Boxplot

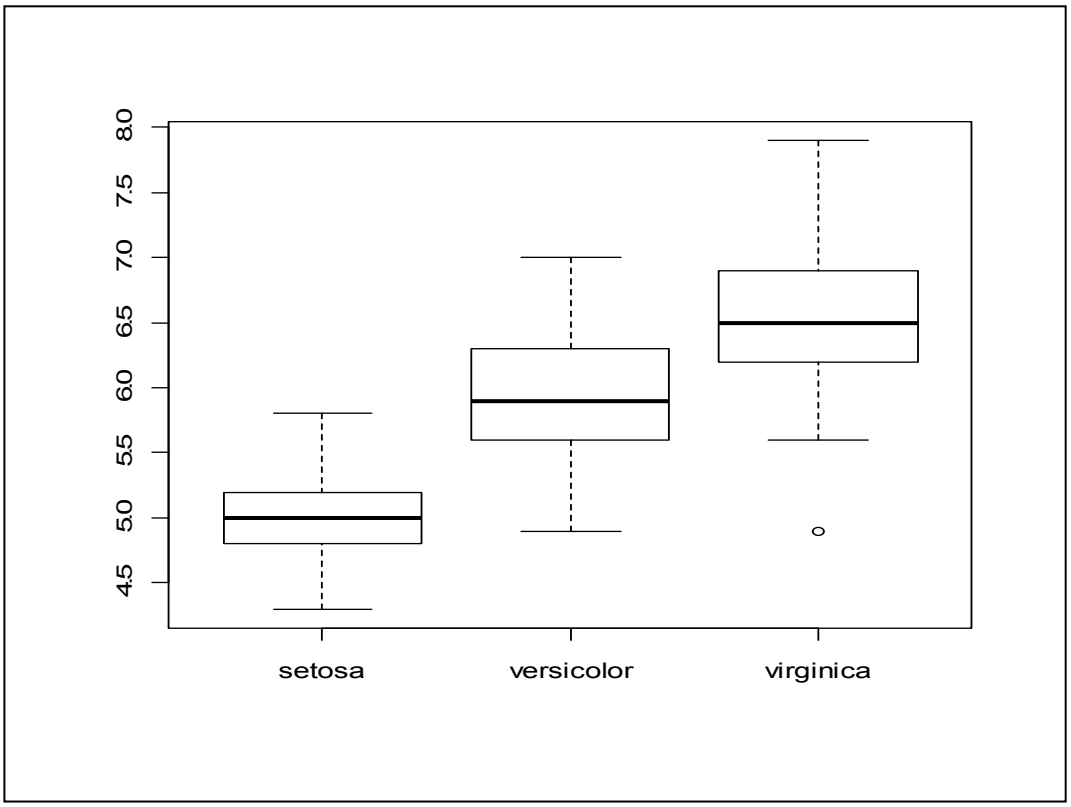
By: Meaghan Petix, Samia Porto & Franco Porto

A **boxplot** is a convenient way of graphically depicting groups of numerical data through their five-number summaries: the smallest observation (sample minimum), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (sample maximum). A boxplot may also indicate which observations, if any, might be considered outliers.

- Use when the explanatory variable is categorical rather than continuous
    - Categorical variables are factors with two or more levels

#SAMPLE BOXPLOT SCRIPT

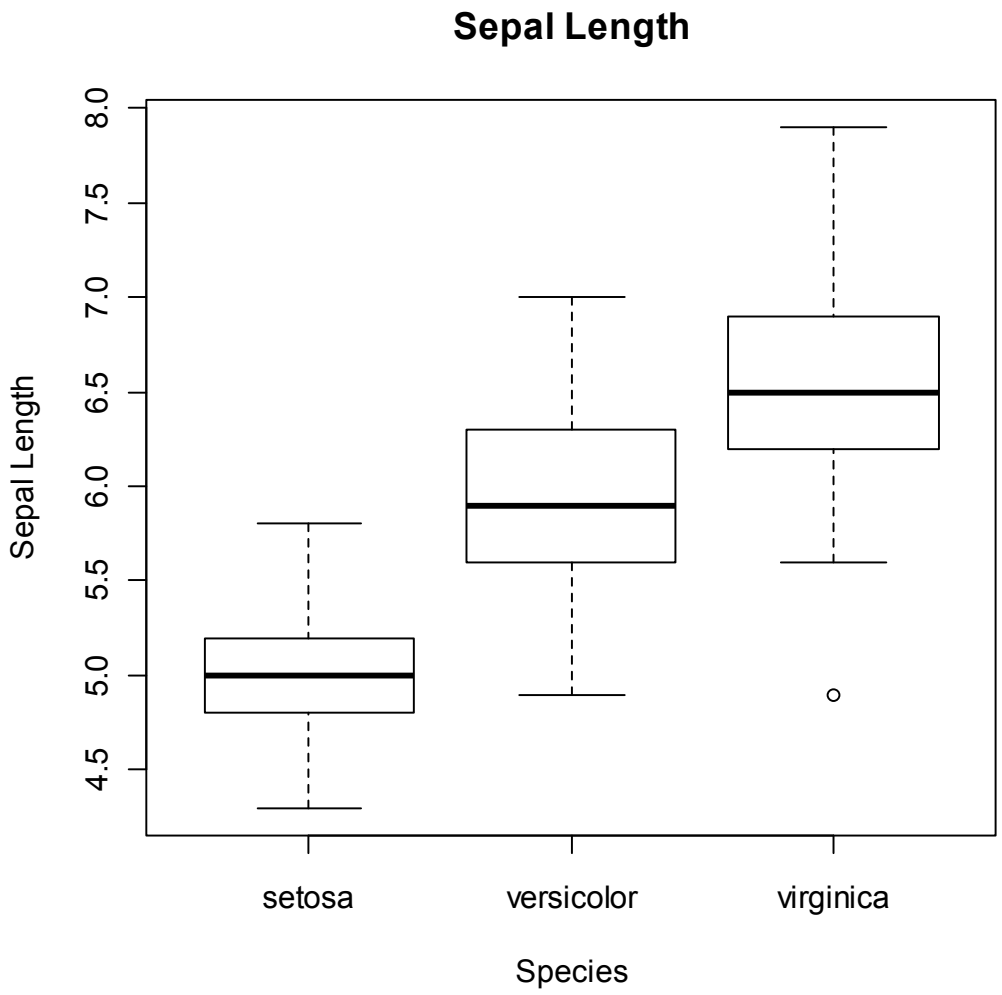
```
iris
attach(iris)
SL=Sepal.Length
plot(Species,SL)
```



#This plot looks at the variable of Sepal Length comparatively between the three distinct Iris species (setosa, versicolor, and virginica). Not only does it portray useful information, it is a good plot because it's easily read and aesthetically pleasing. A thick black line reveals the mean Sepal Length for each of the species, and the range of Sepal Lengths is also shown - from the min to the max value, also including the lower and upper quartiles. Possible outliers are denoted by dots outside of the boxplot.

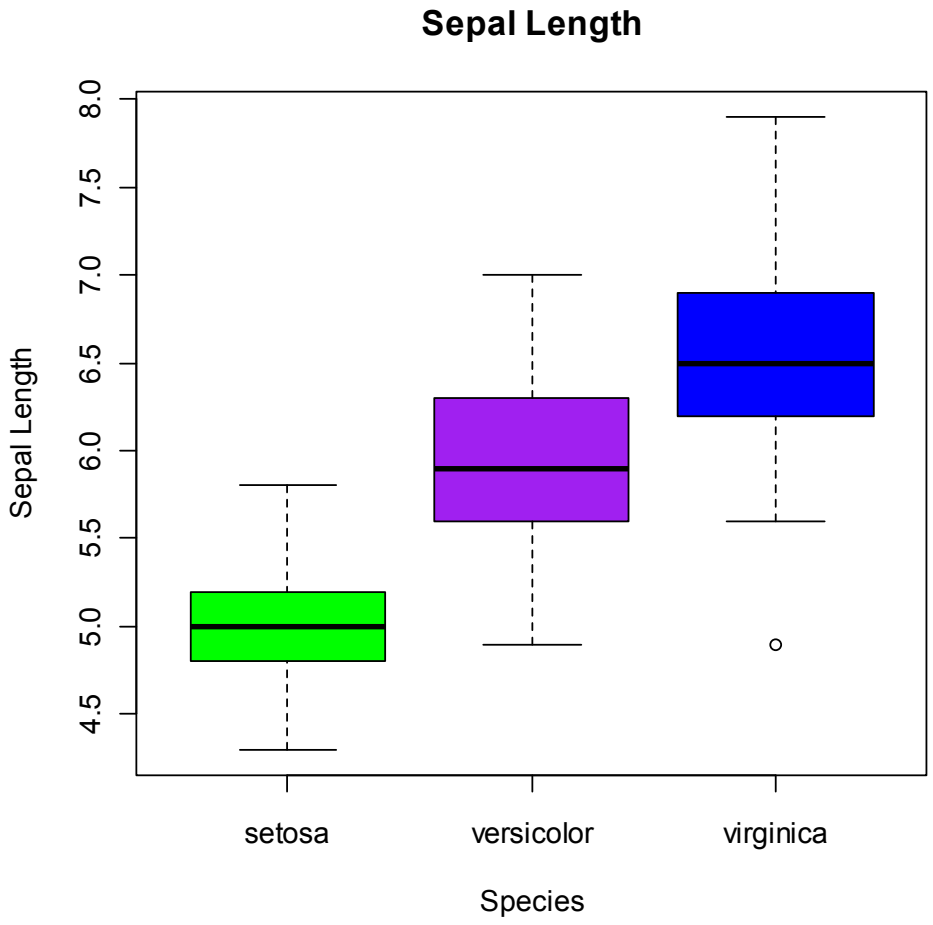
Creating Labels and Adding a Title for Box Plot:

```
#SAMPLE BOXPLOT SCRIPT (w/ axes labels and title)
attach(iris)
plot(Species, Sepal.Length,xlab="Species",ylab="Sepal Length")
title("Sepal Length")
```



The horizontal line shows the **median** sepal length value for each species. The bottom and top of the box show the 25th and 75th **percentiles**, respectively. The vertical dashed lines are called the "whiskers"; they show either the maximum value or 1.5 times the **interquartile range** of the data, whichever is smaller. The quantity '1.5 times the interquartile range of the data' is roughly 2 standard deviations, and the interquartile range is the difference in the response variable between the first and third **quartiles** (also called 25th and 75th percentiles). Points more than 1.5 times the interquartile range *above the first quartile* and points more than 1.5 times the interquartile range *below the first quartile* are defined as **outliers** and plotted individually. Thus, when there are no outliers the whiskers show the maximum and minimum values. Boxplots not only show the location and spread of the data, but also indicate skewness; skewness shows up as asymmetry in the sizes of the upper and lower parts of the box.

#Adding colors:
plot(Species, Sepal.Length,col=(c("green","purple","blue")),main="Sepal Length",xlab="Species",ylab="Sepal Length")
#Notice you can use main="Title" as another way to add a title to your plot



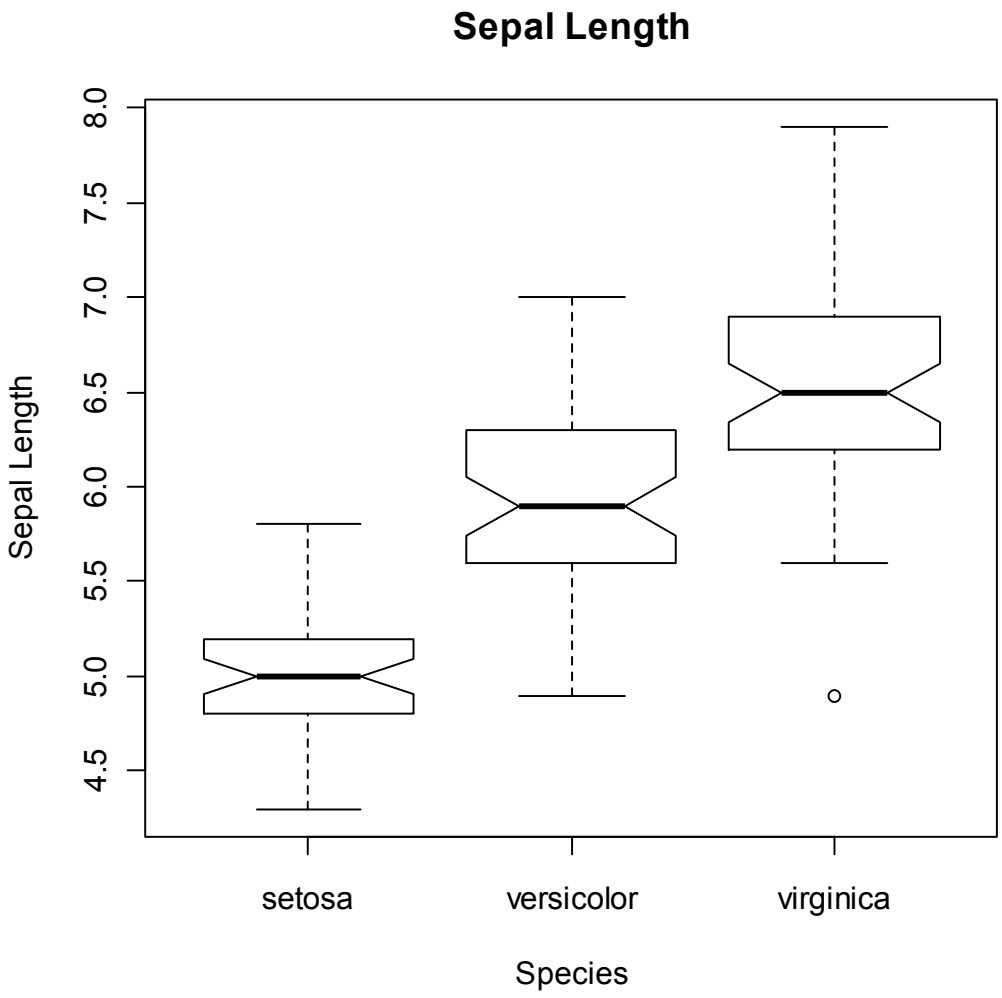
Checking Your Graph for Accuracy:
Built-In Functions for Five-Number Summaries

#To save on typing, assign shorter names:
SL=Sepal.Length
SL.setosa=Sepal.Length[Species=="setosa"]
SL.versicolor=Sepal.Length[Species=="versicolor"]
SL.virginica=Sepal.Length[Species=="virginica"]
quantile(SL)
quantile(SL.setosa)
quantile(SL.versicolor)
quantile(SL.virginica)
#You can also use the following functions for minimum value, median, and maximum value:
min(SL.setosa)
median(SL.setosa)
max(SL.setosa)

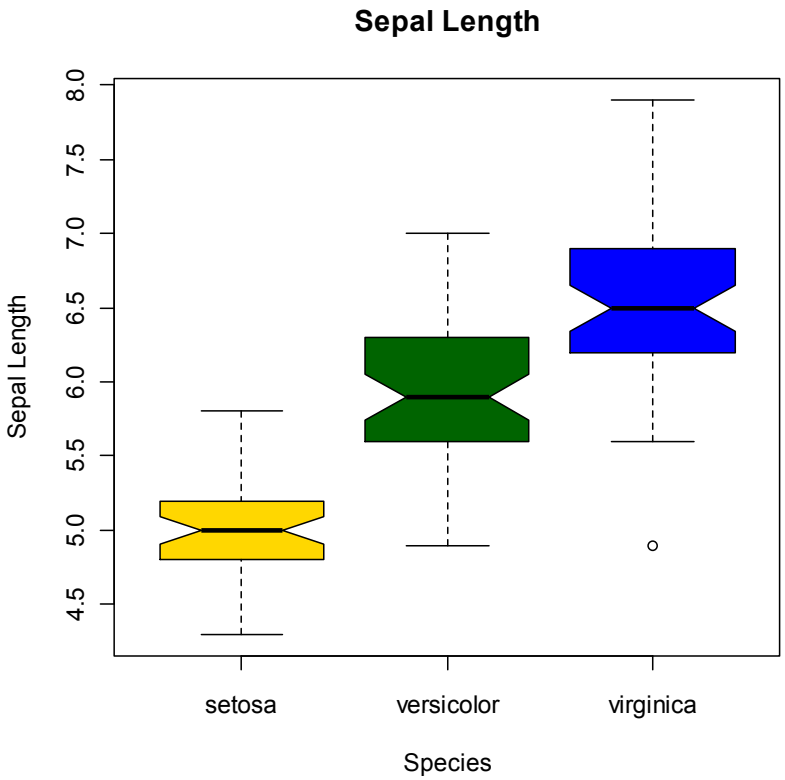
Notched Boxplot

Notched boxplots are useful in showing whether or not median values are significantly different from one another. The **notches** are drawn as a 'waist' on either side of the median and are intended to give a rough impression of the significance of differences between two medians.

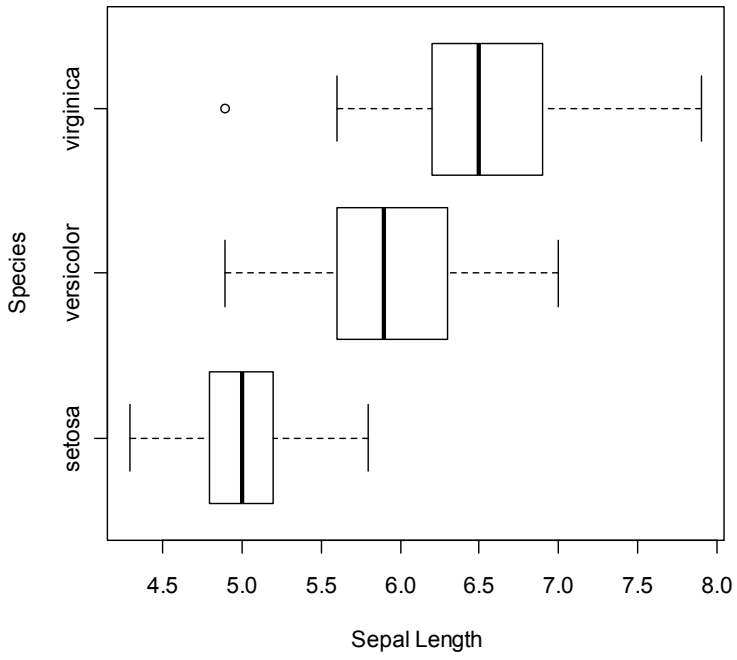
#Notched Boxplot (with axes labeled and title included)
boxplot(Sepal.Length~Species,notch=TRUE,xlab="Species",ylab="Sepal Length")
title("Sepal Length")
#You can change certain aspects of your plot by indicating "TRUE or "FALSE" (which can be simplified to "T" or "F")
#ex. notch=TRUE for the plot to be notched



#Adding colors:
boxplot(Sepal.Length~Species, notch=TRUE, col=(c("gold","darkgreen","blue")),main="Sepal Length", xlab="Species",ylab="Sepal Length")



#For the boxplot to be oriented **horizontally**:
plot(Species, Sepal.Length,xlab="Sepal Length",ylab="Species", horizontal=TRUE)

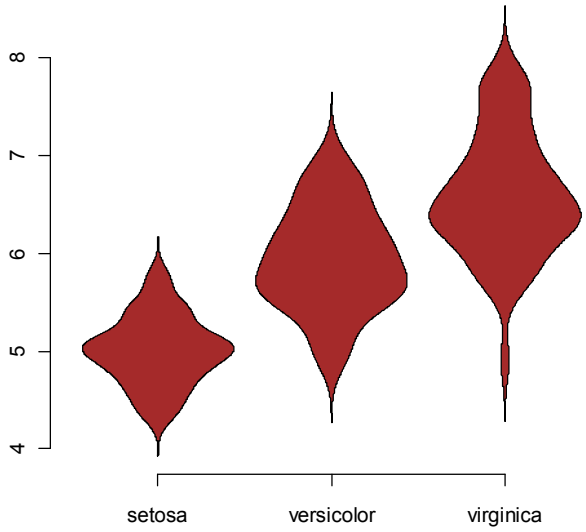


Violin Plot

This function serves the same utility as side-by-side boxplots, only it provides more detail about the different distribution. It plots violinplots instead of boxplots. That is, instead of a box, it uses the density function to plot the density. For skewed distributions, the results look like "violins" (hence the name).

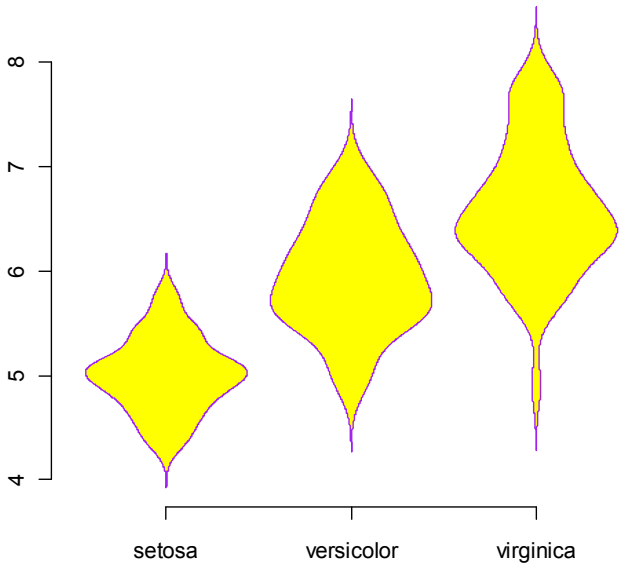
In order to use the violinplot() function you have to install the package "UsingR"

```
?violinplot
violinplot(Sepal.Length~Species, col="brown")
```



#To add a border color:

```
violinplot(Sepal.Length~Species, col="yellow",border="purple")
```

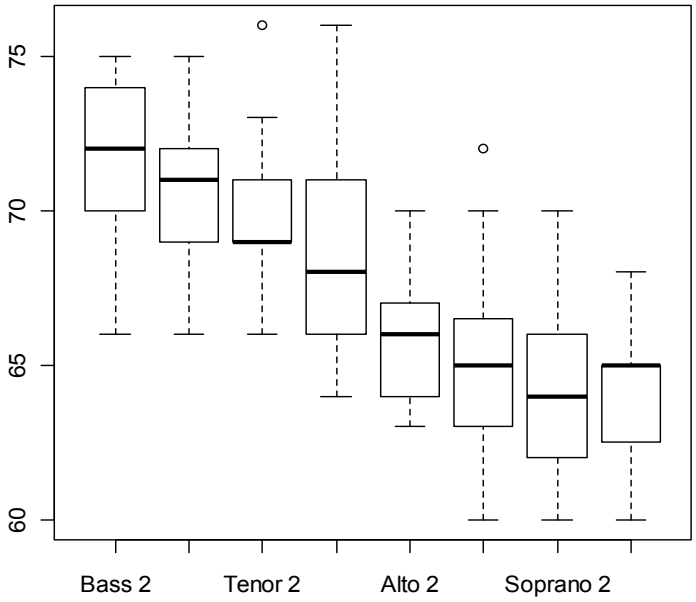


Boxplots with varied widths:

#Boxes are drawn with widths proportional to the square-roots of the number of observations in the group

Install package “lattice” and then load package:

```
singer
attach(singer)
boxplot(height~voice.part)
```




```
boxplot(height~voice.part,varwidth=TRUE)
```