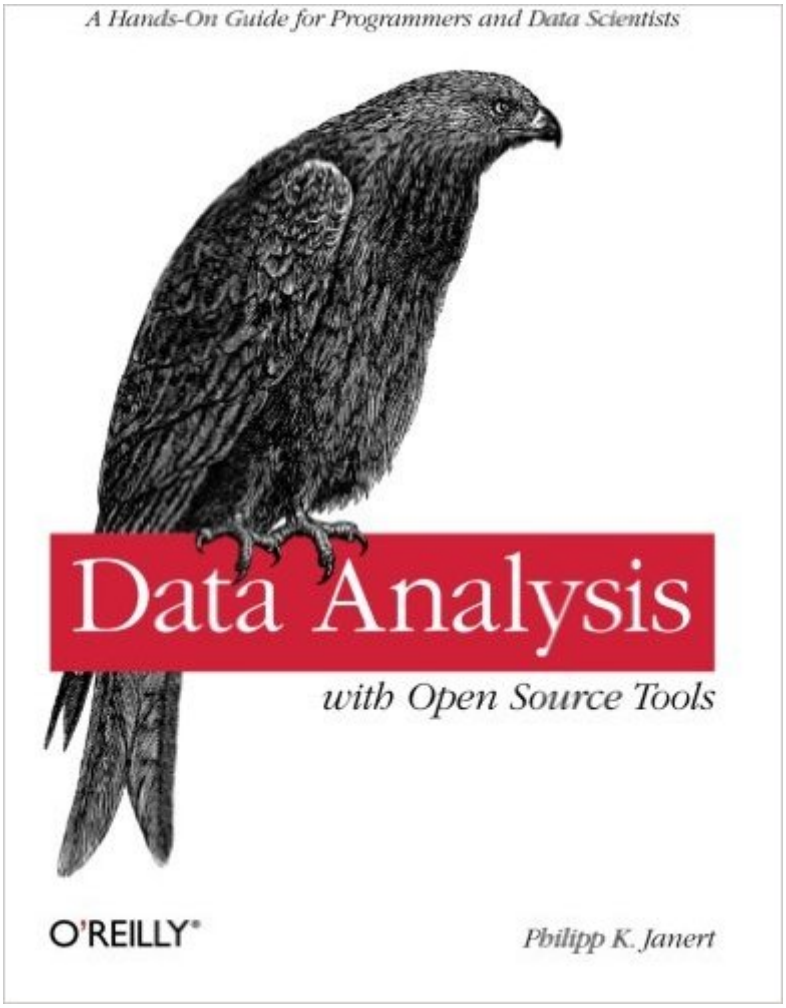
Data Analysis With Open Source Tools



Synopsis

These days it seems like everyone is collecting data. But all of that data is just raw information -- to make that information meaningful, it has to be organized, filtered, and analyzed. Anyone can apply data analysis tools and get results, but without the right approach those results may be useless. Author Philipp Janert teaches you how to think about data: how to effectively approach data analysis problems, and how to extract all of the available information from your data. Janert covers univariate data, data in multiple dimensions, time series data, graphical techniques, data mining, machine learning, and many other topics. He also reveals how seat-of-the-pants knowledge can lead you to the best approach right from the start, and how to assess results to determine if they’re meaningful.

Book Information

Paperback: 540 pages

Publisher: O’Reilly Media; 1 edition (November 28, 2010)

Language: English

ISBN-10: 0596802358

ISBN-13: 978-0596802356

Product Dimensions: 7 x 1.4 x 9.2 inches

Shipping Weight: 2.2 pounds (View shipping rates and policies)

Average Customer Review: 4.2 out of 5 starsÂ Â See all reviewsÂ (42 customer reviews)

Best Sellers Rank: #245,179 in Books (See Top 100 in Books) #165 inÂ Books > Computers & Technology > Databases & Big Data > Data Mining #179 inÂ Books > Computers & Technology > Software > Mathematical & Statistical #241 inÂ Books > Computers & Technology > Programming > Languages & Tools > Python

Customer Reviews

This book is aimed at offering a practical, hands-on introduction to data analysis for pragmatic readers without strong scientific or statistical background. Some basic programming experience is required. The author provides many personal (and sometimes useful) comments about different tools and procedures in data analysis.However, a careful reading reveals many problems, specially an obscure presentation of key concepts. In my opinion, the target audience for this book would be people without previous contact with data analysis. Hence the importance of presenting its core elements correctly. Otherwise, it’s useless for them.In particular:- Few pages are actually dedicated to present open source tools supporting the different graphs and techniques included in the book.

From the title, I expected a more complete tour through available open source tools for data analysis.- No clues about how to obtain most of the graphs and results presented in the book. No related data sets are available for download, either. A book like this is useless if we cannot learn how to replicate all the examples.- The formula of the variance for a sample is just wrong. One must divide by n-1 and not n; see "Applied Statistics and Probability for Engineers" (Montgomery and Runger 2006).- The author presents one of the most obscure explanations for the median I've ever come across. Recurring to an RFC (RFC 2330) to explain such a simple concept is really awkward.- In chapter 3 and Appendix B, natural logarithms (base e) are presented in the text, while graphs plot powers of 10. Definitely, not the right way to transmit correct concepts and methods.- I concur with a previous review in that "Workshop" sections just present an ultra-short overview of some open source tools. A quick search in your favourite engine will display much more informative introductions (even quick start guides).- Today, effective data analysis heavily depends on using the best possible implementation. While I might find educational to learn some of this implementations, in a real situation it is much better to rely on precise implementations of algorithms already available (e.g. libraries in GNU R).All in all, I still recommend "R in a Nutshell" for a gentle introduction to data analysis with an open source tool (GNU R). It also has some inaccuracies and typos, but at least it's much more informative and clear. Besides, it does include an R package with all datasets and examples, ready to be installed and explored.

This book covers such a wide range of topics that it necessarily skims over all of them but it always hits all the major points that an introductory survey should. Each chapter has a straight forward tone, strikes the right balance between developing mathematical rigor and developing an intuitive understanding of data , and undeniably passes on the lessons of hard earned, real world experience. But a reader who is actually working on a real data problem will almost certainly come to the realization that the understanding gained is somewhat superficial - that it's going to take a lot more heavy reading (probably of books, papers, and software tools recommended in this book) to get any real work done! The single biggest problem with this book is its misleading title. This book is not going to teach you how to use open source software to analyze data. There is only minimal information about how one would actually use the software tools being discussed. What you get is a brief commentary about what the author thinks each software package is good for. It's the same story as with the mathematical details: you will not find them here, but this book will give you an excellent idea of what to look for. So in the end it does leave you feeling just a little bit cheated, even though all the advice you got seems extremely well informed. What this book does

astonishingly well is communicate an attitude to data analysis that most textbooks (and nearly all the college courses I took) seem to miss. Nearly every chapter is a stream of stunningly insightful observations on how to approach data, without the mathematical detail that overwhelms most practicing programmers. I would recommend it to any reader who understands that truly useful insights are hard to come by, but detailed algorithms and formulae are easily found in the Internet Age. I wish the book were a few hundred pages shorter, that it corrected a few sloppy mistakes (like confusing revenue and profit), but I'm certainly glad I read it.

The book is very good for the intermediate-to-advanced data analysts. Beginners beware: there are some important prerequisites that are not obvious before you buy it, and there are some organization problems.First, the prerequisites. "I strongly recommend that you make it a habit to avoid all statistical language"..."Once we start talking about standard deviations, the clarity is gone." These are two sentences in the same passage from the Preface. The rest of that passage is similar. However, even the first chapters make heavy use of statistical language. Moreover, they assume that you already know statistics to the level of density estimation, noise, splines, and regression. Page 21 even features a footnote about the Fourier transform and Fourier convolution theorem. Clearly this book is not for the statistically-shy or for mathematically-shy in general, no matter what the Preface suggests. You also need to know Python and R.Second, the chapter organization problems. There's a mismatch between the first part of each chapter, which introduces concepts and techniques, and the Workshop part of the same chapter, which uses software. I was expecting the Workshop to illustrate the implementation of the same concepts and techniques. It's not really so. The Workshop introduces Python and R facilities at a different (lower) speed than the rest of the chapter. One could even wonder why the Workshop is in the same chapter. I'd rather that each chapter consisted of a few detailed case studies that first introduce concepts and techniques and then illustrate them with software libraries.

and Report Builder 3.0 with SQL Server Network Performance Toolkit: Using Open Source Testing Tools Remote Sensing and GIS for Ecologists: Using Open Source Software (Data in the Wild) Juan Ponce de Leon: A Primary Source Biography (Primary Source Library of Famous Explorers) From the Source - Thailand: Thailand's Most Authentic Recipes From the People That Know Them Best (Lonely Planet from the Source) Strunk's Source Readings in Music History: The Nineteenth Century (Revised Edition) (Vol. 6) (Source Readings Vol. 6) Great Source Write Source Texas: SkillsBook Student Edition Grade 3 Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications) Analytics: Data Science, Data Analysis and Predictive Analytics for Business Make: Lego and Arduino Projects: Projects for extending MINDSTORMS NXT with open-source electronics Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information MICO: An Open Source CORBA Implementation (The Morgan Kaufmann Series in Software Engineering and Programming) FreeBSD: An Open-Source Operating System for Your Personal Computer, Second Edition (with CD-ROM) FreeBSD: An Open-Source Operating System for Your Personal Computer