FT PUBLISHING

HOW CHARTS WORK

UNDERSTAND AND EXPLAIN DATA WITH CONFIDENCE

ALAN SMITH

THE FT CHART DOCTOR

Foreword by Tim Harford

Learning point – Correlation and causation

Before we go any further, it's time to flag one of the perennial banana skins in the world of statistics.

A correlation might be *causal* (cold weather *causes* higher heating bills), or the link might be associated with an unseen third variable (sales of ice cream and violent crimes are correlated – but only because both are associated with warmer weather). It might also be that a correlation is completely spurious . . .

At www.tylervigen.com[5], we can see that per capita cheese consumption is strongly correlated with the number of people who died by becoming tangled in their bedsheets, while the marriage rate of Wyoming correlates with the number of domestically produced passenger cars sold in the USA.

Statistics are not necessarily a good determinant of underlying causes, but they can help you spot patterns – just make sure they're helpful ones.

Visualising *correlations* is important because it helps us see the extent to which things are connected.

As with the chart relationships we've already looked at – *change over time* (line chart) and *magnitude* (bar chart) – there is a single chart type that dominates our visual thinking in this relationship. Welcome to the scatterplot.

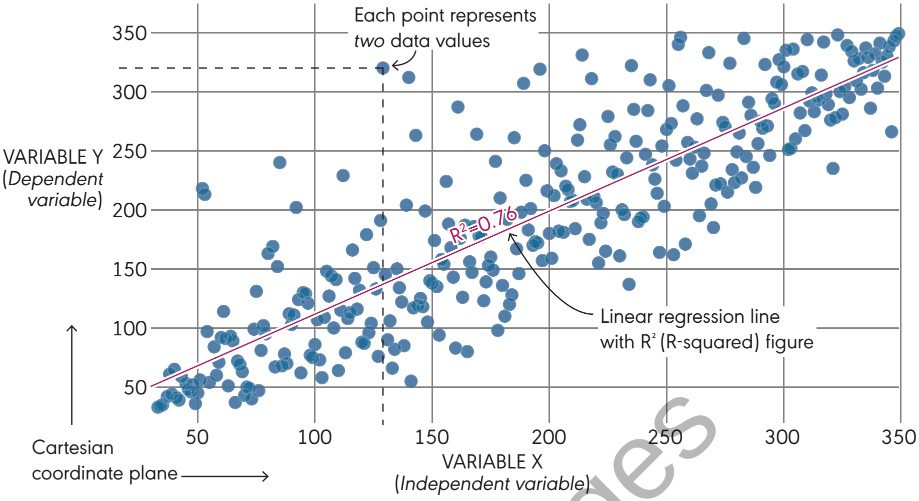
Scatterplot

Data visualisation historians Michael Friendly and Daniel Denis have credited the astronomer John Herschel with the publication of the very first scatterplot in 1833, but it would be a further 50 years before Francis Galton's use of it helped make them a staple of the scientific community.

[5] https://www.tylervigen.com/spurious-correlations.

Anatomy of a scatterplot

Variable X vs Variable Y



The bedrock of a scatterplot is a two-dimensional Cartesian coordinate plane, whose two perpendicular axes (x and y) mean each data point represents two values corresponding to its position relative to each axis.

Conventionally, the horizontal (x) axis is used for the so-called “independent” variable, while the vertical (y) axis is used for the “dependent” variable. These terms are linked to the contentious notion of cause and effect: you can think of the independent variable as “cause” and the dependent variable as “effect”.

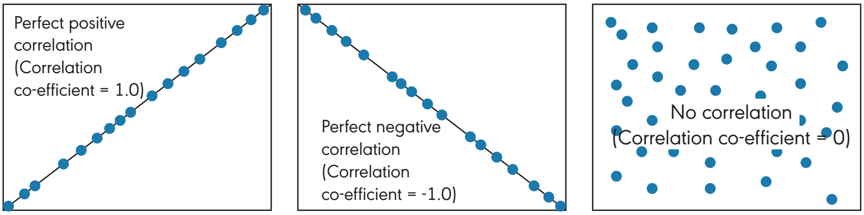
For example, in a plot of data studying hair loss in men, we would put age on the x axis (independent) and the extent of hair loss on the y axis (dependent).

In practice, many scatterplots don’t portray a “causal” relationship, but it’s useful to be aware of the convention.

Finally, when you see scatterplots in the wild (particularly in academic papers), you’ll often see them overlaid with “regression lines”, which are intended to summarise the trend between the two variables. This will be accompanied by a correlation co-efficient value, which describes the strength of that trend. The possible values run between –1 and 1 as follows:

Interpreting scatterplot patterns

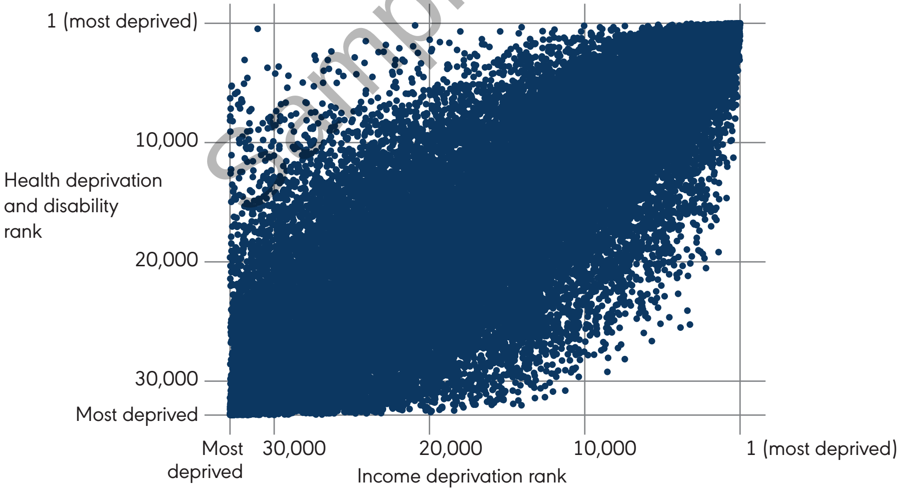
Variable X vs Variable Y



One issue to be aware of with scatterplots is the potential for a simply overwhelming number of dots. For example, take this chart, which correlates income and health deprivation in England. There are 32,844 dots on this chart (don't worry, I won't ask you to count them), each representing data for a small neighbourhood.

Scatterplots: sometimes there are just too many dots...

Income deprivation and health deprivation, England, 2019, by neighbourhood*



* Lower Layer Super Area (LSOA) with a population between 1,000 and 3,000

Source: Ministry of Housing, Local Communities & Local Government

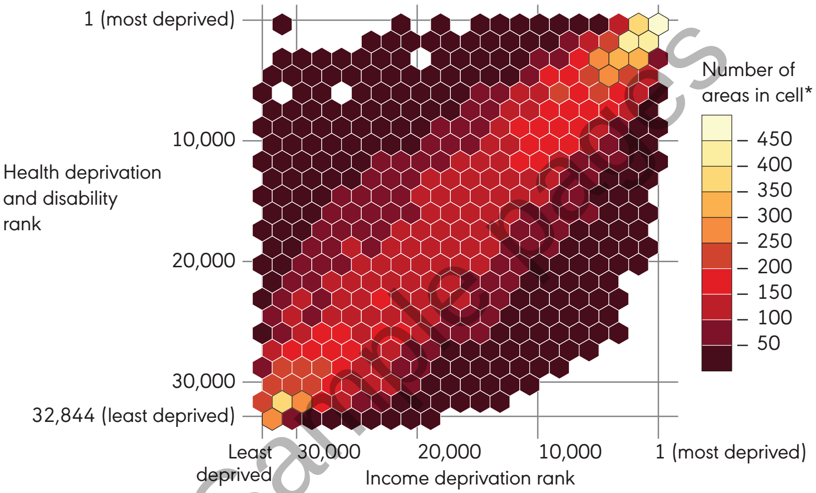
Adding a regression line helps us see the strength and direction of the underlying relationship – but it also reinforces how visually impenetrable the rest of the chart is.

XY Heatmap

An alternative approach is to tesselate the plot area and use colour to represent the number of dots that fall in each "cell". The resulting chart is known as an *XY Heatmap*. In this example, I've used hexagons because they tesselate well, but rectangles can also be used. Notice that the visual pattern depicted by the plot itself means we don't really need the regression line to sense the strength and direction of the relationship (but you could add it, if you wished to).

There is a clear link between income and health deprivation

Income deprivation and health deprivation, England, 2019



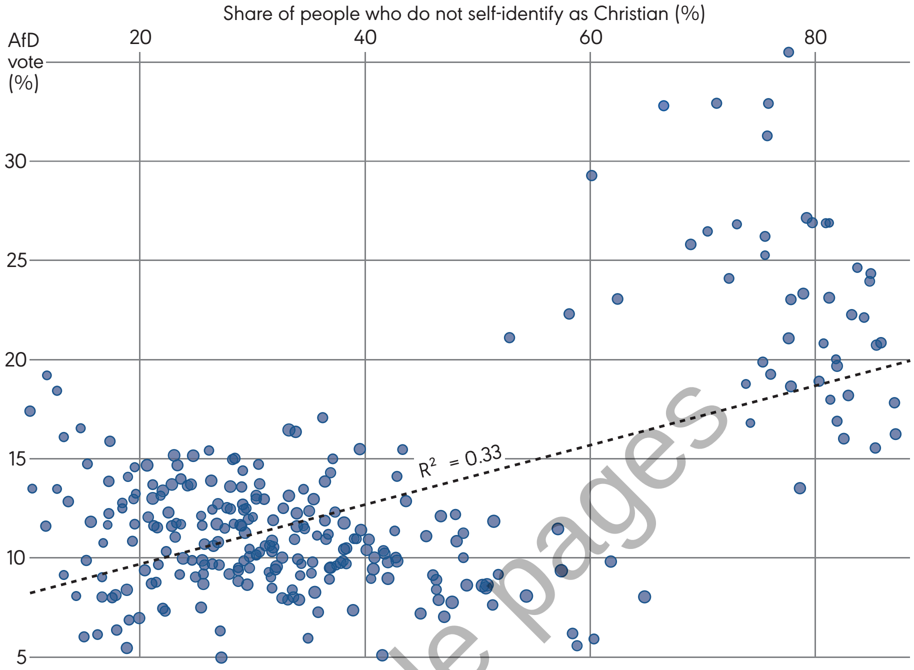
What's the drawback to this approach? Well, we are no longer plotting all the data (our 32,844 points) but aggregated summaries of it. This might mean we miss seeing interesting things at the individual level.

When positive is negative

In fact, even seeing all the data in a scatterplot sometimes isn't enough for us to avoid being misled. The following chart plots the constituency vote share of the Alternative für Deutschland party in the 2017 German federal election with the share of people in each constituency who are non-Christian.

AfD votes: a positive correlation with less Christian constituencies

Each point represents a German constituency



Source: Federal Returning Officer Graphic: John Burn-Murdoch / @jburnmurdoch appeared in Germany's election and the trouble with correlation. Available at: https://www.ft.com/content/94e3acec-a767-11e7-ab55-27219df83c97

It seems an open and shut case: a positive correlation with an R^2 value of 0.33. It suggests that the smaller proportion of Christians there are in an area, the higher the likely vote for AfD (remember our dependent and non-dependent variables).

But let's look at the same data, this time with each point coloured according to whether the constituency is in the east or the west of the country. Suddenly, the trends on the chart look very different – in fact, they are reversed!

This chart from my *FT* colleague John Burn-Murdoch, is an example of a paradox named after Edward Simpson, the Bletchley Park codebreaker who first fully described it. The slightly alarming top line of the paradox is that trends that appear when different groups of data are plotted can reverse when those groups are combined.