THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

WEEK 3

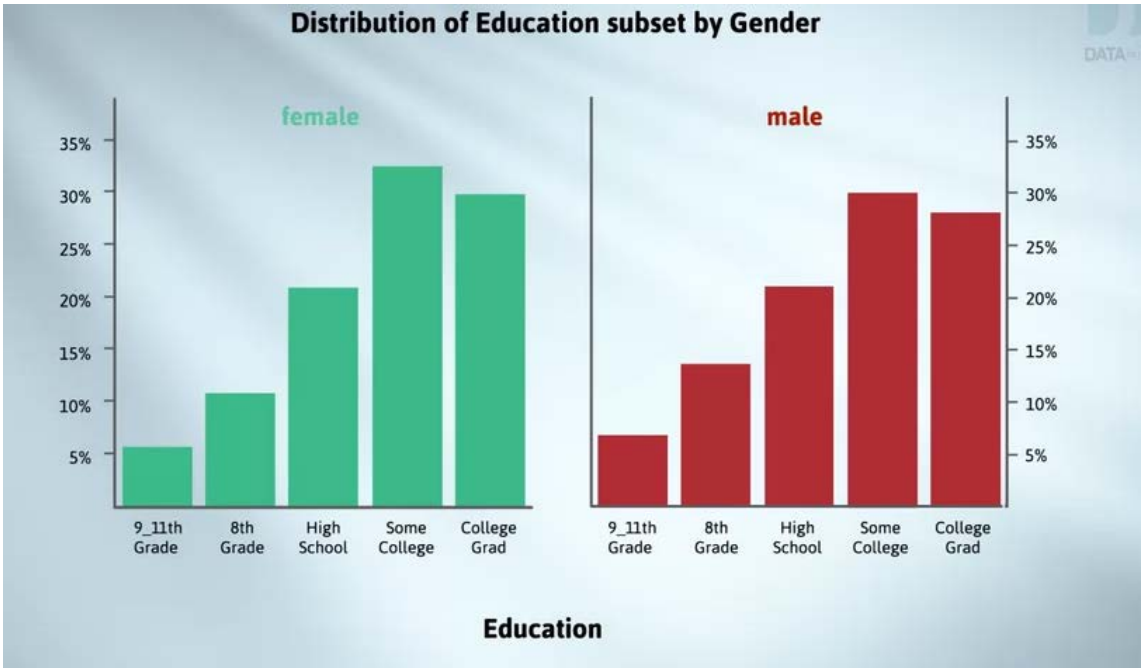
3.2 RELATIONSHIPS BETWEEN CATEGORICAL VARIABLES

Hi again. In this video, you'll learn how to plot data on two categorical variables so that you can look for relationships between them.

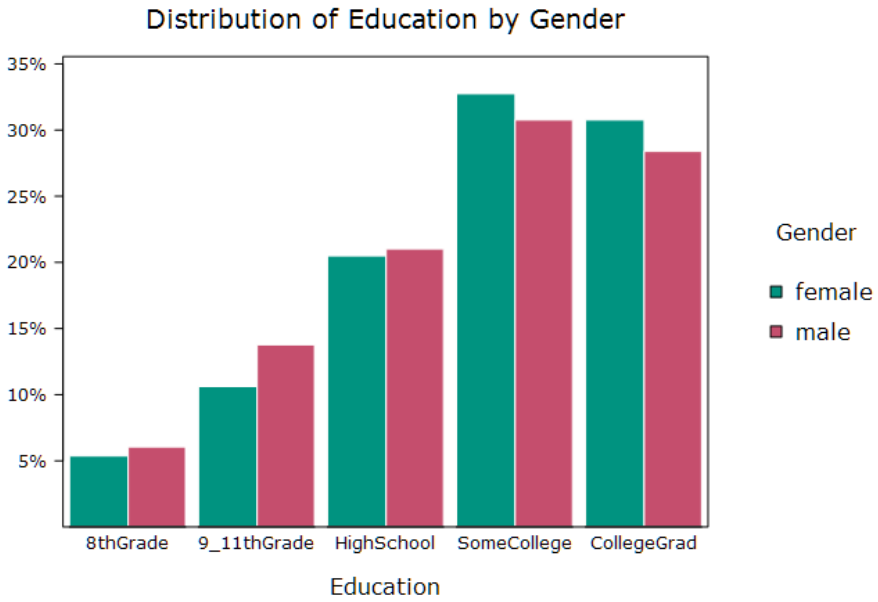
In the week three introductory video, we talked about relationships in terms of a variable of primary interest-- called an outcome variable-- and variables that might help us predict the outcome. These we call predictor variables.

Using the enhanced 2009-2012 data, we'll investigate how age group and gender predict educational achievement levels. Education is our outcome of interest. Age and gender are our predictor variables.

First, using gender as a predictor, we want to see how the distribution of educational attainment differs between females and males. Here we have two separate plots for education -- one for females and one for males. If there was no difference between the female and male distributions, we'd say there was no relationship between education and gender.



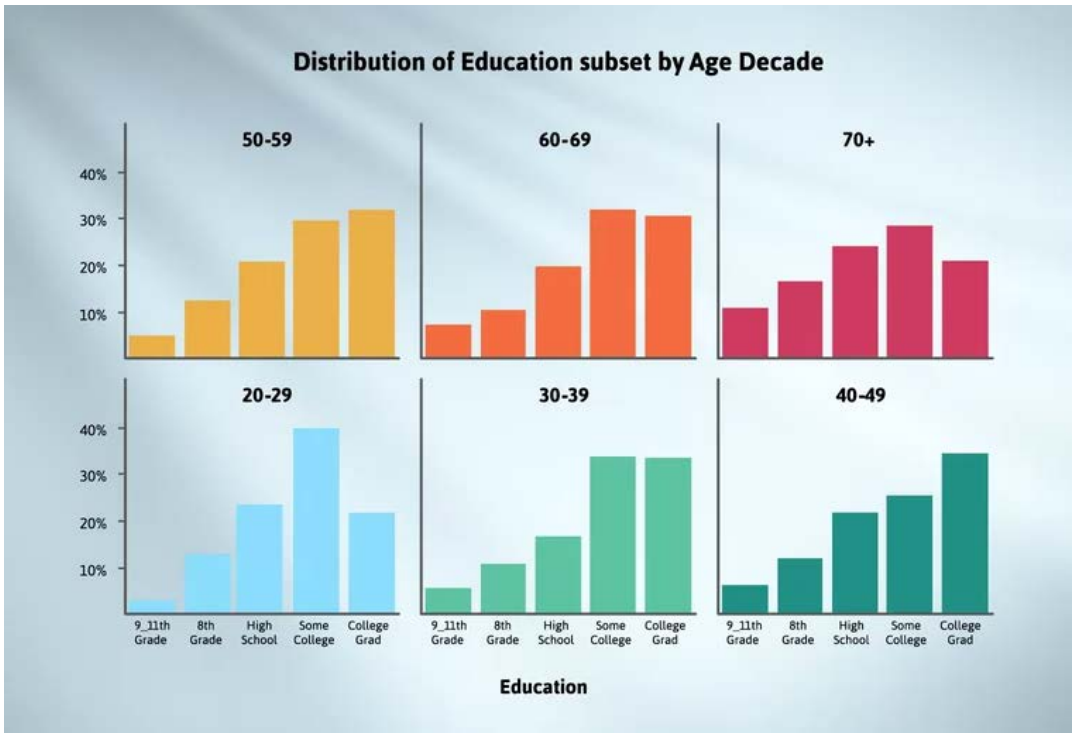
The two graphs here are very similar but slightly different. For example, about 30% of females are college graduates, compared with about 28% of males. It looks as though the two right-hand bars for females are higher than those for males-- females slightly more likely to have college education, while the left-hand bars are slightly shorter-- proportionately more males in the lower levels of educational attainment.



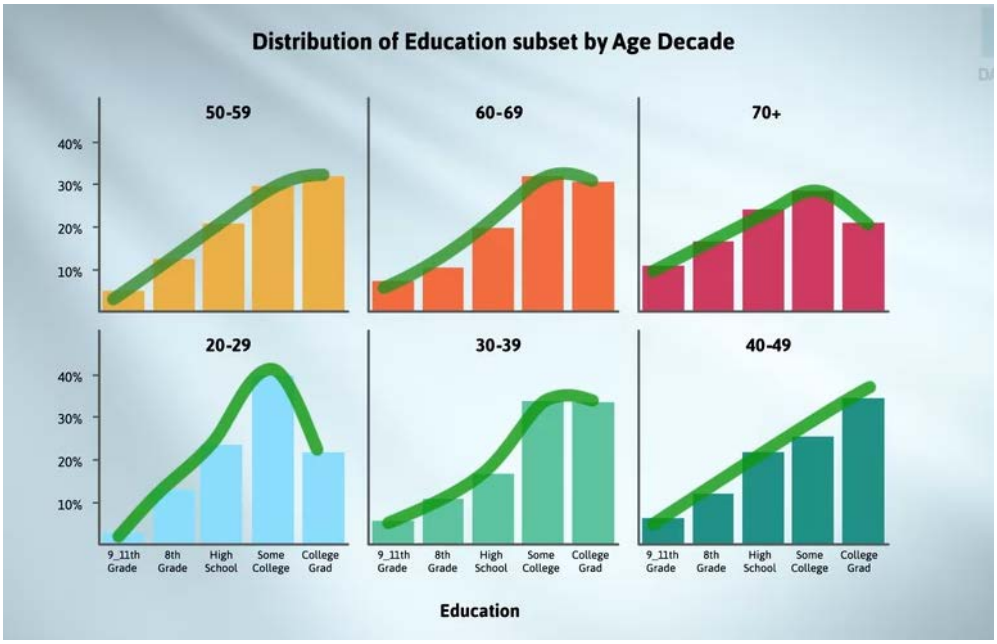
We'll now rearrange the sets of bars so that corresponding bars for females and males are placed beside one another. This is called a side-by-side bar chart. This makes the small differences between the educational attainment outcomes much

more obvious. The colour coding tells us what predictor group we are looking at--green for females, red for males.
Which plot should we use? Both! Both have their strengths, and we should use both.

The first law of using graphics for discovery is that you should look at many types of graphs. Often you'll spot something in one that you missed in another. A separate set of plots of the outcome variable, one for each predictor group, is good for revealing gross differences and overall shape. The side-by-side plot is good for looking at detailed differences.

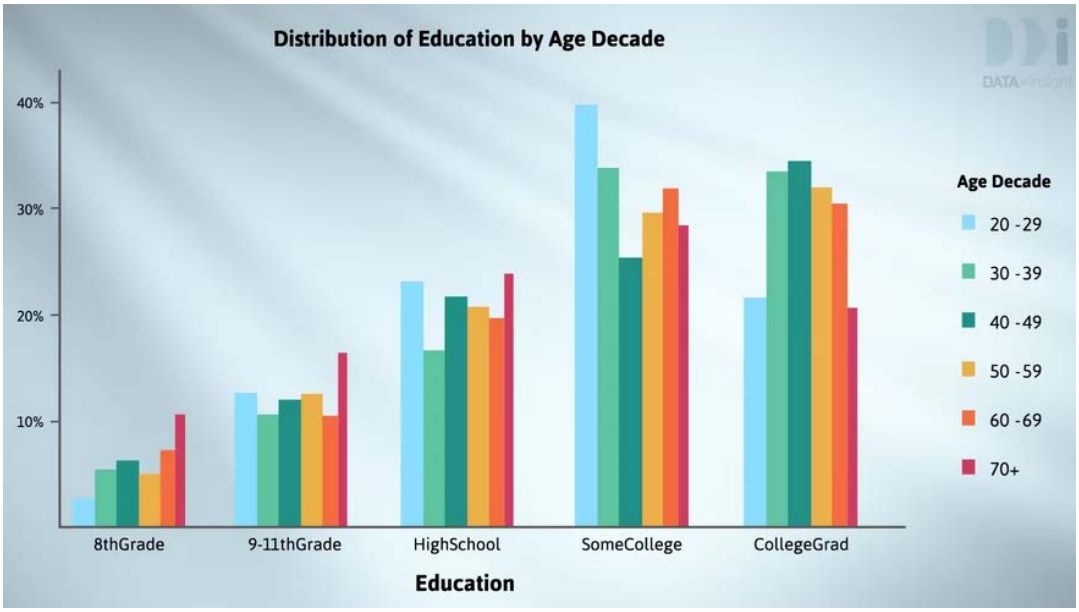


Age in decades is a predictor variable with more categories. Here we have a separate plot for the outcome variable education for each category of age decade. The separateness of the groups has been emphasised by using colour. If there was no relationship between the outcome education and the predictor age decade, all of these plots would be the same. But in fact, there are quite large differences.



These freehand curves emphasise how the plots differ in shape. The shapes of the education distributions for the youngest and the oldest group are very different from the shapes for any of the other age groups. The main difference is the substantially lower percentages of college graduates. We'll think about possible explanations later.

Now we'll use side-by-side bars to highlight the differences. Higher bars correspond to larger percentages and lower bars to lower percentages.



On the right-hand side, we can see the reduction in the percentages of college graduates with each decade from age 50, and the low percentage of college

graduates in the light blue 20-29 age group. The red 70+ group generally has less education, shown by lower than usual bars for college graduates and by higher than usual bars in all of the lower three categories of education. The light blue 20-29 group has unusually large percentages in the high school and some college categories. We should expect this because many of the younger ones in particular would not yet have finished their formal education.

You may have noticed in this plot that some bars are narrower than others. The widths of the bars have been made proportional to the number of people in the group. There are less than half as many people in the 70+ group then there are in the first three age groups. For all their good points, a big disadvantage of the side-by-side arrangement of bars is that people often get confused by these graphs, particularly by getting the "percentages of what for who" the wrong way around.

They may look at the cluster of CollegeGrad bars and think they're being told about the percentages of CollegeGrads who fall into each age group. But they do not. These percentages do not add to 100%. We have to think what is the outcome variable? These percentages add to 100%. And for what groups are we comparing those outcomes? The percentages that add to 100% are those for bars with the same colour. They tell us about the outcome variable results for people in that colour group.

With iNZight graphs, the outcome variable is in the default graph title. If we see distribution of education, we know the graph is telling us about educational outcomes. The colour groups are the age groups. We're looking at the educational outcomes of the various age groups. This is very different from the age group outcomes for the different educational groups.

To keep your bearings, it is best to look at both separate and side-by-side plots. And bear in mind that the side-by-side plot is just a rearrangement of the separate plots' bars.

In summary, our main tools for investigating the relationship between two categorical variables is separate bar charts of the outcome variable for each predictor group and side-by-side bar charts.

Separate bar charts are best for revealing gross differences in our overall shape, while side-by-side bar charts are better for highlighting detailed differences between corresponding categories.

Finally, I'll leave you with these questions to remind you of the ideas we've just covered.