Bespoke Data Visualization using R and ggplot2

Sandy J. J. Gould
University of Birmingham
Birmingham, UK
s.gould@cs.bham.ac.uk

ABSTRACT

Being able to visualize data in consistent high-quality ways is a useful skill for HCI researchers and practitioners. In this course, attendees will learn how to produce high quality plots and visualizations using the ggplot2 library for the R statistical computing language. There are no prerequisites and attendees will leave with scripts to get them started as well as foundational knowledge of free open-source tools that they can build on to produce complex, even interactive, visualizations.

CCS CONCEPTS

• **Human-centered computing** → **Visualization techniques**; **Visualization systems and tools**; **Visualization toolkits**.

KEYWORDS

R; ggplot2; Statistical Computing; Visualization

ACM Reference Format:
Sandy J. J. Gould. 2019. Bespoke Data Visualization using R and ggplot2. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3290607.3298810

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3298810

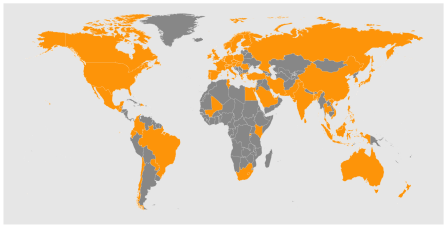

Figure 1: ggplot2 and R have been used by the instructor to create many of the plots for the CHI 2018 and 2019 blogs.

INTRODUCTION

Visualizations are used in HCI research and practice. HCI researchers use visualizations to illustrate publications and in teaching. In practice, visualizations are used to help convey a variety of information to people. The ggplot2 library for the R statistical computing environment allows people to developing bespoke high-quality visualizations that best fit data, rather having to produce 'best effort' visualizations that are constrained by the features of tools like Microsoft Excel and SPSS.

BENEFITS

Attendees of the course will benefit from:

- Gaining an understanding of the advantages of ggplot2 compared to other methods of presenting data
- The ability to combine the basic components of ggplot2 to produce a complete visualization
- The ability to modify template scripts to produce simple graphs, like histograms
- The ability to modify template scripts to produce more sophisticated visualizations like maps

INTENDED AUDIENCE

The intended audience for this course is any CHI attendee who would like to be able to represent their data in their research or products more flexibly and elegantly. It is intended as an introduction, so would not be suitable for experienced users of ggplot2.

PREREQUISITES

There are no knowledge prerequisites for this course, however a laptop capable of running R (e.g., Windows, MacOS or Linux) is essential for engagement with the practical activities in the course. We will be using R Studio [10] to facilitate the exploration of plotting during the session. This is free open-source software. Participants will access the R Studio through a browser — no software will be necessary.

CONTENT

R [8] is a statistical computing language that is widely used to process data. ggplot2 [11] is a popular library for R that is used to create many kinds of visualizations. One of the most useful features of ggplot2 is that it allows for visualizations to be created declaratively. This means that someone using the library can create visualizations by specifying the elements of a visualization and the data that underpins it. The library does the work of creating the 'end product' visualization. This means that high-quality bespoke visualizations can be created quickly and predictably. The goal of the course is

to convince attendees that a declarative approach to the generation of visualizations will save them time and deliver a better product.

The instructor will begin the course with presentation to attendees. The instructor will describe the philosophical approach of ggplot2 to visualization. Attendees will be encouraged to consider how this approach differs from approaches that they may be more used to, such as Microsoft's Excel.

After a high level introduction, the instructor will start to walk attendees through the fundamental components of ggplot2, including aesthetics and graphical primitives. This walk-through will focus on the modularity of ggplot2, and attendees will learn how the basic components of ggplot2 can be combined to create visualizations. The instructor will show attendees how each of these basic elements works alone and then how they combine. We will cover how these components can be broadly influenced by themes to produce a 'house style' for visualizations like those of publications such as the Economist and New York Times. The focus of this section will be on how snippets of code can be plugged together, rather than on creating visualizations from scratch.

The introductory lecture will last for approximately 30 minutes. The remainder of the session will give attendees a chance to build their own visualizations. During this period, attendees will make use of the R Studio development environment to build ggplot2 visualizations by editing template visualizations. Attendees will make guided changes to these templates to produce individual bespoke visualizations. This practical component is detailed below.

PRACTICAL WORK

There is no requirement for participants to be familiar with the R language. Given also that the course is 80 minutes, attendees will not develop R scripts for creating visualizations from scratch. Instead, the session will focus on getting attendees to use the features ggplot2 and to understand how small changes to scripts can have large effects on the visualization that is produced.

Attendees will be guided through editing template scripts. These scripts will start off as minimal working examples. As the practical part of the session progresses, additions will be made to the scripts that increase their complexity and change the appearance of the visualizations. The changes will be guided in such a way that attendees will have choice over how their final visualization appears.

We will work through two example visualizations during the practical part of the session. One histogram and one map. The histogram example will focus on the statistical and categorical operations features build into ggplot2. The map example will focus on how the fundamental elements of ggplot2 can be 'stacked' to build complex visualizations like maps.

At the end of the session, attendees will have customized example scripts that will form the basis of their future efforts with bespoke visualization. In addition to the in-class help provided by the instructor, a full walkthrough of the exercise will be provided on a webpage. This page will be left online for attendees to refer back to in future.

INSTRUCTOR BACKGROUND

Sandy Gould is a Lecturer (Assistant Professor) in Human Computer Interaction in the School of Computer Science at the University of Birmingham, UK. Gould is an experienced classroom practitioner, with experience of teaching students from age 11 through to PhD. Gould has made extensive use of ggplot2 in his publications [6, 7] and as the Analytics Chair for CHI 2018 and 2019.

He has previously run courses on research methods at CHI in 2015 [5], 2016 [3] and 2017 [4]. The 2017 iteration of the course included a demonstration of the ggplot2-based interactive plotting tool, *Shiny* [9].

RESOURCES

Several of Gould's plots made with ggplot2 can be seen in the publications listed below as well as part of the CHI 2018 [1] and 2019 [2] blogs.

REFERENCES

[1] ACM. 2017. ACM CHI 2018 Blog. https://chi2018.acm.org/blog/.

[2] ACM. 2018. ACM CHI 2019 Blog. https://chi2019.acm.org/blog/.

[3] Duncan P. Brumby, Ann Blandford, Anna L. Cox, Sandy J. J. Gould, and Paul Marshall. 2016. Research Methods for HCI: Understanding People Using Interactive Technologies. In *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA. https://doi.org/10.1145/2851581.2856682

[4] Duncan P. Brumby, Ann Blandford, Anna L. Cox, Sandy J. J. Gould, and Paul Marshall. 2017. Understanding People: A Course on Qualitative and Quantitative HCI Research Methods. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 1170–1173. https://doi.org/10.1145/3027063.3027103

[5] Sandy J. J. Gould, Duncan P. Brumby, Anna L. Cox, Geraldine Fitzpatrick, Jettie Hoonhout, David Lamas, and Effie Law. 2015. Methods for Human-Computer Interaction Research. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 2473–2474. https://doi.org/10.1145/2702613.2706691

[6] Sandy J. J. Gould, Anna L. Cox, and Duncan P. Brumby. 2016. Diminished Control in Crowdsourcing: An Investigation of Crowdworker Multitasking Behavior. *ACM Trans. Comput.-Hum. Interact.* 23, 3 (June 2016), 19:1–19:29. https://doi.org/10.1145/2928269

[7] Sandy J. J. Gould, Anna L. Cox, Duncan P. Brumby, and Alice Wickersham. 2016. Now Check Your Input: Brief Task Lockouts Encourage Checking, Longer Lockouts Encourage Task Switching. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3311–3323. https://doi.org/10.1145/2858036.2858067

[8] R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

[9] RStudio, Inc. 2014. *shiny: Easy web applications in R*. URL: http://shiny.rstudio.com.

[10] RStudio Team. 2015. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. http://www.rstudio.com/

[11] Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org