Piotr Fryzlewicz
London School of Economics
Houghton Street, London WC2A 2AE, UK
p.fryzlewicz@lse.ac.uk

Telling Stories with Data: With Applications in R.

Rohan Alexander. Boca Raton: CRC Press, 2023. ISBN 978-1-032-13477-2. xxiii+598 pp.

W. Edwards Deming once famously said "Without data, you're just another person with an opinion". Of course, having an opinion does not have to stop with the arrival of data, and it is the job of a data scientist to have an opinion about best practices in the field. Dr Alexander's book aims to position itself as a vade mecum of modern social data science, but it is also an opinion piece about it – about what it is, and about how to do it in a certain standardised way in the context of the social sciences.

I emphasise opinion here, because it is not something that readers coming to data science from many other fields will necessarily be used to seeing in their sources. Readers with a mathematics or statistics background will be used to books presenting truths rooted in certain paradigms or based on certain assumptions. There is typically little to disagree with in these sorts of works. Readers with a computer science background may be used to manuals, describing tools or advising how to achieve certain effects. Again, there is usually not much to disagree with in a manual. By contrast, opinion pieces, whether good or bad, will invariably divide opinion, and this may be particularly easy for a volume as massive and as broadly pitched as "Telling Stories with Data". Its author, Dr Rohan Alexander, is an assistant professor in the Faculty of Information and the Department of Statistical Sciences at the University of Toronto, PhD economist, experienced applied statistician, engaged data science teacher and advocate for reproducibility and credibility in data science. The book is organised into six parts (*Foundations*, *Communication*, *Acquisition*, *Preparation*, *Modeling*, and *Applications*), further sub-divided into 17 chapters.

Part I, *Foundations*, is made up of Chapter 1, "Telling stories with data", Chapter 2, "Drinking from a fire hose" and Chapter 3, "Reproducible workflows". In Chapter 1, we are given the author's first-draft definition of data science as "humans measuring things, typically related to other humans, and using sophisticated averaging to explain and predict" (this definition will be refined in Chapter 17); we are also given a selection of other definitions of the term from the existing literature. The author introduces his advocated data science workflow, which consists of planning, simulating, acquiring, exploring and sharing. The reader is told of the importance of reproducibility and ethics in

data science, the latter, at a minimum, meaning "considering the full context of the dataset". Some important components of the data scientist's practice are listed; these include, amongst others, asking questions, data collection and cleaning, exploratory data analysis, and modelling. In Chapter 2, we are presented with three case studies involving, respectively, Australian elections, Toronto's unhoused population and neonatal mortality across countries; each case study follows the steps of the workflow advocated earlier. In Chapter 2, the reader gets their first exposure to R with tidyverse, the running programming language of the book. In Chapter 3, the author describes the concept of a reproducible data science workflow, introducing tools and concepts such as Quarto, R projects and version control with Git.

Part II, *Communication*, consists of Chapter 4, "Writing research" and Chapter 5, "Static communication". Chapter 4 is about the different questions that can be asked and answered in a piece of applied quantitative social science research, and how to write an article to summarise the findings of research in this field. Chapter 5 talks about statistical graphics, including maps.

Part III, *Acquisition*, is made up of Chapter 6, "Farm data", Chapter 7, "Gather data" and Chapter 8, "Hunt data". As their titles suggest, the three chapters talk about different modes and aspects of obtaining data, starting with the history of measurement and touching on topics such as census and survey data, sampling, randomised controlled trials and A/B testing, while discussing limitations and ethical aspects of each of these ways of obtaining data. On the more technical side, the reader is told of the practicalities of using various R packages for the purposes of data access though APIs, web scraping and distilling text data from PDFs.

Part IV, *Preparation*, is broken up into Chapter 9, "Clean and prepare" and Chapter 10, "Store and share". In Chapter 9, the author talks about the importance of storing the original dataset as well as various practical approaches to data cleaning and checking whether the cleaned dataset satisfies our requirements for a further analysis. In Chapter 10, the reader is told how to store, document and share datasets, while paying close attention to the topic of privacy (it is here that the book talks about differential privacy). Various data storage formats are discussed, including Apache Parquet.

Part V, *Modeling*, features Chapter 11, "Exploratory data analysis", Chapter 12, "Linear models" and Chapter 13, "Generalized linear models". Chapter 11 includes an extended exploratory data analysis of an Airbnb listing dataset, whereas Chapters 12 and 13 offer brief examples of linear regression, logistic regression, Poisson and negative binomial regression and an introduction to multilevel modelling.

Finally, Part VI, *Applications*, consists of Chapter 14, "Causality from observational data", Chapter 15, "Multilevel regression with post-stratification" and Chapter 16, "Text as data". The book concludes with Chapter 17, "Concluding remarks". Chapter 14 is a brief but comprehensive overview of the main approaches to causal inference from observational data; these include difference-in-differences, propensity score matching, regression discontinuity design and instrumental variables. Simulated and real-data examples are given. Chapter 15 gives an introduction to post-stratification, whereas Chapter 16 talks about the different facets and complexities of dealing with data that arises as text. In Chapter 17, the reader is given the author's final definition of data science as "the process of developing and applying a principled, tested, reproducible, end-to-end workflow that focuses on quantitative measures in and of themselves, and as a foundation to explore questions". The author concludes with advice on further reading, as well as with some inspiring open questions.

There can be no doubt that the book has its strengths. One of them is the breadth of the material it covers. For example, on the technical side, we are taken on a grand tour of incredibly useful R packages for everything from drawing maps, to manipulating census data, to optical character recognition, to differential privacy, not forgetting of course the wonderful ecosystem for tidy data manipulation given by RStudio, the tidyverse suite of packages and Quarto. There can be no such thing as an encyclopaedia of R packages (not least because their list is so dynamic), but for the social data scientist looking for a quick reference on 'how to do it in R', the book will often provide a good starting point.

Another strength is the author's focus on the end-to-end aspect of the data science process. To the uninitiated – e.g. a fresh graduate in theoretical statistics – it can often come as a surprise that so much data work (such as data cleaning and transformation – often referred to as data wrangling, a term surprisingly absent from the book) is required before the actual statistical analysis can take place. The book attempts to normalise this fact and highlight current best practices related to every step of the end-to-end data analytics process. One of these best practices is reproducibility, a cornerstone of credible data science, and the author is right to advertise it in almost sacrosanct terms. Another important concept promoted in the book is that of version control, which many data scientists coming from academic statistics may be unfamiliar with.

Causality can be an important component of "telling stories with data" and from this point of view it is reassuring to see Chapter 14, which provides an overview of modern approaches to causality, in the book. To say even more, I found that chapter mostly well and clearly written, despite its brevity; and thought that the same could be said of the

other chapters in the *Applications* part.

Another strength of the book is its extensive bibliography. The author is impressively well-read and the various bibliography items mentioned at the start of each chapter will no doubt make for fascinating preliminary or further reading.

Going back to the 'opinion piece' angle, the book offers plenty of avenues for fruitful discussion, or perhaps even constructive disagreement, with the author. To start with, the author encourages simulation as part of the standard data science workflow, but does not tell us that simulating a dataset that accurately reflects the intricacies of our real dataset can often be exceedingly difficult (although, reassuringly, some modern references on realistic simulation are given in Section 10.5.2) and does not warn us that simulating an over-simplified version of the data can easily lead the analyst onto the wrong track. An example of how things can go wrong is in Section 7.3.4, in which 50% of the first names in the simulated dataset of Prime Ministers of the United Kingdom are female. This may or may not be desirable, depending on the context and the analyst's purpose, but is historically inaccurate. Equally importantly, many of the simulated datasets throughout the book mainly seem to serve the purpose of helping the reader imagine what the corresponding real dataset might look like – which gave me the impression that the simulation step was encouraged more because it was the author's habit (which not all analysts will share) than an indispensable stage of the process.

Continuing the discussability angle, the concluding Chapter 17 offers plenty of material for debate. The author's view is that data science "barely existed" a generation ago, but some readers, for example, may argue that data science has always existed, to the extent permitted by computing power available in any given era, although it might not always have been called by that name. Similarly, his advice to "be at the intersection of at least a few different areas, rather than hyper-specialised" may not appeal to sought-after experts in one particular area. Finally, definitions of data science are always a fertile ground for discussion and the author's is no exception: according to the author's definition (quoted earlier in this review), is data science a science? What if the principles used in data science (as defined by the author) were flawed – would the definition still hold? Wouldn't a shorter definition, such as "science of learning from data" be better? On the plus side, Chapter 17 concludes with some interesting open questions, including in particular ones relating to the teaching of data science, and to the existing and desired relationship between the data science industry and academia. Also on the debatability front, I was unpersuaded by the author's advice on writing earlier in Chapter 4; I could not see why it was necessarily true that the first draft would be "poorly written".

There are some notable absences in the book. For a book on telling stories with data, I wished there was more on uncertainty quantification; in its absence, how to tell if the signal in the story is real? The limited amount of material on this topic that is there, is not always convincing: for example, in Section 15.3.4, should we be happy to see credible intervals for the proportion of support for President Biden covering 0.5 in states such as California on the one hand, and Texas on the other? More generally, I would have welcomed more focus on topics such as model critiquing, goodness of fit, and being sceptical of the outcomes of data analyses. Also, there is nothing in the book on modern predictive analytics: using R code snippets, it would have been easy to introduce, for example, packages such as keras, xgboost or lightgbm. Similarly, the useful chapter on text as data could have been strengthened further with material on generative AI for text. The side mention of splines in Section 5.2.2 was useful, but I thought it could have been expanded to give the reader a clearer flavour of modern nonparametric statistics. Finally, more general weapons in the ammunition of a solid data scientist, for example the habit of using training, validation and test sets, or being aware of the dangers of using the same dataset to generate hypotheses and to test them, are conspicuously absent from the book, as are some advanced but commonly used tools of exploratory data analysis, such as e.g. principal component analysis.

Most of the good advice on data science writing and presentation, dispensed by Dr Alexander throughout the book, is followed in the book itself, but there are exceptions. To highlight just a few: I did not think the title of the book adequately reflected its focus on social data science; I would have titled the book, perhaps, "Telling Stories with Data: A Handbook/Primer of Social Data Science with R". (Otherwise, readers interested in e.g. astrostatistics might not realise they could be better off looking elsewhere.) In the printed edition, many figures are hard to read; this includes e.g. Figure 3.5, which is critical to understanding the process of setting up GitHub. Some of the data stories presented in the book have an unfinished feel to them and lack depth. For example, the story on Toronto's unhoused population (Section 2.3) is built on the premise that shelter occupancy is higher in the winter months, and the conclusion of the story (Section 2.3.5) attempts to say just that ("there was a steady increase in the daily average number of occupied beds between July and December"); however, it does not comment at all on the fact that the bed occupancy in January and February was lower than in July, which runs counter to the attempted conclusion. Finally, I was struck by the lack of citations in some of the stories. For example, in the story on Australian elections (Section 2.2.5), we read that "in Australia some are systematically excluded from voting", but this is not supported with a literature reference. Strikingly, there are no literature references at all in the example

introduction to a paper given in Section 4.5.3 – in contrast to the generally excellent referencing in the book itself.

The fact that this useful handbook of social data science is not a statistics textbook is clear: for example, linear models in Chapter 12 are only introduced superficially, the formula for the normal density stated in Section 12.2 is not explained or even referred to by this term, there are no attempts to clarify the differences between frequentist and Bayesian statistics or their respective benefits, and many other components of a typical 'Statistics 101' course are missing altogether. This is also not a book to learn R programming from: the handy snippets of R code in the tidyverse paradigm and using the pipe notation are elegant, but cannot replace the first programming or statistical programming course. My impression is that the book will be of the most benefit to readers already familiar with statistics and R who wish to learn best practices and new tools in the modern social data science workflow, expand their knowledge of the vast universe of useful R packages, and be pointed towards interesting further reading. I wholeheartedly recommend this ambitious book to readers who find themselves in these categories.

Piotr Fryzlewicz
London School of Economics