Data visualisation: Gallery of visualisations

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

Manuscript:

In this video we will browse through a gallery of visualisations. I will show typical examples for visualising x-y relationships, amounts, distributions, proportions, geospatial data, trends, and uncertainty. And I will show the main pitfalls for each category. This will not be a complete list, obviously. The goal of doing this is that we get used to the vocabulary and know what to search for when creating our own plots and looking for inspiration.

There are excellent overviews and categorisations and in the slides accompanying this course you find the corresponding links. I will mainly follow the categorisation by Wilke (first reference).

Let's start with perhaps the most common type: x-y relationships. On the left side I show a line graph (different colours, different dash patterns). On the right side we see a scatterplot where lines have been added to represent the trend. We can also see confidence intervals. More about these later.

Bubble plots are popular as well. In this case there are four dimensions encoded as x and y values, size of the bubble, and colour of the bubble. A pitfall of scatterplots can be overplotting if there are too many dots or bubbles and then we can resort to bins or contours. If we need too many colours this can call for a different plot type.

And on the next slide we see an example of a (in this case hexagonal) bin plot. Instead of hexagons we could use squares. And on the right side we see a contour plot.

After x-y relationships, we turn to amounts and I will show a few examples. A bar plot on the left. We remember the principle of proportional ink from another video: bars need to start at zero. Instead of bars we could have used dots. On the right side we group bars. We can do that horizontally or vertically. If there are too many colours in one group it can be a bit hard to decipher, and what we can do then is “faceting”, by plotting individual colours in separate bar charts arranged like tiles.

Bars can be stacked. In this case we arranged them horizontally, which was better for the longer labels. We should try to arrange them in a logical and intuitive order. Stacking can be useful when the sum of amounts represents something meaningful. But note that the inner colours can be difficult to compare across rows. So this chart type works well if there are only two bars in each stack.

The heat map is a popular way of visualising amounts depending on 2 variables. And it is related to the hexagonal and square bins we have seen earlier.

Another category of visualisations is distributions. Here we show histograms, faceted on the left (in other words arranged in tiles) and layered on the right. When plotting histograms, always explore multiple bin widths to check which best represents the data.

Histograms can be stacked (as seen on the left), but this can be difficult to interpret. Much better, on the right side, is the same data represented using a density plot. For density plots, check scaling and boundary conditions. Also remember that both histograms and density plots often require arbitrary parameter choices for binning or convolution. And we should report these parameters together with our plots.

An interesting solution to show several distributions in one plot is a ridgeline plot. In this case we are plotting temperature distributions for the 12 months in Seattle.

Distributions can also be represented by box-plots (showing minimum, maximum, median, and lower and upper quartiles) as well as violin plots. Observe, however, how box-plots can miss representing a changing structure in the data.

Proportions: the classic pie chart on the left, not very space efficient in terms of data to ink ratio. On the right side a stacked bar chart.

Proportions can also be represented by stacked areas or densities (on the left), or a tree map (on the right).

Geospatial data requires extra care as it requires a choice because a projection from a 3D globe onto 2D introduces distortion – we can preserve either angles or areas but not both. On this slide I show a couple of such projections. But many more exist.

Choropleth maps like this one work best when colouring represents a density. They can be problematic if the colours represent a quantity which is not a density. We are used to seeing choropleth maps, especially during election times. Note that in this projection Alaska is too small compared to the mainland.

Finally, trends and uncertainty. We often represent trends by smoothing lines. There are a number of smoothing models to visualise trends: averaging, polynomial fits, splines, LOESS (locally estimated scatterplot smoothing). These are very frequently used, and many visualisation tools can compute these automatically.

Uncertainty can also be visualised with the help of confidence bands like in this example.

We have mentioned box-plots earlier, and also error bars are common in plots. Error bars can extend horizontally or vertically or both. They can also be graded.

It is important to point out that there is no commonly accepted standard for error bars. *Always* indicate what these mean: is this a standard deviation? standard error of the mean? 95% confidence interval? 95% credible interval? Always specify which definition you are using. And once again, box-plots can be problematic. They were invented when figures were still drawn by hand. Today we have computers and can often do better. There are many more ways to visualise data and I encourage you to explore existing galleries which are also linked from the lecture slides.