Journal of Statistical Software

April 2010, Volume 34, Book Review 1. http://www.jstatsoft.org/

Reviewer: Hadley Wickham
Rice University

Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications

Glenn J. Myatt and Wayne P. Johnson
John Wiley & Sons, Cambridge, 2009.
ISBN 978-0-470-22280-5. 291 pp. USD 84.00.
http://www.wiley.com/remtitle.cgi?ISBN=0470222808

Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications had me excited with the title, but let me down with dry content and uninteresting data. The book does a good job of summarizing the range of techniques used in modern data mining, but there is too much focus on describing each procedure, and too little focus on why you should use each one and how you can interpret the results.

Most damningly, the book fails to convey any enthusiasm for data. There is little real data analysis, and the examples tend to be small, dated and uninteresting. The applications chapter (Chapter 5) does provide a couple of larger examples, but these read largely as a description of what was done to each dataset, not why and what we learned. Despite the inclusion of visualization in the title, many results are presented as large tables of numbers, as exemplified by Figure 5.9, which presents over 5,000 digits in a 10×100 table spread over two pages.

The 291 page book is arranged in five chapters: introduction, data visualization, clustering, predictive analysis (also known as modelling) and applications. Two appendices give an introduction to matrices and the authors' software package, which is designed to support the book. Generally, each chapter does a good job of explaining the common techniques used, and provides a couple of small examples. Two sections deserve special mention: Chapter 2, data visualization, and Appendix B, the software description.

The first part of Chapter 2, data visualization, presents a grab bag of good advice, but it fails to form a cohesive whole. The chapter then continues to present a list of different graphic types. While the description of each is adequate, there are a few peculiarities. For example, the whiskers of a boxplot are initially described as spanning the full range of the data. This flaw is later corrected with the introduction of "outlier cutoffs", but there is no reference to Tukey. Unfortunately there are many small errors in the plots in this chapter. For example, Figure 2.19 is missing the grid lines described in the text, Figure 2.34 has incorrect axis labels, and the image plot in Figure 2.48 is unreadable.

Appendix B provides an introduction to the author's Java application that accompanies the book. Unfortunately, the software is closed source, and there is no evidence that the routines have been validated. This would make it difficult to use in a practical setting, and I am unsure why the authors did not just use an existing open source or commercial application.

This book aims high but does not deliver on its promises, and I would hesitate to recommend it to anyone.

Reviewer:

Hadley Wickham
Rice University
Department of Statistics
Houston, TX, United States of America
E-mail: hadley@rice.edu
URL: http://had.co.nz/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 34, Book Review 1 *Published:* 2010-04-08
April 2010