Complex Data Visualisation Made Easy with R and ggplot2

Sandy J.J. Gould
goulds@cardiff.ac.uk
School of Computer Science and Informatics
Cardiff University
Cardiff, Wales, UK

ABSTRACT

Being able to visualise 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 visualisations 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, visualisations. Course information materials can be found at https://www.sjjg.uk/chi22-course.

CCS CONCEPTS

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

KEYWORDS

R; ggplot2; Statistical Computing; visualisation

1 INTRODUCTION

Visualisations are used in HCI research and practice. HCI researchers use visualisations to illustrate publications and in teaching. In practice, visualisations are used in research summaries, news articles and business reports. The ggplot2 library for the R statistical computing environment allows people to developing bespoke high-quality visualisations that best fit data, rather having to produce 'best effort' visualisations that are constrained by the features of tools like Microsoft Excel and SPSS. By taking a *declarative* approach to visualisation, ggplot2 allows us to systematically produce visualisations that are consistent over time and across datasets.

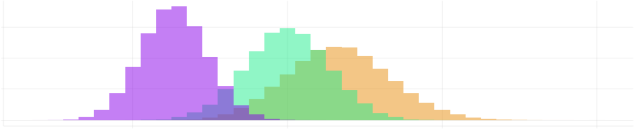
2 BENEFITS

Attendees of the course will benefit from:

- An understanding of the advantages of ggplot2 compared to other methods of presenting data
- Knowledge of how to combine the basic components of ggplot2 to produce a complete visualisation
- The ability to modify template scripts to produce simple graphs, like histograms
- The ability to modify template scripts to produce more sophisticated visualisations like maps

3 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. No prior experience with R or ggplot2 is required.



4 PREREQUISITES

There are no knowledge prerequisites for this course. We will be using R Studio [11] 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 installation will be necessary; an internet connection and a modern graphical browser are the only things required.

5 CONTENT

R [10] is a statistical computing language that is widely used to process data. ggplot2 [12] is a popular library for R that is used to create many kinds of visualisations. One of the most useful features of ggplot2 is that it allows for visualisations to be created declaratively, based on the principles of the Grammar of Graphics [13]. This means that someone using the library can create visualisations by specifying the elements of a visualisation and the data that underpins it. The library does the work of creating the 'end product' visualisation. This means that high-quality bespoke visualisations can be created quickly and *predictably.* The goal of the course is to convince attendees that a declarative approach to the generation of visualisations will save them time and deliver a better product than using visualisation wizards in tools like Microsoft Excel or IBM SPSS.

I will begin the course with presentation to attendees. I will describe the philosophical approach of ggplot2 to visualisation. Attendees will be encouraged to consider how this approach differs from approaches that they may be more used to. After a high level introduction, I 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 visualisations. I will show attendees how each of these basic elements works alone and then how they can be combined. I will cover how these components can be broadly influenced by themes to produce a 'house style' for visualisations like those of publications such as the Economist and New York Times.

The introductory lecture will last for approximately 30 minutes. The remainder of the session will give attendees a chance to build their own visualisations. During this period, attendees will make use of the R Studio development environment to build ggplot2

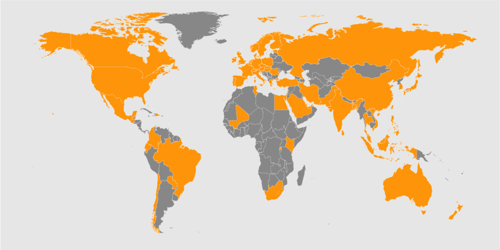


Figure 1: I have used ggplot2 and R to create many of the plots for the CHI 2018, 2019 and 2020 blogs.

visualisations by editing template visualisations. Attendees will make guided changes to these templates to produce individual bespoke visualisations. This practical component is detailed below.

6 PRACTICAL WORK

There is no requirement for participants to be familiar with the R language. Given also that the course is two 75 minute units, attendees will not develop R scripts for creating visualisations from scratch; this would be too much to get through in the time available. 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 visualisation that is produced. Understanding how aesthetics are constructed is critical for being able to make effective use of ggplot2.

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 visualisations. The changes will be guided in such a way that attendees will have control over how their final visualisation appears.

I will work through a few example visualisations during the course, including maps and histograms. The histogram example will focus on the statistical and categorical operations features built into ggplot2. The map example will focus on how the fundamental elements of ggplot2 can be 'stacked' to build complex visualisations 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 visualisation. In addition to the in-class help that I will provide, a full walkthrough of the exercises will be provided for on a webpage. This page will be left online for attendees to refer back to in future.

6.1 Changes from 2019 iteration of this course

I ran a version of this ggplot2 course at CHI 2019 [6], and this proposal is based on the proposal for that course. The course was very well received by participants, with 70% of responding participants strongly agreeing (7 or 6 on a seven-point scale) that the course was worth the time and money. However, 60% of responding participants indicated that the course was too short. None indicated that it was too long. Therefore, for this iteration of the course, I am *expanding the course to cover two 75-minute sessions*. This will permit a more relaxed pace, give me more time to provide individual help and give participants more chance to explore some of the extension activities under supervision.

7 BUDGET AND RESOURCES

To avoid software installation issues, the course will make use of a hosted instance of R Studio that participants can access through a browser. This will require a temporary virtual instance that will be active for the duration of the course. Instances are memory-intensive and so this temporary instance will cost USD $50 for the duration of the course.

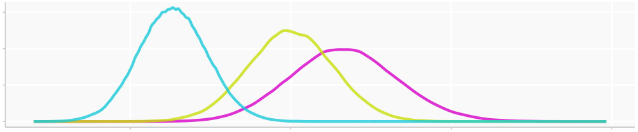
The course overview[1] and detailed course materials[2] for the CHI 2019 version of this course are publicly available. There have not been major changes to the way that ggplot2 works, so the structure and details of the course will be similar to the 2019 iteration. The course notes for the CHI 2022 interaction of the course are available and under development[3].

8 ACCESSIBILITY

The focus of the course is intrinsically visual. However, the declarative nature of ggplot2 means it is somewhat more accessible to people with visual impairments than other means for producing visualisation. I will to do my very best to accommodate anyone with a visual impairment.

9 INSTRUCTOR BACKGROUND

I am a Senior Lecturer (≈ Associate Professor) at the School of Computer Science and Informatics at Cardiff University. I am an experienced teacher across ages and levels. I have made extensive use of ggplot2 in my publications [8, 9]. As Analytics Chair for CHI 2018, 2019 and 2020, I also made extensive used of ggplot2 in the composition of blog posts about the submission process [3–5]. In addition to a previous iteration of this course [6], I have previously run courses on research methods at CHI in 2015 [7], 2016 [1] and 2017 [2].



REFERENCES

[1] 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, 1028–1031. https://doi.org/10.1145/2851581.2856682

[2] 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

[3] CHI 2018 Programme Committee. 2018. Paper Length in the Reformulated Papers Track.

[1] https://www.sjjg.uk/chi19-course/

[2] https://www.sjjg.uk/chi19-course/materials/

[3] https://www.sjjg.uk/chi22-course

[4] CHI 2019 Programme Committee. 2018. What's Been Provisionally Accepted? – CHI 2019.

[5] CHI 2020 Programme Committee. 2019. CHI2020 Paper Reviews and Rebuttals.

[6] Sandy J. J. Gould. 2019. Bespoke Data Visualization Using R and Ggplot2. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3290607.3298810

[7] 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

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

[9] 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

[10] R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[11] RStudio Team. 2015. *RStudio: Integrated Development Environment for r.* RStudio, Inc., Boston, MA.

[12] Hadley Wickham. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[13] Leland Wilkinson. 1999. *The Grammar of Graphics* (1st ed. 1999. ed.). Springer-Verlag, New York, New York State. https://doi.org/10.1007/978-1-4757-3100-2