Tidy Finance with R

Christoph Scheuch (wikifolio Financial Technologies) and Stefan Voigt (University of Copenhagen and I

2022-07-15

Contents

Preface

Why does this book exist?

Financial economics is a vibrant area of research, a central part of all businesses activities, and at least implicitly relevant for our everyday life. Despite its relevance for our society and a vast number of empirical studies of financial phenomenons, one quickly learns that the actual implementation is typically rather opaque. As graduate students, we were particularly surprised by the lack of public code for seminal papers or even textbooks on key concepts of financial economics. The lack of transparent code not only leads to numerous replication efforts (and their failures), but it also constitutes a waste of resources on problems that have already been solved by countless others in secrecy.

This book aims to lift the curtain on reproducible finance by providing a fully transparent code base for many common financial applications. We hope to inspire others to share their code publicly and take part in our journey towards more reproducible research in the future.

Who should read this book?

We write this book for three audiences:

- Students who want to acquire the basic tools required to conduct financial research ranging from undergrad to graduate level. The book's structure is simple enough such that the material is sufficient for self-study purposes.
- Instructors who look for materials to teach in empirical finance courses. We provide plenty of examples and (hopefully) intuitive explanations which can easily be adjusted or expanded.

- Data analysts or statisticians who work on issues pertaining to financial data and need practical tools to do so.

Our book is close in spirit to Regenstein Jr (2018) and Coqueret and Guida (2020) in that we provide fully reproducible code for useful applications and methods in finance.

The book "Reproducible Finance with R" provides an excellent introduction and discussion of different tools (based on the tidyverse, tidyquant, and xts) for standard applications in finance (e.g., how to compute returns and sample standard deviations of a time series of stock returns). Our book, in contrast, has a clear focus on applications of state-of-the-art for academic research in finance. We thus fill a niche that allows aspiring researchers or instructors to rely on a well-designed code base.

The book "Machine Learning for Factor Investing" is a great compendium to our book with respect to applications related to return prediction and portfolio formation. The book primarily targets practitioners and has a hands-on focus. Our book, in contrast, relies on the typical databases used in financial research and focuses on the preparation of such datasets for academic applications. In addition, our chapter on machine learning focuses on factor selection instead of return prediction.

What will you learn?

The book is currently divided into 5 parts:

- Chapter 1 introduces you to important concepts around which our approach to Tidy Finance revolves.
- Chapter 2 provides tools to organize your data and prepare the most common data sets used in financial research: CRSP and Compustat. We reuse the data from this chapter in all following chapters.
- Chapters 3-7 deal with key concepts of empirical asset pricing such as beta estimation, portfolio sorts, and performance analysis.
- Chapters 8-9 apply machine learning methods to problems in factor selection and option pricing.
- Chapters 10-11 provide approaches for parametric, constrained portfolio optimization, and backtesting procedures.

Each chapter is self-contained and can be read individually. Yet the data chapter provides important background necessary for the data management in subsequent chapters. The number of chapters and covered content is subject to change as we will introduce additional material in the near future.

What won't you learn?

This book is about empirical work. While we assume only basic knowledge in statistics and econometrics, we do not provide detailed treatments of the underlying theoretical models or methods applied in this book. Instead, you find references to the seminal academic work in journal articles and to more detailed treatments. We believe that our comparative advantage is to provide a thorough implementation of portfolio sorts, backtesting procedures, machine learning methods, or other related topics in empirical finance and enrich these implementations with discussions of the needy-greedy choices you face while conducting empirical analyses. We hence refrain from deriving theoretical models or discussing the statistical properties of well-established tools.

Why R?

We believe that R is among the best choices for a programming language in the area of finance. Some of our favorite features include:

- R is free and open-source so that you can use it in academic and professional contexts.
- A diverse and active online community works on a broad range of tools.
- A massive set of actively maintained packages for all kinds of applications exists, e.g., data manipulation, visualization, machine learning, etc.
- Powerful tools for communication, e.g., Rmarkdown and shiny, are readily available.
- RStudio is one of the best development environments for interactive data analysis.
- Strong foundations of functional programming are provided.
- Smooth integration with other programming languages, e.g., SQL, Python, C, C++, Fortran, etc.

For more information, we refer to Wickham et al. (2019a).

Why tidy?

As you start working with data, you quickly realize that you spend a lot of time reading, cleaning, and transforming your data. In fact, it is often said that more than 80% of data analysis is spent on preparing data. By *tidying data*, we want to structure data sets to facilitate further analyses. As Wickham (2014) puts it:

> [T]idy datasets are all alike, but every messy dataset is messy in its own way. Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).

In its essence, tidy data follows these three principles:

1. Every column is a variable.

2. Every row is an observation.
3. Every cell is a single value.

Throughout this book, we try to follow these principles as best as we can. If you want to learn more about tidy data principles in an informal manner, we refer you to this vignette[1].

In addition to the data layer, there are also tidy coding principles outlined in the tidy tools manifesto[2] that we try to follow:

1. Reuse existing data structures.
2. Compose simple functions with the pipe.
3. Embrace functional programming.
4. Design for humans.

In particular, we heavily draw on a set of packages called the `tidyverse`[3] (Wickham et al., 2019a). The `tidyverse` is a consistent set of packages for all data analysis tasks, ranging from importing and wrangling to visualizing and modeling data with the same grammar. In addition to explicit tidy principles, the `tidyverse` has further benefits: (i) if you master one package, it is easier to master others, and (ii) the core packages are developed and maintained by the Public Benefit Company RStudio, Inc. The core packages contained in the `tidyverse` (Wickham et al., 2019b) are: `lubridate` (Grolemund and Wickham, 2011), `dplyr` (Wickham et al., 2022a), `tidyr`(Wickham and Girlich, 2022), `readr` (Wickham et al., 2022b), `purrr` (Henry and Wickham, 2020), `tibble` (Müller and Wickham, 2022), `stringr` (Wickham, 2019), and `forcats` (Wickham, 2021).

Throughout the book we use the pipe `|>`, a powerful tool to clearly express a sequence of operations. Readers familiar with the `tidyverse` may be used to the predecessor `%>%` by the `magrittr` package. For simple cases `|>` and `%>%` behave identically, however, we follow Hadley Wickhams note “the main advantage of `|>` is that it does less than `%>%`, i.e. it has what is important about the pipe with less of what is not important.” For a more thorough discussion on the subtle differences, we refer to the second edition[4] of Wickham and Grolemund (2016).

Prerequisites

Before we continue, make sure you have all the software you need for this book:

- Install R and RStudio[5]. To get a walk-through of the installation for every major operating system, follow the steps outlined in this summary[6]. The whole pro-

[1] https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html
[2] https://tidyverse.tidyverse.org/articles/manifesto.html
[3] https://tidyverse.tidyverse.org/index.html
[4] https://r4ds.hadley.nz/workflow-pipes.html
[5] https://rstudio-education.github.io/hopr/starting.html#starting
[6] https://rstudio-education.github.io/hopr/starting.html#starthng

cess should be done in a few clicks. If you wonder about the difference: R is an open-source language and environment for statistical computing and graphics, free to download and use. While R runs the computations, RStudio is an integrated development environment that provides an interface by adding many convenient features and tools. We suggest doing all the coding in RStudio.
- Open RStudio and install the `tidyverse`[7]. Not sure how it works? You find helpful information on how to install packages in this brief summary[8].

If you are new to R, we recommend starting with the following sources:

- A very gentle and good introduction into the workings of R can be found in the form of the weighted dice project[9]. Once you are done setting up R on your machine, try to follow the instructions in this project.
- The main book on the `tidyverse`, Wickham and Grolemund (2016) is available online and for free: R for Data Science[10] by Hadley Wickham and Garrett Grolemund explains the majority of the tools we use in our book.
- If you are an instructor searching for effectively teach R and data science methods, we recommend to take a look on the excellent data science toolbox[11] by Mine Cetinkaya-Rundel[12].
- RStudio provides a range of excellent cheat sheets[13] with abundand information on how to use the `tidyverse` packages.

About the authors

We met at the Vienna Graduate School of Finance[14] from which each of us graduated with a different focus but a shared passion: coding with R. We continue to sharpen our R skills as part of our current occupations:

- Christoph Scheuch[15] is the Director of Product at the social trading platform wikifolio.com[16] where he is responsible for product planning, execution, and monitoring. He also manages a team of data scientists to analyze user behavior and develop new products.
- Stefan Voigt[17] is an Assistant Professor of Finance at the Department of Economics at the University in Copenhagen[18] and a research fellow at the

[7] https://tidyverse.tidyverse.org/
[8] https://rstudio-education.github.io/hopr/packages2.html
[9] https://rstudio-education.github.io/hopr/project-1-weighted-dice.html
[10] https://r4ds.had.co.nz/introduction.html
[11] https://datasciencebox.org/
[12] https://mine-cr.com/about/
[13] https://www.rstudio.com/resources/cheatsheets/
[14] https://www.vgsf.ac.at/
[15] https://christophscheuch.github.io/
[16] https://www.wikifolio.com/
[17] https://voigtstefan.me/
[18] https://www.economics.ku.dk/

Danish Finance Institute[19]. His research focuses on blockchain technology, high-frequency trading, and financial econometrics. Stefan teaches parts of this book in his courses on empirical finance.
- Patrick Weiss[20] is a Post-Doc at the Vienna University of Economics and Business[21]. His research centers around the intersection between asset pricing and corporate finance.

License

Colophon

This book was written in RStudio using bookdown. The website is hosted with github pages and automatically updated after every commit. The complete source is available from GitHub[24]. We generated all plots in this book using `ggplot2` and its classic dark-on-light theme (`theme_bw()`).

This version of the book was built with R version 4.2.1 (2022-06-23, Funny-Looking Kid) and the following packages:

[19] https://danishfinanceinstitute.dk/
[20] https://sites.google.com/view/patrick-weiss
[21] https://www.wu.ac.at/en/
[22] https://creativecommons.org/licenses/by-nc-sa/4.0/
[23] https://creativecommons.org/publicdomain/zero/1.0/
[24] www.github.com/voigtstefan/tidy_finance

Package	Version
bookdown	0.27
dplyr	1.0.9
forcats	0.5.1
ggplot2	3.3.6
jsonlite	1.8.0
kableExtra	1.3.4
knitr	1.39
purrr	0.3.4
readr	2.1.2
renv	0.15.5
rmarkdown	2.14
stringr	1.4.0
tibble	3.1.7
tidyr	1.2.0

Bibliography

Coqueret, G. and Guida, T. (2020). *Machine Learning for Factor Investing: R Version*. Chapman and Hall/CRC.

Grolemund, G. and Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25.

Henry, L. and Wickham, H. (2020). *purrr: Functional Programming Tools*. R package version 0.3.4.

Müller, K. and Wickham, H. (2022). *tibble: Simple Data Frames*. R package version 3.1.7.

Regenstein Jr, J. K. (2018). *Reproducible finance with R: Code flows and shiny apps for portfolio analysis*. Chapman and Hall/CRC.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(1):1–23.

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.

Wickham, H. (2021). *forcats: Tools for Working with Categorical Variables (Factors)*. R package version 0.5.1.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019a). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019b). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wickham, H., François, R., Henry, L., and Müller, K. (2022a). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.9.

Wickham, H. and Girlich, M. (2022). *tidyr: Tidy Messy Data*. R package version 1.2.0.

Wickham, H. and Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".

Wickham, H., Hester, J., and Bryan, J. (2022b). *readr: Read Rectangular Text Data*. R package version 2.1.2.