10th GESIS Summer School in Survey Methodology

[2nd Virtual GESIS Summer School]

28 July – 20 August 2021

Syllabus for Course 3: Introduction to R for Data Analysis

Lecturers:	Dr. Johannes Breuer	Dr. Stefan Jünger
E-mail:	johannes.breuer@gesis.org	stefan.juenger@gesis.org
Homepage:	https://www.johannesbreuer.com/	https://github.com/stefmue

Date: 02-06 August 2021
Time: 10:00-12:30 + 14:00-16:30
Time zone: CEST/CEDT, course starts Monday at 10:30 am
Venue: Online via Zoom

About the Instructors:

Dr. Johannes Breuer is as a senior researcher in the team Data Linking & Data Security at GESIS where his work focuses on data linking and the use of digital trace data. He received his Ph.D. in psychology from the University of Cologne. Before joining GESIS, he worked in several research projects investigating the use and effects of digital media. His other research interests include computational methods, data management, and open science.

Dr. Stefan Jünger is a postdoctoral researcher in the team Data Linking & Data Security at the GESIS Data Archive working on the use of georeferenced data in social science research. He received his Ph.D. in sociology from the University of Cologne. In addition to his substantive research and his work on the use of georeferenced data in the social sciences, he is also interested in topics of data management and reproducible research.

Selected Publications:

- Breuer, J. (2017). R(Software). In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), The International Encyclopedia of Communication Research Methods. Wiley. doi: 10.1002/9781118901731.iecrm0201
- Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. New Media & Society, 22(11), 2058–2080. doi: 10.1177/1461444820924622
- Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. Social Science Computer Review, Advance online publication. doi: 10.1177/0894439319843669
- Jünger, Stefan. 2019. Using Georeferenced Data in Social Science Survey Research. The Method of Spatial Linking and Its Application with the German General Social Survey and the GESIS Panel. Köln: GESIS - Leibniz-Institut für Sozialwissenschaften.
- Müller, Stefan. 2019. Räumliche Verknüpfung georeferenzierter Umfragedaten mit Geodaten: Chancen, Herausforderungen und praktische Empfehlungen. In Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten. Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten, Hrsg. Uwe Jensen, Sebastian Netscher und Katrin Weller, 211–229. Opladen, Berlin, Toronto: Verlag Barbara Budrich.
- Klinger, Julia, Stefan Müller, und Merlin Schaeffer. 2017. Der Halo-Effekt in einheimisch-homogenen Nachbarschaften: Steigert die ethnische Diversität angrenzender Nachbarschaften die Xenophobie? Zeitschrift für Soziologie 46: 402–419.

Short Course Description:

The open source software package R is free of charge and offers standard data analysis procedures as well as a comprehensive repertoire of highly specialized processes and procedures, even for complex applications. In addition to providing an introduction to the basic concepts and functionalities of R, we will go through a prototypical data analysis workflow in the course: import, wrangling, exploration, (basic) analysis, reporting.

Keywords:

R, data wrangling, exploratory data analysis, data visualization, data analysis

Course Prerequisites:

- prior experience with quantitative data analysis, basic statistics, and regression
- experience with using other statistical packages (e.g., SPSS or Stata) is helpful, but not a requirement.

Target Group:

Participants will find the course useful if they want to use R to wrangle, explore, visualize and analyse their data.

Course and Learning Objectives:

By the end of the course participants will:

- Be comfortable with using R and RStudio
- Be able to import, wrangle, and explore their data with R
- Be able to conduct basic visualizations and analyses of their data with R

Organizational Structure of the Course:

The best way to learn R is to try things out and apply the presented concepts. Therefore, we will have a mixture of lectures and hands-on exercises. More specifically, each topic will be introduced in a lecture by the instructors. Participants will then receive a set of exercises on each topic that they work on alone. The solution of the exercises will then be discussed before the start of the next lecture part.

Software and Hardware Requirements:

Course participants will need a computer or laptop with R (https://cran.r-project.org/) and RStudio installed (https://www.rstudio.com/). Both programs are free and open source.

Long Course Description:

Getting started
The first session will cover all preliminary topics. This includes installing and loading packages in R, using the RStudio GUI, basic data structures in R, and where/how to find help.

Data import & export
We will discuss how to import different types of data into R (e.g., CVS, Excel, SPSS and Stata files) as well as how to store data in R-specific formats and how to export them to various other formats.

Data wrangling: Basics
Before researchers can start to analyze their data, they first have to wrangle them (i.e., clean and transform). In this session, we will focus on basic functions for "getting our data in shape" (renaming variables, creating new ones, changing variable types, etc.) from the so-called Tidyverse which is "an opinionated collection of R packages designed for data science" (see https://www.tidyverse.org/).

Data wrangling: Advanced
After making ourselves familiar with the handy tools for data wrangling in R, we will learn about more advanced wrangling methods and processing whole or multiple datasets. For this purpose, we will explore some programming and coding tools available in R, which are especially helpful if we aim to repeat specific tasks.

Exploratory data analysis
In this session, we will learn to explore our data to, e.g., check distributions, missing values or outliers. We will also learn about generating summary statistics for our data.

Data visualization - Part 1
In this first session on data visualization, participants will learn how to create basic visualizations of their data. We will discuss the plotting functions that base R offers. However, the main focus will be on the powerful visualization package ggplot2 (which is also part of the Tidyverse). In this part, we will use some of the functionalities of ggplot2 to visually explore our data.

Confirmatory data analysis
In this part we will give an introduction to basic confirmatory data analysis techniques in R. We will cover basic bivariate and multivariate analyses (e.g., t-tests, correlation, regression) and discuss how model statistics can be transferred to a standard data format which we can further work with.

Data visualization Part 2
This second session on data visualization concentrates on the visualization of statistical models. We will learn about tools for gathering results from (multiple) models and how to visualize them. Again, the main focus of this session will be on the package ggplot2.

Reporting with R Markdown
R Markdown is a combination of a simple markup language (Markdown) and R code. In this part of the course, we will explore how to generate fully reproducible reports with R Markdown and discuss what else you can do with it (e.g., write manuscripts or create presentations or posters).

Outlook, Advanced Use of R, Extended Q&A session
In this last session, we will provide an outlook on some advanced topics that we did not cover in this course as well as current and future developments of (and around) R that might be of interest for participants. We will also provide the opportunity for an extended Q&A session to discuss any open questions.

Day-to-day Schedule and Literature:

Day	Topic(s)
1	**Morning** Getting started with R and RStudio **Afternoon** Data import & export
	Suggested reading: ▪ http://www.r-bloggers.com/why-use-r/ ▪ Fogarty, B. J. (2019). Quantitative social science data with R. Chapter 2. ▪ Wickham, H., & Grolemund, G. (2016). R for data science. Chapters 2, 4, and 8 (= 4, 6, & 11 in the online version).
2	**Morning** Data wrangling: Basics **Afternoon** Data Wrangling: Advanced
	Suggested reading: ▪ Fogarty, B. J. (2019). Quantitative social science data with R. Chapters 4 and 5. ▪ Wickham, H., & Grolemund, G. (2016). R for data science. Chapters 3, 9, 10, and 17 (= 5, 12, 13, & 21 in the online version).
3	**Morning** Exploratory data analysis **Afternoon** Data visualization - Part 1

	Suggested reading: ▪ Fogarty, B. J. (2019). Quantitative social science data with R. Chapters 7 and 8. ▪ Wickham, H., & Grolemund, G. (2016). R for data science. Chapter 1 (= 3 in the online version).
4	**Morning** Confirmatory data analysis **Afternoon** Data visualization - Part 2
	Suggested reading: ▪ Fogarty, B. J. (2019). Quantitative social science data with R. Chapters 9 to 11. ▪ Wickham, H., & Grolemund, G. (2016). R for data science. Chapters 18 to 20 (= 23 to 25 in the online version).
5	**Morning** Reporting with R Markdown **Afternoon** Outlook, Advanced Use of R, Extended Q&A session
	Suggested reading: ▪ Wickham, H., & Grolemund, G. (2016). R for data science. Chapters 21 and 23 (= 27 & 29 in the online version).

Preparatory Reading:

Not necessary.

Additional Recommended Literature:

- Fogarty, B. J. (2019). Quantitative social science data with R. Sage.
- Grolemund, G. (2014). Hands-on programming with R. Write your own functions and simulations. O'Reilly. Also freely available online: https://rstudio-education.github.io/hopr/
- Healy, K. (2019). Data visualization. A practical introduction. Princeton University Press.
- Matloff, N. (2011). The art of R programming: A tour of statistical software design. No Starch Press.
- Muenchen, B. (2011). R for SPSS and SAS Users. Springer Science & Business Media.
- Long, J. D., & Teetor, P. (2019). R Cookbook: Proven recipes for data analysis, statistics, and graphics. 2nd edition. O'Reilly. Also freely available online: https://rc2e.com/
- Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data (First edition). O'Reilly. Also freely available online: http://r4ds.had.co.nz/
- Xie, Y., Allaire, J. J., & Grolemund, G. (2019). R Markdown. The definitive guide. CRC Press. Also freely available online: https://bookdown.org/yihui/rmarkdown/
- Xie, Y., Dervieux, C., & Riederer, E. (2020). R Markdown Cookbook. CRC Press. Also freely available online: https://bookdown.org/yihui/rmarkdown-cookbook/
- Google's R Style Guide. https://google.github.io/styleguide/Rguide.xml
- The Tidyverse Style Guide: https://style.tidyverse.org/