Brad Boehmke & Brandon Greenwell

Hands-on Machine Learning with R

To my son,

without whom I should have finished this book two years earlier

Contents

List of Tables

List of Figures

Preface

Note to readers: this text is a work in progress. It will eventually be published by Chapman & Hall/CRC. Prior to the formal copyediting process, we wanted to open it up to public review to get feedback on the content. Any feedback would be greatly appreciated and can be given at https://github.com/koalaverse/hands-on-machine-learning-with-r/issues. Public reviewers that help improve this book will be recognized in the Acknowledge Section.

Welcome to *Hands-on Machine Learning with R*. This book provides hands-on modules for many of the most common machine learning methods to include:

- Generalized low rank models
- Clustering algorithms
- Autoencoders
- Regularized models
- Random forests
- Gradient boosting machines
- Deep neural networks
- Stacking / super learners
- and more!

You will learn how to build and tune these various models with R packages that have been tested and approved due to their ability to scale well. However, our motivation in almost every case is to describe the techniques in a way that helps develop intuition for its strengths and weaknesses. For the most part, we minimize mathematical complexity when possible but also provide resources to get deeper into the details if desired.

Who should read this

We intend this work to be a practitioner's guide to the machine learning process and a place where one can come to learn about the approach and to gain intuition about the many commonly used, modern, and powerful methods accepted in the machine learning community. If you are familiar with the analytic methodologies, this book may still serve as a reference for how to work with the various R packages for implementation. While an abundance of videos, blog posts, and tutorials exist online, we have long been frustrated by the lack of consistency, completeness, and bias towards singular packages for implementation. This is what inspired this book.

This book is not meant to be an introduction to R or to programming in general; as we assume the reader has familiarity with the R language to include defining functions, managing R objects, controlling the flow of a program, and other basic tasks. If not, we would refer you to R for Data Science[1] (Wickham and Grolemund, 2016) to learn the fundamentals of data science with R such as importing, cleaning, transforming, visualizing, and exploring your data. For those looking to advance their R programming skills and knowledge of the languge, we would refer you to Advanced R[2] (Wickham, 2014). Nor is this book designed to be a deep dive into the theory and math underpinning machine learning algorithms. Several books already exist that do great justice in this arena (i.e. Elements of Statistical Learning[3] (Friedman et al., 2001), Computer Age Statistical Inference[4] (Efron and Hastie, 2016), Deep Learning[5] (Goodfellow et al., 2016)).

Instead, this book is meant to help R users learn to use the machine learning stack within R, which includes using various R packages such as **glmnet**, **h2o**, **ranger**, **xgboost**, **lime**, and others to effectively model and gain insight from your data. The book favors a hands-on approach, growing an intuitive understanding of machine learning through concrete examples and just a little bit of theory. While you can read this book without opening R, we highly recommend you experiment with the code examples provided throughout.

[1] http://r4ds.had.co.nz/index.html

[2] http://adv-r.had.co.nz/

[3] https://web.stanford.edu/~hastie/ElemStatLearn/

[4] https://web.stanford.edu/~hastie/CASI/

[5] http://www.deeplearningbook.org/

Why R

R has emerged over the last couple decades as a first-class tool for scientific computing tasks, and has been a consistent leader in implementing statistical methodologies for analyzing data. The usefulness of R for data science stems from the large, active, and growing ecosystem of third-party packages: **tidyverse** for common data analysis activities; **h2o**, **ranger**, **xgboost**, and others for fast and scalable machine learning; **iml**, **pdp**, **vip**, and others for machine learning interpretability; and many more tools will be mentioned throughout the pages that follow.

Conventions used in this book

The following typographical conventions are used in this book:

- ***strong italic***: indicates new terms,
- **bold**: indicates package & file names,
- `inline code`: monospaced highlighted text indicates functions or other commands that could be typed literally by the user,
- code chunk: indicates commands or other text that could be typed literally by the user

```
1 + 2
## [1] 3
```

In addition to the general text used throughout, you will notice the following code chunks with images, which signify:

Signifies a tip or suggestion

Signifies a general note

✋ Signifies a warning or caution

Additional resources

There are many great resources available to learn about machine learning. Throughout the chapters we try to include many of the resources that we have found extremely useful for digging deeper into the methodology and applying with code. However, due to print restrictions, the hard copy version of this book limits the concepts and methods discussed. Online supplementary material exists at https://github.com/koalaverse/hands-on-machine-learning-with-r. The additional material will accumulate over time and include extended chapter material (i.e., random forest package benchmarking) along with brand new content we couldn't fit in (i.e., random hyperparameter search). In addition, you can download the data used throughout the book, find teaching resources (i.e., slides and exercises), and more.

Feedback

Reader comments are greatly appreciated. To report errors or bugs please post an issue at https://github.com/koalaverse/hands-on-machine-learning-with-r/issues.

Acknowledgments

TBD

Software information

An online version of this book is available at http://bit.ly/HOML_with_R. The source of the book along with additional content is available at https://github.com/koalaverse/hands-on-machine-learning-with-r. The book is powered by https://bookdown.org which makes it easy to turn R markdown files into HTML, PDF, and EPUB.

This book was built with the following packages and R version. All code was executed on 2017 MacBook Pro with a 2.9 GHz Intel Core i7 processor, 16 GB of memory, 2133 MHz speed, and double data rate synchronous dynamic random access memory (DDR3).

```r
# packages used
pkgs <- c(
  "AmesHousing",
  "bookdown",
  "caret",
  "cluster",
  "DALEX",
  "data.table",
  "dplyr",
  "dslabs",
  "e1071",
  "earth",
  "emo",
  "extracat",
  "factoextra",
  "ggplot2",
  "gbm",
  "glmnet",
  "h2o",
  "iml",
  "ipred",
  "keras",
  "kernlab",
  "MASS",
  "mclust",
  "mlbench",
  "pBrackets",
  "pdp",
  "pls",
  "pROC",
  "purrr",
  "ranger",
  "recipes",
  "reshape2",
  "ROCR",
  "rpart",
  "rpart.plot",
  "rsample",
  "tfruns",
  "tfestimators",
  "vip",
  "xgboost"
)

# package & session info
sessioninfo::session_info(pkgs)
#> - Session info ---------------------------------------
#>  setting  value
#>  version  R version 3.6.0 (2019-04-26)
#>  os       macOS Sierra 10.12.6
#>  system   x86_64, darwin15.6.0
#>  ui       RStudio
#>  language (EN)
#>  collate  en_US.UTF-8
#>  ctype    en_US.UTF-8
#>  tz       America/New_York
#>  date     2019-06-25
#>
#> - Packages -------------------------------------------
#>  ! package       * version    date       lib
#>    abind           1.4-5      2016-07-21 [1]
#>    AmesHousing     0.0.3      2017-12-17 [1]
#>    assertthat      0.2.1      2019-03-21 [1]
#>    backports       1.1.4      2019-04-10 [1]
#>    base64enc       0.1-3      2015-07-28 [1]
#>    BH              1.69.0-1   2019-01-07 [1]
#>    bitops          1.0-6      2013-08-17 [1]
#>    bookdown        0.11       2019-05-28 [1]
#>    boot            1.3-22     2019-04-02 [1]
#>    car             3.0-3      2019-05-27 [1]
#>    carData         3.0-2      2018-09-30 [1]
#>    caret           6.0-84     2019-04-27 [1]
#>    caTools         1.17.1.2   2019-03-06 [1]
#>    cellranger      1.1.0      2016-07-27 [1]
#>    checkmate       1.9.3      2019-05-03 [1]
#>    class           7.3-15     2019-01-01 [1]
#>    cli             1.1.0      2019-03-19 [1]
#>    clipr           0.6.0      2019-04-15 [1]
#>    cluster         2.0.8      2019-04-05 [1]
#>    codetools       0.2-16     2018-12-24 [1]
#>    colorspace      1.4-1      2019-03-18 [1]
#>    config          0.3        2018-03-27 [1]
#>    cowplot         0.9.4      2019-01-08 [1]
#>    crayon          1.3.4      2017-09-16 [1]
#>    curl            3.3        2019-01-10 [1]
#>    DALEX           0.3.0      2019-03-25 [1]
#>    data.table      1.12.2     2019-04-07 [1]
#>    dendextend      1.12.0     2019-05-11 [1]
#>    digest          0.6.19     2019-05-20 [1]
#>    dplyr           0.8.1      2019-05-14 [1]
#>    dslabs          0.5.2      2018-12-19 [1]
#>    e1071           1.7-1      2019-03-19 [1]
#>    earth           5.1.1      2019-04-12 [1]
#>    ellipse         0.4.1      2018-01-05 [1]
#>    ellipsis        0.1.0      2019-02-19 [1]
#>    emo             0.0.0.9000 2019-05-03 [1]
#>    evaluate        0.14       2019-05-28 [1]
#>  R extracat        <NA>       <NA>       [?]
#>    factoextra      1.0.5      2017-08-22 [1]
#>    FactoMineR      1.41       2018-05-04 [1]
#>    fansi           0.4.0      2018-10-05 [1]
#>    flashClust      1.01-2     2012-08-21 [1]
#>    forcats         0.4.0      2019-02-17 [1]
#>    foreach         1.4.4      2017-12-12 [1]
#>    foreign         0.8-71     2018-07-20 [1]
#>    forge           0.2.0      2019-02-26 [1]
#>    Formula         1.2-3      2018-05-03 [1]
#>    gbm             2.1.5      2019-01-14 [1]
#>    gdata           2.18.0     2017-06-06 [1]
#>    generics        0.0.2      2018-11-29 [1]
#>    ggplot2         3.1.1      2019-04-07 [1]
#>    ggpubr          0.2        2018-11-15 [1]
#>    ggrepel         0.8.1      2019-05-07 [1]
#>    ggsci           2.9        2018-05-14 [1]
#>    ggsignif        0.5.0      2019-02-20 [1]
#>    glmnet          2.0-16     2018-04-02 [1]
#>    glue            1.3.1.9000 2019-05-03 [1]
#>    gower           0.2.0      2019-03-07 [1]
#>    gplots          3.0.1.1    2019-01-27 [1]
#>    gridExtra       2.3        2017-09-09 [1]
#>    gtable          0.3.0      2019-03-25 [1]
#>    gtools          3.8.1      2018-06-26 [1]
#>    h2o             3.22.1.1   2019-01-10 [1]
#>    haven           2.1.0      2019-02-19 [1]
#>    highr           0.8        2019-03-20 [1]
#>    hms             0.4.2      2018-03-10 [1]
#>    htmltools       0.3.6      2017-04-28 [1]
#>    iml             0.9.0      2019-02-05 [1]
#>    inum            1.0-1      2019-04-25 [1]
#>    ipred           0.9-9      2019-04-28 [1]
#>    iterators       1.0.10     2018-07-13 [1]
#>    jsonlite        1.6        2018-12-07 [1]
#>    keras           2.2.4.1    2019-04-05 [1]
#>    kernlab         0.9-27     2018-08-10 [1]
#>    KernSmooth      2.23-15    2015-06-29 [1]
#>    knitr           1.23       2019-05-18 [1]
#>    labeling        0.3        2014-08-23 [1]
#>    lattice         0.20-38    2018-11-04 [1]
#>    lava            1.6.5      2019-02-12 [1]
#>    lazyeval        0.2.2      2019-03-15 [1]
#>    leaps           3.0        2017-01-10 [1]
#>    libcoin         1.0-4      2019-02-28 [1]
#>    lme4            1.1-21     2019-03-05 [1]
#>    lubridate       1.7.4      2018-04-11 [1]
#>    magrittr        1.5        2014-11-22 [1]
#>    maptools        0.9-5      2019-02-18 [1]
#>    markdown        1.0        2019-06-07 [1]
#>    MASS            7.3-51.4   2019-03-31 [1]
#>    Matrix          1.2-17     2019-03-22 [1]
#>    MatrixModels    0.4-1      2015-08-22 [1]
#>    mclust          5.4.3      2019-03-14 [1]
#>    Metrics         0.1.4      2018-07-09 [1]
#>    mgcv            1.8-28     2019-03-21 [1]
#>    mime            0.7        2019-06-11 [1]
#>    minqa           1.2.4      2014-10-09 [1]
#>    mlbench         2.1-1      2012-07-10 [1]
#>    ModelMetrics    1.2.2      2018-11-03 [1]
#>    munsell         0.5.0      2018-06-12 [1]
#>    mvtnorm         1.0-10     2019-03-05 [1]
#>    nlme            3.1-139    2019-04-09 [1]
#>    nloptr          1.2.1      2018-10-03 [1]
#>    nnet            7.3-12     2016-02-02 [1]
#>    numDeriv        2016.8-1   2016-08-27 [1]
#>    openxlsx        4.1.0.1    2019-05-28 [1]
#>    partykit        1.2-3      2019-01-31 [1]
#>    pbkrtest        0.4-7      2017-03-15 [1]
#>    pBrackets       1.0        2014-10-17 [1]
#>    pdp             0.7.0      2018-08-27 [1]
#>    pillar          1.4.1      2019-05-28 [1]
#>    pkgconfig       2.0.2      2018-08-16 [1]
#>    plogr           0.2.0      2018-03-25 [1]
#>    plotmo          3.5.4      2019-04-06 [1]
#>    plotrix         3.7-5      2019-04-07 [1]
#>    pls             2.7-1      2019-03-23 [1]
#>    plyr            1.8.4      2016-06-08 [1]
#>    polynom         1.4-0      2019-03-22 [1]
#>    prediction      0.3.6.2    2019-01-31 [1]
#>    prettyunits     1.0.2      2015-07-13 [1]
#>    pROC            1.14.0     2019-03-12 [1]
#>    processx        3.3.0      2019-03-10 [1]
#>    prodlim         2018.04.18 2018-04-18 [1]
#>    progress        1.2.2      2019-05-16 [1]
#>    ps              1.3.0      2018-12-21 [1]
#>    purrr           0.3.2      2019-03-15 [1]
#>    quantreg        5.38       2018-12-18 [1]
#>    R6              2.4.0      2019-02-14 [1]
#>    ranger          0.11.2     2019-03-07 [1]
#>    RColorBrewer    1.1-2      2014-12-07 [1]
#>    Rcpp            1.0.1      2019-03-17 [1]
#>    RcppEigen       0.3.3.5.0  2018-11-24 [1]
#>    RcppRoll        0.3.0      2018-06-05 [1]
#>    RCurl           1.95-4.12  2019-03-04 [1]
#>    readr           1.3.1      2018-12-21 [1]
#>    readxl          1.3.1      2019-03-13 [1]
#>    recipes         0.1.5      2019-03-21 [1]
#>    rematch         1.0.1      2016-04-21 [1]
#>    reshape2        1.4.3      2017-12-11 [1]
#>    reticulate      1.12       2019-04-12 [1]
#>    rio             0.5.16     2018-11-26 [1]
#>    rlang           0.3.4      2019-04-07 [1]
#>    rmarkdown       1.13       2019-05-22 [1]
#>    ROCR            1.0-7      2015-03-26 [1]
#>    rpart           4.1-15     2019-04-12 [1]
#>    rpart.plot      3.0.7      2019-04-12 [1]
#>    rsample         0.0.4      2019-01-07 [1]
#>    rstudioapi      0.10       2019-03-19 [1]
#>    scales          1.0.0      2018-08-09 [1]
#>    scatterplot3d   0.3-41     2018-03-14 [1]
#>    sp              1.3-1      2018-06-05 [1]
#>    SparseM         1.77       2017-04-23 [1]
#>    SQUAREM         2017.10-1  2017-10-07 [1]
#>    stringi         1.4.3      2019-03-12 [1]
#>    stringr         1.4.0      2019-02-10 [1]
#>    survival        2.44-1.1   2019-04-01 [1]
#>    TeachingDemos   2.10       2016-02-12 [1]
#>    tensorflow      1.13.1     2019-04-05 [1]
#>    tfestimators    1.9.1      2018-11-07 [1]
#>    tfruns          1.4        2018-08-25 [1]
#>    tibble          2.1.2      2019-05-29 [1]
#>    tidyr           0.8.3      2019-03-01 [1]
#>    tidyselect      0.2.5      2018-10-11 [1]
#>    timeDate        3043.102   2018-02-21 [1]
#>    tinytex         0.13       2019-05-14 [1]
#>    utf8            1.1.4      2018-05-24 [1]
#>    vctrs           0.1.0      2018-11-29 [1]
#>    vip             0.1.2.9000 2019-06-04 [1]
#>    viridis         0.5.1      2018-03-29 [1]
#>    viridisLite     0.3.0      2018-02-01 [1]
#>    whisker         0.3-2      2013-04-28 [1]
#>    withr           2.1.2      2018-03-15 [1]
#>    xfun            0.7        2019-05-14 [1]
#>    xgboost         0.82.1     2019-03-11 [1]
#>    yaImpute        1.0-31     2019-01-09 [1]
#>    yaml            2.2.0      2018-07-25 [1]
#>    zeallot         0.1.0      2018-01-28 [1]
#>    zip             2.0.2      2019-05-13 [1]
#>  source
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  Github (hadley/emo@02a5206)
#>  CRAN (R 3.6.0)
#>  <NA>
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  Github (tidyverse/glue@ea0edcb)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  Github (koalaverse/vip@9d537bb)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>  CRAN (R 3.6.0)
#>
#> [1] /Library/Frameworks/R.framework/Versions/3.6/Resources/library
#>
#>  R -- Package was removed from disk.
```

Bibliography

Efron, B. and Hastie, T. (2016). *Computer age statistical inference*, volume 5. Cambridge University Press.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Wickham, H. (2014). *Advanced R*. Chapman and Hall/CRC.

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