DATA TO INSIGHT: AN INTRODUCTION TO DATA ANALYSIS

THE UNIVERSITY OF AUCKLAND

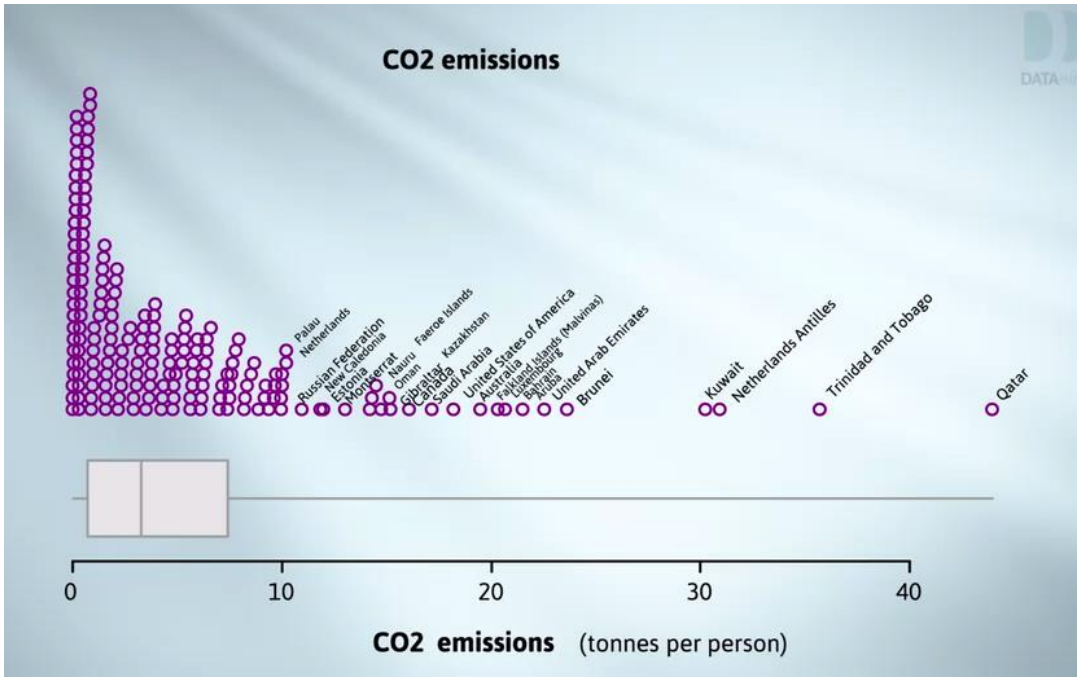
THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

WEEK 3

3.7 RELATIONSHIPS BETWEEN NUMERIC VARIABLES by Chris Wild

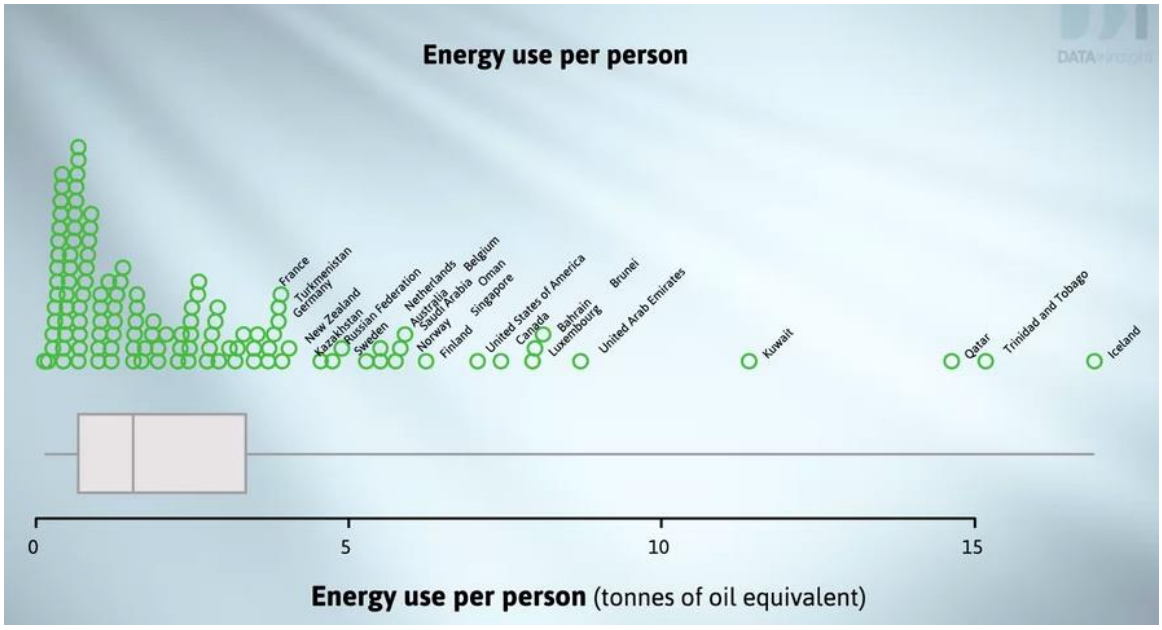
We'll now turn our attention to visualising relationships when both variables are numeric. We're going back to country-level data from Gapminder, and we'll look at 2009 data on some of the variables that figure in climate change discussions. We will look at per-person CO_2 emissions and energy use.

2009 is the most recent year for which this data is reasonably complete.



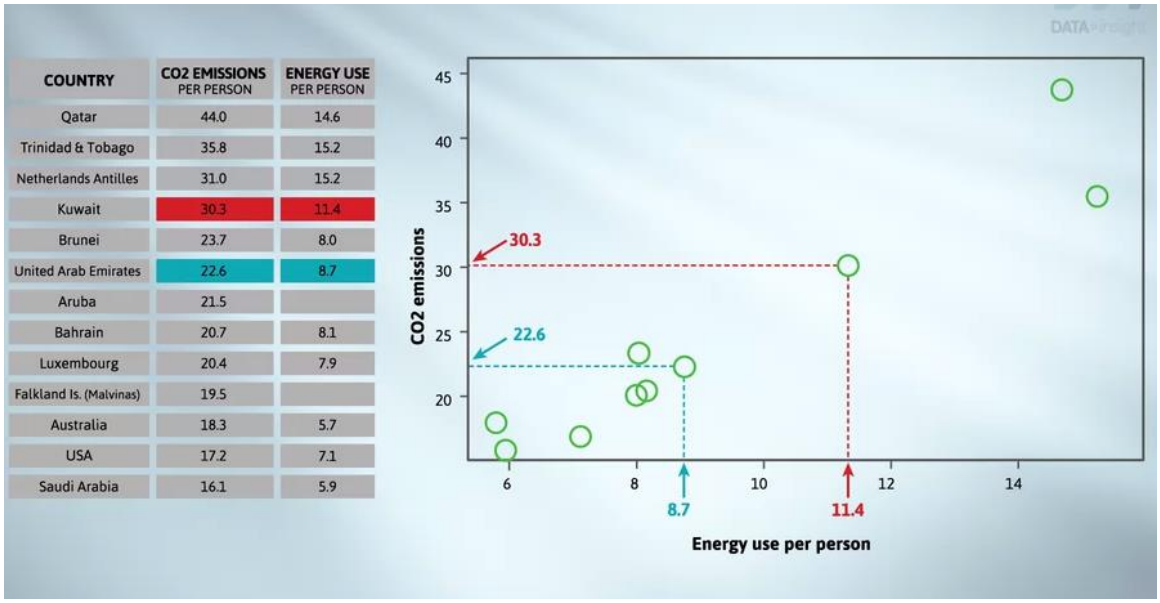
Here is a dot plot of average CO_2 emissions per person. I have labelled the 25 countries with the largest per-person emissions. The results will probably surprise you. Because the media always gives us country-level emissions so that China, which has a huge population, comes to the top of the list.

Our variable comes closer to personal contribution. On these 2009 per-person figures, China only ranked 65th.



This is a dot plot of average energy use per person, again with the top 25 countries identified. Some of the countries at the top are the same. Others are different. So our new question is, "How well does energy use predict CO_2 emissions?"

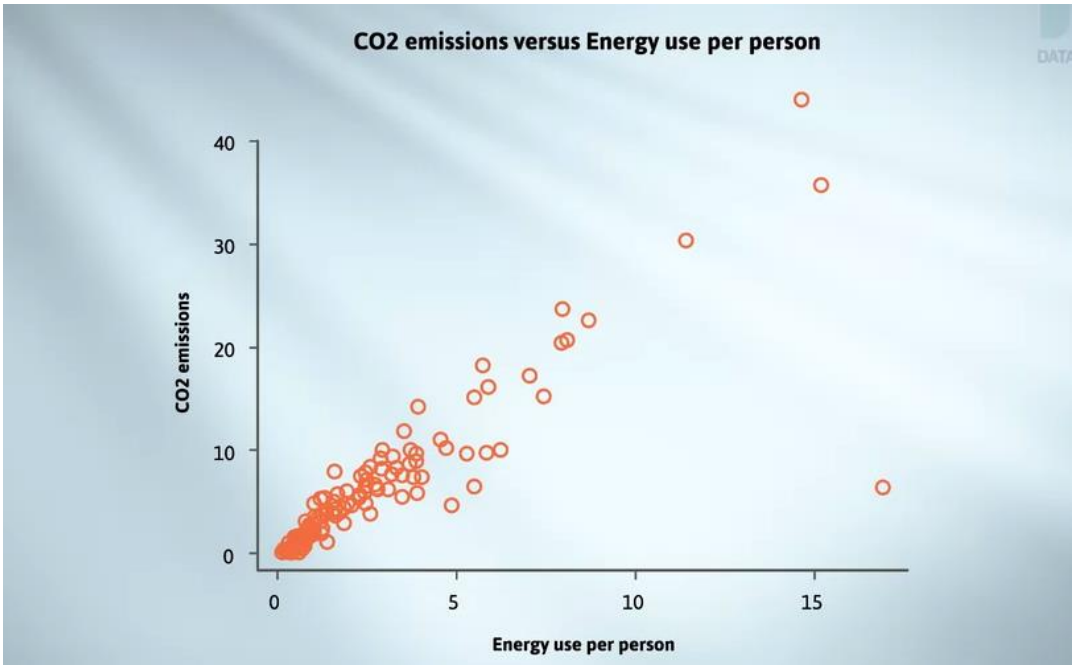
Energy use and CO_2 emissions are both numeric variables. And the standard tool for looking at a relationship between two numeric variables is a scatter plot. Here we have a scatter plot of per-person CO_2 emissions against energy use for the country's hitting our CO_2 emissions list.



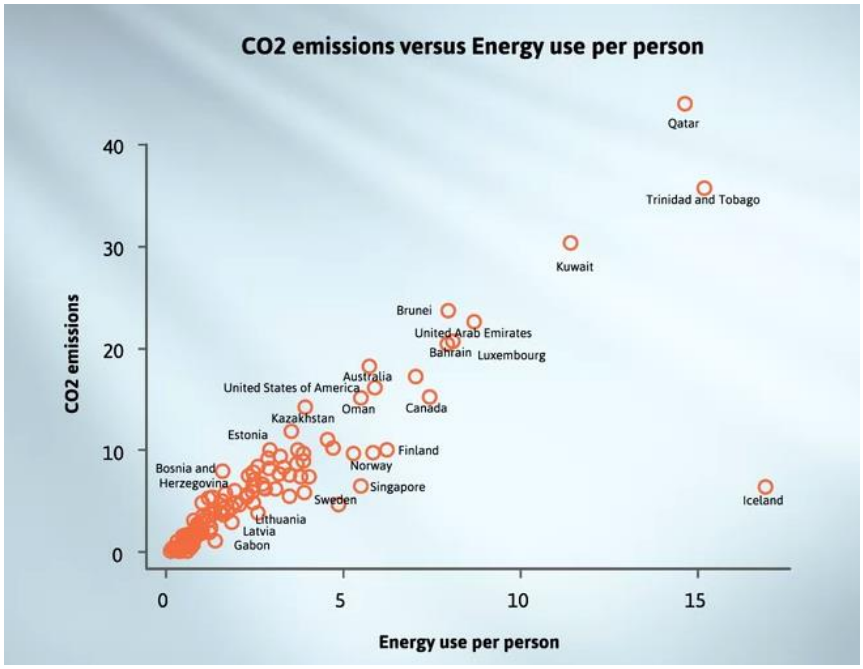
COUNTRY	CO2 EMISSIONS PER PERSON	ENERGY USE PER PERSON
Qatar	44.0	14.6
Trinidad & Tobago	35.8	15.2
Netherlands Antilles	31.0	15.2
Kuwait	30.3	11.4
Brunei	23.7	8.0
United Arab Emirates	22.6	8.7
Aruba	21.5	
Bahrain	20.7	8.1
Luxembourg	20.4	7.9
Falkland Is. (Malvinas)	19.5	
Australia	18.3	5.7
USA	17.2	7.1
Saudi Arabia	16.1	5.9

This plot has a scale for CO_2 emissions running vertically and a scale for energy use running horizontally. We plot each pair of points (same row) against these two scales as shown in the plot using the values for Kuwait in red and the United Arab

Emirates in blue. We use the vertical axis for an outcome variable and the horizontal axis for a predictor variable.
Most people will be familiar with this basic form of plot construction. If you're not, pause the movie and look carefully at how the information for these two countries has been placed on the graph.

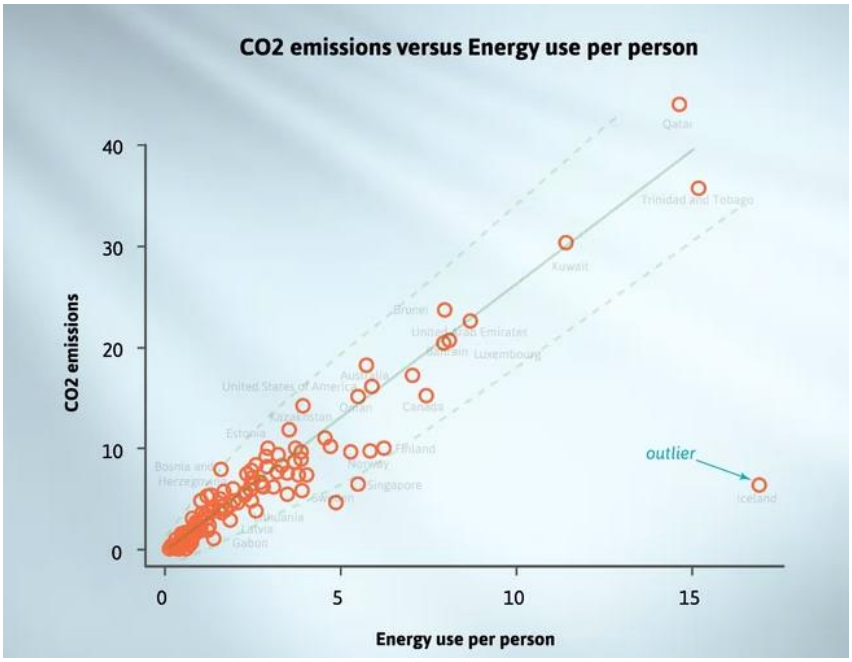


Here's the scatter plot for all of the observations.



This is the same graph, but with the country's that stand out labelled.

What patterns do we see here? The relationship has a clear upward trend, which in this case looks to me like a straight line. Countries with high average energy use tend to have higher CO_2 emissions.

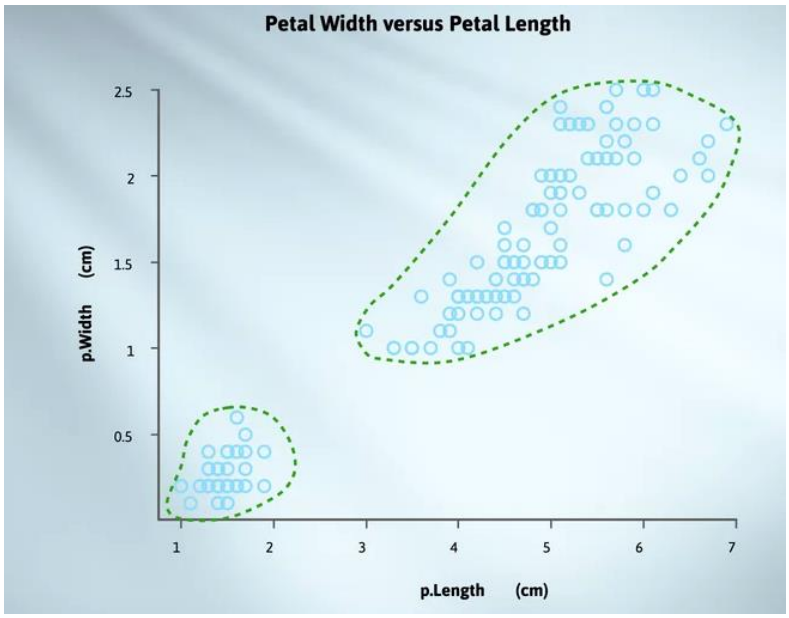


I've put on a green line drawn by eye to reflect what I see. The points are scattered about my trend line. I've put on some green dashed lines, again drawn by eye, to take in the sort of scattering about the trend that I see.

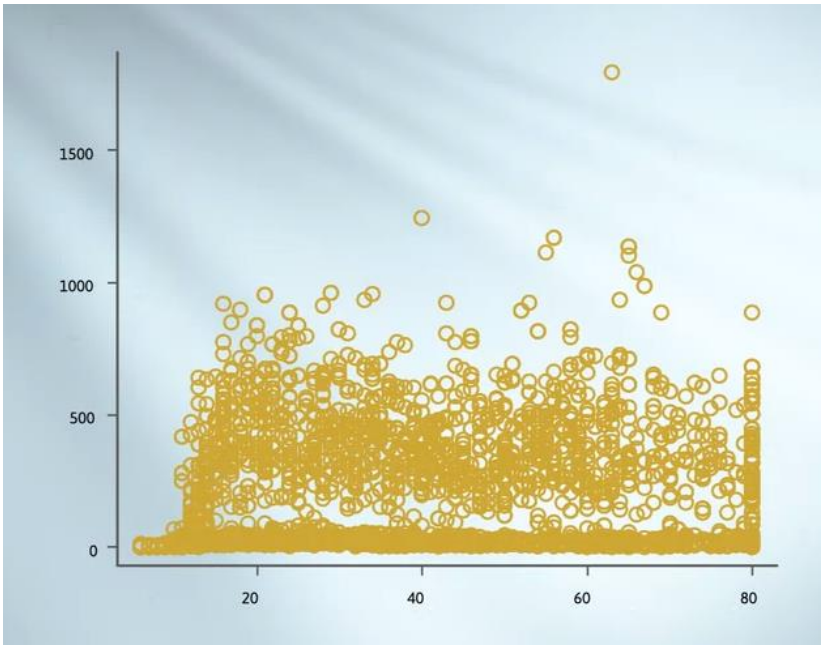
And there is one point that looks very odd, far away from the general pattern-- an outlier-- the sort of observation that is different enough to trigger the thought, "Is that an error, or is there just something very different happening here?" In this case, the outlier is Iceland, and it is very different indeed.

For example, according to Wikipedia, about 65% of total primary energy supply in Iceland is derived from geothermal energy, basically, naturally-produced steam, which can be harnessed directly for things like home heating without needing the sorts of processing that add to CO_2 emissions. Another 20% is hydroelectricity. This is why Iceland has very low CO_2 emissions in comparison to their high energy use.

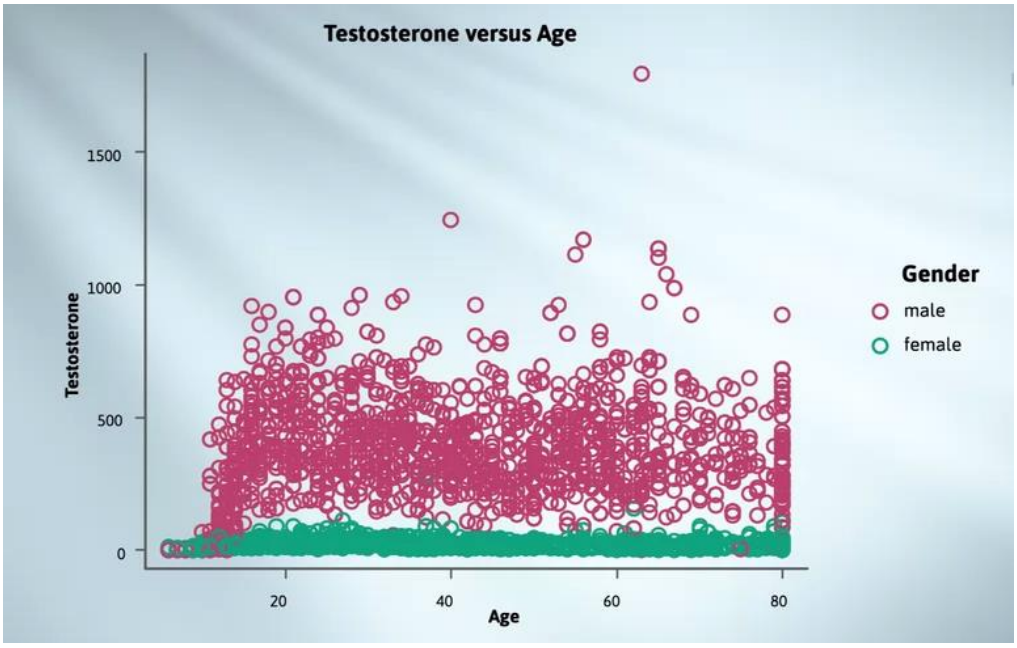
Discounting outlying Iceland, the pattern is strong enough that if we knew a country's per-person energy use, we would have a reasonable idea of its likely CO_2 emissions as we've done here. While most scatter plots can be usefully viewed in terms of trend and scatter with the occasional outlier, there are exceptions such as in the following plots.



This is data on petal widths and petal lengths of iris flowers. It is from a famous data set called Fisher's Iris Data. The most striking feature is the clear separation into two clusters. Clustering points to the existence of distinctly different groups and should set us wondering about what the defining characteristics for those groups might be. Here it turns out that the lower cluster comes from a different species than the upper cluster. The upper cluster is actually a mixture of two species, but they do not separate on these variables.



There's also something rather strange going on here too. It looks like a low dense set of points near the bottom axis with another dispersed set of points above it. When you see the variables, you'll probably guess what is going on.



It is a plot of testosterone levels versus age. And the two groups? You've got it. It's the males and females. Early in life, there is little difference in testosterone levels. And then puberty hits. And for the boys, well, you can see the result. I'll leave you with these questions to remind you of the ideas we've just covered.

QUESTIONS

What is the standard way of displaying the relationship between 2 numeric variables?

What sort of variable is plotted against the vertical scale what against the horizontal scale?

It is often useful to think of the patterns in such plots in terms of 3 components. What are they?

What type of question should separate clusters of dots suggest to us?