## Video: Boxplots in MATLAB (5:36)

**(0:00):**
First, I'm going to load the fisheriris data. In my first plot, I want to look at the distribution of the length data measurements without respect to species, so I'm going to create a new variable, called flowerLens that has just the length measurements in it. Then create a new figure, and tell MATLAB to create a boxplot on my new variable, and that the first data column is for the sepal values and the 2nd column is for the petal values. And this is the figure that is generated. In looking at the sepal values, they are fairly evenly distributed in that the distance between the minimum, the 1st quartile, the median and the 3rd quartile are all about the same size. But look at the distance between the 3rd quartile and the maximum. It is almost twice the distance of the other values, so those values are much more spread out than the lower 75% of the data. To contrast, look at the petal lengths. Here, the lower 25% is about the same as the distance between the median and the 75% data mark, so that data is distributed evenly, but the upper whisker is much longer, but still not as big as the distance between the 25% data and the median. What could this mean? Remember that we are looking at 3 different species, and petal length might be different between the three, but it doesn't seem that sepal length is that different.
**(1:25):**
Let's dig into the sepal length by species. I create a variable that has just the sepal length, create a new figure, and tell MATLAB to create a boxplot on this new variable, and break out the data by the variable species, which contains the species name for each row of the data. This data set is grouped by each species, but if the data was not grouped, this command would still work. Let's analyze this new figure. When comparing the y-axis from the previous boxplot, it maps correctly, with a range of about 4.5 to 8mm.. But with an expanded y axis, the three species have very different sepal lengths. First, look at the setosa species. Very tightly grouped interquartile range, and the whiskers are about the same length. The entire data set is between about 4.4 to 5.8mm for about 1.4 mm range – very small... The versicolor species has about the same distribution – even interquartile range and even whiskers, but the box is larger with an overall range of almost 2 mm.. The third species has about the same box structure as the versicolor, but the upper whisker is much larger and the lower whisker is about right and there is an outlier. Notice that the boxes on the 2 right species overlap, so there are many similarities, and there is no box overlap on the first two. Going back to our first boxplot of the consolidated data, the larger upper whisker makes sense, most of that is probably the virginica species measurements. What does this mean? Remember that these plants require pollinators, and so similarities allow the same pollinators to visit both flowers.

**(3:32):**
What about the sepal widths? How do they compare? First, I need to create a variable with just the sepal widths, and then have MATLAB plot it, again, based on species, but this time, I want to add a 'notch'. The notch marks the 95% confidence level for the median values. What that means

is that we are 95% certain that the actual median for the underlying population is within the interval marked by the notches. If the notches from two box plots do NOT overlap, we can assume that the medians are different, to a 0.05 significance value. In this plot, we see that the setosa is much wider than the other two, and it's notches do NOT overlap either of the other two notches, so we are confident that it really is different. The notches for the other two species overlap, so those medians might be the same. – we're just not sure.

**(4:52):**

In Example 7, we create a boxplot for the Daphne Island and Santa Cruz data sets. Remember that they are very different sized arrows – Daphne island has 751 data points, and Santa Cruz has only 43. These birds have different beak sizes as shown in this boxplot because the notches do NOT overlap. Boxplots are powerful in that they show the data distribution, and make no assumption about the underlying statistics of the data. If you cannot assume a normal data set, mean and standard deviation should not be used, but a boxplot, showing the quartiles, shows you a lot about the data and it's distribution.