DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

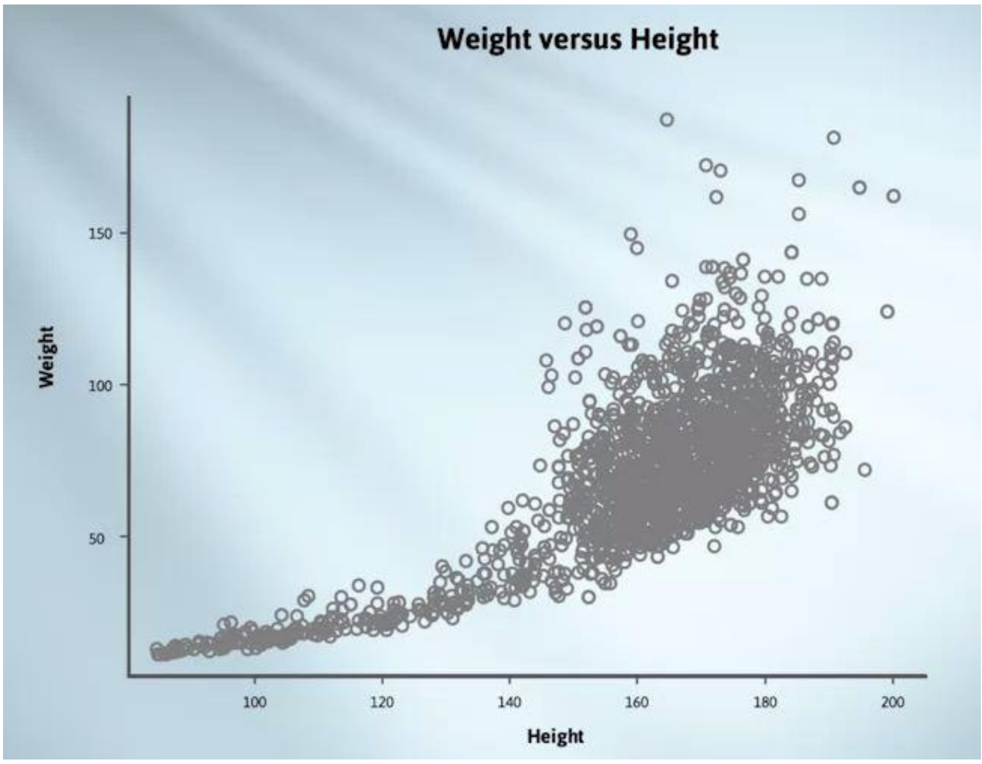
THE UNIVERSITY OF AUCKLAND

WEEK 4

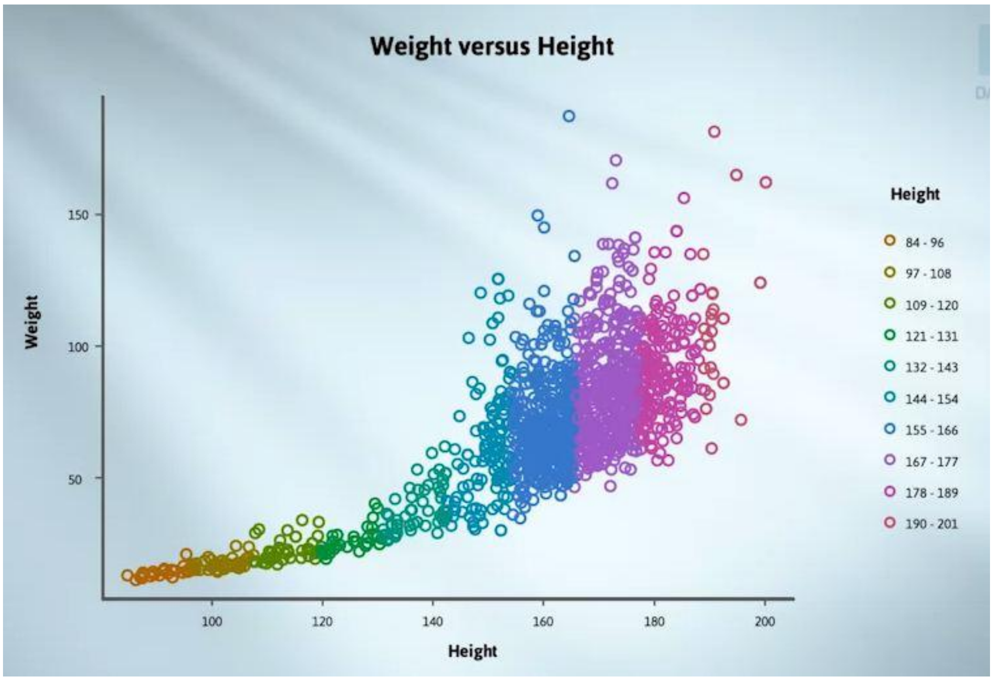
DIVING DEEPER WITH MORE VARIABLES by Chris Wild

We've been investigating relationships between two numeric variables using scatter plots. We'll now see how colour and subsetting can help us to answer more subtle questions involving more variables.

We'll work with the NHANES 2000 data set and variables you understand well. Our tools will reveal things you already know, but seeing how they do that will help you understand how the tools work and their potential for revealing features in new situations.

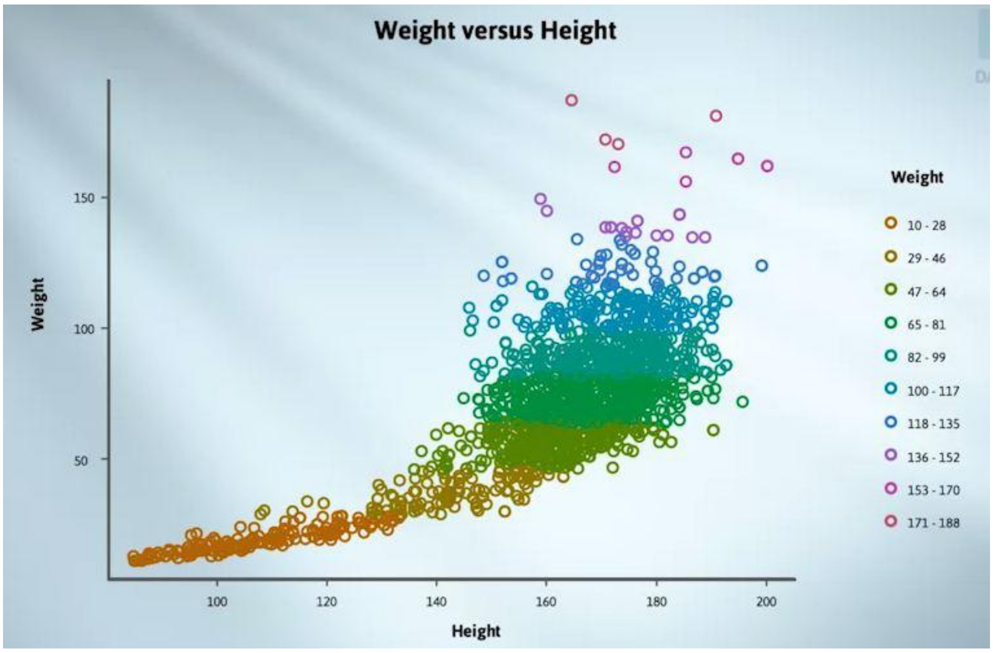


Let's start with relating weight to height. One of the things we'll try to understand is why this relationship looks the way it does.



Coloured by height

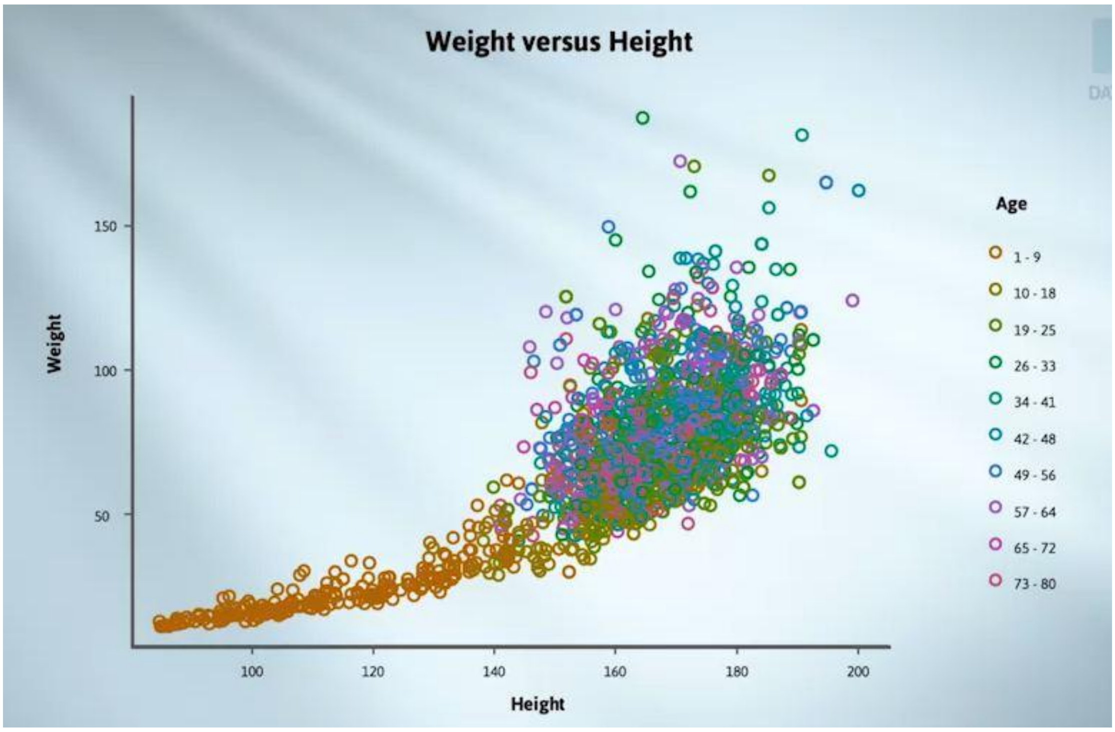
Here we've coloured by height (the horizontal scale). The height range has been broken up into intervals. Reading the right-hand legend from the top, they're 84-96, 97-108, 109-120, and so on. The points are then coloured accordingly. This makes vertical bands of colour across the plot as we might expect.



Coloured by weight

Now we've coloured by weight (the vertical scale) resulting in horizontal bands. I've shown these two graphs to give the idea of a colour gradient. The ones we're seeing are perfect colour gradients, going from left to right with increasing height when we used height, and bottom to top with increasing weight when we used weight. Of course, we don't need colour to see the effects of height or weight because that's what's shown directly on the graph.

In reality, we use colour to investigate the effect of a new variable.



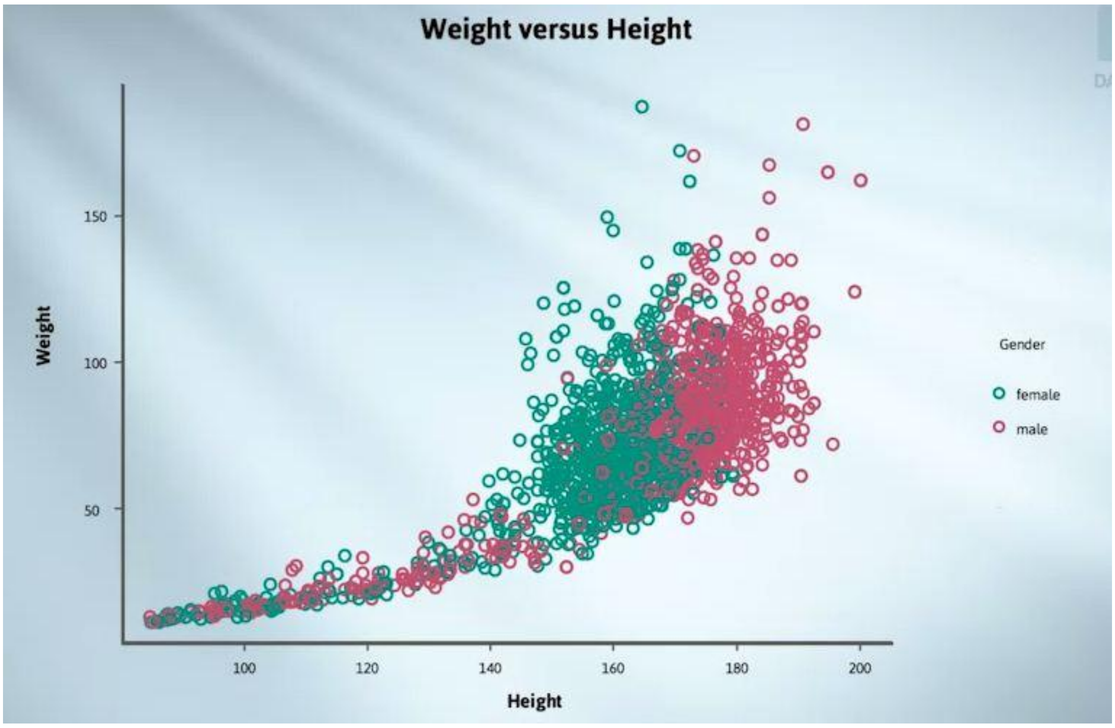
Coloured by age

Here I've used colour to see what's happening with age. The age range has been broken into bands. Reading from the legend, they're 1-9, 10-18, 19-25. The colours are shown alongside.

What does the graph show?

That lower left-hand tail shape up to a height of about 140 centimetres is brown, belonging to children under 10. Then up to a height of about 150 centimetres, we're mainly seeing the dull brown-green colour of the next stage band. After a height of about 150 centimetres, we're seeing the colours relating to all of the age bands over 10 all mixed up. And there is some suggestion of fewer people from the older two age bands in the higher weights and heights.

These are all things we might expect from our general knowledge. Early in life, people grow a lot. Later in life, they tend to shrink a little. But this example shows us how colour can alert us to these types of behaviour.

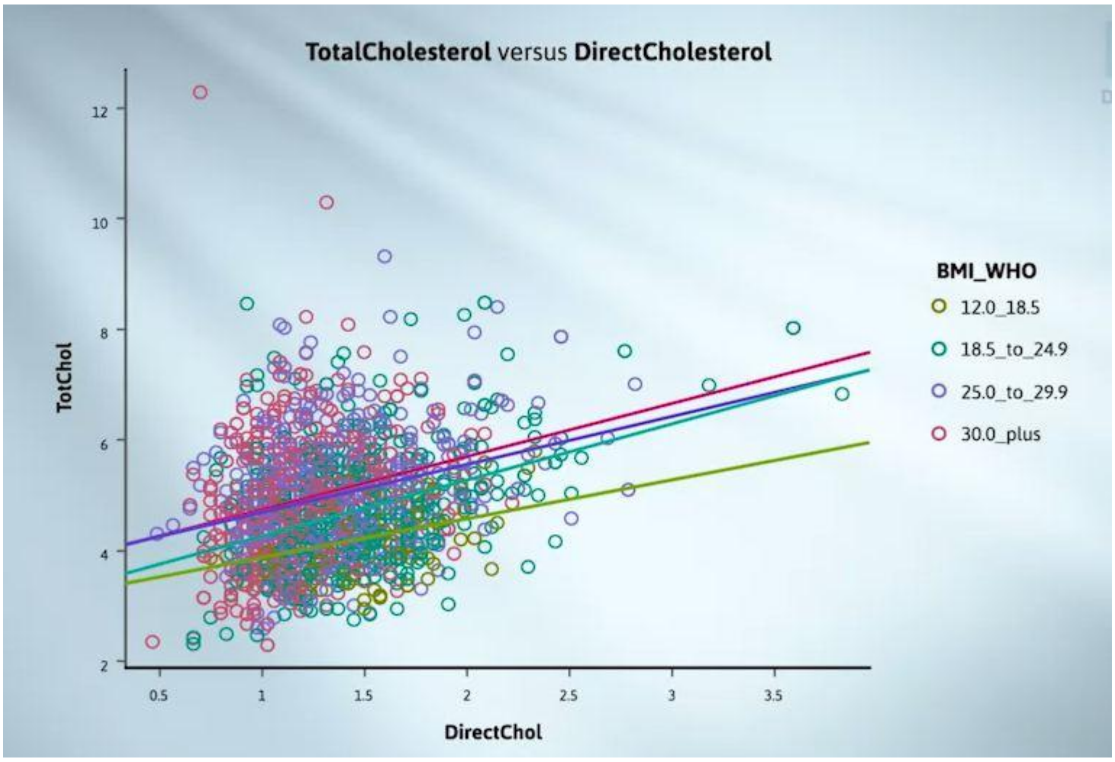


Coloured by gender

Here we've coloured by gender. The females are green, the males pink. In the bottom tail, which we saw was mainly made up of children, the green females and pink males are all mixed up. For heights above 150 centimetres (teenagers and adults) there's a great deal of separation, with green females towards the lower left (smaller weights and heights) and pink males towards the upper right (larger weights and heights).

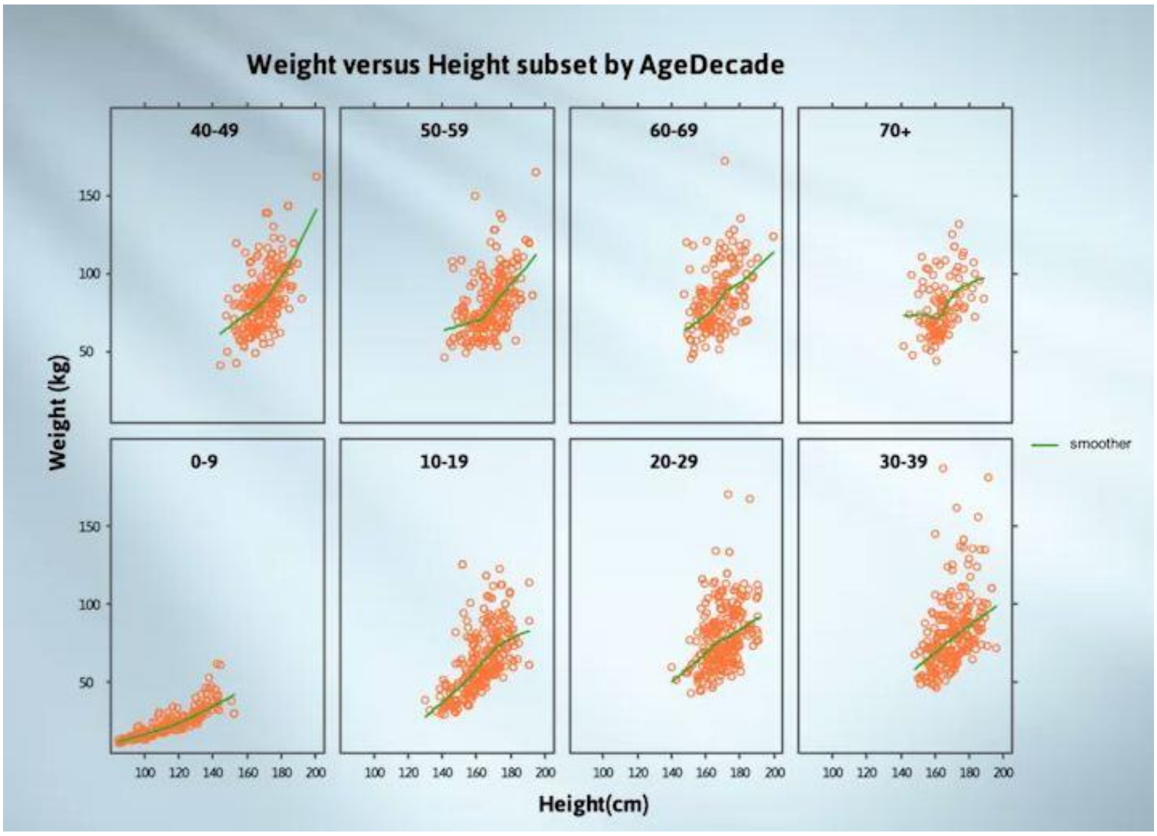
To summarise, we can use colour to see where points from different groups sit in the scatterplot. What we see might lie anywhere on a spectrum ranging from completely separated to totally mixed up. If our colouring variable is a numeric variable or ordered categories, we look for the extent to which there is a colour gradient, reflecting increasing or decreasing values of the colouring variable.

It can also be useful to ask for different trends for each colour group.



This is a graph of total cholesterol versus direct cholesterol, coloured by a World Health Organisation obesity measure. A different trend line has been added for each of these groups.

The trend lines for the higher BMI values are displaced upwards, telling us that for a given direct cholesterol, total cholesterol increases with increasing obesity. There are also small differences in the slopes of the lines.



We'll now move on to subsetting. These are different weight versus height graphs for the people in each decade of life. We should start reading from the 0 to 9-year-olds at the bottom left. As age increases (moving right) we see the point cloud initially moving up and changing in shape until we get to about 20-29. After that, everything stays pretty much the same. We can see this with the set of plots or by playing through the plots like a movie.

I'm not paying attention to the slight differences in where the trend is going because there's a lot of uncertainty about the trend. Thus subsetting gives us an alternative to colour coding as a way of investigating the effect of a third variable. For data exploration, it's not a matter of which is best. We often use both in the hopes of triggering insights.