Box-and-Whisker Plot

Summary

The **Box-and-Whisker Plot** procedure creates a plot designed to illustrate important features of a numeric data column. It was first described by John Tukey (1977) in his box Exploratory Data Analysis. The box-and-whisker plot summarizes a data sample through 5 statistics:

1. minimum
2. lower quartile
3. median
4. upper quartile
5. maximum

It can also indicate the presence of outliers.

Sample StatFolio: *boxplot.sgp*

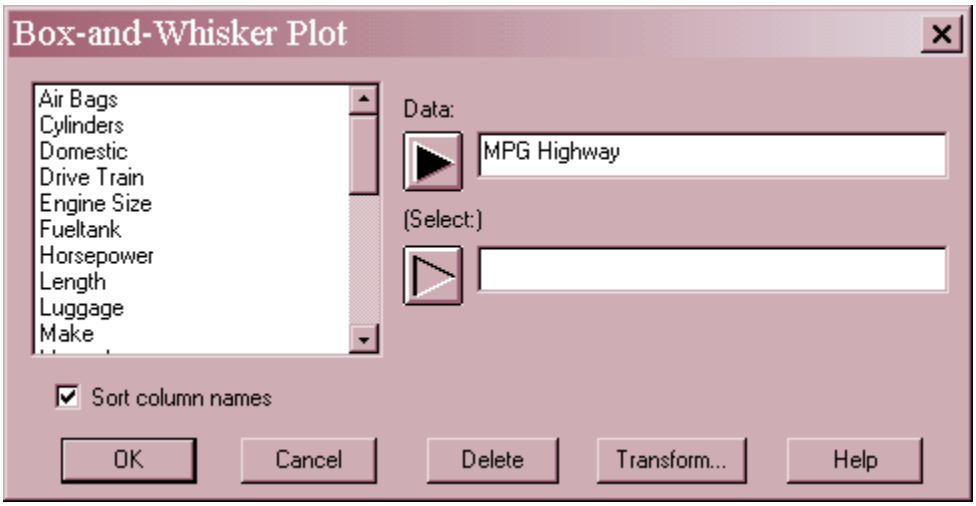
Sample Data:

The file *93cars.sf3* contains information on 26 variables for $n = 93$ makes and models of automobiles, taken from Lock (1993). The table below shows a partial list of 3 columns from that file:

Make	*Model*	*MPG Highway*
Acura	Integra	31
Acura	Legend	25
Audi	90	26
Audi	100	26
BMW	535i	30
Buick	Century	31
Buick	LeSabre	28
Buick	Roadmaster	25
Buick	Riviera	27
Cadillac	DeVille	25
Cadillac	Seville	25
Chevrolet	Cavalier	36

Data Input

The data to be analyzed consist of a single numeric column containing *n* = 2 or more observations.



- **Data :** numeric column containing the data to be summarized.
- **Select:** subset selection.

Analysis Summary

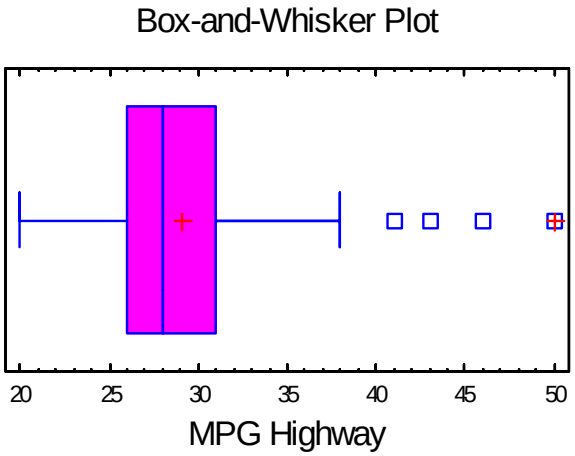
The Analysis Summary shows the number of observations in the data column.

Box-and-Whisker Plot - MPG Highway
Data variable: MPG Highway
93 values ranging from 20.0 to 50.0

Also displayed are the largest and smallest values.

Box-and-Whisker Plot

This pane displays the box-and-whisker plot.



The plot is constructed in the following manner:

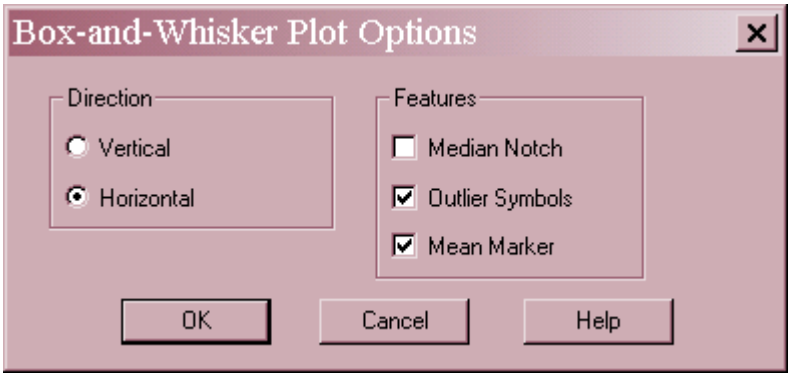
- A box is drawn extending from the *lower quartile* of the sample to the *upper quartile*. This is the interval covered by the middle 50% of the data values when sorted from smallest to largest.
- A vertical line is drawn at the *median* (the middle value).
- If requested, a plus sign is placed at the location of the sample mean.
- Whiskers are drawn from the edges of the box to the largest and smallest data values, unless there are values unusually far away from the box (which Tukey calls *outside points*). Outside points, which are points more than 1.5 times the interquartile range (box width) above or below the box, are indicated by point symbols. Any points more than 3 times the interquartile range above or below the box are called *far outside points*, and are indicated by point symbols with plus signs superimposed on top of them. If outside points are present, the whiskers are drawn to the largest and smallest data values which are not outside points.

The above plot for the data on highway miles per gallon is notable for several reasons. First, it indicates a somewhat asymmetric data sample. This can be seen by the fact that the sample mean lies somewhat to the right of the median line. In addition, the data extend out further above the median than they do below the median.

Second, there are 4 outside points. When sampling 100 observations from a normal distribution, outside points can be expected to occur just by chance about half the time, but usually only one or two. Far outside points, of which there is 1, occur extremely rarely.

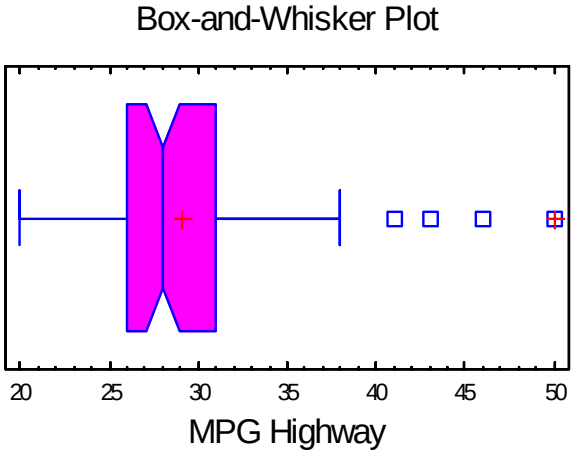
To identify the far outside point at the far right, hold the mouse down on it. A small box will appear indicating that the data is from row 39 of the datasheet and that the value on the X axis is 50 mpg. If you now select the datasheet, row 39 will have been highlighted. The automobile in that row is a Geo Metro, which also happens to be the car with the lowest weight in the dataset.

Pane Options



- **Direction**: the orientation of the plot, corresponding to the direction of the whiskers.
- **Median Notch**: if selected, a notch will be added to the plot showing an approximate 100(1-α)% confidence interval for the median at the default system confidence level (set on the *General* tab of the *Preferences* dialog box on the *Edit* menu).
- **Outlier Symbols**: if selected, indicates the location of outside points.
- **Mean Marker**: if selected, shows the location of the sample mean as well as the median.

Example – Notched Box-and-Whisker Plot
The following plot shows the addition of a median notch at the 95% confidence level.



The notch covers the interval

$$sample\ median \pm z_{\alpha/2}\frac{1.25(IQR)}{1.35\sqrt{n}} \qquad (1)$$

where *IQR* is the sample interquartile range, *n* is the sample size, and $z_{\alpha/2}$ is the upper (α/2)% critical value of a standard normal distribution. The notch, which ranges from approximately 27 to 29, provides an indication of the potential sampling error in the median, assuming that the data are a random sample from a normal population (a dubious assumption in this case).