Informatics 2 – Foundations of Data Science: Visualisation principles and guidance

Principle 1: Show the data

Aim to show as much of the data as possible without leading to a confusing visualisation. There are often multiple ways of representing the same dataset, and no "right" answer. The following guidance should help show as much of the data as possible:

- **Choose an appropriate plot type.** Basic types include:
  - Bar charts: for plotting numeric variables associated with categorical or ordinal variables, e.g. the mean weight (numeric variable) of male and female (categorical variable) squirrels.
  - Line charts: for showing trends of numerical variables over time (a numerical variable).
  - Scatter plots: show the relationship between two numeric variables.
  - Box plots: represent the distribution of a numeric variable for multiple categories, e.g. the weights of male and female squirrels.
  - Histograms and density plots: good for showing the distribution of a single variable.
- **Show multiple variables by using length, shape, size and colour:**
  - Use shape and colour to create extra dimensions for categorical variables. E.g. in a scatter plot of squirrel weight versus length, indicate sex using colour, thus displaying 3 variables. In addition, indicate age categories (ordinal) by changing the size or the shape of the markers (4 variables). But take care that the plot is not too complex to read.
  - Bar charts can be extended to two categorical variables and one numerical variable by using colour.
- **Use colour effectively.** (Wexler et al., 2017, pp. 14–18)
  - Choose an appropriate colour scale, depending on if the data is sequential (numeric), diverging (numeric with a zero point in the the scale) or categorical.
  - Colour can also be used to highlight features in the plot, e.g. the largest two bars in a bar plot.
- **Encourage the eye to compare several pieces of data**, e.g. by using multiple plots with the same scale.
  - Wexler et al. (2017), p. 31, is a nice example of how this can work better than using multiple symbols on a plot (p. 30).
- **Present many numbers in a small space**
  - A box plot takes up as much space as a barplot, but conveys more information. For example, a box plot of the squirrel's weight versus sex shows information about the distribution of the weight as well as the median weight.
- **Choose appropriate transforms**
  - Transforming data can make features of it clearer. For example, plotting the value of Bitcoin over time shows very little detail about the early history of the currency, when it was not valuable. However, plotting the log of the value of Bitcoin on the y-axis allows this detail to be seen.

Principle 2: Make the meaning of the data clear

A visualisation is meaningless if it's not labelled. Every plot should have:

- Title or caption
- Axis labels as English words
- Units given, where appropriate (e.g. "Length (mm)" *not just* "Length")
- All variables labelled – e.g. a legend indicating the colours used to represent squirrel sex
- Graphical and textual annotation where appropriate – e.g. it can be helpful to highlight a time series with events that you know about

Principle 3: Avoid distorting what the data have to say

Choices in visualisation design can lead to the instant impression given by preattentive processing of the visualisation being quite different to the numbers in the dataset. Tufte (1982) measures the level of distortion in a visualisation by the "Lie factor":

$$\text{Lie factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

The following guidelines help to avoid distorting the data:

- **Use appropriate scales and baselines**
  - A very common problem is that the baseline (i.e. the lowest point on the $y-$axis) in a bar chart is not zero. This can lead to small differences appearing large.
- **Be aware of limitations of our perception of size**
  - Although marker area can be useful for indicating categories, humans are not very good at relating the area to a quantity – we are much better at comparing lengths.

Principle 4: Make the data accessible

A visualisation is meaningless if it's illegible and loses impact if it's difficult to read. To ensure data is accessible:

- **Make sure text is legible**, i.e. font size of minimum 8 points in a PDF, or about 20 points in a presentation. (It is surprising how often talks are given in which it's impossible to read the labels on plots even from the front row.)
- **Use colours that work for people with colour-vision deficiency.** Wexler et al. (2017), Chapter 1 has an excellent introduction to using colour in visualisations.

Principle 5: Focus on the content

Give the viewer's brain as little work to do as possible.

- Avoid chartjunk – e.g. colours that don't have any meaning.
- Reduce clutter
- Use consistent colours for plots in the same study
- Use an appropriate number of decimal places
- Check spelling is correct

References

Tufte, E. (1982). *The visual display of quantitative information*. Graphics Press, Cheshire, Connecticut.

Wexler, S., Shaffer, J., and Cotgreave, A. (2017). *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*. Wiley.