The Journal of Open Source Education

# R for Data Analysis: An open-source resource for teaching and learning analytics with R

**Trevor French**[1]

**1** Independent Researcher, USA

**DOI:** 10.21105/jose.00202

**Software**

- Review
- Repository
- Archive

**Submitted:** 16 February 2023
**Published:** 20 May 2023

## Summary

R for Data Analysis is a sequential learning system designed to teach anyone how to use R to analyze data. The materials can be used in aggregate to learn how to analyze data programmatically or broken into modules and used to supplement an existing educational program. Each lesson contains a conceptual overview, practical examples, resources for further study, and exercises designed to reinforce understanding.

The materials have been made publicly available at: https://trevorfrench.github.io/R-for-Data-Analysis/ and licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 (CC BY-NC-ND 4.0) License.

## Statement of Need

In the article titled "An empirical study of the rise of big data in business scholarship," the authors suggest that the amount of data that exists in our current society creates a "constant flow of potential new insights for business, government, education and social initiatives" (Frizzo-Barker et al., 2016). This presents an opportunity to educate practitioners in both industry and academia on programmatic data analysis techniques. These practitioners may have historically relied on specialists and/or methodologists to perform analyses, but it is important to ensure that analysis tools are as accessible as the data has become.

There are plenty of resources aimed at teaching specialists how to apply advanced analytics techniques to their chosen discipline; however, there is a notable lack of resources which aim to educate the general public on programmatic data analysis. This phenomenon was observed in an article titled "What is Statistics?" when the authors proclaimed "statistical education has not been sufficiently accessible." (Brown et al., 2009). Furthermore, the contents of R for Data Analysis are centered around the idea of the "process of data analysis" broadly applied to any discipline. This differs from other high-quality resources, such as "R for Reproducible Scientific Analysis" (Zimmerman et al., 2019), which teaches similar topics in the context of the scientific process.

## Instructional Design

The learner is guided through programming concepts which progress in difficulty while simultaneously applying these concepts to the process of data analysis. The process of data analysis is broken into five steps:

1. **Gathering Requirements** - Before one embarks on an analysis, it's important to make sure the requirements are understood. Requirements include the questions which one's stakeholders are hoping to answer as well as the technical requirements of how the analysis will be performed.
2. **Data Acquisition** - As one might imagine, data must be acquired before conducting an analysis. This may be done through methods such as manual creation of data sets, importing pre-constructed data, or leveraging APIs.
3. **Data Preparation** - Most data will not be received in the precise format one needs. The process of data preparation is where features and structure will be added to the data.
4. **Developing Insights** - Once the data is prepared, one can begin to make sense of the data and develop insights about its meaning.
5. **Reporting** - Finally, it's important to report on the data in such a way that the information is able to be digested by the people who need to see it when they need to see it.

No prior knowledge is required to begin using these materials. The content starts at the very beginning by showing learners how to set up their R environment and the basics of programming in R. By the end of the materials, learners will be able to perform intermediate analytics techniques such as linear regression and automatic report generation. The materials are structured as follows:

- **Part I (Fundamentals)** will introduce learners to the basics of programming in the context of R.
- **Part II (Data Acquisition)** will teach learners how to create, import, and access data.
- **Part III (Data Preparation)** will show learners how to begin preparing data for analysis.
- **Part IV (Developing Insights)** goes through the process of searching for and extracting insights from data.
- **Part V (Reporting)** demonstrates how to wrap an analysis up by developing and automating reports.

Each part contains several chapters which cover specific ideas related to the overarching topic. At the end of each of these chapters the learner will find additional resources to use to dive deeper into the ideas. Each part is then concluded with practical exercises for learners to test their skills.

# References

Brown, E. N., Kass, R. E., Johnstone, I., Gibbs, A., Reid, N., Madigan, D., Gelman, A., Hedayat, S., Stufken, J., Nolan, D., & Lang, D. T. (2009). What is statistics? [with comments and rejoinder]. *The American Statistician*, *63*(2), 105–123. http://www.jstor.org/stable/25652239

Frizzo-Barker, J., Chow-White, P. A., Mozafari, M., & Ha, D. (2016). An empirical study of the rise of big data in business scholarship. *International Journal of Information Management*, *36*(3), 403–413. https://doi.org/10.1016/j.ijinfomgt.2016.01.006

Zimmerman, N., Wilson, G., Silva, R., Ritchie, S., Michonneau, F., Oliver, J., Dashnow, H., Boughton, A., Teucher, A., Mawdsley, D., MacDonald, A., Rice, T., Emonet, R., Daigle, R., Mills, B., Bolker, B., Penrose, S., Sloggett, C., Blischak, J., … Takemon, Y. (2019). *swcarpentry/r-novice-gapminder: Software Carpentry: R for Reproducible Scientific Analysis, June 2019* (Version v2019.06.1). Zenodo. https://doi.org/10.5281/zenodo.3265164