

# Preface

Welcome to *Data Wrangling with R!* In this book, I will help you learn the essentials of preprocessing data leveraging the R programming language to easily and quickly turn noisy data into usable pieces of information. Data wrangling, which is also commonly referred to as data munging, transformation, manipulation, janitor work, etc., can be a painstakingly laborious process. In fact, it has been stated that up to 80% of data analysis is spent on the process of cleaning and preparing data (cf. Wickham 2014; Dasu and Johnson 2003). However, being a prerequisite to the rest of the data analysis workflow (visualization, modeling, reporting), it's essential that you become fluent *and* efficient in data wrangling techniques.

This book will guide you through the data wrangling process along with giving you a solid foundation of the basics of working with data in R. My goal is to teach you how to easily wrangle your data, so you can spend more time focused on understanding the content of your data via visualization, modeling, and reporting your results. By the time you finish reading this book, you will have learned:

- How to work with the different types of data such as numerics, characters, regular expressions, factors, and dates.
- The difference between the various data structures and how to create, add additional components to, and how to subset each data structure.
- How to acquire and parse data from locations you may not have been able to access before such as web scraping or leveraging APIs.
- How to develop your own functions and use loop control structures to reduce code redundancy.
- How to use pipe operators to simplify your code and make it more readable.
- How to reshape the layout of your data, and manipulate, summarize, and join data sets.

Not only will you learn many base R functions, you'll also learn how to use some of the latest data wrangling packages such as `tidyr`, `dplyr`, `httr`, `stringr`, `lubridate`, `readr`, `rvest`, `magrittr`, `xlsx`, `readxl` and others. In essence, you will have the data wrangling toolbox required for modern day data analysis.

## Who This Book Is for

This book is meant to establish the baseline R vocabulary and knowledge for the primary data wrangling processes. This captures a wide range of programming activities which covers the full spectrum from understanding basic data objects in R to writing your own functions, applying loops, and web scraping. As a result, this book can be beneficial to all levels of R programmers. Beginner R programmers will gain a basic understanding of the functionality of R along with learning how to work with data using R. Intermediate and advanced R programmers will likely find the early chapters reiterating established knowledge; however, these programmers will benefit from the mid and latter chapters by learning newer and more efficient data wrangling techniques.

## What You Need for This Book

Obviously to gain and retain knowledge from this book, it is highly recommended that you follow along and practice the code examples yourself. Furthermore, this book assumes that you will actually be performing data wrangling in R; therefore, it is assumed that you have or plan to have R installed on your computer. You will find the latest version of R for Linux, Mac OS, and Windows at <https://cran.r-project.org>. It is also recommended that you use an integrated development environment (IDE) as it will simplify and organize your coding environment greatly. There are several to choose from; however, I highly recommend the RStudio IDE which you can download at <https://www.rstudio.com>.

## Reader Feedback

Reader comments are greatly appreciated. Please send any feedback regarding typos, mistakes, confusing statements, or opportunities for improvement to [wranglingdata@gmail.com](mailto:wranglingdata@gmail.com).

## Bibliography

- Dasu, T., & Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning* (Vol. 479). John Wiley & Sons.
- Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59 (i10).



<http://www.springer.com/978-3-319-45598-3>

Data Wrangling with R

Boehmke, B.

2016, XII, 238 p. 24 illus., 10 illus. in color., Softcover

ISBN: 978-3-319-45598-3