/// DETAILED CONTENTS

/// PREFACE

1 The goals of this book

The main goal of this book is to prepare students and other readers for the messy and exciting reality of working with data. As data on many aspects of life accumulate at unprecedented rates, there is an increasing need for people who can clean, manage, and analyze information. Secondarily, the book aims to encourage women and other underrepresented groups to consider data science careers. Representation of women has decreased in computer science and math fields since 2006, and gaps persist in data science fields by race and ethnicity, limiting the perspectives contributing to the growth of this field. Finally, this book aims to improve the quality of social science through the promotion of reproducible research practices. Science has been increasingly under scrutiny for use of questionable research practices, some of which can be overcome through well-formatted and documented code for data importing, cleaning, and analysis.

To reach all three of these goals, I employed several strategies, such as a narrative writing style, diversity in the main and supporting characters, a focus on social problems, use of publicly available data and open-source R statistical software, and demonstrations of reproducible research practices.

2 The audience for the book

This book was written with first-year statistics courses in the social sciences in mind. Often, these courses start with descriptive statistics and probability theory and continue through general and generalized linear modeling. Others who may find this book useful include people working with data who are interested in learning R for the first time, learning more R, or reinforcing and improving their statistical skills. General readers interested in data science might find this book to be an accessible introduction to one of the primary software packages and many of the foundational concepts used in data science.

3 The features of the book

3.1 A NARRATIVE APPROACH

The book includes an underlying storyline about three characters working together to learn statistics and R: Nancy, Kiara, and Leslie. Nancy and Kiara are data scientists, and Leslie is a student. The first chapter describes their initial meeting at an R-Ladies community event where they discuss the benefits and challenges of using R and decide to work together to learn R and statistics. The remaining chapters each start in a different setting with a conversation between the three introducing the statistical method of the chapter and the social problem they will address while

learning the method. The use of narrative serves at least two purposes. The first purpose is to provide an accessible and relatable way to introduce statistical topics that are often perceived as some combination of difficult and boring. Students start by reading a casual conversation among friends that gently transitions into more technical concerns. The second purpose is to show women in the roles of experts and learners of coding and statistics. Through dialogue, character development, and images, I portray scenarios of women collaborating to learn and apply data management and analysis skills to important social issues.

Each of the three main characters has expertise or interest in a different aspect of learning and applying R code and statistical concepts. Their expertise and interests are highlighted in three different types of boxed features that appear throughout the chapters:

- **Nancy's Fancy Code:** Nancy is interested in writing code and has expertise in more advanced or quirky R coding, which she shares with the other characters in this feature.
- **Kiara's Reproducibility Resource:** Kiara is interested in improving data science by ensuring her work and the work of her colleagues is reproducible through the use of good coding practice; she shares tips for improving reproducibility in this feature.
- **Leslie's Stats Stuff:** Leslie has interest and expertise in statistical theory and adds detail to their discussions through her explanations of statistical concepts in this feature.

3.2 A FOCUS ON SOCIAL PROBLEMS

Each chapter of the book focuses on a different problem from a social science or related field using publicly available data sources. One reason the social sciences are so interesting is because they help us understand and advocate for the world—the one we live in and the one we want to create. I've tried to choose compelling topics, including several with moral dilemmas that are in the news and in students' lives. Most readers should be able to find at least some of the chapters intriguing and relevant. This approach contrasts with textbooks that focus exclusively or predominantly on statistical theory, use simulated data, or choose data sources specifically to avoid controversial topics.

The topics for the chapters are as follows:

- Chapter 1: Marijuana legalization policy
- Chapter 2: Cancer screening for transgender patients
- Chapter 3: Gun deaths, gun use, gun manufacturing, and funding for gun research
- Chapter 4: Opioid deaths, opioid policy, and opioid treatment facilities
- Chapter 5: Perceptions about voter registration and mandatory voting
- Chapter 6: Blood pressure
- Chapter 7: Time spent using technology
- Chapter 8: Global access to clean water, sanitation, and education
- Chapter 9: Distance to needle exchanges
- Chapter 10: The digital divide and library use
- Chapter 11: Representation of women in data science careers

3.3 USE OF PUBLICLY AVAILABLE DATA SETS

Each chapter uses one or more publicly available data sets, and most chapters include instructions for importing the data directly into R from the original online location to encourage the use of reproducible research practices.

Many textbooks in statistics use simulated data or data that have been pre-cleaned. This book takes a different approach in order to provide the audience experience with data situations they are likely to encounter outside of the learning environment. Simulated and pre-cleaned data sources have advantages, including being useful for clearly demonstrating what it means to meet assumptions for statistical models or fail assumptions in specific ways. However, use of pre-cleaned data can set unrealistic expectations of how most data actually look outside the classroom. Likewise, simulated data can reinforce the stereotype that learning and using statistics are not only difficult but also disconnected from the real world. My book tries to overcome these challenges by using real data that address compelling social problems.

Admittedly, there are challenges to my approach. It is decidedly more difficult to demonstrate some concepts and to meet statistical model assumptions with real-world data. That's life. The challenges with this approach mimic the challenges of data science. Moreover, readers will be introduced to strategies for thinking like a data scientist to identify and deal with common obstacles.

The data sets used are as follows:

- 2016 General Social Survey (Chapter 1)
- 2014 Behavioral Risk Factor Surveillance Survey (Chapter 2)
- 2016 Federal Bureau of Investigation homicide data table (Chapter 3)
- 2011–2012 National Health and Nutrition Examination Survey (Chapter 3)
- 2017 Kaiser Family Foundation state opioid policy data (Chapter 4)
- 2017 American Foundation for AIDS Research distance to substance use treatment facility data (Chapter 4)
- 2017 Pew Research Center voting perceptions data (Chapter 5)
- 2015–2016 National Health and Nutrition Examination Survey (Chapter 6)
- 2018 General Social Survey (Chapter 7)
- 2015 World Health Organization data on access to water and sanitation (Chapter 8)
- 2015 United Nations Educational, Scientific, and Cultural Organization data on education (Chapter 8)
- 2017 American Foundation for AIDS Research distance to needle exchange data (Chapter 9)
- 2016 Pew Research Center library use data (Chapter 10)
- 2017 National Science Foundation Scientists and Engineers Statistical Data System (Chapter 11)

3.4 USE OF R STATISTICAL SOFTWARE

R is a coding language used to conduct statistical analyses and to create visual displays of data. The options available in R for analysis and graphing are competitive with, and in some cases surpass, the Statistical Package for the Social Sciences (SPSS) and Statistical Analysis System (SAS) software packages. However, R statistical software is free and open source. Anyone can contribute a "package" to R and, if it is accepted to the Comprehensive R Archive Network (CRAN) (https://cran.r-project.org/

submit.html), it becomes available for all R users worldwide to access and use. As of July 2019, there are approximately 15,000 user-contributed packages available on the CRAN, with most packages being designed to conduct some specific type of analysis. The use of R has been growing and, by some metrics, has caught up to or surpassed SPSS and SAS when it comes to data-related jobs (Muenchen, 2019). Because R is free, it allows people and organizations with fewer resources to access and use a high-quality data science tool (Krishnaswamy & Marinova, 2012; Sullivan, 2011). Consequently, the software is more egalitarian and inclusive, creating opportunities for collaborations and contributions to emerge from less privileged nations and individuals.

Among the unique qualities of R is the community that has formed around its use. On any given day, the #rstats hashtag on Twitter includes several hundred tweets from people making suggestions, asking questions, and posting R memes and jokes. The R community is also highly focused on diversity. For example, groups like R-Ladies Global support and encourage underrepresented voices in R (Daish, Frick, LeDell, de Queiroz, & Vitolo, 2019).

3.5 INCLUSION OF DIVERSE CHARACTERS, AUTHORS, AND ARTWORK

The book emphasizes diversity and inclusion in several ways. First, to address the underrepresentation of women in math and computer science fields, the underlying story features three women. In addition, the final chapter of the book examines data related to job type, job satisfaction, and sex, with a focus on computer science and math. The chapter portrays the student character coming to terms with data on employment and job satisfaction as she ponders a data science career. Third, when relevant and available, I cited women and authors from countries underrepresented in the statistical literature. Specifically, when two equally citable sources supporting the same concept were available, I opted for the underrepresented authors as determined by a commonsense reading of names and affiliations and, in some cases, searching for and reading of online profiles. Finally, the main and supporting characters in the artwork included throughout the text include diverse representation of race, ethnicity, and sex. If women and students of color can literally "see" themselves in the characters, they may find data jobs more appealing and feasible (Herrmann et al., 2016; Johnson, 2011).

3.6 EMPHASIS ON REPRODUCIBLE RESEARCH PRACTICES

Science is increasingly under scrutiny for the use of questionable research practices that threaten the quality of the evidence underlying decisions that impact lives (Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016; Steen, Casadevall, & Fang, 2013). Among the many strategies that could improve the quality of the science is the use of reproducible practices during data management and analysis that allow other scientists to reproduce the work (Harris et al., 2019; Harris, Wondmeneh, Zhao, & Leider, 2019). This text suggests and demonstrates choices that contribute to reproducibility when importing, cleaning, managing, and analyzing data. Reproducibility is mentioned throughout the text by one of the characters, and most chapters have a resource box that offers one or more strategies for ensuring that data management and analysis tasks are performed in a reproducible way. For example, one of the sections in the first chapter describes and demonstrates how to pick a consistent and clear way to name and format constants, variables, and functions. This strategy results in code that is more readable for humans, which improves code comprehension and facilitates the reuse of code to verify results with the same data or new data.

4 Book website

The book website at **edge.sagepub.com/harris1e** includes the following resources:

INSTRUCTOR TEACHING SITE: edge.sagepub.com/harris1e

SAGE edge for instructors supports your teaching by making it easy to integrate quality content and create a rich learning environment for students with

- A **password-protected site** for complete and protected access to all text-specific instructor resources;
- **Test banks** that provide a diverse range of ready-to-use options that save you time—you can also easily edit any question and/or insert your own personalized questions;
- **Tutorial videos** produced exclusively for this text that demonstrate **how to use R** to conduct key statistical tests using real-world data;
- **Editable, chapter-specific PowerPoint® slides** that offer complete flexibility for creating a multimedia presentation for your course;
- **Downloadable Coder (beginner/intermediate) and Hacker (advanced) exercises** from the book that can be used as homework or labs—students can take the **multiple choice pre-test questions** electronically first to check their level;
- **Downloadable data files and R code available** for use with the book and exercises;
- **Solutions** to selected in-text exercises;
- **Instructor Ideas for Gamification** compiled by the author, offered for those who want to gamify their course; and
- **Full-color figures** from the book available for download.

STUDENT STUDY SITE: edge.sagepub.com/harris1e

SAGE edge for students enhances learning, it's easy to use, and it offers

- An **open-access site** that makes it easy for students to maximize their study time, anywhere, anytime;
- **Tutorial videos** produced exclusively for this text that demonstrate **how to use R** to conduct key statistical tests using real-world data;
- **Downloadable Coder (beginner/intermediate) and Hacker (advanced) exercises** from the book—students can take the **multiple choice pre-test questions** electronically first to check their level; and
- **Downloadable data files and R code available** for use with the book and exercises.

5 Acknowledgments

Many people helped in major and minor and in obvious and subtle ways throughout the development of this text. Leanne Waugh read well over half of the book, found typos, and identified places where an explanation or the storyline was incomplete or unclear. Amy Sklansky suggested major improvements to the narrative and chapter openings and taught me the difference between writing a story and writing stage directions. Shelly Cooper read several chapters and offered suggestions for fixing and improving the code, including suggesting a useful figure that is a combination of a boxplot, a violin plot, and a scatterplot (check it out in Chapter 9). Scott Harris used his expert copyediting skills to help me with grammar, punctuation, and an unusual amount of repeated words, even for me me. Bobbi Carothers offered

helpful suggestions on very early drafts of the first few chapters, providing a more solid foundation for the rest of the work. Leslie Hinyard answered too many text messages to count. Chelsea West, Chris Prener, Angelique Zeringue, Sarah Van Alsten, Bryan Newman, Alexis Duncan, Kristen Ruckdashel, Joe Steensma, Kim Johnson, Robert Singer, Paaige Turner, Doug Luke, Ellen Mrazek, and some people I do not know on Twitter also gave helpful opinions and suggestions. Thank you all!

At SAGE, Helen Salmon and Chelsea Neve were the epitome of patience, enthusiasm, and professionalism. The reviewers in the following list had great ideas and enthusiastic suggestions that improved the work immensely:

Jennifer Bachner, *Johns Hopkins University*

Matthew C. Bell, *Santa Clara University*

Patrick Bolger, *Texas A&M University*

William J. Bosl, *University of San Francisco*

Joseph Nathan Cohen, *City University of New York–Queens College*

Daniel Conroy-Beam, *University of California, Santa Barbara*

Gabriel I. Cook, *Claremont McKenna College*

James J. Cortright, *University of Wisconsin–River Falls*

Jacqueline S. Craven, *Delta State University*

Todd Daniel, *Missouri State University*

Michael Erickson, *Hawaii Pacific University*

Marte Fallshore, *Central Washington University*

Sylvain Fiset, *Université de Moncton, Edmundston*

Jonathan S. Hack, *Harvard Law School*

Johannes Karreth, *Ursinus College*

George Kikuchi, *California State University, Fresno*

Brandon LeBeau, *University of Iowa*

Michael S. Lynch, *University of Georgia*

Michael E. J. Masson, *University of Victoria*

Avery McIntosh, *Boston University*

Matthew R. Miles, *Brigham Young University, Idaho*

Maura J. Mills, *University of Alabama*

Mary Moore, *Colorado State University–Global*

Derek Mueller, *Carleton University*

David A. M. Peterson, *Iowa State University*

Matthew D. Phillips, *UNC Charlotte*

Darrin L. Rogers, *State University of New York at Fredonia*

Samantha Seals, *University of West Florida*

Yi Shao, *Oklahoma City University*

Ches Thurber, *Northern Illinois University*

Drew Tyre, *University of Nebraska–Lincoln*

Mary Beth Zeni, *Ursuline College*

The artists, Rob Schuster and Rose Storey, did an excellent job producing graphics that fit with the story and goals of the work. My thanks to Hadley Wickham for permission to include one of his tweets as a figure, and to Hao Zhu and Matt Dowle for permission to include their `kableExtra` and `data.table` hex stickers, respectively, in the cover art. Thank you to the data teams at the General Social Survey; the National Health and Nutrition Examination Survey; the World Health Organization; the United Nations Educational, Scientific and Cultural Organization; the Federal Bureau of Investigation; the Pew Research Center; the Kaiser Family Foundation; the Foundation for AIDS Research; and the National Science Foundation. Special thanks go to Dr. David Stark and Dr. Nigam Shaw, who answered my emails, sent me a spreadsheet of the data from their 2017 article (Stark & Shah, 2017), and sent the GitHub location of the data and R code for the paper that was the inspiration for Chapter 3 (https://github.com/davidestark/gun-violence-research/).

Finally, all the hard work in the world would not have resulted in much had I not had the great fortune of having a math-teaching mom with an unfathomably big heart; too many smart, fun, and supportive friends to count (you know who you are, or, if you don't, you'll be receiving a text to confirm!); and a spouse who is an inspiration and true partner.