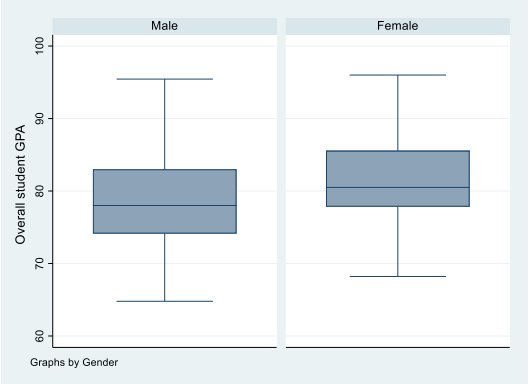
Visualizing data in Stata – Section 3.4
Najib Mozahem

In this section, we will see how we can use the box plot in order to compare different groups. Just like histograms, we can use the **by()** option with the box plot:

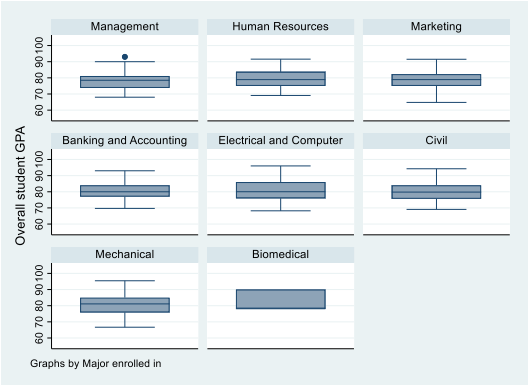
graph box gpa, by(gender)



In this case, the box plot is a better tool than the histogram for comparison. We can easily see that all five values (the lower limit, the 0.25 quantile, the median, the 0.75 quantile, and the upper limit) for females are higher than for males.

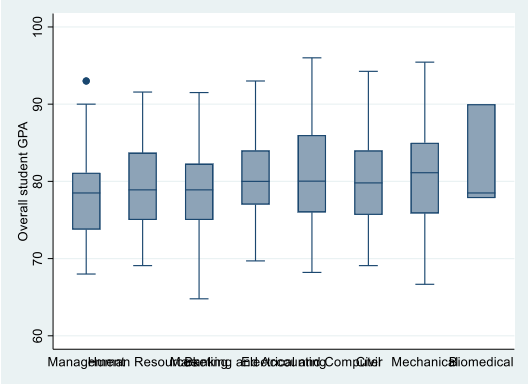
It is possible to divide the graphs when there are more than two groups in a variable:

graph box gpa, by(major)



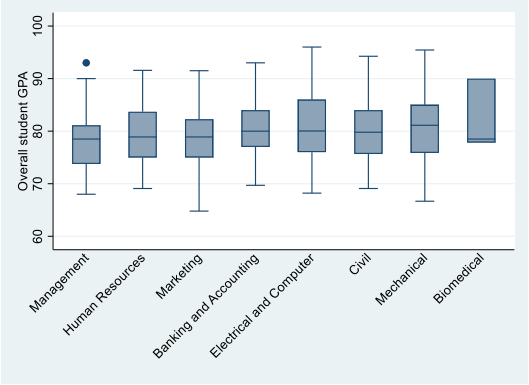
Since there are eight majors, the command produces eight plots. Notice that in the case of the biomedical group that the graph is missing some information. The reason is that there are four students in our dataset that are enrolled in this major. Four observations is not enough to calculate all five statistics used in the box plot. Unfortunately, the graph is not very readable. It is difficult to compare the different groups. In the last section, we used the **twoway** command in order to draw the graphs together, instead of producing separate graphs. This was done by using the *if* qualifier. In the case of the box plot, we do not need to do this. There is an easier solution, and it is to use the **over()** option instead of the **by()** option:

graph box gpa, over(major)



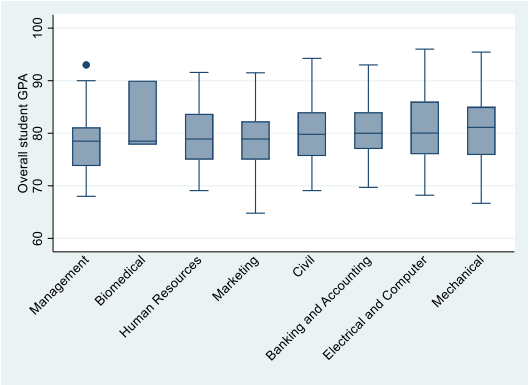
It is now easier to compare the GPAs of the different majors. Unfortunately, the names of the majors overlap. We can solve this problem by controlling the angle in which the labels are draw using the **label()** option:

graph box gpa, over(major, label(angle(45)))



We told Stata that we wanted to produce a box plot for each major, and that the label for each major should be written in a 45 degree angle. This way, long labels did not overlap. To make it even easier to compare, we can ask Stata to sort the graphs:

graph box gpa, over(major, sort(1) label(angle(45)))



We added the **sort(1)** option. This option tells Stata to sort the boxes on the median of variable 1. In our case we have only one variable which is *gpa*, so Stata sorts the boxes in ascending order.