O_3	0.378998	0.009475	0.270103	0.823554
FeO	0.382709	0.001089	0.038889	-.161182
MgO	0.381895	-.026583	0.046328	-.436321
CaO	0.381893	-.026649	0.126107	-.004294
Na_2O	-.350470	0.151212	0.899098	-.126504
K_2O	-.380275	-.011861	-.260233	0.296739
H_2O	0.066310	0.987666	-.138561	0.003283

The two eigenvectors corresponding to the first two eigenvalues listed under the columns PRIN1 and PRIN2 respectively are

a_1=(-0.382587, 0.378998,0.382709,0.381895,0.381893,-0.350470,-0.380275,0.066310)
a_2= (0.001793, 0.009475,0.001089,-0.026583,-0.026649,0.151212,-0.011861,0.987666)

The principal component for this data will be
$Z_1 = -0.382587x_1 + 0.378998x_2 + 0.382709x_3 + 0.381895x_4 + 0.381893x_5 - 0.350470x_6 - 0.380275x_7 + 0.066310x_8$
$Z_2 = 0.001793x_1 + 0.009475x_2 + 0.001089x_3 - 0.026583x_4 - 0.026649x_5 + 0.151212x_6 - 0.011861x_7 + 0.987666x_8$

Hence, in further analysis, the first or first two principal component a_1 and a_2 could replace eight variables by sacrificing negligible information about the total variation in the data.

The scores of principal components can be obtained by substituting the values of x_i's in the equations of Z_i's.

For the above data, the first two principal components for first observation i.e for sample one can be worked out as.
Z_1=-0.3825878*51.64+0.378998*16.25+0.382709*10.41+0.381895*7.44+0.381893*10.53-0.350470*2.77-0.380275*0.52+0.066310*0.44
Z_2=0.001793*54.33+0.009475*16.06+0.001089*9.49-0.026583*6.70+0.026649*8.98+0.151212*2.87-0.011861*1.04+0.987666*0.53

Thus the whole data with eight variables can be converted to a new data set with two principal components.

PRIN1	PRIN2
3.56383095	-0.8400575
2.68217263	0.08515931
2.73175962	-0.3565932
2.49478137	-1.2114953
2.71582327	-0.1336364
1.30359039	2.05244678
0.97256842	-0.6309611

0.09773204	0.11791149
-0.8738431	0.64709644
-0.8942863	1.01998015
-1.4264947	0.30400231
-1.9603425	0.48504632
-2.9902733	1.16390668
-4.2462187	-1.6061165
-4.1708001	-1.0966895

16.13. Factor Analysis

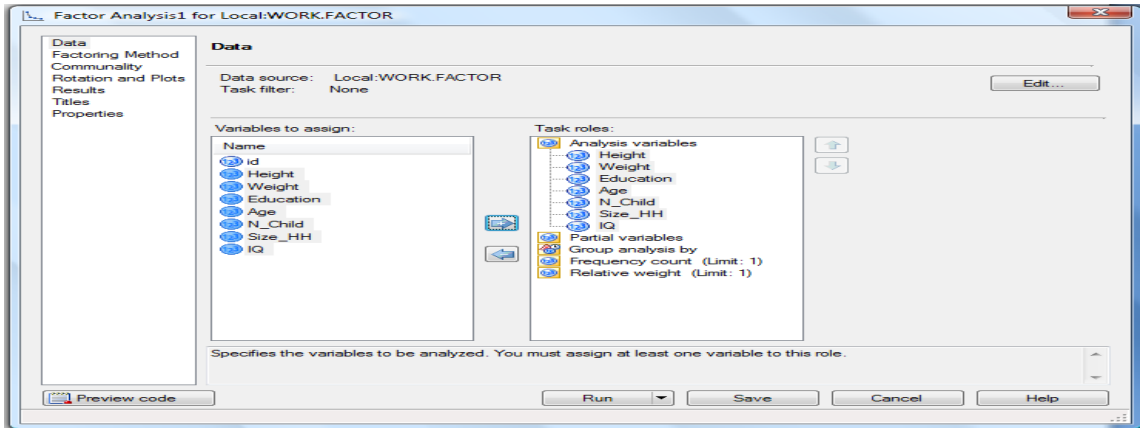
Exercise 14: A manufacture of fabricating parts is interested in identifying the determinants of a successful salesperson. The manufacturer has on file the information shown in the following table. He is wondering whether he could reduce these seven variables to two or three factors, for a meaningful appreciation of the problem.

sales person	Height(x_1)	Weight(x_2)	Education(x_3)	Age(x_4)	N_Child(x_5)	Size_HH(x_6)	IQ(x_7)
1	67	155	12	27	0	2	102
2	69	175	11	35	3	6	92
3	71	170	14	32	1	3	111
4	70	160	16	25	0	1	115
5	72	180	12	30	2	4	108
6	69	170	11	41	3	5	90
7	74	195	13	36	1	2	114
8	68	160	16	32	1	3	118
9	70	175	12	45	4	6	121
10	71	180	13	24	0	2	92
11	66	145	10	39	2	4	100
12	75	210	16	26	0	1	109
13	70	160	12	31	0	3	102
14	71	175	13	43	3	5	112

Can we now collapse the seven variables into three factors ? Intution might suggest the presence of three primary factors : maturity revealed in age/children/size of household, physical size as shown by height and weight, and intelligent or training as revealed by education and IQ.

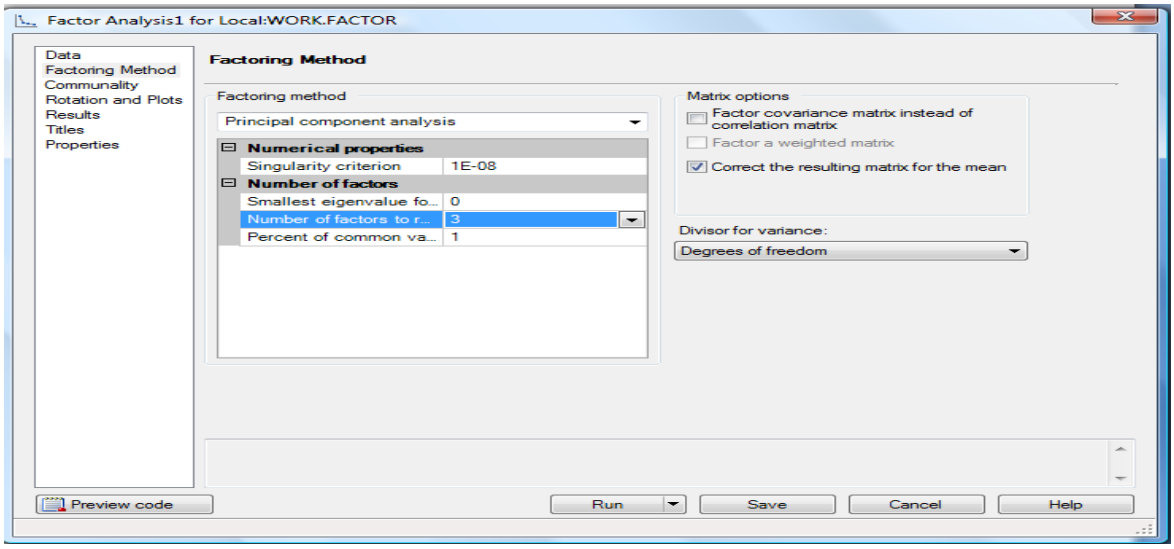
Setting up the factor analysis

From the main menu select **Tasks→Analyze→Multivariate→Factor Analysis**. As shown in the **Task Roles** screen below, drag all of the variables except for **id** to the icon for **Analysis variables**.



The **Task Roles** screen of the **Factor Analysis** procedure

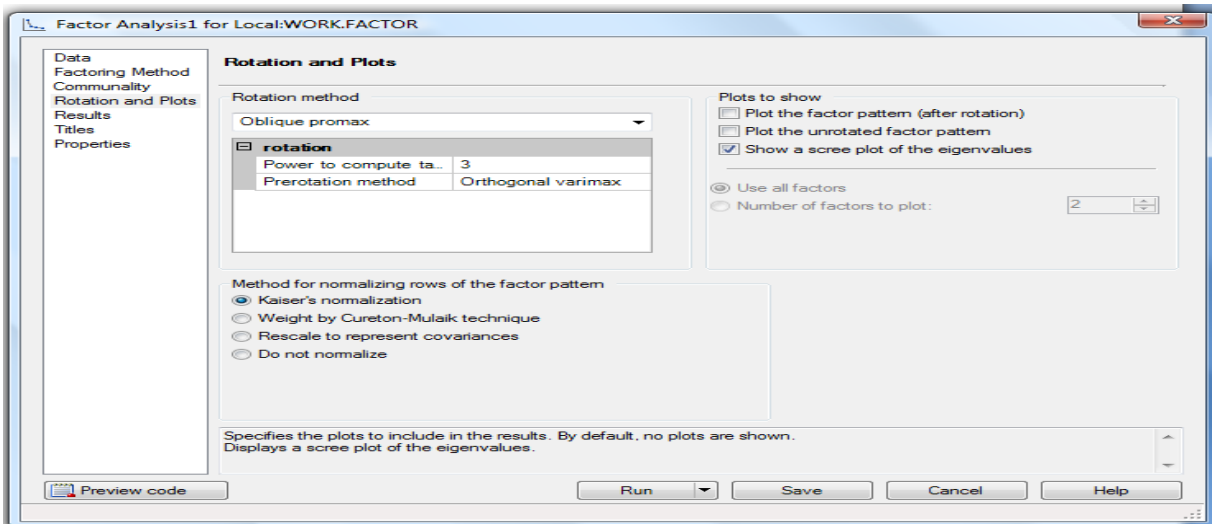
Select **Factoring Method** in the navigation panel. This brings us to the screen shown below. In the **Factoring method** panel, there are several methods available in the dropdown menu, including **Principal components analysis**, **Maximum likelihood factor analysis**, **Iterated principal factor analysis**, and **Unweighted least squares factor analysis**. For illustration purposes, one can select **Principal component analysis**, the simplest of the factoring methods. In the panel for **Number of factors**, click the choice for **Number of factors to r . . .** (the "**r . . .** " stands for the word rotate) to obtain the drop-down menu. From that drop-down menu, select the value of **3**. This instructs SAS to rotate the first four extracted components.



The **Factoring Method** screen of the **Factor Analysis** procedure

The **Rotation and Plots** screen, presented below, is where one can specify the rotation strategy one may wish to use and the plot(s) to be obtained. The scree plot is available in the upper right panel labeled **Plots to show**; we have checked the box corresponding to **Show a scree plot of the eigenvalues**. Rotation is addressed in the panel labeled **Rotation method**. The panel below the place where **Oblique promax** is displayed allows us to specify some details of the promax rotation. Very briefly, a promax rotation is performed in three stages:

- First, the correlation matrix is subjected to an orthogonal rotation. SAS Enterprise Guide gives us a choice of rotation strategies, and one can keep the default of **Orthogonal varimax**.
- Second, the varimax-generated coefficients are raised to an exponential power, typically between 2 and 4. SAS Enterprise Guide uses the power **3** as the default, and one may opt to keep it as well.
- Third, an oblique rotation is performed on the new values of the coefficients following our raising them to the specified exponential power.



The **Rotation and Plots** screen of the **Factor Analysis** procedure

Under the **Method for normalizing rows of the factor pattern**, one may keep the default of **Kaiser's normalization**. The sum of squares of the coefficients (weights) for a factor or component must sum to 1.00 during the rotation, and Kaiser's procedure accomplishes this. Click **Run** to perform the analysis.

The factor analysis output

Component extraction output

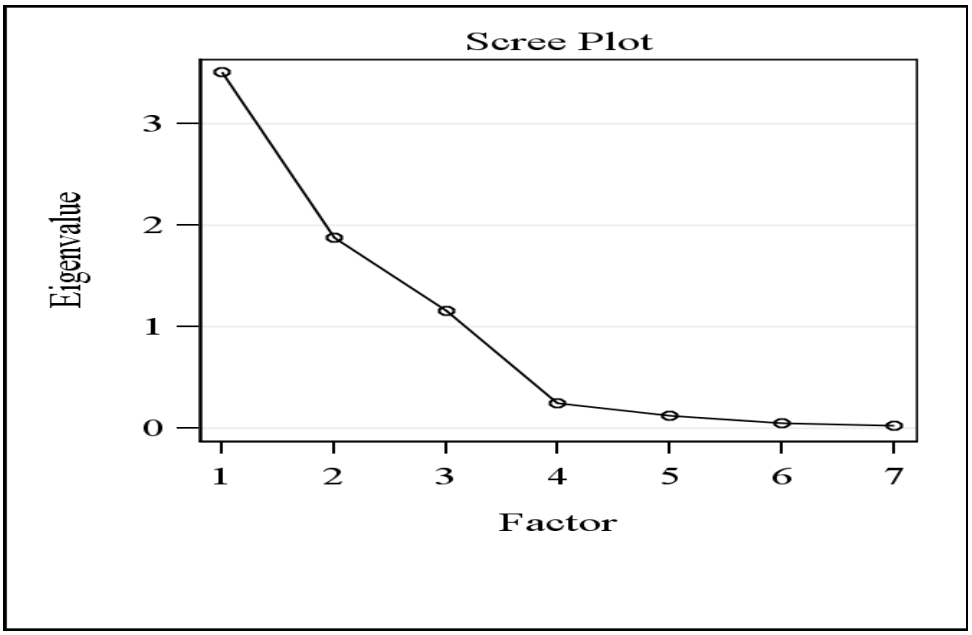
Eigenvalues of the Correlation Matrix: Total = 7 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	3.50869752	1.62728188	0.5012	0.5012
2	1.88141565	0.72367462	0.2688	0.7700
3	1.15774102	0.90981380	0.1654	0.9354
4	0.24792722	0.12283565	0.0354	0.9708
5	0.12509157	0.07267548	0.0179	0.9887
6	0.05241609	0.02570516	0.0075	0.9962
7	0.02671093		0.0038	1.0000

3 factors may be retained by the NFACTOR criterion.

The principal components extraction process results are shown above in the table labelled **Eigenvalues of the Correlation Matrix**. The columns in the table, from left to right, represent the following:

- The first column represents the **component number**. The column is not labelled but each row represents a component in the order it was extracted. The numbers down the column thus start at 1 and end at the number of variables in the analysis, in this case 7.
- The second column represents the **eigenvalue**. This is the amount of variance accounted for by the component. It is computed as the sum of the squared correlations between the variables and the component.
- The third column represents the **difference**. This is the difference between successive eigenvalues. It gives us a sense of how much more variance is accounted for by the next extracted component. For example, the difference between the 1st and 2nd eigenvalues is 3.50869752−1.88141565or 1.62728188..
- The fourth column represents **proportion**. This is the proportion of the total variance accounted for by the component. In principal components analysis, the total variance is equal to the number of variables in the analysis; here, there is a total of 7 units of variance. The first component accounts for approximately 3.50 of those units (that is its eigenvalue), which is approximately 50.12% of the variance (3.50 divided by 7). It is shown as a proportion of **0.5012** in the table.
- The fifth column represents cumulative proportion. This is the cumulative proportion of the variance accounted for by the first k components.

From the above table, one can see that the first three factors cumulatively accounted for 93.54% or approximately 93% of the variance. All else being equal, a three-factor solution accounting for this much variance would be considered reasonably good.



The scree plot

In the scree plot, the *X*- axis is the component number and corresponds to the first column of the **Eigenvalues of the Correlation Matrix** table. The *Y*-axis represents the eigenvalues and corresponds to the values in the second column of the table. For example, the data point identified as **1** is the eigenvalue of 3.50 for the first component, the data point identified as **2** is the eigenvalue of 1.88 for the second component, and so on. The scree plot exhibits the traditional backward- J-shaped function.

Factor Pattern			
	Factor1	**Factor2**	**Factor3**
Height	-0.58911	0.71888	-0.31374
Weight	-0.44726	0.75722	-0.45119
Education	-0.80909	0.18488	0.41644
Age	0.81381	0.45657	0.19480
N_Child	0.84701	0.48734	0.05440
Size_HH	0.91869	0.28862	-0.02536
IQ	-0.28858	0.47730	0.80048

The component matrix at the completion of the extraction phase

The above table presents the factor or component matrix, named **Factor Pattern**, for the four-component solution. This is the last portion of the extraction process and anticipates the number of factors one may rotate; it is the structure that will be rotated in the next phase of the analysis. The numerical entries in the matrix are the coefficients for the variables on the components.

There are two types of coefficients that are represented in this matrix: pattern coefficients and structure coefficients

- For **pattern coefficients**, each component is a weighted linear combination (a variate) composed of the 7 variables. The pattern coefficients are the standardized regression weights in this variate. The different configurations of weights differentiate the components from each other. For example, in the first component (**Factor 1** in the table), the first item is weighted as -0.58911, the second item is weighted as -0.44726, and so on. In the second component (**Factor 2** in the table), the first item is weighted as 0.71888, the second item is weighted as 0.75722, and so on.
- For **structure coefficients**, each variable is correlated to a certain extent with each component. The structure coefficients are these correlations. The different configurations of correlations differentiate the components from each other. For example, in the first component (**Factor 1** in the table), the first item is correlated -0.58911with the component, the second item is correlated -0.44726, with the component, and so on. In the second component (**Factor 2** in the table), the first item is correlated 0.71888 with the component, the second item is correlated 0.75722 with the component, and so on.

Sum of squares & Variance Explained by Each Factor		
Factor1	**Factor2**	**Factor3**
3.5086975	1.8814156	1.1577410
0.585478	0.313569	0.192956

Variance explained by each factor

Factor I accounts for 58.5% of the total variance. Factor II for 31.3% and Factor III for 19.2%. Together the three "explain" almost almost 95 % of the variance.

Final Communality Estimates: Total = 6.547854						
Height	**Weight**	**Education**	**Age**	**N_Child**	**Size_HH**	**IQ**
0.96227270	0.97698731	0.86222837	0.90868425	0.95788639	0.92793926	0.95185592

The communality is over 85% for every variable. Thus, the three factors seem to capture the underlying dimension involved in these variables.

Factor Structure (Correlations)			
	Factor1	**Factor2**	**Factor3**
Height	-0.27074	0.97748	0.36715
Weight	-0.13396	0.98120	0.21232
Education	-0.64121	0.42014	0.78582
Age	0.93856	-0.17876	-0.04049
N_Child	0.97276	-0.12021	-0.14870
Size_HH	0.95416	-0.27689	-0.33203
IQ	-0.02511	0.20678	0.95594

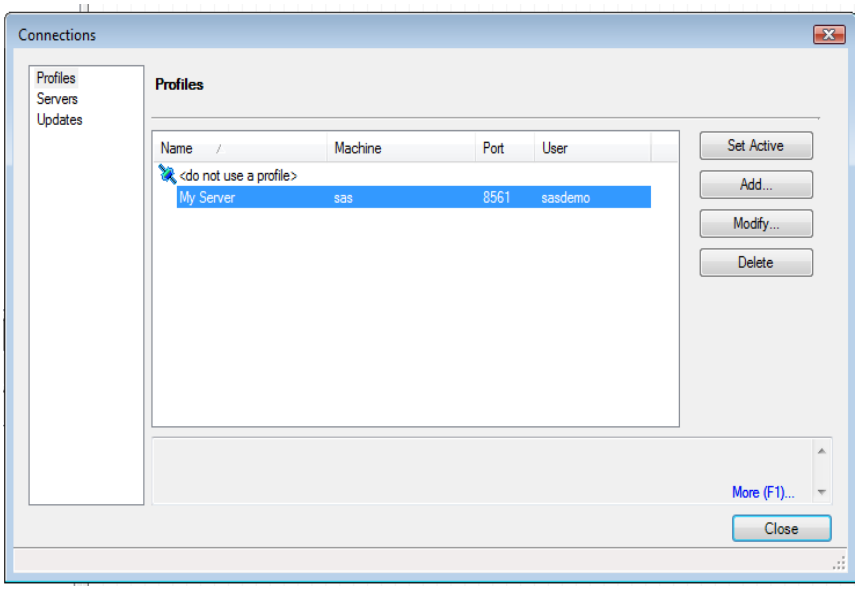
The promax rotated component structure matrix

Examining the rotated factor matrix shown above, one finds that **Height & weight** correlates ("loads") to an acceptable degree on the second component (**Factor 2** in the matrix with a correlation of 0.97748 & 0.98120). One may also note that **Education & IQ** correlates to an acceptable degree on the fourth component (**Factor 3** in the matrix with a correlation of 0.78582 & 0.95594), **Age,N_Child,Size_HH** correlates to an acceptable degree on the first component (**Factor 1** in the matrix with a correlation of 0.93856,0.97276 &0.95416).

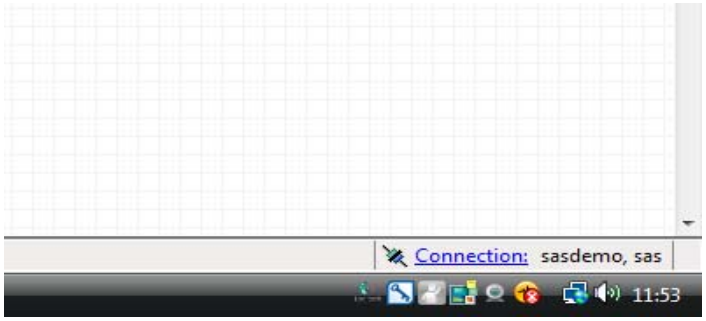
17. How to Connect SAS EBI Server through Enterprise Guide?

Click on the tab "No connection" at the right bottom of the process flow

Select the Sever ie My Sever or IASRI Sever or any other name defined by administrator and click Set Active



Close the connection window we find "No connection tab" change into "Connection"



Now we import the data file from server as well as from Local Computer. It depends on us from where we want to import the file.

Importing the data file from server:
File →Import Data →select Server →Local →Files →D: (Drive) →NAIP_training →NAIP_training →sscnars.xls →Next → Select Sheet "Training" →Finish. Then perform the desired task on data file.

Importing the data file from Local Computer:
File →Import Data →Select Local computer → Select the location say "Desktop" and double click → Select the folder Say "Training Data" → select the file say "sscnars.xls" →Next → Select Sheet "Training" →Finish. Then perform the desired task on data file.

Remarks:

1. Discrete frequency table of a single variable can be generated **Tasks** →Describe → One-way Frequencies… From the variables to assign, select the analysis variable and Click on Run.
2. 2-way frequency tables can be generated **Tasks** →Describe → Table Analysis…. From the variables to assign, select the variables on which frequency table need to be prepared and

assign them to Table Variables. Then Select Tables Pane and Drag and Drop the variable in rows and columns of the Table Preview. If it is desired to test the independence of attributes of the above contingency table, then Select the Pane Table Statistics and then select appropriate test options.

For details on SAS Enterprise Guide may be seen from SAS Enterprise Help:
Help → Contents → Using SAS Enterprise Guide →….

To change the format and style of output, use Tools→ Option → Results → Viewer ….

References:

Parsad, R., Gupta, V.K., umrao, AK and Kole, B. Analysis of Data (http://iasri.res.in/design/Analysis%20of%20data/Contact%20us.html). *Design Resources Server*. Indian Agricultural Statistics Research Institute (ICAR), New Delhi 110 012, India. www.iasri.res.in/design (Accessed lastly on 14.08.2010).

Slaughter, S.J. and Delwiche, LD (2010*). The Little SAS Book for Enterprise Guide 4.2*. SAS Institute Inc., Cary, NC, USA

www.sas.com

www.support.sas.com

http://www.sas.com/technologies/bi/query_reporting/guide/