OXFORD

INSIGHTS FROM DATA WITH R

AN INTRODUCTION FOR THE LIFE AND ENVIRONMENTAL SCIENCES

OWEN L. PETCHEY

ANDREW P. BECKERMAN

NATALIE COOPER

DYLAN Z. CHILDS

Preface

Overview

Many activities, including experiments, surveys, clinical trials, and fieldwork, generate data. These data provide insights—intuitions and conclusions that come from identifying patterns in data. Insights are critical for answering questions, solving problems, guiding decisions, and formulating strategy. But getting insights from data, and doing so efficiently, reliably, and confidently, does not come easily. Yet getting insights from data is a foundational skill for all scientists.

Insights from Data with R is for life and environmental science undergraduates (though may also help anyone beginning in their learning about data analysis), and for their instructors to teach alongside. It is not about statistics per se, but about that initial transition from having collected data as part of a project to that first, and so satisfying, realization that there is a pattern in your data. It combines the elements of the successful undergraduate data analysis courses of Petchey at the University of Zürich and of Childs at the University of Sheffield, the 'Introduction to R' courses taught internationally for 15 years by all four authors, and the book *Getting Started with R: An Introduction for Biologists*, second edition, by Beckerman, Childs, and Petchey (2017), all using R with the RStudio platform.

Insights (*from Data with R*) first covers what insights are and why they're so important, and moves on to discuss features of data that can make it hard or easy to gain insights. It then describes how to obtain insights

Insights from data with R: An Introduction for the Life and Environmental Sciences.
Owen L. Petchey, Andrew P. Beckerman, Natalie Cooper and Dylan Z. Childs, Oxford University Press (2021).
DOI: 10.1093/oso/9780198849810.001.0001

from data. Obtaining them involves knowing what you are aiming for, and then a whole lot of preparation, importing, cleaning, tidying, checking, double-checking, manipulating, and ultimately summarizing and visualizing the data.

It is common to hear people who work a lot with data say that about 80% of effort and time during real-world data analysis is spent on these kinds of tasks (and only about 20% on making statistical inference). Yet many books about data analysis ignore this 80%. They also overlook that the skills involved in this 80% are valuable in their own right. We are of the opinion that these skills alone go a long way towards allowing you to gain robust, informative insights from your data.

Insights will help you develop an efficient, reliable, and confidence-inspiring workflow for managing your data and drawing those initial insights out of them, and at the same time introduce you to core R skills for data management and visualization. Efficiency comes from learning methods of analysis that are transferable between problems and their associated datasets, and putting these methods together into an equally transferable workflow. Reliability—the ability to avoid, identify, and correct mistakes, and to reproduce work—comes from being able to evaluate multiple methods and functions and use a system of checks and balances throughout your workflow. Confidence comes from practice, encouragement, and achievement. We seek confidence that our workflows successfully generate insight.

Given our expertise and its ever-growing importance, we use R and RStudio throughout *Insights*. We use RStudio to interact with R, as it makes working with R a more pleasurable experience for the user. As in our undergraduate courses, and in the second edition of *Getting Started with R*, we teach an approach to using R based on the 'tidyverse' packages that have revolutionized data exploration and analysis in R. This approach provides a very consistent, efficient, and transferable workflow that is easily taught and learned. It is also usable with online data sources and scalable to large datasets, particularly by interfacing well with various database systems. Getting to grips with the tools to manage, summarize, and visualize small

datasets like the ones we use here for insights will inspire you with confidence for much bigger ones.

Although we are biologists, and the demonstrations of getting real insights from data in *Insights* are from the biological and environmental sciences, we imagine *Insights* will be appropriate for anyone seeking to gain insights from data, and at the beginning of their journey in doing so.

The learning 'curve'

It's worth knowing what's coming. The learning curve (Figure 0.1) for this book is not a curve! It is a continual upward line, hopefully not too steep at the beginning, and never too steep, but also not so shallow that you get bored. As you work through the book you will learn more and more, while building on what came before. You should feel continually challenged (which may get a bit tiring), and perhaps at times feel a little overwhelmed, but always be clear that you, with our help, have the ability to make progress.

There will likely be some tough times, perhaps even times when you feel like you can't continue. You will be learing new vocabulary, new ways of using your computer, working with data that has problems, fixing these problems, and ultimately developing summaries and graphs to develop

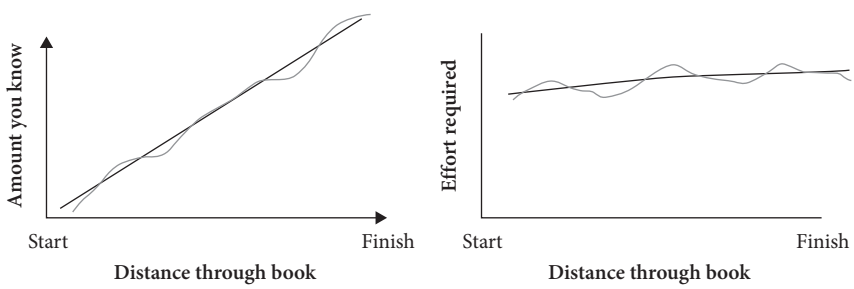

Figure 0.1 The *Insights* learning curve (left) and the effort-required curve (right). We try to make the beginning of the learning curve not steep, and then to keep you learning and learning, such that a reasonable and relatively constant effort is required.

insight. If you get stuck or frustrated, don't be afraid to take a break, have a drink and a cookie/biscuit, go for a walk, and then try again, perhaps with some help.

Untidy and dirty data

The data used in *Insights* are different from those associated with many other data analysis courses and books. The data are deliberately *disorganized*. This is quite different from many data analysis courses and books, where datasets are supplied ready for analysis. But it is also more like what you might start with from lab books, machines, or online data sources. A consequence is that the data are not visualization or analysis 'ready'. One might say the data are *untidy*. Also, the data are not provided by us; rather, you will download them from websites where the data are available to the public. Expect to spend time working with the data to get them 'research ready', getting to know the data, and learning the tricks and tips of how to do so efficiently and confidently.

No statistical tests or models

As we noted above, insights are intuitions and conclusions that come from identifying patterns in data. This does not formally require statistics. It does, however, require have developed an understanding of what the question is we (you) are trying to answer before making data summaries and graphs.

This book does not include any statistical tests, such as null hypothesis significance tests (or any other statistical tests or models), for a few reasons. (i) There is enough to be learned and gained from data analysis without such tests. We believe that the first steps in an introductory data analysis course should focus around the content of *Insights*; statistical tests can wait their turn. (ii) Statistical tests can be quite daunting and difficult, so we leave them until we have a solid hold on identifying patterns with respect to our questions that ultimately form the basis for developing appropriate statistical models and making statistical inferences. (iii) There is a risk

that early learning of statistical tests encourages a rather one-dimensional view of data analysis (e.g. the dimension of a *p*-value), whereas in reality we need to take into account many features of the data, including why they were collected, how they were collected, and even who they were collected by. (iv) Avoiding statistics at this initial stage of data analysis forces you to focus on the questions motivating the collection of the data and expectations of patterns in the data rather than focusing on *p*-values and statistical significance. The great success of Hans Rosling in publicizing and explaining issues in global health and development, via brilliant and simple data visualization, is a great example of how clear messages can (sometimes) be conveyed without statistical tests.

Perhaps you are of the opinion that statistics and hypothesis testing are required for objectivity, and that without them we are just subjectively looking for patterns. If so, perhaps take a look at the article 'Many analysts, one dataset: Making transparent how variations in analytical choices affect results.'[1] There are many rather subjective choices involved in doing statistics. To be clear, we do think there is a very important, even necessary, place for statistical models and tests, but that an introduction-to-data-analysis course is not that place.

Exploratory data analysis

Exploratory data analysis (EDA) was promoted by the statistician John Tukey in his 1977 book *Exploratory Data Analysis*. The broad aim of EDA is to help us formulate and refine hypotheses that will lead to informative analyses or further data collection. The core objectives of EDA are:

- to suggest hypotheses about the causes of observed phenomena;
- to guide the selection of appropriate statistical tools and techniques;
- to assess the assumptions on which statistical analysis will be based;
- to provide a foundation for further data collection.

[1] https://psyarxiv.com/qkwst/

EDA involves a mix of both numerical and visual methods. Statistical methods are sometimes used to supplement EDA, but its main purpose is to facilitate understanding before diving into formal statistical modelling. Even if we think we already know what kind of analysis we need to pursue, it's always a good idea to *explore a dataset before diving into the analysis*. At the very least, this will help us to determine whether or not our plans are sensible. Very often it uncovers new patterns and insights. In a sense, this book concerns EDA. But this book is also about answering questions, including assessing the weight of evidence in support of (or against) a hypothesis. Therefore it perhaps goes a little further than EDA.

Zen and the art of 'data science'

The emergence of ever more data about ever more things, and of more and more methods, techniques, and tools for looking at these data has led to the emergence of 'data science': the science of analysing complex and large data resources. Included in data science are activities such as data collection, storage, archiving, distribution, analysis, modelling, communication, and ethics. The book *Data Science for Undergraduates: Opportunities and Options*[2] states that 'all undergraduates will benefit from a fundamental awareness of and competence in data science.' It's probably OK to think of *Insights* as a book for learning the foundations of data science, but it's also important to know that *Insights* doesn't cover lots of data science aspects (such as data archiving or ethics).

Where does Zen come into this? To gain the deepest, most robust, most interesting, most valuable insights from data we need to be 'at one with the data'. How do we achieve this heady state of mind? We need to know the *details* of the data while maintaining broad awareness of *why* we're working with the data. We must have awareness of the big picture of why we're working on the data. We need to anticipate missing values and be prepared to ask why there are missing values when one might not expect any. We need to be keen to explore the distribution of the data and perhaps ask why

[2] https://www.nap.edu/catalog/25104/data-science-for-undergraduates-opportunities-and-options

there are a few extreme-looking values. And we need to be OK with getting warning messages from R. Put another way, we must get stuck deeply into the details and also see the big picture. We must see every detail of every tree, and the whole forest. An article along these lines discusses how data scientists with this ability can be very competitive business consultants.[3]

Open-science trends

There is increasing movement towards making science a more open process. Part of this movement involves making data more findable, accessible, interoperable, and reusable (the FAIR guiding principles of data management and stewardship).[4] When working with the datasets in the Workflow Demonstrations in *Insights*, you might take a moment to think whether they are particularly findable, accessible, interoperable, and reusable. However, *Insights* is not about teaching you how to adhere to the FAIR guidelines—that is a story for another place, and one that is being increasingly told. *Insights* does focus on data analysis methods that are repeatable, shareable, and reliable… if there are guiding principles for data analysis, then *Insights* adheres to them!

Put another way, *Insights* teaches data analysis methods that result in high *reproducibility* (a study is reproducible if someone can take the same data and reproduce the same results as reported in the original study). Another fashion in which *Insights* assists with open science is that it teaches methods that make collaborative work rather easier than it might otherwise be, such as making our work easy for other people to understand and implement themselves, hopefully without breaking it.

Intended readers

Insights is aimed at first- or second-year undergraduates in the life and environmental sciences, to accompany their first course in 'data analysis',

[3] http://www.programmingr.com/content/zen-and-the-art-of-competing-against-mbas/
[4] https://www.nature.com/articles/sdata201618

and at their instructors. As far as we are aware, there is no equivalent book available (though we describe in some detail the numerous related books on the *Insights* companion website (http://insightsfromdata.io)). *Insights* purposely excludes statistical methods, so students can first master the valuable and prerequisite skills of working with data, such as manipulating, summarizing, and visualizing data. It teaches an approach to using R based on the tidyverse of add-on packages, providing efficient, reliable, and confidence-inspiring methods and workflows. Our approach to learning and teaching has developed over more than two decades and proven successful in both undergraduate courses and training programmes.

Some competencies required for beginning with this book:

- You should know your way around your computer (e.g. how to find files, make folders, install applications).
- You should know how to look at and enter data into a spreadsheet (e.g. in Excel).
- You should know how to use the internet, download files, find them on your computer, and move them to a specific folder on your computer.

How is the book organized?

Figure 0.2 shows the organization of this book, and the arrows show how you could (probably should) work through it. Nothing is very special about the organization of the first two chapters.

Chapter 1. An introduction to insights, to data, and to the demonstrations in the book and on the *Insights* companion website.[5]

Chapter 2. Getting acquainted with R and RStudio, including installing them, doing some basic calculations, and getting help.

[5] http://insightsfromdata.io

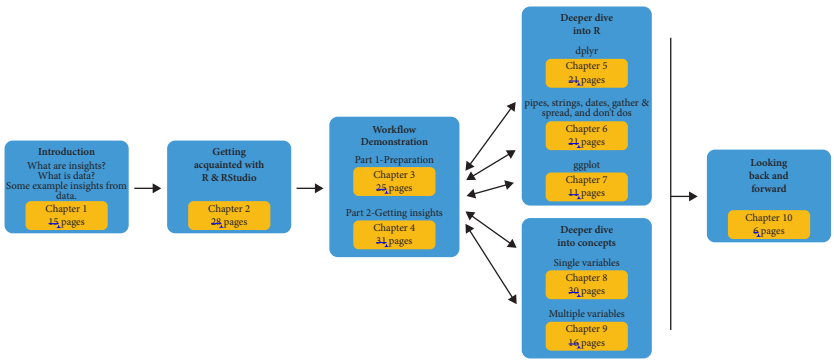

Figure 0.2 How this book is organized, and how you should work through it. This is explained in detail in the text.

Then, with *Chapters 3* and *4*, the organization of the book shifts. Chapters 3 and 4 walk through *getting insights* using an example dataset. *Chapters 5–7* contain more in-depth, complete, and detailed explanations of the mechanics of what you are doing with R and with tidyverse functions in Chapters 3 and 4. *Chapters 8* and *9* return to a focus on the example dataset and further develop core skills for insight around the various types of data in the example.

Hence, as you work through Chapters 3 and 4, you may, or may not, choose to dip into a section of Chapters 5–9. All of this is reflected in the bidirectional arrows joining Chapters 3 and 4, and Chapters 5–9 in Figure 0.2. It will be up to you how you work with these chapters; each of you is different and will probably do it differently. It will, however, likely be worth all of you being organized, for example by keeping notes about what you understood during the workflows in Chapters 3 and 4 and what you did not, and then checking this off when working through Chapters 5–9.

Here is a quick summary of Chapters 3–10.

Chapter 3 demonstrates preparation tasks, such as preparing your question, study, data, and computer, and getting data into R and ready for making insights. All of this provides a solid foundation for developing a robust workflow to gain insights from data.

Chapter 4 demonstrates getting insights, including constructing new variables, graphing data, calculating summaries (e.g. means), and evaluating patterns in the graphs and tables to gain insights.

Chapter 5 provides a deeper dive into data manipulation using tools in the **dplyr** package, including subsetting datasets, and making summaries of these subsets.

Chapter 6 provides a deeper dive into other data manipulation requirements that often arise in the life and environmental sciences. These include working with strings (words) and dates, and rearranging data from being across columns to within columns of a dataset. We also consider some formal dos and don'ts.

Chapter 7 gives an in-depth and guided explanation of how to make multiple types of graphs and enhance their capacity to provide insights using the **ggplot2** package. This builds on the introduction in Chapter 4.

Chapter 8 provides a deeper dive into evaluating features of specific variables in your data, including visualizing sample distributions and estimating numeric descriptors of central tendency (means vs medians), data dispersion, and asymmetry (variation, interquartile ranges).

Chapter 9 shifts the focus to examining patterns between two variables. The chapter includes sections on examining relationships between two numeric/continuous variables, two categorical variables (factors), and one numeric and one categorical variable. It finishes with a flurry, looking at relationships among three or more variables (including potential interactions).

Chapter 10 is the final chapter of the book, offering congratulations and some information and advice about reproducibility, an equally important subject when getting insights from data.

So, overall, you'll be learning a language of data management and visualization using R, you'll be working with example data, and you'll develop robust numerical summaries and classy visualizations of data. You certainly won't learn everything you want to know, but we can guarantee that you'll develop some excellent autonomy in learning, a platform on which to develop your *Insights from Data with R* skillset.

Online companion material

The *Insights* companion website[6] contains supplementary material including:

- an online overview of the *Insights* workflow;
- more topics in R;
- additional data analysis concepts;
- three additional Workflow Demonstrations;
- complete Workflow Demonstration R scripts;
- details of a live data analysis demonstration we often use in our introductory undergraduate classes;
- exercises and questions for each section of the book;
- more study questions and datasets that could be developed into new Workflow Demonstrations (perhaps for students to practise with and/or instructors to use);
- some related/suggested reading;

Boxes

Throughout the book are four types of box:

Efficiency and reliability. In these, we describe practices and methods for achieving higher efficiency and reliability in our journey from data to insights. They contain information about how to make our work more robust and reliable, such that it can still function if we get or add some new data, or otherwise make some changes in our work. And information to help ensure that conclusions/insights are robust.

[6] http://insightsfromdata.io

Be aware. These contain instructions about an opportunity/need to carefully consider an issue, for example a way to work that reduces the potential for mistakes, such as including appropriate checks and balances. These can also concern a warning or a common 'gotcha'. There are a number of common pitfalls that trip up new users of R (and more experienced users too!). We aim to highlight these and show you how to avoid them.

Action. A box containing instructions for you to do something important. Now!

Information. These aim to offer a not-too-technical discussion of how or why something works the way it does. You do not have to understand everything in these boxes to use R, but the information will help you understand how it works.

Box icon attributions:

- 'Action' by Icons Producer from the Noun Project (https://thenounproject.com/icon/1899450/).
- 'Information' by SELicon from the Noun Project (https://thenounproject.com/icon/2119887/).
- 'Efficiency and reliability' by BomSymbols from the Noun Project (https://thenounproject.com/icon/1555215/).
- 'Be aware' is the 'Warning' icon by Kristin Hogan from the Noun Project (https://thenounproject.com/icon/77514/).

All icons are licensed as Creative Commons CCBY (https://creativecommons.org/licenses/by/3.0/). Colours and sizes have been altered.

Some ideas for instructors using this book

As mentioned, we have good experiences teaching introduction-to-data-analysis undergraduate classes of 200+ students using the approach and methods in this book. Students tell us that the learning is challenging, represents a relatively high workload, is valuable, and is enjoyable. Our aim is that all students pass the course, and so far over 95% do. Here are some recommendations based on our experiences:

- The material is suitable for undergraduates with little or no prior experience of working with data, of programming, or of statistics.
- The amount of material is suitable for a six-week course of five to six hours per week (reading, practicals, and homework).
- In the first class of the first week we lead a *live data analysis demonstration*. Within one hour we go from question to answer, including collecting some data about each of the students. We believe this demonstration helps students connect with the importance and fun of the content of the course. Details of the demonstration are on the *Insights* companion website.[7]
- We have four activities each week: a lecture (sometimes in person, sometimes video lectures), before practical reading or viewing (e.g. a chapter or section of this book, or some video tutorials), a practical session (e.g. see the material at http://insightsfromdata.io), and a weekly graded assessment (administered in an automated online learning platform).
- All decisions about the course, e.g. organization, schedule, content, requirement, assessment, are taken in the context of maximizing student autonomy, purpose, and mastery, in order to nurture and stimulate students' intrinsic motivation. We treat the students like the adults they are.

[7] http://insightsfromdata.io

- All decisions are also taken with efficiency in mind . . . achieving a combination of great learning outcomes and reasonable instructor effort.
- Heavy use of automated feedback and grading leaves time for instructors to give students 1:1 support, even in a class of over 200 students.
- Owen Petchey uses the **exams** package[8] to organize a library of questions and to create examinations from these with (almost) the click of a button. The **rexams** package has options for output format, including pdf and various ones compatible with many online learning platforms.

If you have any questions about using this book as a coursebook for your undergraduate introduction-to-data-analysis course, please get in touch. We are very happy to share expertise and experiences.

Relationship with *Getting Started with R* (*GSwR*), second edition, Beckerman, Childs, and Petchey (2017)

Insights is a completely different book from *Getting Started with R*. Here are the most important differences, provided with the aim of helping you know which book to work with. If you have any uncertainty after looking at these differences, don't hesitate to contact one of us.

- What differentiates the audiences of *GSwR* and *Insights*? *GSwR*: folk who already do data work and statistics and want to learn to use R. *Insights*: folk who haven't done any data work before.
- *GSwR* motivates people who already have reasonable knowledge of getting insights from data with non-R tools to learn to use R and to implement data management, visualization, and statistical analysis with R and the tidyverse set of packages. *Insights* motivates people to

[8] http://www.r-exams.org

learn how to get insights from data with R and the tidyverse set of packages.
- *GSwR* assumes some prior knowledge of statistics. *Insights* assumes no prior experience of working with data or of statistics.
- *Insights* is designed as a textbook for an undergraduate 'introduction to getting knowledge from data' course. *GSwR* was not designed for this, and seems to not work very well for such purposes (though selected chapters from it combine well with chapters from other books).
- For the small amount of overlapping content, *Insights* provides more detail about how and why (rather than providing an overview tour).
- *Insights* and the *Insights* companion website[9] contain more of the content often associated with undergraduate courses than does *GSwR*, such as exercises and quizzes.

Acknowledgements

R is a product of the efforts of many individuals. RStudio is also the work of many individuals, organized by the vision of the company RStudio, whose mission is to create open source software for data analysis and statistical computing. The tidyverse collection of add-on packages was initiated by Hadley Wickham and has many contributors. We are extremely grateful to these individuals for making our data analysis and research so much more reliable, efficient, and fun. This book was written using the **bookdown** package created by Yihui Xie, which provides a suitable environment for R and RStudio users to author documents, from simple to complex.

We each have been teaching R for nearly 20 years, and in that time it is our experiences with interested, bored, critical, and all other types of student that have allowed us to become better at teaching R. We thank all the students for putting in the effort and giving their feedback about what

[9] http://insightsfromdata.io

works, about what does not, and what might work better. And we apologize to the bored and critical students for our oversights and mistakes.

We are honoured to publish with Oxford University Press and to work with its staff, particular Ian Sherman, Charles Bath, and Lucy Nash. Douglas Meekison very skilfully copyedited the manuscript. Several reviewers commented on the original book proposal, including making suggestions for improvements that were implemented.

Vanessa Mata was kind enough to place on Dryad the data used in her and her colleagues' study of bat diets. This made it possible for us to use the data and questions from the study as the basis of the Workflow Demonstration in this book.

Finally, thanks to our families for letting us have the time during evenings and holidays to work on *Insights*. We love you all, lots.