Principles of effective visualisation

Sam Clifford

2022-01-27

Introduction

Here we provide a little more detail on the principles (Tufte 1983, 1997) of creating effective visualisations.

For further guidance on the use of ggplot2 in R, there are online versions of Chang (2018) and Wickham and Grolemund (2020) (specifically chapter 3) as well as the RStudio cheatsheets (RStudio 2012). The BBC has used ggplot2 to develop an in-house style of data visualisation for their data journalism (BBC data team 2019).

Graphical integrity

Visual representations of data must tell the truth

1. The visual representation of numbers, as physically measured on the graph, should be directly proportional to the numerical quantities represented

   - Lie Factor = the size of the effect shown in the graphic divided by the size of the effect in the data.
   - Values over 1 overstate the effect & under 1 understate.
   - Don't make points bigger to emphasise them as important

2. Clear, detailed and thorough labelling helps avoid graphical distortion and ambiguity. Write explanations of data on the graph. Label important events from data.

3. Show variation through data, not through design.

4. Ensure appropriate standardization & comparisons are used, e.g. CPI-adjusted or seasonally-adjusted.

5. The number of information carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.

   - In practice, this means not mapping the same variable to two graphical elements (e.g. a gradient on point colour based on value in y axis)

6. Graphics must not quote data out of context.

Maximising `data:ink` ratio

Good graphical representations maximize data-ink and erase as much non-data-ink as possible

1. Above all else, show data
2. Maximise the data-ink ratio
   - ink on a graph that represents data
   - data-ink ratio = 1 minus the proportion of the graph that can be erased without loss of data-information
3. Erase non-data-ink
4. Erase redundant data-ink
5. Revise and edit

For example, the boxplot is a simple five number summary of a continuous variable that draws multiple rectangles, lines and points to show the quantiles of the data and any outliers (Frigge, Hoaglin, and Iglewicz 1989). Tufte considers that the outliers are not important in showing the typical spread of the data and that the box is unnecessary as the quantiles are shown with the 'whiskers' of the plot and the median line. A standard boxplot from ggplot2 is shown in Figure 1 and a version that removes extraneous lines and points in Figure 2.

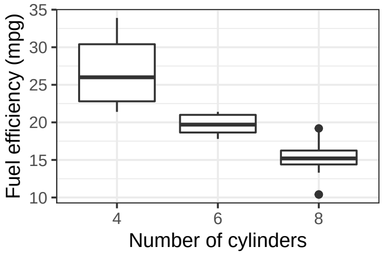


Figure 1: Standard boxplot

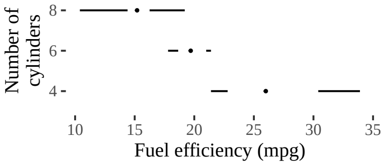


Figure 2: Tufte's interpretation of boxplot that maximises data ink

A lot of the elements of the boxplot can be removed without reducing our ability to distinguish the relevant information.

Avoid Chartjunk

If an element of a graph is to be included, it should be because it aids understanding and reveals information, not because it looks pretty

- Examples of chartjunk include but are not restricted to:
  - hatching (patterns instead of colours)
  - heavy grids
  - equally spaced lines, too close together, as in bar charts, histograms, boxplots
  - self-promoting graphs that demonstrate the graphical ability of the designer rather than displaying the data
  - 3D graphics that distort perspective (e.g. 3D pie charts)

Maximising data density

We must ensure that when we are plotting we don't end up with large amounts of white space when avoidable. Small multiples can be attractive, but leaving empty facets leads to unnecessary amounts of white space that reduce the amount of space in the figure dedicated to showing the data in existing combinations of faceting variables.

1. Maximise data density and the amount of data shown, within reason
2. Apply the shrink principle
   - most graphs can be shrunk down very far without losing legibility or information
   - this is because most graphs are quite sparse
3. Exploit **small multiples** to provide for comparisons across groups
   - series of the same small graph repeated in one visual
   - can compare a main relationship across one or more grouping variables
   - a great way to visualise large quantities of data, or when there are a high number of dimensions

Few's rules on usage of colour

Few (2008) suggests the following rules for effective use of colour in figures and tables

1. If you want different objects of the same color in a table or graph to look the same, make sure that the background (the color that surrounds them) is consistent. (A gradient background just adds confusion).
2. If you want objects in a table or graph to be easily seen, use a background color that contrasts sufficiently with the object.
3. Use color only when needed to serve a particular communication goal.
4. Use different colors only when they correspond to differences of meaning in the data.
5. Use soft, natural colors to display most information, and bright colors and/or dark colors to highlight information that requires greater attention.

6. When using color to encode a sequential range of quantitative values, stick with a single hue (or a small set of closely related hues) and vary intensity from pale colors for low values to increasingly darker and brighter colors for high values.

7. Non-data components of tables and graphs should be displayed just visibly enough to perform their role, but not more so, for excessive salience could cause them to distract attention from the data.

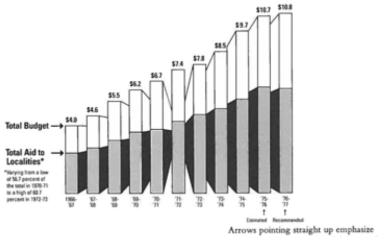
8. To guarantee that most people who are colorblind can distinguish groups of data that are color coded, avoid using a combination of red and green in the same display.

9. Avoid using visual effects in graphs. (A plain bar chart is preferable to one with three-dimensional rods).

Examples of visualisation for critique

The following visualisations are found in Tufte (1983) and we discuss them here with a brief critique each.

Bar plot



Arrows pointing straight up emphasize recent growth. Compare with horizontal arrows at left.

Figure 3: Total aid to localities as a fraction of total budget, p83 Tufte (1983)

In Figure 3, we see an example of a bar plot with additional chart junk. Specifically, we have: a 3D effect; too many additional labels, such as the total budget for each year; and arrows for labelling the estimate for the current year and recommendation for the next year.

By removing the elements in the left hand panel of Figure 4, we retain the key information and are more easily able to make comparisons without being distracted by 3D effects.

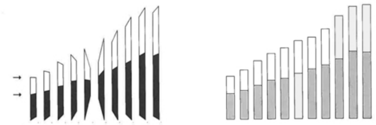


Figure 4: Elements of bar plot

Line plot

In Figure 5 we see thermal conductivity as a function of temperature. Each experiment is a point, and data from the same lab are joined with a line. The curve varies across labs due to impurities in copper, and this grouping enables comparisons across labs *and* against the overall behaviour. The scales are logarithmic on both the x axis (temperature in Kelvin) and y axis (Watts per metre-Kelvin). While this may have been an effective plot when it was created, it is difficult to tell which lab a given data point and curve comes from; today this might be drawn as an interactive plot.

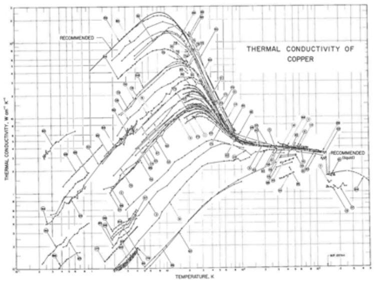


Figure 5: Thermal conductivity of copper across laboratories, p49 Tufte (1983) citing Ho et al. (1974).

Ribbon plot

The coloured ribbon plot shows the age split of college enrolments between 1972-1976, for under-25s and 25 and over. This is called a ribbon plot because the plot is a ribbon, a shaded area, between some minimum and maximum lines.

One issue with this ribbon plot is that the two ribbons sum to 100% but not in an obvious way, as the ribbons are separate rather than stacked and a doubly broken axis is used without showing the zero point or the top of the axis. Four colours are used, to show each age group and to give a visually distracting 3D effect. No colouring is necessary given that only one set of numbers can be used to represent these mutually exclusive and completely exhaustive age groups. That is, this could be a line plot or even a table, e.g.

Year	1972	1973	1974	1975	1976
Enrolments aged <25	72.0%	70.8%	67.2%	66.4%	67.0%

N.B. Guidance on the layout of tables is available in Mori (2007) and Fear (2005).

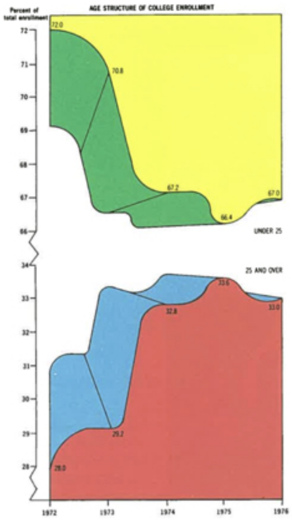


Figure 6: Proportion of college enrolments under or over 25 years old. p43 Tufte (1983)

Artist collaboration

The diagram of the life cycle of the Japanese beetle (Figure 7) is rich with information and visually appealing - a highly impactful graph. The result of a collaboration with an artist, we see the depth of the grub, the life stage of grubs and larvae over time and which foliage is available as a food source. As this is an annual cycle, the graphs wraps around at each end and annotations provide additional detail which may not be easily captured graphically but in such a way that they do not hijack the viewer's attention.

References

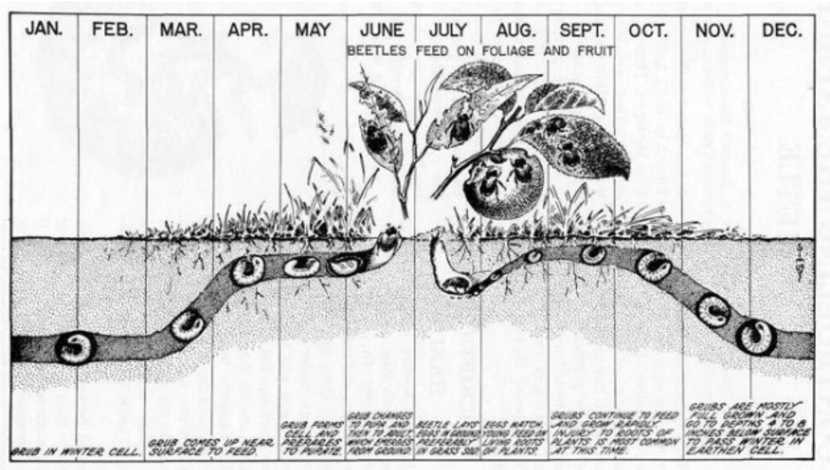
BBC data team. 2019. "BBC Visual and Data Journalism cookbook for R graphics." BBC. https://bbc.github.io/rcookbook/.

Chang, Winston. 2018. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. 2nd ed. O'Reilly Media. https://r-graphics.org/.

Fear, Simon. 2005. "Publication Quality Tables in LaTeX."

Few, Stephen. 2008. "Practical Rules for Using Color in Charts." *Visual Business Intelligence Newsletter* 11. http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf.

Frigge, Michael, David C. Hoaglin, and Boris Iglewicz. 1989. "Some


JAN.
FEB.
MAR.
APR.
MAY
JUNE
JULY
AUG.
SEPT.
OCT.
NOV.
DEC.
BEETLES FEED ON FOLIAGE AND FRUIT
GRUB IN WINTER CELL.
GRUB COMES UP NEAR SURFACE TO FEED.
GRUB FORMS CELL AND PREPARES TO PUPATE
GRUB CHANGES TO PUPA AND THEN TO ADULT WHICH EMERGES FROM GROUND
BEETLE LAYS EGGS IN GROUND PREFERABLY IN GRASS SOD
EGGS HATCH. YOUNG FEED ON LIVING ROOTS OF PLANTS.
GRUBS CONTINUE TO FEED AND GROW RAPIDLY. INJURY TO ROOTS OF PLANTS IS MOST COMMON AT THIS TIME.
GRUBS ARE MOSTLY FULL GROWN AND GO TO DEPTHS 4 TO 8 INCHES BELOW SURFACE TO PASS WINTER IN EARTHEN CELL.
FR.I.2789

Seasonal life cycle of the Japanese beetle in the vicinity of Philadelphia.

Figure 7: Life cycle of Japanese beetle, p43 Tufte (1983)

Implementations of the Boxplot." *The American Statistician* 43 (1): 50. https://doi.org/10.2307/2685173.

Ho, C. Y., R. W. Powell, and P. E. Liley. 1974. "Thermal Conductivity of the Elements: A Comprehensive Review." *Journal of Physical and Chemical Reference Data, Supplement 1* 3 (December): 1–244.

Mori, Lapo Filippo. 2007. "Tables in LaTeX2ε: Packages and Methods." *The PracTeX Journal* 1: 2007–1.

RStudio. 2012. "RStudio Cheat Sheets." 2012. https://www.rstudio.com/resources/cheatsheets/.

Tufte, Edward R. 1983. *The Visual Display of Quantitative Information*. Graphics Press.

———. 1997. "Visual and Statistical Thinking." In *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press. https://www.edwardtufte.com/tufte/books_textb.

Wickham, Hadley, and Garrett Grolemund. 2020. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. http://r4ds.had.co.nz.