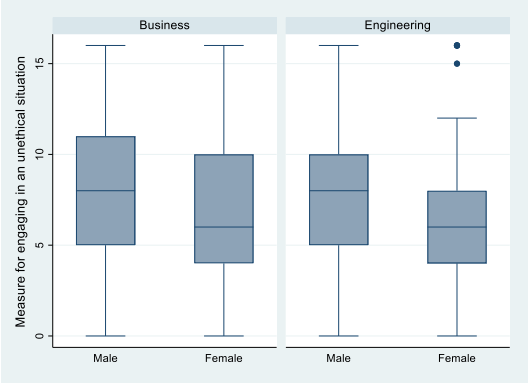
Visualizing data in Stata – Section 4.2
Najib Mozahem

We start with the box plots. In the previous section we saw how we can use the **by()** and **over()** options in order to look at group differences. The difference between the two is in the fact that the **over()** option draws both boxes on the same graph thus making it easier to compare. However, it is also possible to use both together:

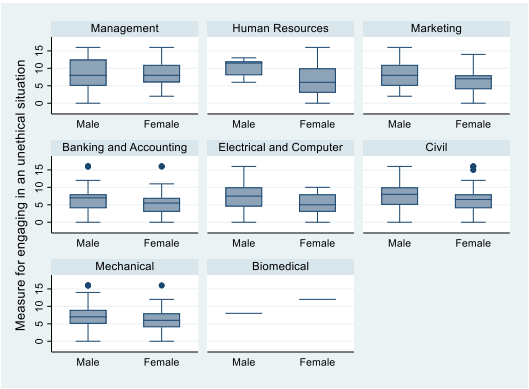
graph box engage, by(college, note("")) over(gender)



What this graph does is that it compares the level of engagement in cheating behavior of both genders in each of the two colleges. What we actually see is that the median (the line in between) is lower for females than it is for males in both colleges. An interesting finding is that female engineers cheat the least out of all group, since the 0.75 quantile point is much lower than that of other groups. We also see that there are two outliers for the case of female engineers.

We can produce the same graph but choose to divide it by major instead of college:

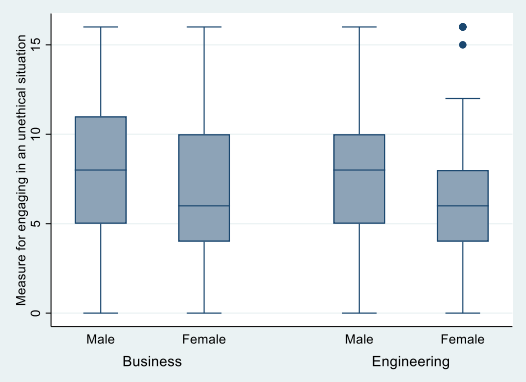
graph box engage, by(major , note("")) over(gender)



Again we see that in every major except biomedial, females engage in cheating less than males, or at least this is what they report.

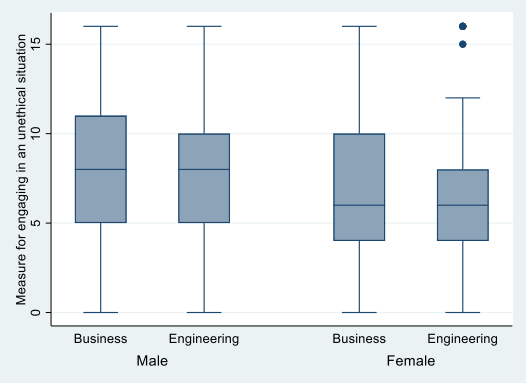
Stata also allows us to specify the **over()** option more than once. So instead of using both the **by()** and the **over()** option, we can use the **over()** option twice:

graph box engage, over(gender) over(college)



Notice that since the **over(gender)** option was specified first, Stata divides the graphs first by gender and then by college. We can reverse this order:

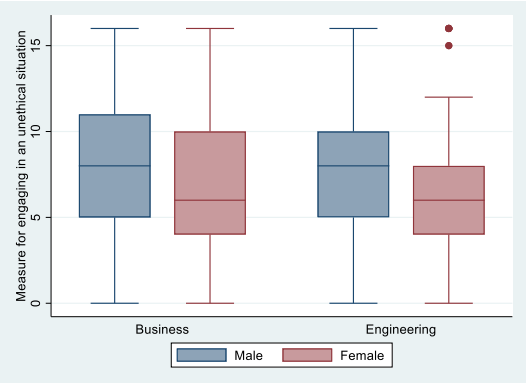
graph box engage, over(college) over(gender)



The choice of graph depends on what you are looking for.

At this point, you might comment that the graph colors are not very intuitive. Perhaps it would be better to use different colors for different groups in order to make things clearer. This can be accomplished by using the **asyvars** option:

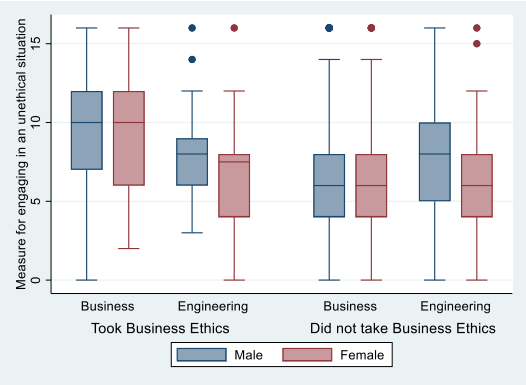
graph box engage, over(gender) over(college) asyvars



The **asyvars** option tells Stata to treat the first grouping, which in our case is the *gender* variable, as if it were different multiple y variables. Stata does not use different colors for different groups, but it uses different colors for different variables. The result is that we have a graph that is more readable. In addition, when there is more than one y variable, Stata produces a legend in order to tell us which color corresponds to which graph. This is why when we used the **asyvars** option, Stata produced a legend.

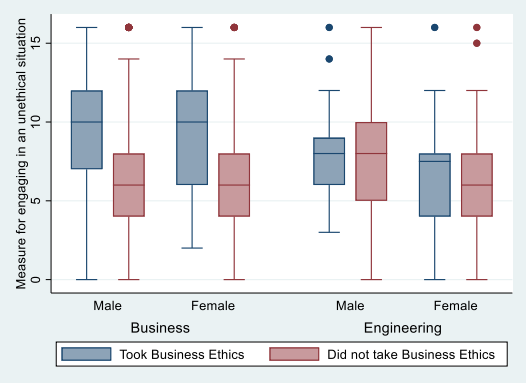
Let us now introduce a third variable, which is *course*:

graph box engage, over(gender) over(college) over(course) asyvars



This graph is not easy to read because the way the figures are divided is not intuitive. What I would like is to compare students who have taken the course business ethics with those who haven't. Although I can do that in the graph produced above, it is not easy because the two groups are not side-by-side. We can produce a graph that is better suited to our research question by changing the order of the group division:

graph box engage, over(course) over(gender) over(college) asyvars



Now we see that the different colors correspond to whether the student has taken the course or not. Since the different colors are placed side-by-side, it is not easier to visually compare the two groups. Interestingly, we see that students who have taken the course tend to engage in more cheating than those who hadn't, and this seems to be true for both gender in both colleges.