9th GESIS Summer School in Survey Methodology
Cologne, August 2020

Syllabus for course: “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: 03-07 August 2020
Time: 10:00-12:30 + 14:00-16:30
Course starts Monday at 10:00
Venue: Online via Zoom

About the Lecturers:
Dr. Johannes Breuer works as a senior researcher in the team Data Linking & Data Security at the GESIS Data Archive. He received his Ph.D. in psychology from the University of Cologne in 2013. Before joining GESIS, he worked in several research projects investigating the use and effects of digital media at the universities of Cologne, Hohenheim, and Münster, and the Leibniz-Institute für Wissensmedien (Knowledge Media Research Center). His other research interests include computational methods, data management, and open science.

Dr. Stefan Jünger (né Müller) 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. While he has a Dr. in sociology from the University of Cologne, he research interests also include general topics, such as research data management and reproducible research.

Selected Publications:

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 data analysis, basic statistics, and regression;
• basic familiarity with the use of a computer
• 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.

Programming with R
In this session we will discuss programming basics in R, focusing on functions and loops. We will also cover alternatives to loops, such as functions from the apply family as well as the purrrr package.

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: Base R vs. the Tidyverse
Before researchers can start to analyze their data, they first have to wrangle (i.e., clean and transform). In this sessions we will compare the options that base R and the Tidyverse - "an opinionated collection of R packages designed for data science" (see https://www.tidyverse.org/ ) - offer for getting data into formats that we can work with when we want to visualize and analyze them.

Data visualization
In the two sessions on data visualization, participants will learn how to create visualizations of data. We will discuss the plotting functions that base R offers, but the main focus will be on the powerful visualization package ggplot2 (which is also a part of the Tidyverse).
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 use some of the visualization techniques discussed in the previous sessions to 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 how model statistics can be transferred to a standard data format with the broom package.

Reporting with RMarkdown
RMarkdown 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 RMarkdown and discuss what else you can do with it (e.g., write manuscripts or create presentations or posters).

Application example “Geospatial data analysis with R” or Extended Q&A session
Based on the preferences of the participants we will either go through the complete process of an exemplary geospatial data analysis or provide the opportunity for an extended Q&A session to discuss any open questions or provide pointers to what to do or explore next.

Day-to-day Schedule and Literature:

<table>
<thead>
<tr>
<th>Day</th>
<th>Topic(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Morning</td>
</tr>
<tr>
<td></td>
<td>Getting started with R and RStudio</td>
</tr>
<tr>
<td></td>
<td>Afternoon</td>
</tr>
<tr>
<td></td>
<td>Programming with R</td>
</tr>
<tr>
<td></td>
<td>Suggested reading:</td>
</tr>
<tr>
<td></td>
<td><a href="http://www.r-bloggers.com/why-use-r/">http://www.r-bloggers.com/why-use-r/</a></td>
</tr>
<tr>
<td>2</td>
<td>Morning</td>
</tr>
<tr>
<td></td>
<td>Data import &amp; export</td>
</tr>
<tr>
<td></td>
<td>Afternoon</td>
</tr>
<tr>
<td></td>
<td>Data wrangling: Base R vs. the Tidyverse</td>
</tr>
<tr>
<td></td>
<td>Suggested reading:</td>
</tr>
<tr>
<td></td>
<td>• Fogarty, B. J. (2019). Quantitative social science data with R. Chapter 4 – Data management.</td>
</tr>
<tr>
<td>3</td>
<td>Morning</td>
</tr>
<tr>
<td></td>
<td>Data visualization Part 1</td>
</tr>
<tr>
<td></td>
<td>Afternoon</td>
</tr>
<tr>
<td></td>
<td>Data visualization Part 2</td>
</tr>
<tr>
<td></td>
<td>Suggested reading:</td>
</tr>
<tr>
<td></td>
<td>• Fogarty, B. J. (2019). Quantitative social science data with R. Chapter 8 – Visualising data.</td>
</tr>
<tr>
<td>4</td>
<td>Morning</td>
</tr>
<tr>
<td></td>
<td>Exploratory data analysis</td>
</tr>
<tr>
<td></td>
<td>Afternoon</td>
</tr>
<tr>
<td></td>
<td>Confirmatory data analysis</td>
</tr>
<tr>
<td></td>
<td>Suggested reading:</td>
</tr>
<tr>
<td></td>
<td>• Fogarty, B. J. (2019). Quantitative social science data with R. Chapters 9 to 11.</td>
</tr>
</tbody>
</table>
Morning
Reporting with R Markdown

Afternoon
Application example "Geospatial data analysis with R" or Extended Q&A session

Suggested reading:

Preparatory Reading:
- Not necessary

Recommended Literature:
- The Tidyverse Style Guide: https://style.tidyverse.org/