Journal of Marketing Analytics (2024) 12:465–470
https://doi.org/10.1057/s41270-023-00264-0

SOFTWARE REVIEW

Marketing analytics with RStudio: a software review

David Dege[1] · Philipp Brüggemann[1]

Revised: 28 September 2023 / Accepted: 29 September 2023 / Published online: 3 November 2023

Introduction

Empirical research heavily depends on the use of statistical software for editing, analyzing, and visualizing the data. One widely used statistical software is R (https://cran.r-project.org) and its integrated development environment (IDE) RStudio (https://posit.co/downloads/). Originally developed by Ross Ihaka and Robert Gentleman, R is especially noteworthy for being an open-source software published under the GNU General Public License, which therefore can be freely used, even for commercial purposes (Hornik and the R Core Team, 2022). Though R being developed for statistical computation as well as for graphic visualizing it achieved high popularity among programming languages according to several indexes, regularly reaching the top ten of most popular languages (Carbonnelle 2023; Cass 2022). Among published research papers, R is the second most used software (the most frequently used software is SPSS). Moreover, papers requiring R show on average a higher impact factor than published papers using SPSS (Comeau et al 2019). Perhaps most demonstrative of R's popularity is the number of questions posted on websites exclusively focusing on the statistical questions like the Q&A community "CrossValidated" which is part of Stack Overflow. Among the commonly used statistical software, questions concerning R are most often, outranking the second placed software (SPSS) by four.

This software review provides an overview of R and its commonly used IDE RStudio, its structure, features, and functionality. Furthermore, basic commands are presented and an outlook on advanced applications is given. This software overview is particularly relevant for researchers and practitioners who want to start using R, or for advanced users who desire to learn more about the broad range of potential R applications.

Basics of using RStudio

Structure of RStudio

The following review is based on RStudio, which we highly recommend, especially because of its greater ease of use. By default, RStudio is divided into four areas of which elements can be detached to independent movable windows. Figure 1 illustrates the standard structure of RStudio.

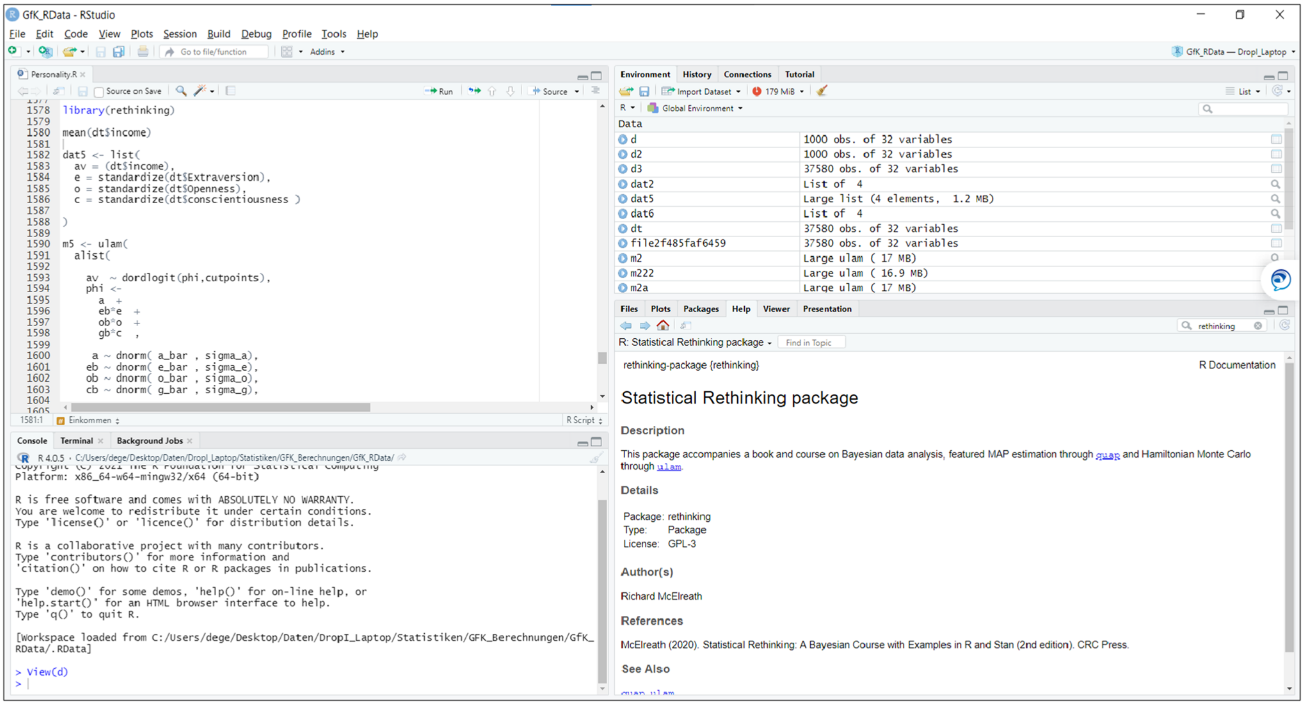
In the following, we will explain and briefly describe the different areas step by step.

1. Top left area: This area has several tabs that can be switched manually. In doing so, users can call imported data frames or R-script files, in which R-code is written and saved. For example, an R-code can be used to call a data frame ("View(data.fame.name)"). This data frame opens after executing the command (ctrl + enter) in the top left area.
2. Top right area: The top right area contains several tabs of which the most important is the environment tab. This tab contains all data frames, lists, data variables, and arrays which were created and/or imported to RStudio. By clicking on a data frame, this data frame will open in the top left area.
3. Bottom right area: This area also contains several tabs. For example, produced plots can be viewed here, an overview of installed packages is given, and the help tab, which contains documentation about installed packages, can be found.
4. Bottom left area: At the bottom left, among some other things, is the console tab that displays the output. If you perform a calculation via the R-script (e.g., calculate a mean value), the result will be displayed in this bottom left area.

✉ David Dege
daviddege@gmail.com

Philipp Brüggemann
philipp.brueggemann@fernuni-hagen.de

[1] Fernuniversitat Hagen Fakultat fur Wirtschaftswissenschaft, Hagen, NRW, Germany



```
library(rethinking)

mean(dt$income)

dat5 <- list(
  av = (dt$income),
  e = standardize(dt$Extraversion),
  o = standardize(dt$Openness),
  c = standardize(dt$conscientiousness )

)

m5 <- ulam(
  alist(

    av  ~ dordlogit(phi,cutpoints),
    phi <-
      a  +
      eb*e  +
      ob*o  +
      gb*c   ,

     a ~ dnorm( a_bar , sigma_a),
    eb ~ dnorm( e_bar , sigma_e),
    ob ~ dnorm( o_bar , sigma_o),
    cb ~ dnorm( g_bar , sigma_g),
```

```
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from C:/Users/dege/Desktop/Daten/DropI_Laptop/Statistiken/GFK_Berechnungen/GfK_
RData/.RData]

> View(d)
> 
```

```
Data
d                  1000 obs. of 32 variables
d2                 1000 obs. of 32 variables
d3                 37580 obs. of 32 variables
dat2               List of  4
dat5               Large list (4 elements,  1.2 MB)
dat6               List of  4
dt                 37580 obs. of 32 variables
file2f485faf6459   37580 obs. of 32 variables
m2                 Large ulam ( 17 MB)
m222               Large ulam ( 16.9 MB)
m2a                Large ulam ( 17 MB)
```

rethinking-package {rethinking} R Documentation

Statistical Rethinking package

Description

This package accompanies a book and course on Bayesian data analysis, featured MAP estimation through quap and Hamiltonian Monte Carlo through ulam.

Details

Package: rethinking
Type: Package
License: GPL-3

Author(s)

Richard McElreath

References

McElreath (2020). Statistical Rethinking: A Bayesian Course with Examples in R and Stan (2nd edition). CRC Press.

See Also

quap, ulam

Fig. 1 Structure of RStudio

Operation and basic commands

This section details operations and basic commands of RStudio. R itself is operated by code. Therefore, all commands need to be written. Using RStudio, some commands are accessible via drop-down menu. However, a menu interface similar to SPSS does not exist. Furthermore, not all statistical operations are accessible from the very start. R relies on add-ons called packages which need to be downloaded and invoked manually (Hornik and the R Core Team, 2022). Packages are largely created and maintained by fellow researchers and often published with research papers. To use a package, it must be installed (e.g., using "install.package(package.name)") and activated (using "library(package.name)"). Currently, at least 19.749 packages do exist (Hornik and the R Core Team, 2022). Hereinafter, we present some basic commands to illustrate the functionality and use of RStudio.

The operation with RStudio usually involves several steps, which are performed simultaneously or iteratively. Figure 2 gives a schematic overview of these steps. It should be emphasized that this is only an exemplary overview and does not claim to be exhaustive. We explain this exemplarily procedure of using RStudio in the next section step by step.

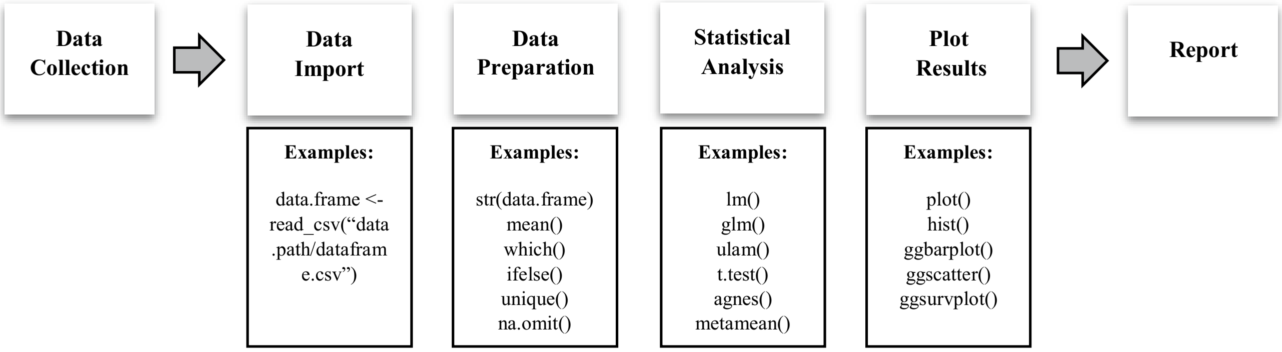


Fig. 2 Procedure of using RStudio

Procedure of using RStudio

Data import

Following the usual process of data analysis, at first data need to be *imported*. RStudio provides the option to import data via a drop-down menu (top right area). Especially when working with many different data sets, an import via code is much more comfortable. Via packages like “readr” (Wickham et al 2023) or “Hmisc” (Harrell and Dupont 2019), data can easily be imported (e.g., data.frame <—read_csv(“data.path/dataframe.csv”)).

Data preparation

Next, data need to be *prepared*. The possibilities for data preparation are huge in R. First checking the data structure and data classes is advisable. Here the “str(data.frame)”-command is one helpful solution. Applying the “ifelse”-command, new variables can be inserted in the data frame, which are conditioned on other variables’ values. Since data frames contain columns and rows, every data point within a R data frame is accessible via the following command: data.frame [row.number, column.number]. Addressing specific columns or rows is very advantageous in combination with numerous other commands like, e.g., the “which()”-command: sum(data.frame[which(data.frame$variableA = = 0),]$variableB). In this case the values of the variable “variableB” are summed, but only if within the same row the variable “variableA” is equal to zero. Deleting duplicate cases (“unique(data.frame)”), merging data frames (“merge(data.frame1, dataframe2, by = “ID”)”), deleting empty rows (“na.omit(data.frame)”) , and many other functions are available in RStudio. Finally, R contains typical programming features like looping, dynamic variable creation or nested variables.

Since R can be used to implement extremely comprehensive data preparation and analysis, users should comment their R-code for better readability. To not execute a comment, every line starting with “#” will be skipped from execution. In general, if not several lines are highlighted, code execution ends within one line, except a starting symbol (e.g., opening quotation mark or brackets) was not closed, yet. Furthermore the arrow (“ <-“) assigns a value (e.g., array, list, data frame) to a variable. A variable, respectively, a column, within a data frame is addressable via the “$”-sign (data.frame$variableA). One recommendation, especially when working with large data sets, is the “data.table”-package, which was especially developed to facilitate and accelerate calculations (Dowle et al 2023).

Statistical analysis

After finishing the preparations, the data can be *analyzed*. Generally speaking, there is a R-function, respectively, a package, for every need in Marketing Analytics and beyond. The preinstalled “stats”-package already contains popular methods like linear regression modeling (“lm()”) or general linear modeling (“glm()”). For more sophisticated methods, e.g., like ridge regression or multi-level regression modeling, additional packages may be needed (e.g., “MASS”-package (Verzani 2012) or “lme4”-package (Bates et al 2015)). Getting an overview of existing packages is not always easy, especially since oftentimes several packages provide solutions for similar statistical tasks. For example, the “sem”-package (Fo et al 2022) and the “lavaan”-package (Rosseel 2012) are both specialized on Structural Equation Modeling. Though, in general, every package holds unique features, among which users can choose the ones most fitting for their research.

Plot results

Finally, it may be necessary to *plot* analytical findings. This is again one strength of R. Already the preinstalled “graphics”-package is suitable for a wide array of plots (e.g., scatterplots, histograms, or pie plots). Though more display possibilities are offered by the “ggplot2”-package (Wickham 2016). This package does not only contain single commands, but the opportunity to combine different components using a grammar based approach to create customized graphics (Wickham 2016). Hereby, every plot starts with the basic data, e.g., a scatterplot, then step-by-step additional graphic features can be added. In this way, a publication quality plot and more can easily be created.

Advanced applications using RStudio

The following paragraphs provide an introducing overview of advanced applications in RStudio. We selected applications that we believe are particularly relevant to marketing and marketing analytics. Further information on the application of R in marketing analytics is provided by Yildirim and Kübler (2023).

Create user-defined functions

If a particular combination of statistical methods or procedures is to be run repeatedly, user-defined R functions can be created. An R-function is an object which allows for modular programming (Hornik and the R Core Team, 2022) and thus automation. A user-defined function is in a sense a personalized command. Within the “function()”-command, parameters are appointed, which are then called for calculations. Using the vast repertoire of R-packages, different statistical operations can be combined and automated in a function. For example, if a complex value transformation for a variety of variables is necessary, a function can be created and applied to all variables. This way for implementing the transformation, it is only needed to call the self-defined function instead of writing the formula for each variable. User-defined functions are enormously helpful when repeatedly complex and exhaustive analyses are required, as it is often the case for marketing analyses based on Big Data.

Create new packages

If a research project leads to a new statistical method, which shall be provided to the public, it is possible to create a R-package and upload it, e.g., to the "Comprehensive R Archive Network" (CRAN), where most of the R-packages are stored. The documentation on how to create a R-package is stored within the R software folder (doc). Additionally, research papers (Leisch 2008) or even a textbook (Wickham 2023) provides supplemental assistance in creating R-packages. Furthermore, RStudio simplifies the process of creating such packages by its menu-options as well. For example, Lu et al. (2023) recently developed the R-package DEPART to decompose prices in terms of regular prices and promotional prices. This R-package can now be used free of charge by other researchers or practitioners.

Cluster analysis

Cluster analytical procedures to identify groups of high similarity are conveniently implemented using the "cluster"-package (Maechler et al 2021). The package provides a wide range of clustering methods with regard to hierarchal clustering and partitioning methods. The package includes also clustering methods meant for large data sets. Other packages provide two-step clustering (Rosenberg et al 2020) or Bayesian clustering (Liverani et al 2015). Further information is provided by Kassambara (2017) textbook about clustering in R.

Conjoint analysis

Conjoint analysis is widely used to measure preferences, e.g., product preferences. Via the R-package "conjoint," necessary functions for creating a data matrix as well as the profile simulations are given (Bak and Bartlomowicz 2012). The package also includes a probabilistic model and a logit model approach for simulation.

Text mining

Working with several packages in R is very common. For text mining, five or more packages are required depending on the need. Via the "tidytext"-package (Silge and Robinson 2016) text is split up and rows are created. Using the "stringr"-package (Wickham 2019) working with regular expressions and text manipulation like transforming capital letters is made easy. Finally for stemming and latent semantic analysis, the packages "SnowballC" (Bouchet-Valat and Bouchet-Valat 2020) and "lsa" (Wild 2015) are needed. A deeper look into Text Mining in R provides the textbook Silge and Robinson (2017).

Machine learning

For R, various machine learning algorithms exist. Since every algorithm demands a different syntax and parameters, one easy solution for applying different approaches is provided by the "caret"-package (Kuhn et al 2020). "caret" standardizes the model training process by unitizing the syntax. The package itself draws on several different other R-packages. Though, by only loading the necessary packages, the installation is made time and resource efficient.

Meta-analysis

A meta-analysis sums up the insights of many different studies regarding one research question (Harrer et al 2022). Therefore, the researched sample does not contain single observations, but studies. In R, several packages provide different functions for analyzing study findings. The "meta"-package (Schwarzer 2007) provides more basic approaches whereas the "metaphor"-package includes multi-level modeling (Viechtbauer 2010) and the "netmeta"-package network meta-analysis methods (Balduzzi et al 2023). Guidance, which method suits the research question the most, offers Harrer et al (2022). Here, an introduction to Bayesian meta-analysis can even be found as well.

Bayesian statistic

Bayesian statistics differs from frequentistic statistics, among other things, by defining variables as random; therefore, it is possible to fit a probability distribution to parameter by applying the Bayes Theorem (Matsuura 2023). Like this, Bayesian statistics allows the use of probability values and the integration of prior knowledge into research. However, Bayesian calculations depend strongly on Markov Chain Monte Carlo simulations. For this purpose, the R-package "rstan" provides access to the probabilistic programming language Stan (Stan Development Team 2022). Furthermore, packages like "brms" (Bürkner 2017) or "rethinking"

(McElreath 2016) simplify using Stan. Therefore, R enables calculating complex multi-level models on Bayesian basis using state-of-the-art algorithms.

Conclusion

R, respectively, RStudio provides a comprehensive selection of statistical methods for marketing analytics and beyond, making the use of other statistical software almost obsolete. Especially the big number of active user willing to develop further applications for R and supporting R-beginner as well as advanced users makes R a good choice for all scientific researchers and anyone interested in statistical analysis. In addition, it must be positively emphasized that the software is free and based on the open-source principle. However, R's code-based approach tends to discourage new users and requires a steep learning curve. Furthermore, the high number of available packages, which often include similar methods, may confuse especially R-beginners. In the end, however, we draw a very positive feedback on R, respectively, RStudio. Everyone willing to take the first steps in working with R will be rewarded by seemingly never-ending statistical possibilities. RStudio's versatility is especially helpful in the area of marketing analytics with Big Data and therefore highly recommended for researchers and practitioners here.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Bak, A., & Bartlomowicz, T. (2012) Conjoint analysis method and its implementation in conjoint R package. Data Analysis Methods and Its Applications: 239–248.

Balduzzi, S., G. Rücker, A. Nikolakopoulou, T. Papakonstantinou, G. Salanti, O. Efthimiou, and G. Schwarzer (2023) netmeta: An R package for network meta-analysis using frequentist methods. *Journal of Statistical Software* 106: 1–40.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v067.i01.

Bouchet-Valat, M. (2020) Package 'SnowballC'. R Package Version 0.7. 0.

Bürkner, P.-C. (2017) brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80: 1–28.

Carbonnelle, P. (2023) PYPL Popularity of Programming Language https://pypl.github.io/PYPL.html. Accessed 20 Aug 2023.

Cass, S. (2022) Top Programming Languages 2022: Python's still No. 1, but employers love to see SQL skills https://spectrum.ieee.org/top-programming-languages-2022. Accessed 15 Aug 2023.

Comeau, D.C., C.-H. Wei, R. IslamajDoğan, and Z. Lu. (2019) PMC text mining subset in BioC: About three million full-text articles and growing. *Bioinformatics* 35 (18): 3533–3535.

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Parsonage, H. (2023) Package 'data. table' https://cran.r-project.org/web/packages/data.table/data.table.pdf. Accessed 25 August 2023.

Fo, J., Nie, Z., & Byrnes, J. (2022) sem: Structural Equation Models https://CRAN.R-project.org/package=sem. Accessed 20 August 2023.

Harrer, M., P. Cuijpers, T.A. Furukawa, and D.D. Ebert. (2022) *Doing meta-analysis with R: A hands-on guide*, 1st ed. Boca Raton: CRC Press.

Harrell Jr, F. E., & Dupont, C. (2019) Package 'hmisc'. CRAN2018, 2019: 235–236.

Hornik, K. & the R Core Team (2022) The R FAQ https://CRAN.R-project.org/doc/FAQ/R-FAQ.html. Accessed 20 August 2023.

Kassambara, A. (2017) Multivariate Analysis I: Practical guide to cluster analysis in R: Unsupervised machine learning, 1st ed., Online: Sthda.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Team, R. C. (2020) Package 'caret'. The R Journal, 223(7).

Leisch, F. (2008) Creating R Packages: A Tutorial Universitätsbibliothek der Ludwig-Maximilians-Universität München, https://doi.org/10.5282/ubm/epub.6175

Liverani, S., D.I. Hastie, L. Azizi, M. Papathomas, and S. Richardson (2015) Premium: An R package for profile regression mixture models using dirichlet processes. *Journal of Statistical Software* 64 (7): 1–30. https://doi.org/10.18637/jss.v064.i07.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021) cluster: Cluster Analysis Basics and Extensions. R package version 2.1. 2. R Package Version R Foundation for Statistical Computing: Vienna, Austria, 1: 56.

Matsuura, K. (2023) *Bayesian Statistical Modeling with Stan, R, and Python*. Singapore: Springer.

McElreath, R. (2016) *Statistical rethinking: A Bayesian course with examples in R and Stan. A Chapman & Hall book:*, vol. 122. Boca Raton, London, New York: CRC Press.

Rosenberg, J.M., Schmidt, J. A., & Beymer P. N. (2020) prcr: Person-Centered Analysis. R package version 0.2.1, https://CRAN.R-project.org/package=prcr. Accessed 21 Aug 2023.

Rosseel, Y. (2012) lavaan : An R Package for structural equation modeling. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v048.i02.

Schwarzer, G. (2007) meta: An R package for meta-analysis. *R News* 7 (3): 40–45.

Silge, J., and D. Robinson (2016) tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software* 1 (3): 37.

Silge, J., and D. Robinson. (2017) *Text mining with R: A tidy approach (First edition) Beijing*. Boston, Farnham, Sebastopol, Tokyo: O'Reilly.

Stan Development Team (2022) RStan: the R interface to Stan. R package version 2.21.7. Online: https://mc-stan.org/.

Verzani, J. (2012) *Getting started with RStudio (2. release) Sebastopol*. California: O'Reilly.

Viechtbauer, W. (2010) Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 36: 1–48.

Wickham, H. (2016) Ggplot2: Elegant graphics for data analysis (Second edition). Use R! Switzerland: Springer. https://doi.org/10.1007/978-3-319-24277-4.

Wickham, H. (2019) Package 'stringr'. Website: https://stringr.tidyverse.org, https://github.com/tidyverse/stringr.

Wickham, H. (2023) R PACKAGES: Organize, test, document, and share your code Erscheinungsort nicht ermittelbar: O'Reilly Media, https://learning.oreilly.com/library/view/-/9781098134938/?ar.

Wickham, H., Hester, J., Francois, R., Bryan, J., Bearrows, S., & Posit, P. B. (2023) Package 'readr'. Read Rectangular Text Data. Available Online: https://cran.R-Project.org/web/packages/readr/readr.Pdf. Accessed 28 June 2022.

Wild, F. (2015) lsa: Latent semantic analysis. R Package Version 0.73, 1.

Yildirim, G., and R. Kübler. (2023) *Applied marketing analytics using R*. California: SAGE Publications Ltd.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

David Christian Dege is a PhD student at the Chair of Business Administration, especially Marketing, at FernUniversität in Hagen, Germany. He studied consumer psychology and has been working for various online marketing agencies since 2017 and is also self-employed in this field. His research interests include therefore digital marketing focusing especially on psychological aspects such as personality. Further fields of interests are retailing, research methods and Bayesian Statistics. He is working on several projects related to market share analysis, consumer purchasing behavior and the impact of weather on sales.

Philipp Brüggemann is Post-Doctoral Researcher at the Chair of Business Administration, specialization Marketing, at FernUniversität in Hagen, Germany. He studied business administration and worked for several years in the field of finance and controlling. His research interests include among others e-commerce and digital marketing, research methods and retailing. He is working on several projects related to technology acceptance, market share analysis and online-offline grocery shopping. Furthermore, he is the initiator of the Marketing Scholars initiative on LinkedIn, that provides information on current conference and journal calls. His research has been published in various journals and he is an EB Member at Journal of Marketing Analytics and an ERB member at International Journal of Consumer Behavior.