Methods: How to Do Data Visualization Using R—Even If You Don't Use R

June 30, 2022

Statisticians Ross Ihaka and Robert Gentleman created the programming language R in 1993 (see R Core Team, 2021), and growing numbers of researchers across disciplines have used it ever since for data processing and statistical analysis. Between 2006 and 2018, citations of the R Core Team averaged a yearly growth of 87% (Barrett, 2019).

There's much to like about R. Among other advantages, it gives users the ability to augment the functions of the R language by selecting and creating R packages (extensions to R's statistical programming language that contain code, data, and documentation in a standardized format that can be installed by the users to do different analyses or data treatments), and the official R software environment is an open-source free software that anyone can use. However, many researchers might perceive R coding skills as difficult to learn and thus opt for point-and-click software such as SPSS and SAS.

In a 2022 article in Advances in Methods and Practices in Psychological Science, Emily Nordmann and colleagues at the University of Glasgow challenged the perception that coding skills needed to use R are difficult to learn. Their tutorial on how to use R for data visualization, which specifically targets researchers who have little or no experience using R, is available via OSF at https://osf.io/bj83f/.[1]

Using R contributes to reproducibility and transparency

One advantage of using R is its potential benefits to reproducibility and transparency. Because R makes the code used to process data, run statistical analysis, and create data visualizations easily available, other researchers can replicate data analysis procedures and directly test the code for errors. R also gives researchers "a much larger range of fully customizable data visualization options than are typically available in point-and-click software because of the open-source nature of R," Nordmann and colleagues wrote.

In addition, writing code to produce data visualizations can save you time later because you can reuse and adapt that code instead of starting from scratch. Moreover, data visualizations in R can look more attractive than the ones created with point-and-click software or Excel and are highly customizable, because you have more control over each element of each visualization.

Using layers within each visualization maximizes usability

In their tutorial, Nordmann and colleagues use the package *ggplot2* (Wickham, 2016), part of a larger collection of packages that provide functions for data processing. To construct visualizations, *ggplot2* uses a layered grammar of graphics (i.e., a standardized way to describe the components of a graphic; Wilkinson et al., 2005), in which plots are built up in a series of layers. For example, a scatterplot can be built with six layers:

Layer 1: Build the **plot space**

Layer 2: Specify the **variables**

Layer 3: Specify the **type of visualization** that is desired for these variables

Layer 4: Add individual **data points** and a **line of best fit**

Layer 5: Edit the **axis labels** for readability

Layer 6: Apply a **theme** to change the overall appearance of the plot

Each layer is independent of the others and can be customized independently. For instance, you can adjust the size, color, and position of each component of each layer and individually remove each layer. "The use of layers makes it easy to build up complex plots step-by-step and to adapt or extend plots from existing code," explained Nordmann and colleagues.

Using R for data visualization

Nordmann and colleagues provide a simulated data set for 100 participants and seven variables of interest. The simulated data come from a 2×2 mixed-design lexical decision task (i.e., participants

decide whether a string of letters is a word or a nonword). They then instruct readers on using [RStudio](#) (a development environment that makes working with R easier; RStudio Team, 2021) to write code in R. You'll learn how to best prepare and format the data; load packages in RStudio; load data; and summarize data with some common customizable visualizations, such as bar charts of counts, plots of aggregates and percentages, and histograms.

[Learn more about the APS journal *Advances in Methods and Practices in Psychological Science*](#).

The authors guide readers in formatting a data set to be used by R—transforming the usual wide format (one row of values per participant) into a long format (several rows per participant, one for each stimulus). With data in this format, *ggplot2* can create visualizations. Nordmann and colleagues provide detailed instructions on creating:

- Histograms (e.g., of reaction times and accuracy)
- Density plots
- Scatterplots

- Box plots
- Violin plots
- Bar charts
- Violin box plots
- Interaction plots

- Combines interaction plots
- Facets (separate plots for each level of a variable)

Besides being able to customize each type of plot, you'll learn how to store and save plots and how to combine multiple plots to be displayed together. In additional online resources (available at https://psyteachr.github.io/introdataviz/), Nordmann and colleagues provide additional advanced plots—including split-violin plots, rain-cloud plots, ridge plots, alluvial plots, and maps—with customization options.

Feedback on this article? Email apsobserver@psychologicalscience.org *or login to comment.*

Related content we think you'll enjoy

[Visualize Data to Communicate Science With Students, the Public, and Policymakers](#)

The latest issue of Psychological Science in the Public Interest explores the good and the bad of data visualizations and how public understanding of science can improve if researchers adopt better visualization techniques.

[Mapping the Moods of COVID-19: Global Study Uses Data Visualization to Track Psychological Responses, Identify Targets for Intervention](#)

More than 60,000 participants have participated in a global study to investigate the psychological implications of the COVID-19 pandemic.

[Presenting Science: Best Practices for Making Data "Pop"](#)

When creating data visualizations, these guiding principles can promote accuracy and engagement.

References

Barrett, T. S. (2019). *Six reasons to consider using R in psychological research.* PsyArXiv. https://doi.org/10.31234/osf.io/8mb6d

Nordmann, E., McAleer, P., Toivo, W., Paterson, H., & DeBruine, L. M. (2022). Data visualization using R for researchers who do not use R. *Advances in Methods and Practices in Psychological Science, 5*(2). https://doi.org/10.1177/25152459221074654

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

RStudio Team. (2021). *RStudio: Integrated development environment for R.* RStudio. http://www.rstudio.com/

Wickham, H. (2016a). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag. https://ggplot2.tidyverse.org

Wilkinson, L., Anand, A., & Grossman, R. (2005). Graph-theoretic scagnostics. In J. Stasko & M. Ward (Eds.), *IEEE Symposium on Information Visualization (InfoVis 05): Proceedings* (pp. 157–158). IEEE Computer Society.