Book Review

Kieran Healy

Data Visualization: A Practical Introduction

Reviewed by: Paul Cuffe

The reviewer is an assistant professor with the School of Electrical and Electronic Engineering, University College Dublin e-mail: (paul.cuffe@ucd.ie).

Book publisher: New Jersey, Princeton University Press, 2019 (272 pages, including index)

ISBN: 978-0-691-18161-5

Index terms: data graphics, data visualisation, visual communication, graphic design

WITH *"Data Visualization: A Practical Introduction"*, Kieran Healy delivers a hands-on masterclass in building effective data graphics using the `ggplot` package for the `R` language. While the implied audience is researchers in the social sciences, there is a lot of practical wisdom here for anyone who works with numerical data. This book treats both the theoretical principles of effective data visualisation design (in the mode of Edward Tufte, Stephen Few, Alberto Cairo and others) as well as concrete guidance on how to integrate such wisdom into a slick data analysis workflow in the `R` ecosystem. This practitioner-oriented approach is a very welcome addition to the literature: by covering both the *whys* and *hows* of data visualisation, this single volume swiftly equips researchers to build compelling graphics from their numerical data.

It is clear that Healy is passionate about his topic: you really feel that he is fed-up with seeing sloppy data graphics in published academic work. Likewise, his commitment to freeing researchers from laborious scutwork is evident in the many productivity tips he is eager to provide for the `R` environment. His prose is engaging and chatty, and the style of instruction is unpretentious and practical. The manuscript itself is attractively typeset: one commendable feature is the direct integration of the many graphics alongside the corresponding passage of text, by setting them as sidenotes in the wide margins. This richly graphical approach makes the lessons engaging, tangible and enjoyable to read.

Chapter 1 discusses why some data graphics are more effective than others. It opens with the famous example of *Anscombe's quartet*, a set of four bivariate datasets which have equivalent summary statistics, but which look strikingly different when inspected using scatterplots. Such a set of plots is provided as the first figure in the book, but, disappointingly, this particular graph is not very well executed: it has needlessly inconsistent ranges & tickmarks between the vertical and horizontal axes, and the physical aspect ratio is slightly off-square for no obvious reason. Aside from this quibble, this is an excellent chapter which gives a rapid survey of the modern consensus on data visualisation. Amongst other topics, this chapter covers the choice of appropriate graph type; the perceptual challenges with different data encodings; minimalism versus chart junk; gestalt rules; and a little colour theory. The goal is to equip the reader with a framework for critiquing and interrogating data graphics: to allow design decision to be unpacked and examined.

Chapter 2 gets us up-and-running in `RStudio` itself. The treatment here is suitable for a complete novice, and introduces the `R` language as well as the `ggplot` package which is central to the book. The workings of `R` should be familiar to users of other high-level languages such as `MATLAB` or `Python`. This chapter starts at a fairly basic level, and gradually builds up enough scripting concepts to allow a novice to produce a simple graph.

Chapter 3 is a lesson on building a series of scatterplots within `ggplot`. This chapter also introduces the concept of *geoms* in `ggplot`, which handle data encoding options in a disciplined way. As Healy says on p. 55: *"the central activity of visualising data with* `ggplot` *more or less always involves the same sequence of steps"* This is a direct consequence of `ggplot`'s well thought-out and consistent structure. This package tries to force you to think about the *grammar* and *structure* of the graphic you're creating. The `ggplot` package is both a paradigm for thinking about graphics and a tool for building them. Crucially, `ggplot` aims to automate low-level graphical tasks so the analyst can operate at a higher level of abstraction, simply stipulating the desired connections between data features and graphical encodings. The worked examples in this chapter make these concepts tangible.

Chapter 4 is about showing the right numbers: it's about making wise data processing decision before the visualization stage. Editorialising data (by grouping, transforming, agglomerating etc.) is an essential step in any data visualisation workflow so it is good to see it treated explicitly here. This chapter also puts a welcome emphasis on the practise of *faceting*, where multiple views of a dataset are provided as a series of adjacent *small multiple* graphs plotted on harmonised axes. A nice example is given on page 77, where we see timeseries line graphs of GDP per capita for the many nations of the world, organised over five panes grouped by continent. This is a sophisticated and powerful graphic and it is a testament to `ggplot` that it can be created with just a few lines of code. Deft use of faceting is a vital skill for effective data visualisation, and Healy is to be commended for giving the topic such a comprehensive treatment with this chapter.

In chapter 5, we're shown various ways to enhance our graphs. There is good coverage here about using *pipes* in R to streamline complex and repetitive data transformation tasks: this is all about equipping the reader with a fast and flexible workflow for maximum productivity. This chapter also introduces a few more *geoms*, to create exotic figures like Cleveland dot plots, box plots and jittered graphs. There's also a bit of tinkering here with scales, guides and themes, to show how to exert granular control over graph elements if `ggplot`'s defaults are not suitable. While this chapter covers a lot of ground, it is practical throughout and will be a lot of fun for readers who are working through the examples in R. It's nice to see the `ggrepel` library introduced here: this is a powerful tool for annotating points within a plot, which automates label placement to avoid clutter and overlap. This is used to slick effect in an impressive example figure on page 120.

Chapters 6 and 7 may be less directly relevant for engineers. Chapter 6 is quite a comprehensive treatment of working with, and plotting, statistical models. While this is vital in fields such as sociology and econometrics, it will not typically be a major concern in engineering fields. The discussion of fitting and plotting smoothers will be useful to many, however. Chapter 7 is about drawing thematic maps. There's lots of lovely colourful examples in this chapter and it is truly impressive what can be achieved with just a few lines of code. There's also lots of useful discussion and caveats about when to represent data on a choropleth map: *"The first thing to remember about spatial data is that you don't have to represent it spatially"* (page 176) While it's a fun chapter, in practice these mapping techniques will not be a core part of an engineer's toolkit for graphing data.

Chapter 8 is similar to chapter 5: it's a further discussion of tweaks and tips to refine and perfect graphs. There's some useful advise here on using colour effectively, emphasising the famous and versatile `ColorBrewer` palettes. The use of themes, layering and titling is also discussed. The chapter ends with a series of case studies, where ineffective graphs are reassembled in a more sensible way.

The book ends with a generous Appendix, which is a compendium of various productivity-oriented tips for working in the R and `tidyverse` ecosystems.

This single volume represents an excellent entry point for those wishing to upskill their abilities in data visualization. Much of the relevant theoretical work is covered explicitly, and the rest is clearly signposted. The software tools presented are mainstream, open-source and powerful, and the author's enthusiasm for the ecosystem is infectious. There's only one topic that felt slightly overlooked: the important of choosing appropriate ranges and divisions for the axes of a graph. Indeed, many of the figures in the book seemed a bit too cluttered with excessive gridlines and tickmarks. These are minor gripes, though: this is a welcome and timely volume which is full of practical wisdom for producing attractive and effective technical graphics.