# Design principles for data visualisation

## Transcript

Full resource: <https://www.ncrm.ac.uk/resources/online/all/?id=20844>

Hello and welcome to the principles of data visualisation with me, Sophie Lee. Data visualisation is one of the most important parts of any analysis journey. This is because it has many different purposes. For example, we can use it to explore our data so it can be important to identify potential outliers in the data, which may also be potential errors. They can be used to check assumptions if that is required for certain tests or models. A common assumption that’s required for tests or models of numeric variables is that the data follows a normal distribution. That looks like this, with a single peak in the centre and then two symmetric tails going either side. A histogram can allow us to check whether this assumption is valid or not. For example, here we can quite clearly see that the variable does not follow a normal distribution and therefore any tests that require that assumption to be true would not be appropriate.

 They allow us to generate hypotheses about the data, so understand potential trends or relationships between variables, for example temporal trends. Here we’re looking at the number of violent crimes over a ten-year period and we can see quite clearly there’s been a steep decline in that number. And finally, they allow us to convey important messages in a clear, concise manner. Often displaying our messages using visualisations can be a much more powerful way of getting our message across than simply writing it into the text.

 But what makes a good data visualisation? Well, data visualisation is part art and part science. Edward Tufte described graphical excellence as a matter of substance, statistics and design, but finding the balance of these elements can be quite a challenge, that’s why I’ve developed four key principles to creating compelling, informative data visualisations. First, we must show the data. Second, we must choose an appropriate design in which to show this data with. Number three, ensure data integrity. And number four, ensure the visualisation is accessible and inclusive.

 First, we must make sure that we show the data. We want to show as much information as possible, maximising the amount of information given in the smallest possible space, but don’t get me wrong, that doesn’t mean show all of the data. In fact, overloading a plot can actually lead to confusion. Here we have four variables all shown on the same graph and it is hard for the reader to kind of understand what the important message is that they’re supposed to be getting from this plot. So, actually adding too much data can make it less informative than a simpler plot that shows this data much more clearly. Multiple simpler graphs where the message is clear and the reader can see exactly what we’re trying to convey with the plot is much better than a single confusing graph that hides that message we wish to impart. We have to have a clear goal in mind that we want our readers to understand and gain from our visualisations and then choose the most appropriate data that will allow us to do this.

 That brings us on to the second principle, to have a good data visualisation, we must choose an appropriate design. The choice of visualisation should first and foremost be driven by the context, the audience and the goal of the visualisation. Why are we trying to do this and what is the message we’re trying to convey. The decision is also driven by the type of data that we have, so the number and the types of variables that we have available to us. If you wanted to have a list of visualisation choices based on the type of variables you have, I’d really recommend going from data-to-viz website, it’s a really good resource for this.

 We must remove any unnecessary clutter or design choices that do not enhance the intended message of the visualisation. This is also known as chart junk. This includes unnecessary colours that are not related to the data or the message that we’re trying to convey, any overbearing gridlines which are too bold and distract from the data, or any distracting patterns that draw the audience’s eyes away from the data and the message that we’re trying to convey with the visualisation. However, do not allow the pursuit of a minimalist design detract from the interpretation of our graph. For example, if we remove all gridlines on a scatter plot, this actually makes the points quite hard to interpret because they’re just simply floating in space, so good visualisation requires a kind of middle ground. We want to ensure that the data is the most important part of the visualisation but that we also give sufficient context to the data as well.

 Next, we must ensure integrity of our visualisations. Data visualisations must not distort the data in any way. They mustn’t mislead the reader or tell anything other than the truth. To achieve this, we must ensure that the visual representation of our data is consistent with the numeric representation of it. One of the main obstacles to doing this is that people actually perceive things differently depending on their experiences and on the context of this visualisation. It’s our job to be mindful of where these interpretations can be interfered with due to perceptions and, to account for this, make sure that we’re making allowance for it in our visualisations. For example, some research has shown that people’s perceptions of dots, lines and bars are far more accurate that distinguishing between angles, proportions and colours hues. This does not mean that we can’t use things like angles and colour to determine differences, but it just means we might need to make things clearer with an annotation or something else to provide more context to our message.

 A really good example of where these differences in perceptions can be a challenge is the pie chart. Here we have the number of recorded crimes in 2023 in the East Midlands in England, and this is separated by different police areas. The proportion of crimes which were recorded in Derbyshire, Leicestershire and Nottinghamshire were all very similar. Due to the way that humans perceive proportions of a whole, it’s really hard to identify these differences and make inferences about where the highest or the lowest levels of crimes were in this region.

 This bar chart shows exactly the same information just in a different way, which is easier for the human brain to perceive. Here we have ordered the bars in length as well, which means that we can distinguish between the highest levels of crimes and the lowest and see those small differences that were hidden previously by a bar chart.

 Another way in which we want to ensure data integrity is to follow the principle of proportional ink. That means that the ink used for a certain data or point in the data is proportional to the quantities that it represents. So, a common violation of this is where axes do not begin at zero, like we can see here. This is the same information as the previous slide, but because we’ve changed the axes, it now exaggerates the differences in the amount of recorded crimes between the areas with the lowest and the highest. It makes the amount of crime in Lincolnshire and Northamptonshire look far smaller than they actually were, and it looks like the differences between those and, say, Nottinghamshire was much larger than it actually was.

 Finally, we must ensure that visualisations are accessible. Visualisations are useless if they’re inaccessible and uninterpretable, so we have to make sure that all our design choices are inclusive. Some considerations we can make to ensure this include ensuring all text is legible. That means that the text is large enough to be read and choosing a font size that is accessible to everyone, including those who have visual impairment or learning difficulties. Guidelines suggest that text should be at least 12 points in size when the plots are printed in a report, for example, and 36 when they’re used in presentations to allow them to be seen from further away.

 We also need to make sure that the font family is inclusive as well. Make sure it’s accessible for different impairments. The are specialist fonts that have been developed to aid particular impairments, but Arial and Verdana are often given as free fonts that are cited as inclusive and accessible.

 We also want to ensure that we’re using inclusive colours, so colour palettes must be chosen depending on the type of data and each colour in that palette should be distinct to everyone, including those with colour vision deficiency. One way to check this is to use a colour blindness simulator to ensure that they are still distinct even with these different deficiencies. So, this is an example of where a colour palette at first might appear to be quite distinct but if people have a certain type of colour blindness they now can’t distinguish between the high and the low values.

 We want to avoid colour palettes that are cyclical, this means where the top and the bottom kind of merge, so things such as the rainbow colour palette, although it looks quite easy to distinguish here, if we have a plot with very high and very low values, they can be quite hard to distinguish between because they’re quite close together.

 And finally, we want to avoid any colour schemes that include potentially harmful stereotypes, so a common example of this is using pink for female and blue for male. There are lots of guidelines online of alternatives that we can use where there is potential for stereotypes and you should check those resources out if that’s something that you’re displaying.

 And finally, as with all science, visualisations should be reproducible, so where it’s possible we want to make sure all data and code that we use to create our visualisations are accessible and open source. Obviously, there are limitations to this with certain types of data, but where that isn’t the case we want to consider hosting the data and the code in a free online repository such as Github.

 Thank you for listening and I hope that you employ these principles of data visualisation to ensure all your visualisations are inclusive, accessible and compelling.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK