Data Insight Foundations

Step-by-Step Data Analysis with R

Nikita Tkachenko

Apress®

Data Insight Foundations

Nikita Tkachenko

Data Insight Foundations

Step-by-Step Data Analysis with R

Apress®

Nikita Tkachenko
San Francisco, CA, USA

ISBN-13 (pbk): 979-8-8688-0579-0 ISBN-13 (electronic): 979-8-8688-0580-6
https://doi.org/10.1007/979-8-8688-0580-6

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Shaul Elson
Development Editor: Laura Berendson
Editorial Assistant: Gryffin Winkler

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media LLC, 1 New York Plaza, Suite 4600, New York, NY 10004. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.
For information on translations, please e-mail booktranslations@springernature.com; for reprint, paperback, or audio rights, please e-mail bookpermissions@springernature.com.
Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.
Any source code or other supplementary material referenced by the author in this book is available to readers on the Github repository. For more detailed information, please visit https://www.apress.com/gp/services/source-code.

If disposing of this product, please recycle the paper

I dedicate this book to my late grandfather, Vladimir Zamashikov.


IF AIN'T
MACHO BUT IT'S
HONEST WORK

Contents

About the Author

Nikita Tkachenko leads Evalyn, a consulting agency specializing in AI-driven customer service audits and data analytics solutions. He helps organizations of all sizes harness AI and data to optimize decision-making, streamline operations, and enhance customer experiences. With a strong foundation in research and analytics, Nikita also teaches courses on research tools, mentors students, and conducts academic research at the University of San Francisco.

About the Technical Reviewer

Sijo Valayakkad Manikandan is an accomplished artificial intelligence and data science leader. He has extensive experience managing large-scale data science projects for Fortune 500 corporations. Sijo holds a Master of Science degree in Business Analytics from the renowned University of Texas at Austin and a Bachelor of Engineering degree from the Birla Institute of Technology and Science in Pilani, India. He combines his academic excellence and practical expertise to deliver exceptional results.

Sijo has made significant contributions to the field of data science and research, not only through his professional and academic achievements but also through his dedication to the community. He is a member of several distinguished organizations, such as the American Statistical Association, and has conducted independent research. Sijo has also actively reviewed academic and professional books, mentored junior data scientists, and guided early-stage startups. Moreover, his expertise has led him to serve on the jury of several prestigious awards, including The Webby Awards.

As a dedicated researcher and thought leader, Sijo continues to profoundly impact the field of data science, inspiring a new generation of data scientists. His contributions have advanced the field of data science and research and helped shape the future of this rapidly evolving industry, making him an invaluable asset to the community and a visionary in his field.

Acknowledgements

I dedicate this book to my late grandfather, Vladimir Zamashikov.

I extend my deepest gratitude to Thomas Weinandy, who convinced me to transform my rough draft into a published book and guided me through the process. Without him, this book would not have been possible.

I am deeply indebted to Parsa Rahimi for his unwavering support, meticulous review of all my chapters, edits, and ideas, and for providing artworks featuring SF.

My thanks also go to Alessandra Cassar and Michael Jonas for their support throughout my academic journey. Special thanks go to Peter Lorentzen, Ker Gibbs, and Jesse Anttila-Hughes for their invaluable guidance and support.

I appreciate Shivani Shukla, Bruce Wydick, Misha Gipsman, Mario Lim, Steve Trettel, Mehmet Emre, Andrew Hobbs, Arman Khachiyan, Robizon Khubulashvili, and Konrad Posch for reviewing chapters and providing their support. Additionally, I am grateful to the faculty, librarians, writing center staff, and the Center for Business Studies and Innovation (CBSI) at the University of San Francisco for their guidance.

Gordon Getty deserves thanks for inspiring me to write more confidently about my skills.

The team at APress, especially Shaul Elson, Commissioning Editor, Laura Berendson, Development Editor, and Gryffin Winkler, Editorial Assistant, guided me through the publishing process. My technical reviewer, Sijo Manikandan, provided comments that significantly enriched the book with additional examples and explanations. Special thanks go to everyone at APress whose efforts made this book a reality.

I gratefully acknowledge John Chetwynd, Jake Consgrove, Timofei Lopukhov, and Alexander Rom, whose support, curiosity, and confidence in my work are deeply appreciated.

I must acknowledge my family, including my mother Ekaterina Tkachenko, whose belief in me inspired me to launch this project, my father Anton Tkachenko, stepmother Daria Tkachenko, and grandparents Valentina and Alexander Trachenko and Raisa Zamashikova.

A special note of thanks goes to Anastasia Ternovskaya for her unwavering belief in me.

I very much appreciate Maria Aksutenko for making the artwork of the “researcher-cat.”

Special thanks also go to the Posit team for their invaluable contributions and support of the community.

Finally, my heartfelt appreciation goes to the open-source community and everyone who has generously shared their knowledge online. You made this possible, and I am thrilled to contribute to the conversation. I hope you find this book useful in your journey toward mastering research.

Introduction

This book was born from my frustrations and experiences in higher education and professional work. It originated from notes and materials from a Spring 2023 course in survey design, inspired by the enthusiastic response and insightful questions from my students.

Young data professionals typically learn about models, experiments, and theories during their classes, frequently returning to that knowledge. However, executing high-quality research and analysis requires a deeper understanding of the tools and the "how" rather than just the "what" and "why." This knowledge often transcends what is taught within the constraints of a standard curriculum. In this book, I aim to bridge that gap, helping you move from knowing what you want to do to understanding how to do it. I have distilled hundreds of hours of frustration into these chapters, so you won't have to traverse that path yourself.

This book is not a comprehensive guide; if that's what you're seeking, you may want to look elsewhere. Instead, the book can serve as a map, outlining the necessary tools and topics for your research journey. The goal is to build your intuition and provide pointers for where to find more detailed information. The chapters are deliberately concise and to the point, aiming to reveal and enlighten rather than bore. You'll learn about efficient data management, reproducible research, literature review and writing practices, and effective data visualization.

Initially inspired by my journey through graduate school in economics, this book offers value across disciplines. It contains essential insights for anyone engaged in data-related work, from the humanities to data analytics and the sciences. Whether you are refining your expertise or new to data analytics, this book promises to offer something of value.

Examples provided are primarily in R, making a basic understanding of the language advantageous but not essential. Several chapters, especially those focusing on theory, require no programming knowledge at all. A diverse audience, including web developers, mathematicians, data analysts, and economists, has found the material beneficial. The book is designed to be inclusive, offering insights irrespective of your programming proficiency or professional background.

Its structure allows for flexible reading paths; you may explore the chapters in sequence for a systematic learning experience or navigate directly to the topics most relevant to you.

Setting Up R and RStudio

Welcome to the exciting world of data analysis with R, a language crafted specifically for statistical analysis and data visualization. R's user-friendly syntax and reproducibility make it an ideal choice for both novices and professionals. However, before diving into data exploration and modeling, it's essential to differentiate between R, the programming language, and RStudio, the integrated development environment (IDE) that enhances R's functionality.

Download and Install R

R is maintained and distributed through the Comprehensive R Archive Network (CRAN), ensuring your access to the latest version and resources.

For macOS Users:

1. Navigate to the CRAN website.
2. Click on "Download R for macOS."
3. Select the appropriate version:
   - For Apple Silicon (e.g., M1, M2), download the version with "-arm64" (e.g., R-4.2.2-arm64.pkg) in its name.
   - For Intel-based Macs, select the version without "-arm64" (e.g., R-4.2.2.pkg).
4. Follow the installation wizard. The default settings are typically sufficient.

For Windows Users:

1. Visit the CRAN website.
2. Choose "Download R for Windows."
3. Select "base" and then the first link at the top of the page (e.g., Download R-4.2.2 for Windows).
4. The installer will guide you through the process. Stick with the default settings for a smooth installation.

5. Additionally, Windows users should download Rtools, which are crucial for compiling packages from sources. Visit Rtools, match the Rtools version with your R version, and follow the installer instructions.

Download and Install RStudio

RStudio provides a user-friendly interface for R, akin to what Microsoft Word offers for text but tailored to R scripting.

- To download RStudio, head to the RStudio download page.
- Click on "DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS" or select your operating system for detailed instructions.

Configure RStudio

Enhance your RStudio experience with these initial setup tips:

- **Change Theme**: Shift from the default theme to a dark theme for improved readability. Go to `Tools` > `Global Options` > `Appearance`, and select "Dracula." Click "Apply."
- **Install Fira Code Font**: For a modern coding aesthetic, install the "Fira Code" font, which supports programming ligatures. Instructions can be found at Fira Code on GitHub. After installation, apply this font in RStudio under `Appearance`.

Install Packages

Packages extend R's functionality. Install them easily with commands in the RStudio console:

```
# Install a single package
install.packages("tidyverse")

# Install multiple packages
install.packages(c("tidyverse", "gapminder"))
```

To use installed packages, load them into your session:

```
library(tidyverse)
```

Now that we have R and RStudio up and running, let's dive into some fundamental data manipulation techniques in R.

Part I
Working with Data

Chapter 1
Data Manipulation

In data analysis, visualization and manipulation are essential for understanding and communicating complex information. R, a powerful programming language for data analysis, offers a variety of packages that enable the creation of visually compelling plots and the streamlining of data manipulation. One of the most user-friendly and widely used collections of R packages[1] is tidyverse, developed by Hadley Wickham, chief scientist at Posit (RStudio). Tidyverse includes packages that cover all common tasks and can be installed with `install.packages("tidyverse")` and activated using `library("tidyverse")`. In this introduction, we will cover the basics of tidyverse using `readr` for data reading, `dplyr` for data manipulation, `tidyr` for data tidying, and, later in the book, `ggplot2` for data visualization. For more information about tidyverse, visit their website https://www.tidyverse.org/.

1.1 Basics

Let's kick off with some fundamental concepts! R can be employed as a simple calculator.

```
# A "#" is used to annotate comments!
2 + 2
```

```
#> [1] 4
```

```
2 * 4
```

```
#> [1] 8
```

[1] A package is a collection of prewritten functions, data, and documentation that enhances the capabilities of the R programming language for specific tasks.

N. Tkachenko, *Data Insight Foundations*, https://doi.org/10.1007/979-8-8688-0580-6_1

```
2^8
```

```
#> [1] 256
```

```
(1 + 3) / (3 + 5)
```

```
#> [1] 0.5
```

```
log(10) # Calculates the natural log of 10!
```

```
#> [1] 2.302585
```

R allows for defining variables and performing operations on them. Both = and <- can be used for assigning values to a variable name, though <- is preferred to avoid confusion and certain errors.

```
x <- 2 # Equivalent to x = 2
x * 4
```

```
#> [1] 8
```

The command `x <- 2` assigns the value `2` to `x`. Thus, when we subsequently type `x * 4`, R replaces `x` with `2` to evaluate `2 * 4` and obtain `8`. The value of `x` can be updated as needed using = or <-. Bear in mind that R is case sensitive, so `X` and `x` are recognized as different variables.

```
x
```

```
#> [1] 2
```

```
(x <- x * 5) # Wrapping with (...) prints the variable
```

```
#> [1] 10
```

To further explore operations in R, the following table presents a comprehensive overview of basic arithmetic, comparison, and logical operators you might need.

Operator	Description	Example	Result
+	Addition	`3 + 2`	`5`
-	Subtraction	`5 - 2`	`3`
*	Multiplication	`3 * 2`	`6`
/	Division	`6 / 2`	`3`
^	Exponentiation	`2 ^ 3`	`8`
%%	Modulus (remainder)	`5 %% 2`	`1`
%/%	Integer division	`5 %/% 2`	`2`
==	Equal to	`2 == 2`	`TRUE`
!=	Not equal to	`2 != 3`	`TRUE`
<	Less than	`2 < 3`	`TRUE`
>	Greater than	`3 > 2`	`TRUE`
<=	Less than or equal to	`2 <= 2`	`TRUE`
>=	Greater than or equal to	`2 >= 2`	`TRUE`
&	Logical AND	`TRUE & FALSE`	`FALSE`
|	Logical OR	`TRUE | FALSE`	`TRUE`
!	Logical NOT	`!TRUE`	`FALSE`

1.1.1 Data Types

R possesses a multitude of data types and classes, including `data.frames`, which are akin to Excel spreadsheets with columns and rows. Initially, we'll examine vectors. Vectors can store multiple values of the same type, with the most basic ones being numeric, character, and logical.

```
x
```

```
#> [1] 10
```

```
class(x)
```

```
#> [1] "numeric"
```

```
(true_or_false <- TRUE)
```

```
#> [1] TRUE
```

```
class(true_or_false)
```

```
#> [1] "logical"
```

```
(name <- "Parsa Rahimi")
```