Chapter 450

Clustered Heat Maps (Double Dendrograms)

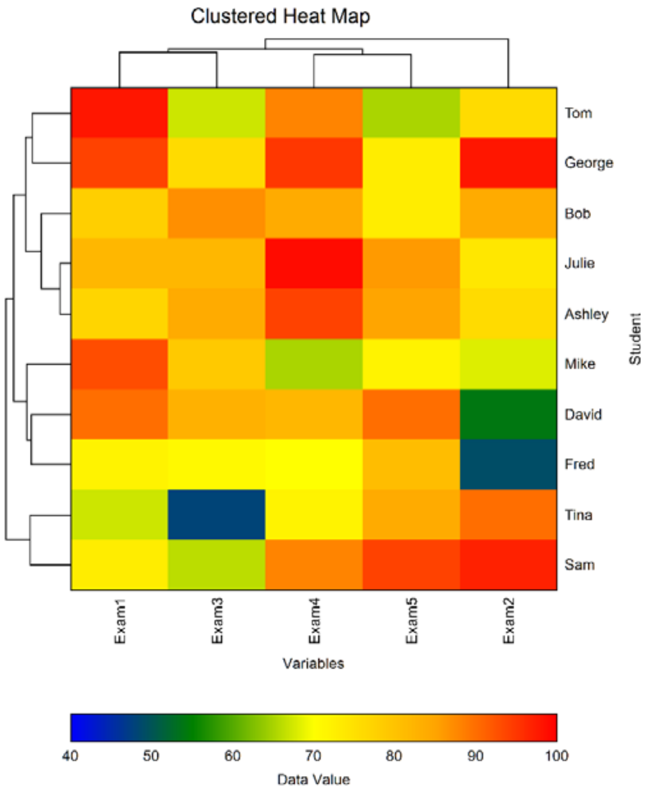
Introduction

This chapter describes how to obtain a clustered heat map (sometimes called a double dendrogram) using the Clustered Heat Map procedure.

Similar to a contour plot, a heat map is a two-way display of a data matrix in which the individual cells are displayed as colored rectangles. The color of a cell is proportional to its position along a color gradient. Usually, the columns (variables) of the matrix are shown as the columns of the heat map and the rows of the matrix are shown as the rows of the heat map, as in the example below. The order of the rows is determined by performing hierarchical cluster analyses of the rows. This tends to position similar rows together on the plot. The order of the columns is determined similarly.

Usually, a clustered heat map is made on variables that have similar scales, such as scores on tests in the example below. If the variables have different scales, the data matrix must first be scaled using a standardization transformation such as z-scores or proportion of the range.

Here is an example of a clustered heat map.



© NCSS, LLC. All Rights Reserved.

Hierarchical Clustering Algorithms

The cluster algorithms used for the rows and columns of the data matrix must be specified. NCSS allows you to select from eight possible hierarchical algorithms. The clustering method selected for the columns need not be the same as the method selected for the rows. Chapter 445 of the ***NCSS*** documentation gives an introduction to hierarchical clustering. Only highlights from that chapter are presented here.

We will first give brief comments about each of the eight hierarchical clustering techniques.

Group Average

Also called the unweighted pair-group method, this is perhaps the most widely used of all the hierarchical cluster techniques. The distance between two groups is defined as the average distance between each of their members.

Single Linkage

Also known as *nearest neighbor* clustering, this is one of the oldest and most famous of the hierarchical techniques. The distance between two groups is defined as the distance between their two closest members. It often yields clusters in which individuals are added sequentially to a single group.

Complete Linkage

Also known as furthest neighbor or maximum method, this method defines the distance between two groups as the distance between their two farthest-apart members. This method usually yields clusters that are well separated and compact.

Simple Average

Also called the weighted pair-group method, this algorithm defines the distance between groups as the average distance between each of the members, weighted so that the two groups have an equal influence on the final result.

Centroid

Also referred to as the unweighted pair-group centroid method, this method defines the distance between two groups as the distance between their centroids (center of gravity or vector average). The method should only be used with Euclidean distances.

Backward links may occur with this method. These are recognizable when the dendrogram no longer exhibits its simple tree-like structure in which each fusion results in a new cluster that is at a higher distance level (moves from right to left). With backward links, fusions can take place that result in clusters at a lower distance level (move from left to right). The dendrogram is difficult to interpret in this case.

Median

Also called the weighted pair-group centroid method, this defines the distance between two groups as the weighted distance between their centroids, the weight being proportional to the number of individuals in each group. Backward links (see discussion under Centroid) may occur with this method. The method should only be used with Euclidean distances.

Ward's Minimum Variance

With this method, groups are formed so that the pooled within-group sum of squares is minimized. That is, at each step, the two clusters are fused which result in the least increase in the pooled within-group sum of squares.

© NCSS, LLC. All Rights Reserved.

Flexible Strategy

Lance and Williams (1967) suggested that a continuum could be made between single and complete linkage. The program lets you try various settings of these parameters which do not conform to the constraints suggested by Lance and Williams.

One interesting exercise is to vary these values, trying to find the set that maximizes the cophenetic correlation coefficient.

Goodness-of-Fit

Given the large number of techniques, it is often difficult to decide which is best. One criterion that has become popular is to use the result that has largest *cophenetic correlation coefficient*. This is the correlation between the original distances and those that result from the cluster configuration. Values above 0.75 are felt to be good. The Group Average method appears to produce high values of this statistic. This may be one reason that it is so popular.

A second measure of goodness of fit called *delta* is described in Mather (1976). These statistics measure degree of distortion rather than degree of resemblance (as with the cophenetic correlation). The two delta coefficients are given by

$$\Delta_A = \left[\frac{\sum_{j<k}^{N} |d_{jk} - d^*_{jk}|^{1/A}}{\sum_{j<k} (d^*_{jk})^{1/A}} \right]^A$$

where A is either 0.5 or 1 and d^*_{ij} is the distance obtained from the cluster configuration. Values close to zero are desirable.

Mather (1976) suggests that the Group Average method is the safest to use as an exploratory method, although he goes on to suggest that several methods should be tried and the one with the largest cophenetic correlation be selected for further investigation.

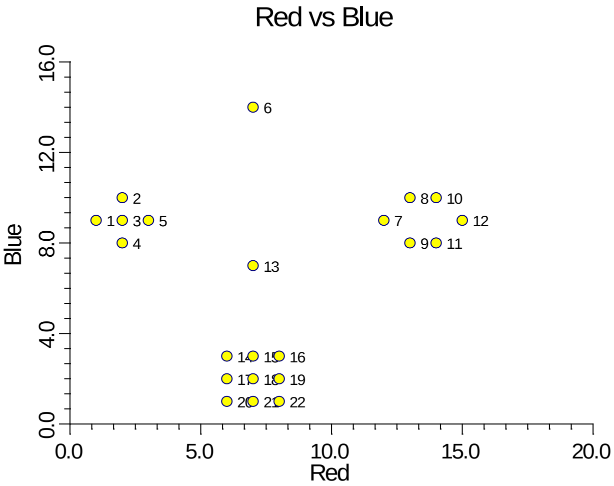
Dendrograms

The *agglomerative hierarchical clustering* algorithms available in this program module build a cluster hierarchy that is commonly displayed as a tree diagram called a *dendrogram*. The algorithm begins by placing each object in a separate cluster. Then, at each step, the two clusters that are most similar (according to a specific definition of similarity) are joined into a single new cluster. Once fused, objects are never separated. The eight clustering methods that are available represent eight methods of defining the similarity between clusters.

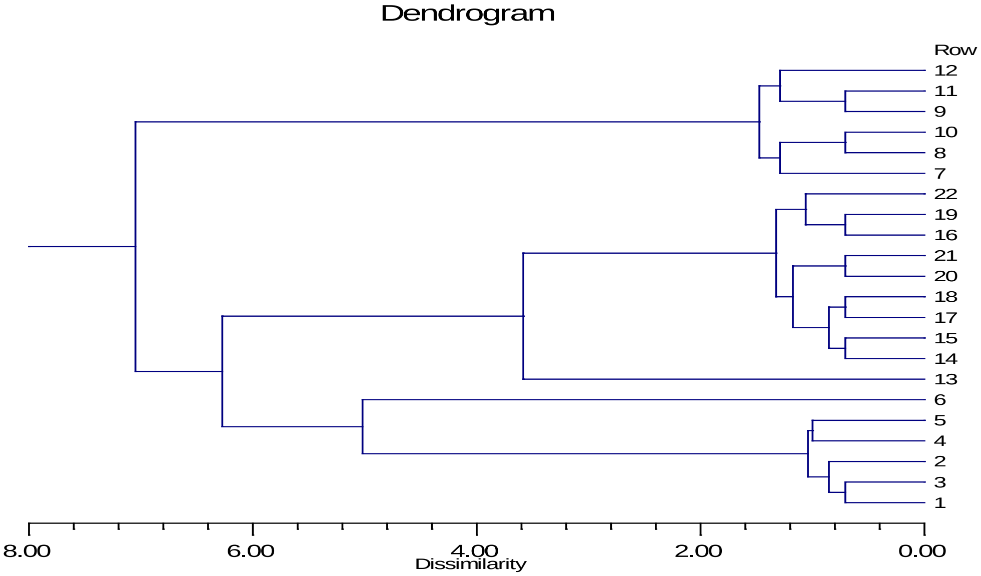
To help understand the dendrogram, consider the following example that has only two variables. Note that if we had only two variables, we could perform the cluster analysis visually. The technique becomes useful once we have three or more variables to consider.

© NCSS, LLC. All Rights Reserved.

Suppose we wish to cluster the bivariate data shown in the following scatter plot. In this case, the clustering may be done visually. The data seem to exhibit three clusters and two singletons, 6 and 13.



The following dendrogram was produced from the above data using popular the Group Average clustering algorithm.



The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters. The vertical axis represents the objects and clusters. The dendrogram is fairly simple to interpret. Remember that our main interest is in similarity and clustering. Each joining (fusion) of two clusters is represented on the graph by the splitting of a horizontal line into two horizontal lines. The horizontal position of the split, shown by the short vertical bar, gives the dissimilarity between the two clusters.

Looking at this dendrogram, you can see the three clusters as three branches that occur at about the same horizontal distance. The two outliers, 6 and 13, are added in rather arbitrarily at much higher distances. This is the interpretation.

In this example we can compare our interpretation with an actual plot of the data. Unfortunately, this usually will not be possible because our data will consist of more than two variables.

© NCSS, LLC. All Rights Reserved.

Missing Values

When an observation has missing values, appropriate adjustments are made so that the average dissimilarity across all variables with non-missing data is computed. Hence, rows with missing values are not omitted unless all variables have missing values. Note that the distances require that at least one variable have non-missing values for each pair of rows.

Procedure Options

This section of the manual describes the function of each of the options on the panel windows.

Variables Tab

This panel specifies the variables and the clustering options.

Variables

These options specify the variables used in the analysis.

Cluster Variables

Specify the numeric variables (columns) to be clustered. These variables must contain numeric values from a linear scale. Examples include test score, count, height, weight, age, and temperature.

The heat map requires all variables to be on a similar scale. If they are not, the data must be standardized before the heat map is drawn. You can perform the standardization manually using NCSS's transformation capabilities, or you can use one of the built-in scaling methods.

Row Label Variable

This optional variable contains labels that can be used for each row to aid in the interpretation. If this value is left blank, the row numbers will be used to identify the rows on the reports and dendrogram.

Variable Scaling Method

Select the scaling method to be applied down the data columns. The possible choices are

- **None**
  The original data values are used without any scaling.

- **Z-Scores**
  The values in each column are standardized by subtracting their mean and dividing by their standard deviation. If all values are identical, zeros are used for the z-scores.

- **Proportions**
  The values in each column are standardized by subtracting their minimum and dividing by their range. If all values are identical, zeros are used for the proportions.

© NCSS, LLC. All Rights Reserved.

Row Scaling Method

Usually, if scaling is needed, it is applied to variables. This option is provided for when you want to scale across the rows. Note that this scaling is applied after the *Variable Scaling*, if any, is applied.

The possible choices are

- **None**
  No scaling is made across the values in a row.

- **Z-Scores**
  The values in each row are standardized by subtracting their mean and dividing by their standard deviation. If all values are identical, zeros are returned.

- **Proportions**
  The values in each row are standardized by subtracting their minimum and dividing by their range. If all values are identical, zeros are returned.

Cluster Options for Variables Dendrogram

Clustering Method

This option specifies which of the eight possible hierarchical techniques is used. These methods were described earlier.

Alpha

Only displayed when the *Flexible Strategy* method is selected. Specifies the values of α_i and α_j. You may enter a number or the letters “NI/NK.” The “NI/NK” will cause this constant to be calculated and used as it is in the Centroid and Group Average methods.

Beta

Only displayed when the *Flexible Strategy* method is selected. Specifies the values of β. You may enter a number between -1 and 1 or the letters “NIJ/NK.” The “NIJ/NK” will cause this constant to be calculated and used as it is in the Centroid method.

Gamma

Only displayed when the *Flexible Strategy* method is selected. Specifies the values of γ. You may enter any number.

Distance Method

This option specifies whether Euclidean or Manhattan distance is used by the clustering algorithm. Euclidean distance may be thought of as straight-line (or as the crow flies) distance. Manhattan distance is often referred to as city-block distance since it is analogous to walking along an imaginary sidewalk to get from point A to B. Most users will use Euclidean distance.

Cluster Options for Rows Dendrogram

Clustering Method

This option specifies which of the eight possible hierarchical techniques is used. These methods were described earlier.

Alpha

Only displayed when the *Flexible Strategy* method is selected. Specifies the values of α_i and α_j. You may enter a number or the letters “NI/NK.” The “NI/NK” will cause this constant to be calculated and used as it is in the Centroid and Group Average methods.

Beta

Only displayed when the *Flexible Strategy* method is selected. Specifies the values of β. You may enter a number between -1 and 1 or the letters “NIJ/NK.” The “NIJ/NK” will cause this constant to be calculated and used as it is in the Centroid method.

Gamma

Only displayed when the *Flexible Strategy* method is selected. Specifies the values of γ. You may enter any number.

Distance Method

This option specifies whether Euclidean or Manhattan distance is used by the clustering algorithm. Euclidean distance may be thought of as straight-line (or as the crow flies) distance. Manhattan distance is often referred to as city-block distance since it is analogous to walking along an imaginary sidewalk to get from point A to B. Most users will use Euclidean distance.

Clustered Heat Map

Format

Click the format button to change the heat map settings.

Edit During Run

Checking this option will cause the clustered heat map format window to appear when the procedure is run. This allows you to modify the format of the graph with the actual data.

Max Variables Clusters

This option specifies the number of variable clusters used in the reports and the heat map.

Max Rows Clusters

This option specifies the number of row clusters used in the reports and the heat map.

Reports Tab

The options on this panel control which reports and plots are generated.

Reports

Clustered Heat Map – Distance Reports

Specify whether to display the indicated plot and reports.

© NCSS, LLC. All Rights Reserved.

Report Options

These options limit the cluster reports.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Precision

Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while double precision will display thirteen-place accuracy.

Max Distance Items

This option specifies the maximum size of a distance matrix that will be displayed in the Distance Section report. Distance matrices with more items than this will not be displayed.

This option is here because for large datasets, the distance matrix may be very large.

Max Linkage Clusters

The Linkage Report can be long if the results for all links are printed. This parameter allows you to limit the number of links displayed so that only meaningful values are printed.

Storage Tab

The cluster id number for each row can be stored on the spreadsheet for further analysis. This option designates the column of the spreadsheet in which the cluster id's are stored.

Store the Cluster Id Number of Each Row in this Variable

The cluster id number for each row can be stored on the spreadsheet for further analysis. This option designates the column of the spreadsheet in which the cluster id's are stored.

WARNING: Existing data in this column will be replaced with the new values automatically when the procedure is run.

Store Cluster Id in Variable

The cluster id number for each row is stored in this column. To omit the automatic storage of the cluster id's, leave this option blank.

© NCSS, LLC. All Rights Reserved.

Example 1 – Creating a Clustered Heat Map

This section presents an example of how to create a clustered heat map from a set of exam score data. The data are found in the Exams database.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Clustered Heat Maps (Double Dendrograms) window.

1 Open the Exams dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Exams.NCSS**.
- Click **Open**.

2 Open the Clustered Heat Map (Double Dendrograms) window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Clustered Heat Map (Double Dendrograms)** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Clustered Heat Map window, select the **Variables tab**.
- Double-click in the **Cluster Variable(s)** box. This will bring up the variable selection window.
- Select **Exam1** through **Exam5** from the list of variables and then click **Ok**. “Exam1, Exam2, Exam3, Exam4, Exam5” will appear in the Cluster Variables box.
- Double-click in the **Row Label Variable** box. This will bring up the variable selection window. Select **Student** from the list of variables and then click **Ok**. “Student” will appear in the Row Label Variable box.
- Set **Variable Scaling Method** to **None**.
- Set **Row Scaling Method** to **None**.

4 Specify the reports.

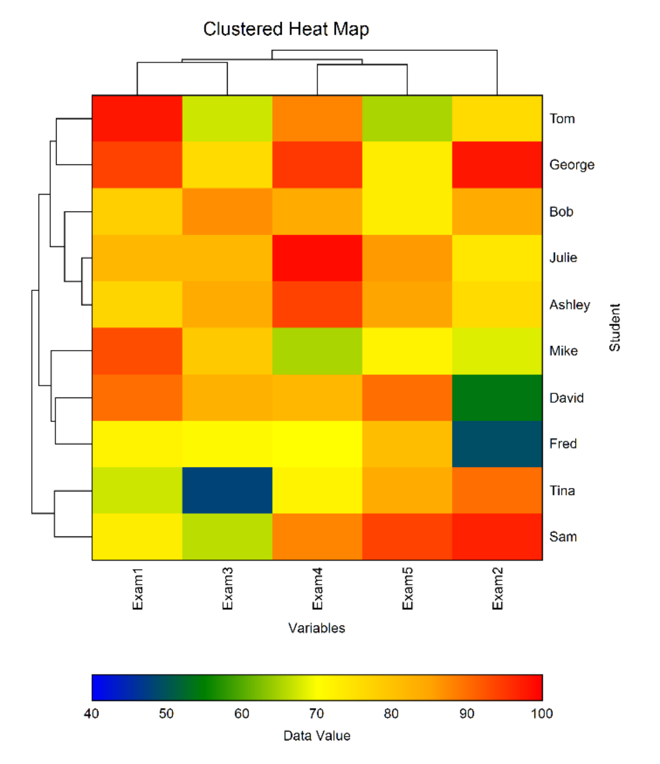
- On the Clustered Heat Map window, select the **Reports tab**.
- Check the heat map and all reports.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).

© NCSS, LLC. All Rights Reserved.

Heat Map Section



This report displays the heat map for these data with the above settings. Note that the rows and columns have been sorted in the order specified by the clustering.

Cluster Detail Section

```
Cluster Detail Report when Clustering Variables
Clustering Method      Group Average
Distance Type          Euclidean
Scale Type             None

Cluster    Variables in this Cluster
1          Exam1, Exam3
2          Exam4, Exam5
None       Exam2

Cluster Detail Report when Clustering Rows
Clustering Method      Group Average
Distance Type          Euclidean
Scale Type             None

Cluster    Rows in Cluster
1          Tom, Bob, Julie, Ashley, George
2          Tina, Sam
3          Mike, David, Fred
```

This report displays the items contained in the row and variable clusters. Those items that cannot be classified are listed in the ‘None’ cluster.

© NCSS, LLC. All Rights Reserved.

Linkage Section

```
Linkage Report when Clustering Variables
Clustering Method        Group Average
Distance Type                Euclidean
Scale Type                        None

          Number             Distance    Distance
Link      Clusters              Value    Bar
4         1                 19.413274    |IIIIIIIIIIIIIIIIIIIIIIIIIIIIII
3         2                 15.493208    |IIIIIIIIIIIIIIIIIIIIIIII
2         3                 14.390969    |IIIIIIIIIIIIIIIIIIIIII
1         4                 13.408952    |IIIIIIIIIIIIIIIIIIIII

Cophenetic Correlation       0.829879
Delta(0.5)                   0.067471
Delta(1.0)                   0.090232

Linkage Report when Clustering Rows
Clustering Method        Group Average
Distance Type                Euclidean
Scale Type                        None

          Number             Distance    Distance
Link      Clusters              Value    Bar
9         1                 19.763410    |IIIIIIIIIIIIIIIIIIIIIIIIIIIIII
8         2                 17.446811    |IIIIIIIIIIIIIIIIIIIIIIIIII
7         3                 13.832703    |IIIIIIIIIIIIIIIIIIIII
6         4                 13.428924    |IIIIIIIIIIIIIIIIIIII
5         5                 12.369317    |IIIIIIIIIIIIIIIIIII
4         6                 11.983322    |IIIIIIIIIIIIIIIIII
3         7                 11.781341    |IIIIIIIIIIIIIIIIII
2         8                  9.159521    |IIIIIIIIIIIIII
1         9                  3.435113    |IIIII

Cophenetic Correlation       0.765894
Delta(0.5)                   0.125960
Delta(1.0)                   0.178253
```

This report displays the number of clusters that exist at each link. The links are displayed in reverse order so that you can quickly determine an appropriate number of clusters to use. It displays the distance level at which the fusion took place. It will let you precisely determine the best value of the number of clusters.

The cophenetic correlation and two delta goodness of fit statistics are reported at the bottom of this report. As discussed earlier, these values let you compare the fit of various cluster configurations.

Link

This is the sequence number of the fusion.

Number Clusters

This is the number of clusters that would result if the cluster cutoff value were set to the corresponding Distance Value or higher. Note that this number includes outliers.

Distance Value

This is distance value between the two joining clusters that is used by the algorithm. Normally, this value is monotonically increasing. When backward linking occurs, this value will no longer exhibit a strictly increasing behavior.

Distance Bar

This is a bar graph of the Distance Values. Choose the number of clusters by finding a jump in the decreasing pattern shown in this bar chart.

© NCSS, LLC. All Rights Reserved.

Cophenetic Correlation

This is the Pearson correlation between the actual distances and the predicted distances based on this particular hierarchical configuration. A value of 0.75 or above needs to be achieved in order for the clustering to be considered useful.

Delta (0.5, 1)

These are the values of the goodness of fit deltas. When comparing to clustering configurations, the configuration with the smallest delta value fits the data better.

Distance Section

```
Distance Report when Clustering Variables
Clustering Method    Group Average
Distance Type        Euclidean
Scale Type           None


First  Second     Actual       Dendrogram       Actual      Percent
Row    Row      Distance         Distance   Difference   Difference
1      2       20.386270        19.413274     0.972996         4.77
1      3       14.390969        14.390969     0.000000         0.00
1      4       13.479614        15.493208    -2.013593       -14.94
1      5       16.991174        15.493208     1.497966         8.82

2      3       22.074873        19.413274     2.661599        12.06
2      4       16.555966        19.413274    -2.857308       -17.26
2      5       18.635987        19.413274    -0.777287        -4.17

3      4       15.678010        15.493208     0.184802         1.18
3      5       15.824032        15.493208     0.330825         2.09

4      5       13.408952        13.408952     0.000000         0.00
.      .               .                .            .            .
.      .               .                .            .            .
.      .               .                .            .            .
```

This report displays the actual and predicted distance for each pair of variables (and, later, rows). It also includes their difference and percent difference. Since the report grows very long for even a modest number of rows, it is often omitted.

Example 2 – Creating a Clustered Heat Map with only Two Colors

This section presents an example of how to create a clustered heat map of the exam score data with a gradient of just two colors. The data are found in the Exams database.

You may follow along here by making the appropriate entries or load the completed template **Example 2** by clicking on Open Example Template from the File menu of the Clustered Heat Maps (Double Dendrograms) window.

1 Open the Exams dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Exams.NCSS**.
- Click **Open**.

2 Open the Clustered Heat Map (Double Dendrograms) window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Clustered Heat Map (Double Dendrograms)** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Clustered Heat Map window, select the **Variables tab**.
- Double-click in the **Cluster Variable(s)** box. This will bring up the variable selection window.
- Select **Exam1** through **Exam5** from the list of variables and then click **Ok**. “Exam1, Exam2, Exam3, Exam4, Exam5” will appear in the Cluster Variables box.
- Double-click in the **Row Label Variable** box. This will bring up the variable selection window. Select **Student** from the list of variables and then click **Ok**. “Student” will appear in the Row Label Variable box.
- Set **Variable Scaling Method** to **None**.
- Set **Row Scaling Method** to **None**.

4 Specify the reports.

- On the Clustered Heat Map window, select the **Reports tab**.
- Uncheck all reports except the heat map.

5 Enable the interactive Clustered Heat Map Format window.

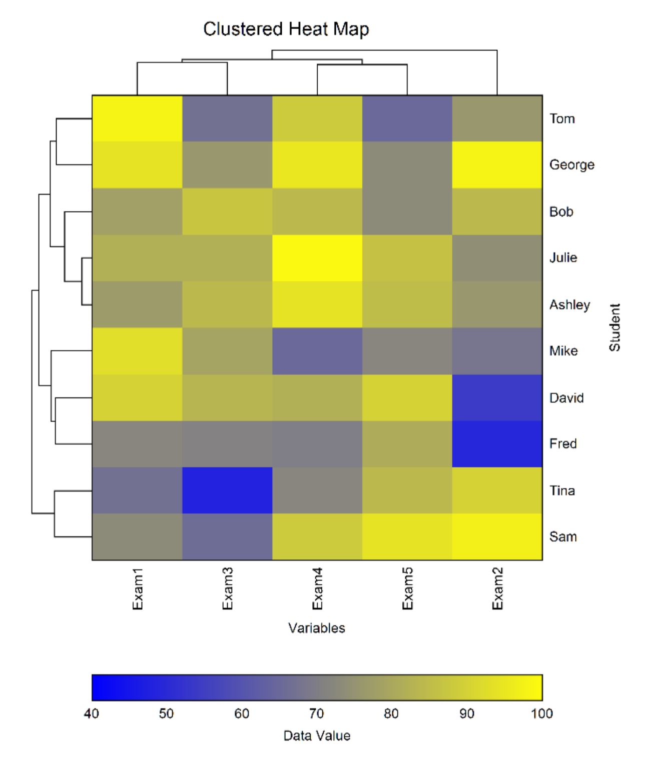
- On the Clustered Heat Map panel, select the **Heat Map tab**.
- Check the Edit During Run box on the upper-right corner of the Heat Map icon.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).
- This will display the Heat Map Format window.
- Click the **Heat Map** button on the left.
- Click the left portion of the **Data Gradient Fill** button. This will display the Gradient Window.
- Remove the center three stops so that there are only a blue and a red stop showing. Stops are removed by selecting them and then pressing the **Remove** button.
- Select the Red (right) button. Change the color to yellow by clicking on the **Color** button which is just below the **Remove** button.
- Click **OK** to hide the Gradient Window.
- Click **OK** again to hide the Heat Map Format window and display the following report.

© NCSS, LLC. All Rights Reserved.

Output



This report displays the heat map for these data with the above settings.

© NCSS, LLC. All Rights Reserved.

Example 3 – Creating a Clustered Heat Map with Slanted Labels

This section presents an example of how to create a clustered heat map of the exam score data with a gradient of just two colors and with slanted row labels down the right side of the plot. The data are found in the Exams database.

You may follow along here by making the appropriate entries or load the completed template **Example 3** by clicking on Open Example Template from the File menu of the Clustered Heat Maps (Double Dendrograms) window.

1 Open the Exams dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Exams.NCSS**.
- Click **Open**.

2 Open the Clustered Heat Map (Double Dendrograms) window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Clustered Heat Map (Double Dendrograms)** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Clustered Heat Map window, select the **Variables tab**.
- Double-click in the **Cluster Variable(s)** box. This will bring up the variable selection window.
- Select **Exam1** through **Exam5** from the list of variables and then click **Ok**. “Exam1, Exam2, Exam3, Exam4, Exam5” will appear in the Cluster Variables box.
- Double-click in the **Row Label Variable** box. This will bring up the variable selection window. Select **Student** from the list of variables and then click **Ok**. “Student” will appear in the Row Label Variable box.
- Set **Variable Scaling Method** to **None**.
- Set **Row Scaling Method** to **None**.

4 Specify the reports.

- On the Clustered Heat Map window, select the **Reports tab**.
- Uncheck all reports except the heat map.

5 Enable the interactive Clustered Heat Map Format window.

- On the Clustered Heat Map panel, select the **Heat Map tab**.
- Check the Edit During Run box on the upper-right corner of the Heat Map icon.

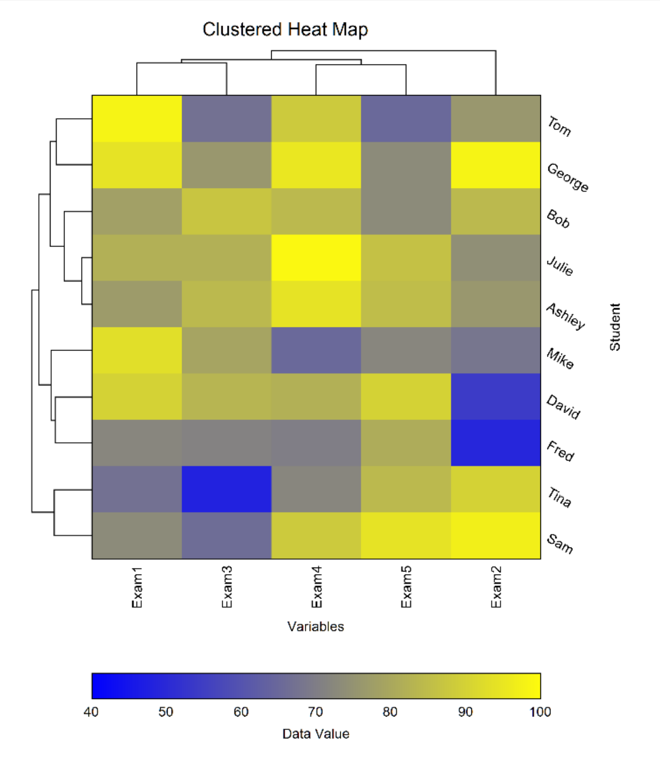
6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the Run button (the left-most button on the button bar at the top).
- This will display the Heat Map Format window.
- Click the **Heat Map** button on the left.
- Click the left portion of the **Data Gradient Fill** button. This will display the Gradient Window.
- Remove the center three stops so that there are only a blue and a red stop showing. Stops are removed by selecting them and then pressing the **Remove** button.
- Select the Red (right) button. Change the color to yellow by clicking on the **Color** button which is just below the **Remove** button.
- Click OK to hide the Gradient Window.

© NCSS, LLC. All Rights Reserved.

- Still on the Heat Map Format window, select the **Tick Labels – Rows** layout button at the bottom.
- Change the **Rotation Angle** to **-30.**
- Click **OK** to hide the Layout of Tick Labels window.
- Click **OK** again to remove the Heat Map Format window and display the plot.

Output



This report displays the heat map for these data with the above settings.

© NCSS, LLC. All Rights Reserved.