Preface

Background and motivation

The increasing volume and sophistication of data poses new challenges for analysts, who need to be able to transform complex data sets to answer important statistical questions. The widely-cited McKinsey & Company report stated that "by 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions." There is a pressing need for additional resources to train existing analysts as well as the next generation to be able to pose questions, suggest hypotheses, collect, transform, and analyze data, then communicate results. According to the online company ratings site *Glassdoor*, "data scientist" was the best job in America in 2016 [142].

Statistics can be defined as the science of learning from data [203]. Michael Jordan has described data science as the marriage of computational thinking and inferential thinking. Without the skills to be able to "wrangle" the increasingly rich and complex data that surround us, analysts will not be able to use these data to make better decisions.

New data technologies and database systems facilitate scraping and merging data from different sources and formats and restructuring it into a form suitable for analysis. State-of-the-art workflow tools foster well-documented and reproducible analysis. Modern statistical methods allow the analyst to fit and assess models as well as to undertake supervised or unsupervised learning to extract information. Contemporary data science requires tight integration of these statistical, computing, data-related, and communication skills.

The book is intended for readers to develop and reinforce the appropriate skills to tackle complex data science projects and "think with data" (as coined by Diane Lambert). The ability to solve problems using data is at the heart of our approach.

We feature a series of complex, real-world extended case studies and examples from a broad range of application areas, including politics, transportation, sports, environmental science, public health, social media, and entertainment. These rich data sets require the use of sophisticated data extraction techniques, modern data visualization approaches, and refined computational approaches.

It is impossible to cover all these topics in any level of detail within a single book: Many of the chapters could productively form the basis for a course or series of courses. Our goal is to lay a foundation for analysis of real-world data and to ensure that analysts see the power of statistics and data analysis. After reading this book, readers will have greatly expanded their skill set for working with these data, and should have a newfound confidence about their ability to learn new technologies on-the-fly.

Key role of technology

While many tools can be used effectively to undertake data science, and the technologies to undertake analyses are quickly changing, R and Python have emerged as two powerful and

extensible environments. While it is important for data scientists to be able to use multiple technologies for their analyses, we have chosen to focus on the use of R and RStudio to avoid cognitive overload. By use of a "Less Volume, More Creativity" approach [162], we intend to develop a small set of tools that can be mastered within the confines of a single semester and that facilitate sophisticated data management and exploration.

We take full advantage of the RStudio environment. This powerful and easy-to-use front end adds innumerable features to R including package support, code-completion, integrated help, a debugger, and other coding tools. In our experience, the use of RStudio dramatically increases the productivity of R users, and by tightly integrating reproducible analysis tools, helps avoid error-prone "cut-and-paste" workflows. Our students and colleagues find RStudio an extremely comfortable interface. No prior knowledge or experience with R or RStudio is required: we include an introduction within the Appendix.

We used a reproducible analysis system (`knitr`) to generate the example code and output in this book. Code extracted from these files is provided on the book's website. We provide a detailed discussion of the philosophy and use of these systems. In particular, we feel that the `knitr` and `markdown` packages for R, which are tightly integrated with RStudio, should become a part of every R user's toolbox. We can't imagine working on a project without them (and we've incorporated reproducibility into all of our courses).

Modern data science is a team sport. To be able to fully engage, analysts must be able to pose a question, seek out data to address it, ingest this into a computing environment, model and explore, then communicate results. This is an iterative process that requires a blend of statistics and computing skills.

Context is king for such questions, and we have structured the book to foster the parallel developments of statistical thinking, data-related skills, and communication. Each chapter focuses on a different extended example with diverse applications, while exercises allow for the development and refinement of the skills learned in that chapter.

Intended audiences

This book was originally conceived to support a one-semester, 13-week upper-level course in data science. We also intend that the book will be useful for more advanced students in related disciplines, or analysts who want to bolster their data science skills. The book is intended to be accessible to a general audience with some background in statistics (completion of an introductory statistics course).

In addition to many examples and extended case studies, the book incorporates exercises at the end of each chapter. Many of the exercises are quite open-ended, and are designed to allow students to explore their creativity in tackling data science questions.

The book has been structured with three main sections plus supplementary appendices. Part I provides an introduction to data science, an introduction to visualization, a foundation for data management (or 'wrangling'), and ethics. Part II extends key modeling notions including regression modeling, classification and prediction, statistical foundations, and simulation. Part III introduces more advanced topics, including interactive data visualization, SQL and relational databases, spatial data, text mining, and network science.

We conclude with appendices that introduce the book's R package, R and RStudio, key aspects of algorithmic thinking, reproducible analysis, a review of regression, and how to set up a local SQL database.

We have provided two indices: one organized by subject and the other organized by R function and package. In addition, the book features extensive cross-referencing (given the inherent connections between topics and approaches).

Website

The book website at `https://mdsr-book.github.io` includes the table of contents, subject and R indices, example datasets, code samples, exercises, additional activities, and a list of errata.

How to use this book

The material from this book has supported several courses to date at Amherst, Smith, and Macalester Colleges. This includes an intermediate course in data science (2013 and 2014 at Smith), an introductory course in data science (2016 at Smith), and a capstone course in advanced data analysis (2015 and 2016 at Amherst). The intermediate data science course required an introductory statistics course and some programming experience, and discussed much of the material in this book in one semester, culminating with an integrated final project [20]. The introductory data science course had no prerequisites and included the following subset of material:

- Data Visualization: three weeks, covering Chapters 2 and 3
- Data Wrangling: four weeks, covering Chapters 4 and 5
- Database Querying: two weeks, covering Chapter 12
- Spatial Data: two weeks, covering Chapter 14
- Text Mining: two weeks, covering Chapter 15

The capstone course covered the following material:

- Data Visualization: two weeks, covering Chapters 2, 3, and 11
- Data Wrangling: two weeks, covering Chapters 4 and 5
- Ethics: one week, covering Chapter 6
- Simulation: one week, covering Chapter 10
- Statistical Learning: two weeks, covering Chapters 8 and 9
- Databases: one week, covering Chapter 12 and Appendix F
- Text Mining: one week, covering Chapter 15
- Spatial Data: one week, covering Chapter 14
- Big Data: one week, covering Chapter 17

We anticipate that this book could serve as the primary text for a variety of other courses, with or without additional supplementary material.

The content in Part I—particularly the `ggplot2` visualization concepts presented in Chapter 3 and the `dplyr` data wrangling operations presented in Chapter 4—is fundamental and is assumed in Parts II and III. Each of the chapters in Part III are independent of each other and the material in Part II. Thus, while most instructors will want to cover most (if not all) of Part I in any course, the material in Parts II and III can be added with almost total freedom.

The material in Part II is designed to expose students with a beginner's understanding of statistics (i.e., basic inference and linear regression) to a richer world of statistical modeling and statistical inference.

Acknowledgments

We would like to thank John Kimmel at Informa CRC/Chapman and Hall for his support and guidance. We also thank Jim Albert, Nancy Boynton, Jon Caris, Mine Çetinkaya–Rundel, Jonathan Che, Patrick Frenett, Scott Gilman, Johanna Hardin, John Horton, Azka Javaid, Andrew Kim, Eunice Kim, Caroline Kusiak, Ken Kleinman, Priscilla (Wencong) Li, Amelia McNamara, Tasheena Narraido, Melody Owen, Randall Pruim, Tanya Riseman, Gabriel Sosa, Katie St. Clair, Amy Wagaman, Susan (Xiaofei) Wang, Hadley Wickham, J. J. Allaire and the RStudio developers, the anonymous reviewers, the Spring 2015 SDS192 class, the Fall 2016 STAT495 class, and many others for contributions to the R and RStudio environment, comments, guidance, and/or helpful suggestions on drafts of the manuscript.

Above all we greatly appreciate Cory, Maya, and Julia for their patience and support.

Northampton, MA and St. Paul, MN
December 2016