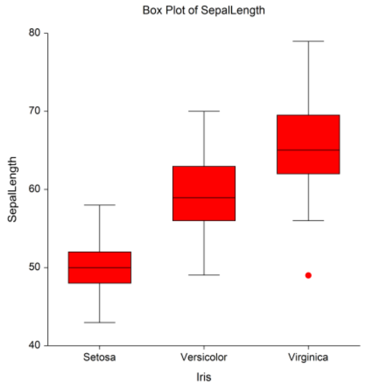
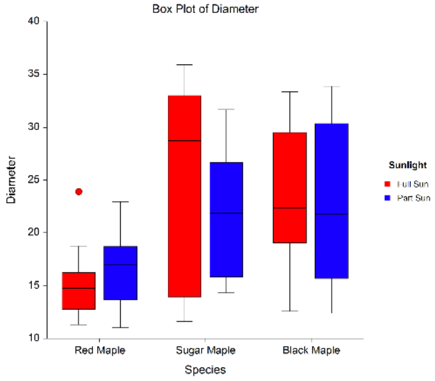
## Chapter 152

# Box Plots

## Introduction

When analyzing data, you often need to study the characteristics of a single group of numbers, observations, or measurements. You might want to know the center and the spread about this central value. You might want to investigate extreme values (referred to as outliers) or study the distribution or pattern of the data values. Several plots are available to allow you to study the distribution. One such plot is the box plot.





## Box Plot Definition

The box plot is defined by five data-summary values and also shows the outliers.

### Median and Box

The box portion of the box plot is defined by two lines at the 25th percentile and 75th percentile. The 25th percentile is the value at which 25% of the data values are below this value. Thus, the middle 50% of the data values fall between the 25th percentile and the 75th percentile. The distance between the upper (75th percentile) and lower (25th percentile) lines of the box is called the inter-quartile range (IQR). IQR is a popular measure of spread.

A line is drawn inside the box at the median (the 50th percentile). The median is a popular measure of the variable’s location (center).

### Whisker and Outlier Boundaries

A box plot whisker is a line that goes out from the box to the whisker boundaries. Often a crossbar line is drawn at the whisker boundary. Points outside the whisker boundaries are considered outliers. An additional boundary is sometimes used for severe outliers, although there is no line drawn at the severe outlier boundaries.

In NCSS there are two ways to define these boundaries. One way uses a multiplier of the inter-quartile range. The other uses percentiles.

### Boundaries using the Inter-Quartile Range (IQR)

This is the traditional method for determining the boundaries. In this method, the whisker boundary is found by multiplying a value (usually 1.5) times the IQR, and then going out that distance from the edge of the box. The whisker boundary is then brought back in to the first data value that is reached. In technical terms (for the upper whisker boundary), it is the largest observation that is less than or equal to the upper edge of the box plus the multiplier times IQR.

The severe outlier boundary is defined similarly, but the multiplier is larger (usually 3).

### Boundaries using Percentiles

The whisker boundary may also be defined in terms of percentiles, similarly to the edges of the box. For example, the two whisker boundaries might be the 10th percentile and 90th percentile (or 5th and 95th).

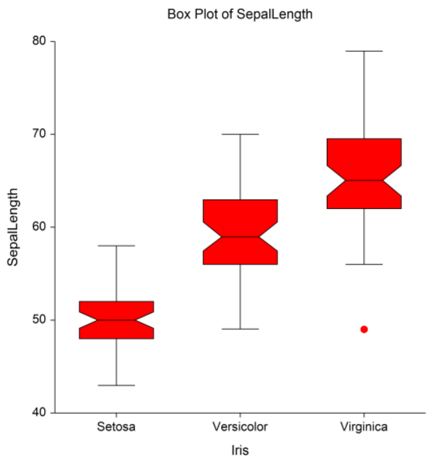
The severe outlier boundaries might be the 1st and 99th percentiles, for example.

## Multiple Comparisons

Box plots are often used for comparing the distributions of several groups of data, since they summarize the center and spread of the data very nicely.  When making comparisons among the locations (medians) of various batches, a modified box plot called the *notched box plot* is useful. The notches are constructed using the formula:

$$Median \pm 1.57 \times \left(IQR\right)/\sqrt{n}$$

Notched box plots are used to make multiple comparisons among the batches. If the notches of two boxes do not overlap, we may assume that the medians are significantly different (the centers are statistically significant). The 1.57 is selected for the 95% level of significance. The box plot below is an example of a notched box plot.



Note that when making comparisons among several batches, the notched box plots do not make any adjustment for the multiplicity of tests being conducted. Numeric testing with multiple comparison adjustments is recommended when multiple tests occur.

## Data Structure

A box plot is constructed from a numeric variable. A second variable may be used to divide the first variable into groups (e.g., age group or gender). In the two-factor procedure, a third variable may be used to divide the groups into subgroups.

# Procedure Options

This section describes the options available in this procedure.

# Variables Tab

This panel specifies which columns are used to create the box plot.

## Variables – Simple Box Plots

### Data Variable(s)

This option lets you designate which variables are plotted. If more than one Data Variable is designated and no Horizontal (Group) Variable is selected, a set of box plots will be displayed on a single chart, one box for each variable. If more than one variable is designated and a Horizontal (Group) Variable is selected, a separate box plot will be drawn for each variable.

### Horizontal (Group) Variable

Designates an optional variable used to separate the observations into groups. The groups defined by this variable will all appear on the same plot.

### Frequency Variable

Specify an optional frequency (count) variable. This variable contains integers that represent the number of observations (frequency) associated with each observation. If left blank, each observation has a frequency of one.

### Data Label Variable

A data label is text that is displayed beside each outlier point. This option designates the variable containing the data labels. The values may be text or numeric.

### Break Variable

Select an optional break (categorical) variable. A separate plot will be generated for each unique value of this variable. If you specify more than one break variable, a separate plot will be generated for each unique combination of the values of the variables specified.

## Variables – Two-Factor Box Plots

### Data Variable(s)

This option lets you designate which variables are plotted. If more than one Data Variable is designated and no Legend (Subgroup) Variable is selected, the variables will become the legend (subgroup) levels. If more than one variable is designated and a Legend (Subgroup) Variable is selected, a separate plot will be drawn for each variable.

### Horizontal (Group) Variable

Designates an optional variable used to separate the observations into groups. An individual box will be displayed for each unique combination of this variable with the Legend (Subgroup) Variable.

### Legend (Subgroup) Variable

Designates an optional variable used to separate the observations into subgroups. An individual box will be displayed for each combination of this variable with the Horizontal (Group) Variable. The levels of this variable will be shown in the legend.

### Frequency Variable

Specify an optional frequency (count) variable. This variable contains integers that represent the number of observations (frequency) associated with each observation. If left blank, each observation has a frequency of one.

### Data Label Variable

A data label is text that is displayed beside each outlier point. This option designates the variable containing the data labels. The values may be text or numeric.

### Break Variable

Select an optional break (categorical) variable. A separate plot will be generated for each unique value of this variable. If you specify more than one break variable, a separate plot will be generated for each unique combination of the values of the variables specified.

## Format Options

### Variable Names

This option specifies whether the column names or column labels are used on the chart.

### Value Labels

This option specifies whether the actual values or the labels from the Data Label Variable are used to label the points, and whether the values or the value labels are used for the group level labels of the plot.

## Box Plot Format

### Format

Click the format button to change the plot settings (see Box Plot Window Options below).

### Edit During Run

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

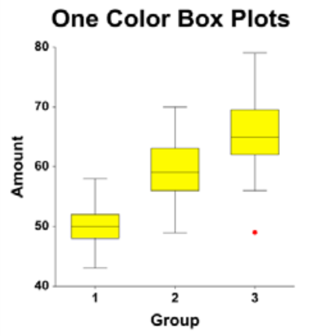
# Box Plot Window Options

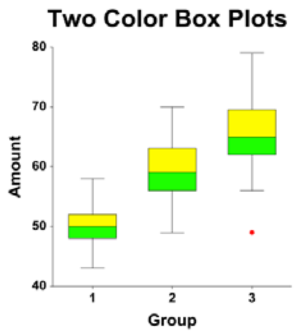
This section describes the specific options available on the Box Plot window, which is displayed when the Box Plot button is clicked. Common options, such as axes, labels, legends, and titles are documented in the Graphics Components chapter.

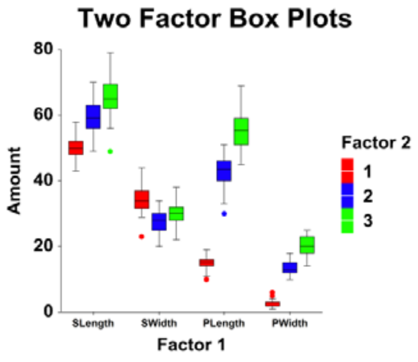
## Box Plot Tab

### General Section

This option specifies whether the same box colors are used above and below the median.

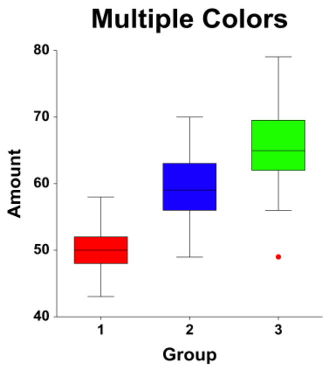


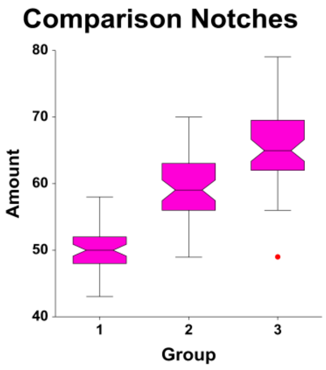




### Boxes Section

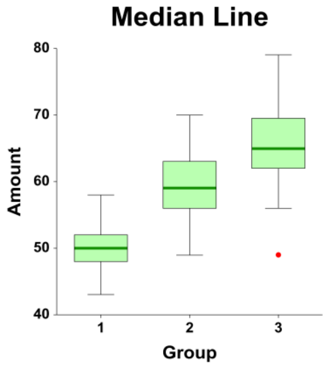
You can modify the colors of the boxes and their outline using the options in this section. You can also add special notches to the box so visual multiple comparisons to be made.

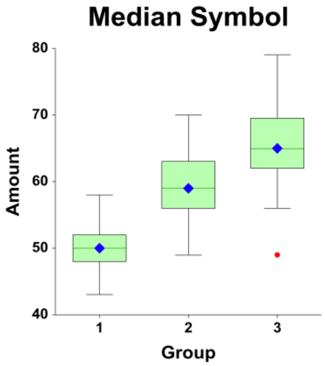




## Median Section

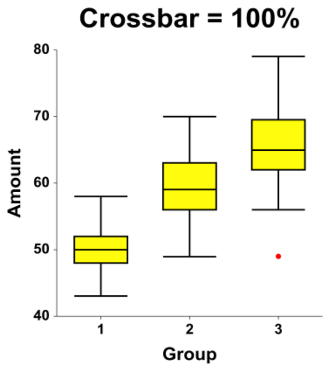
You can modify the color of the median line and/or symbol using the options in this section.

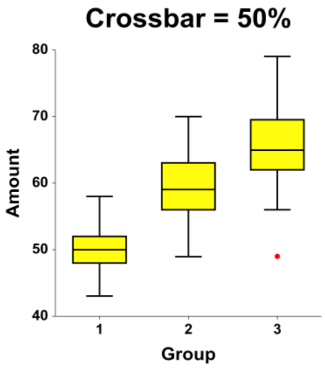

Median Line

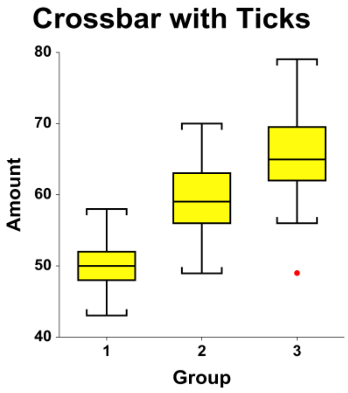

Median Symbol

## Whiskers Section

You can modify the format of the whiskers (the lines extending from the boxes) and the crossbars using the options in this section.

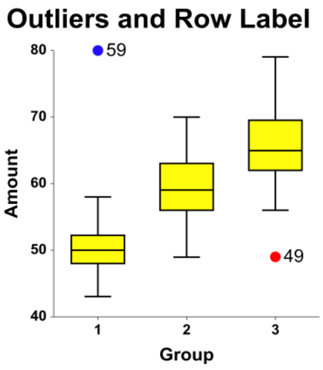

Crossbar = 100%


Crossbar = 50%


Crossbar with Ticks

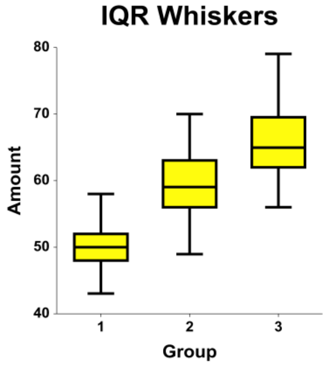
## Outliers and Outlier Labels Sections

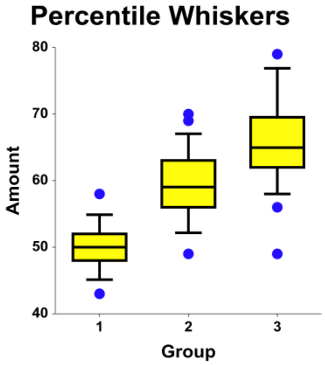
You can modify the outlier symbols and labels using the options in these sections.


Outliers and Row Label

## Whisker and Outlier Boundaries Section

You can modify the way in which the box and whisker boundaries are calculated using the option in this section. You can also change the multipliers used to calculate the regions for the outliers. The technical details are given above.

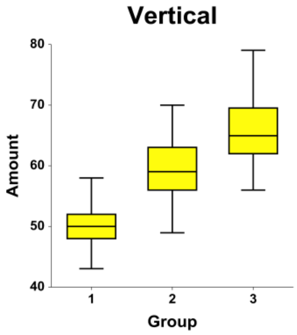

IQR Whiskers

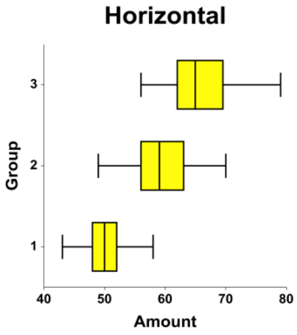

Percentile Whiskers

# Layout Tab

## Orientation Section

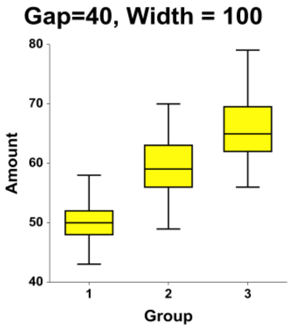
You can orient the box plots horizontally or vertically.

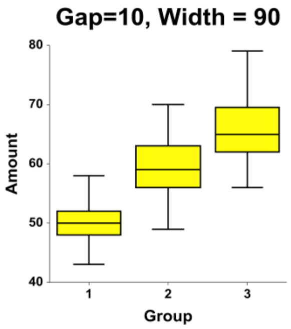

Vertical


Horizontal

## Object Spacing and Size Section

You can change the size of, and the gap between, individual box plots.

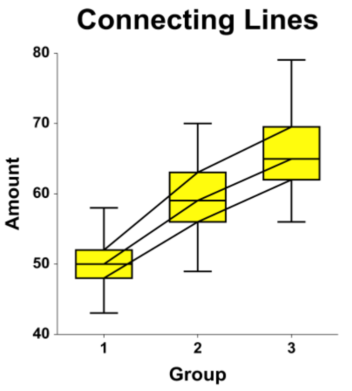

Gap=40, Width = 100


Gap=10, Width = 90

# Connecting Lines Tab

## Connect Between Groups Section

You can add reference lines at group means and percentiles.



# Titles, Legend, Numeric Axis, Group Axis, Grid Lines, and Background Tabs

Details on setting the options in these tabs are given in the Graphics Components chapter.

# Example 1 – Creating a Box Plot

This section presents an example of how to generate a box plot. The data used are from the Fisher dataset. We will create box plots of the *SepalLength* variable, grouping on the type of iris.

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 Box Plots window.

1 **Open the Fisher dataset.**

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

2 **Open the Box Plots window.**

- Using the Graphics menu or the Procedure Navigator, find and select the **Box Plots** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

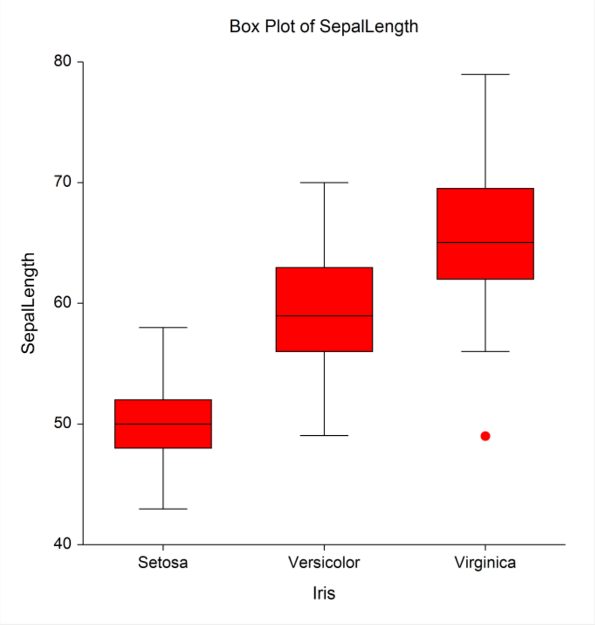
3 **Specify the variables.**

- On the Box Plots window, select the **Variables tab**.
- Double-click in the **Data Variable(s)** text box. This will bring up the variable selection window.
- Select **SepalLength** from the list of variables and then click **Ok**. "SepalLength" will appear in the Variable(s) box.
- Double-click in the **Horizontal (Group) Variable** text box. This will bring up the variable selection window.
- Select **Iris** from the list of variables and then click **Ok**. "Iris" will appear in the Horizontal (Group) Variable box.
- Set **Value Labels** to **Value Labels**.

4 **Run the procedure.**

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

## Box Plot Output



## Example 2 – Creating a Box Plot with Subgroups

This section presents an example of how to generate a box plot with subgroups. The data used are from the fictitious Tree dataset. We will create box plots of the *Diameter* variable, grouping on *Species*, with subgroups according to *Sunlight*.

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 Box Plots (2 Factors) window.

### 1 Open the Tree dataset.

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

### 2 Open the Box Plots (2 Factors) window.

- Using the Graphics menu or the Procedure Navigator, find and select the **Box Plots (2 Factors)** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

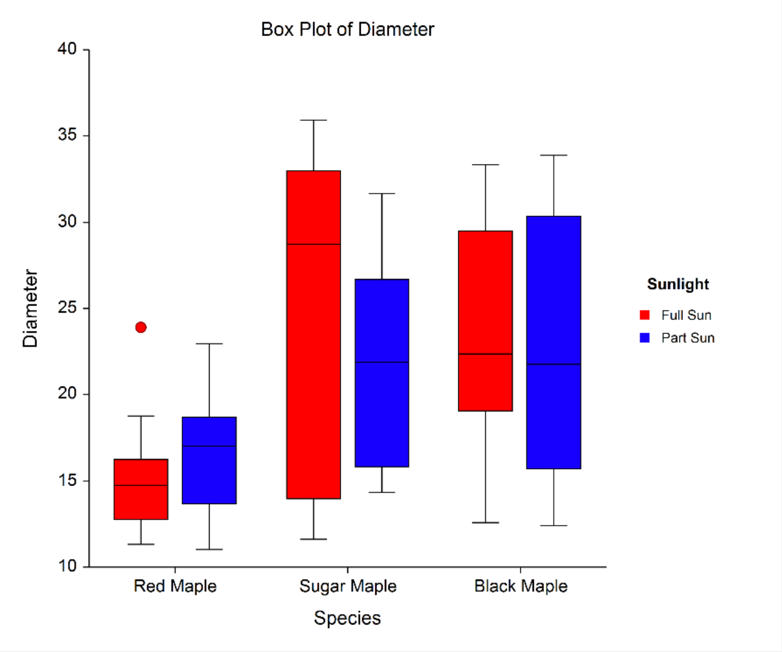
### 3 Specify the variables.

- On the Box Plots (2 Factors) window, select the **Variables tab**.
- Double-click in the **Data Variable(s)** text box. This will bring up the variable selection window.
- Select **Diameter** from the list of variables and then click **Ok**. “Diameter” will appear in the Variable(s) box.
- Double-click in the **Horizontal (Group) Variable** text box. This will bring up the variable selection window.
- Select **Species** from the list of variables and then click **Ok**. “Species” will appear in the Horizontal (Group) Variable box.
- Double-click in the **Legend (Subgroup) Variable** text box. This will bring up the variable selection window.
- Select **Sunlight** from the list of variables and then click **Ok**. “Sunlight” will appear in the Legend (Subgroup) Variable box.
- Set **Value Labels** to **Value Labels**.

### 4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

## Box Plot Output


Box Plot of Diameter
Diameter
Red Maple
Sugar Maple
Black Maple
Species
Sunlight
🟥 Full Sun
🟦 Part Sun