Data visualisation: Figure design, design process, and fundamentals

By Radovan Bast, UiT The Arctic University of Norway
ORCID: 0000-0002-7658-1847
Date: October 2020

Manuscript:

In this video I will talk about figure design, design process, and fundamentals. Let's start by defining data visualization (DV): it is a visual representation and presentation of data to facilitate understanding (from Wilke's book). DV maps data values onto aesthetics or channels. Examples of channels: position, length, shape, size, colour, line width, line type.

Everybody will have their own approach, this is how I do it. I often start by sketching with a pen and paper. Then I browse directories and galleries of existing libraries for inspiration and look for an example that is close to what I want. I first try to reproduce the example with the example data, then I try to run it with my own data. Later I refine.

Here is an example from the gapminder dataset where I print the life expectancy as a function of the GDP. The dots represent different countries. I started with a rough plot (top left), then made the dots a bit transparent and added axis titles, then switched to a log axis (bottom left), then increased the font size and arrived at an OK looking plot where we could add a smoothing line to emphasise the trend that the countries with a higher gross domestic product tend to have longer life expectancy.

Three design principles I took from Andy Kirk's book: good DV is trustworthy, accessible, and elegant. I will show some examples of reliability and usability. A good recipe for elegant plots is to maximise the data/ink ratio, within reason.

We should obey the principle of proportional ink. The length of bars or the area under a curve should be proportional to the values that are plotted. Here I show two examples with a disproportional data/ink ratio. Note how the bars do not start at 0. Also the plot on the right does not start at zero and the representation is skewed.

Also avoid 3D plots like this example. Very hard to interpret the data. A 3D plot can make sense when plotting something inherently 3D (molecules, enzymes, relief of a terrain).

Coordinates: sometimes we need to choose between linear and log coordinates. In this example a log scale is more useful because the dataset contains numbers of very different magnitudes. There exist also polar and spherical coordinates and their projections (more later).

Finally, I would like to talk about one important output channel: the colour.

Colours can be used to distinguish groups of data, represent data values, and to highlight. This is how many of us start when we need colours for a plot: black, red, green, ... but this can be problematic.

The reason for this are colour vision deficiencies. 4% of the population is affected and cannot distinguish certain sets of colours and I highly recommend you check your colour figure under CVD

simulations. Also you don't need to invent your own colour scales, but use colour scales designed to be CVD-friendly.

There are at least 3 types of colour scales: Discrete (designed to distinguish), continuous (designed to represent data values), and diverging colour scales when we need to visualise deviation relative to a neutral point.

Discrete colour scales are great for scatter plots. This is a scale by Okabe and Ito designed for colour vision deficiencies and contains 8 colours. What if you need more than 8? Use direct labelling or another plot type.

Continuous colour scales are great for choropleth plots such as the one on the right. CVD is less of a concern for this type of plot.

Finally, here is an example of a diverging colour scale used in a heat map. This is the ColorBrewer pink to yellow-green scale which is suitable for CVD. Take home message: use existing colour scales designed for CVD and for the DV type at hand.