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

Principle 1: Show the data

Show as much of the data as possible without making a confusing visualisation. There are multiple ways of representing the same dataset, and no "right" answer.

- **Identify the variables and their type.** For tabular data in a tidy (long) format, each column corresponds to a variable, and is numeric, categorical, ordinal or unstructured.
- **Choose an appropriate plot type to show one or two variables.** The `https://www.data-to-viz.com/` tool helps. E.g.:
  - One numeric variable → histogram or density plot shows distribution
  - One categorical variable → bar plot shows counts
  - One categorical variable and one numeric variable → bar plot can show mean of numeric variable for each category; box plot or violin plot show distribution.
  - Two unordered numeric variables → scatter plot shows relationship
  - One ordered and one ordered numeric variable (e.g. time series) → line plot
- **Consider showing extra variables by using length, shape, size and colour.**
  - E.g. In a scatter plot (two numeric variables), the colour and shape of each marker can represent two categorical variables, thus displaying four variables. Size can represent ordinal variables.
  - But assess whether the plot is too complex to read.
- **Consider using a table.** data patterns are clearer in tables than graphics. However, tables are a form of visualisation, and good for conveying raw data or dealing with large numbers of variables.
- **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 visualisation, e.g. the largest two bars in a bar plot or the largest values in each column of a table.
- **Encourage the eye to compare several pieces of data**,
  - E.g. use multiple plots with the same scale ("small multiples"), which can work better than using large numbers of symbols or colours on a single plot
- **Present many numbers in a small space**
  - E.g. A box plot of a numeric variable uses as much space as a bar plot, but conveys more information.
- **Choose appropriate transforms**
  - E.g. for a positive variable that varies over many orders of magnitude (e.g. the value of Bitcoin) a log transform can show changes when the variable is both small and large.

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 on axes or legends
- Graphical and textual annotation where appropriate – e.g. it can be helpful to highlight a time series with events that you know about
- Description of what any error bars represent, e.g. 95% confidence interval, standard deviation, or standard error

Principle 3: Avoid distorting what the data have to say

Design choices can mean the instant impression given by preattentive processing of the visualisation is quite different to the numbers in the dataset.

- **Use appropriate scales and baselines**
  - A common problem is that the baseline (i.e. the lowest point on the y−axis) in a bar chart is not zero, leading to small differences appearing large.
- **Be aware of limitations of human perception**
  - Humans are better at comparing lengths than areas → consider whether area represents a given numeric variable well
  - Humans are better at comparing lengths than angles → consider alternatives to pie charts, especially with many categories

Principle 4: Make the data accessible

Make visualisations accessible so that they are meaningful for everyone.

- **Make sure text is legible**, i.e. font size of minimum 8 points in a PDF, or about 20 points in a presentation. (Surprisingly often in talks, it's impossible to read plot labels, 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

Minimise distractions for the viewer's brain.

- Avoid chartjunk – e.g. colours that don't have any meaning, or 3D bar charts
- Reduce clutter – e.g. vertical grid lines with a categorical x-axis
- Use consistent colours for plots in the same study
- Use an appropriate number of decimal places
- Check spelling is correct

References

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