Best Practices In Data Cleaning A Complete Guide To Everything You Need To Do Before And After Collecting Your Data

The contributors to Best Practices in Quantitative Methods envision quantitative methods in the 21st century, identify the best practices, and, where possible, demonstrate the superiority of their recommendations empirically. Editor Jason W. Osborne designed this book with the goal of providing readers with the most effective, evidence-based, modern quantitative methods and quantitative data analysis across the social and behavioral sciences. The text is divided into five main sections covering select best practices in Measurement, Research Design, Basics of Data Analysis, Quantitative Methods, and Advanced Quantitative Methods. Each chapter contains a current and expansive review of the literature, a case for best practices in terms of method, outcomes, inferences, etc., and broad-ranging examples along with any empirical evidence to show why certain techniques are better. Key Features: Describes important implicit knowledge to readers: The chapters in this volume explain the important details of seemingly mundane aspects of quantitative research, making them accessible to readers and demonstrating why it is important to pay attention to these details. Compares and contrasts analytic techniques: The book examines instances where there are multiple options for doing things, and make recommendations as to what is the "best" choice—or choices, as what is best often depends on the circumstances. Offers new procedures to update and explicate traditional techniques: The featured scholars present and explain new options for data analysis, discussing the advantages and disadvantages of the new procedures in depth, describing how to perform them, and demonstrating their use. Intended Audience: Representing the vanguard of research methods for the 21st century, this book is an invaluable resource for graduate students and researchers who want a comprehensive, authoritative resource for practical and sound advice from leading experts in quantitative methods.

The book shows you how to view data from multiple perspectives, including data frame and column attributes. You will cover common and not-so-common challenges that are faced while cleaning messy data for complex situations. You will learn to manipulate data and get them down to a form that can be useful for making the right decisions.

How do you take your data analysis skills beyond Excel to the next level? By learning just enough Python to get stuff done. This hands-on guide shows non-programmers like you how to process information that's initially too messy or difficult to access. You don't need to know a thing about the Python programming language to get started. Through various step-by-step exercises, you'll learn how to acquire, clean, analyze, and present data efficiently. You'll also discover how to automate your data process, schedule file- editing and clean-up tasks, process larger datasets, and create compelling stories with data you obtain. Quickly learn basic Python syntax, data types, and language concepts Work with both machine-readable and human-consumable data Scrape websites and APIs to find a bounty of useful information Clean and format data to eliminate duplicates and errors in your datasets Learn when to standardize data and when to test and script data cleanup Explore and analyze your datasets with new Python libraries and techniques Use Python solutions to automate your entire data-wrangling process

Written in Ron Cody's signature informal, tutorial style, this book develops and demonstrates data cleaning programs and macros that you can use as written or modify which will make your job of data cleaning easier, faster, and more efficient. --

The process of developing predictive models includes many stages. Most resources focus on the modeling algorithms but neglect other critical aspects of the modeling process. This book describes techniques for finding the best representations of predictors for modeling and for finding the best subset of predictors for improving model performance. A variety of example data sets are used to illustrate the techniques along with R programs for reproducing the results.

Many researchers jump from data collection directly into testing hypothesis without realizing these tests can go profoundly wrong without clean data. This book provides a clear, accessible, step-by-step process of important best practices in preparing for data collection, testing assumptions, and examining and cleaning data in order to decrease error rates and increase both the power and replicability of results. Jason W. Osborne, author of the handbook Best Practices in Quantitative Methods (SAGE, 2008) provides easily-implemented suggestions that are evidence-based and will motivate change in practice by empirically demonstrating—for each topic—the benefits of following best practices and the potential consequences of not following these guidelines.

Data use in the library has specific characteristics and common problems. Data Clean-up and Management addresses these, and provides methods to clean up frequently-occurring data problems using readily-available applications. The authors highlight the importance and methods of data analysis and presentation, and offer guidelines and recommendations for a data quality policy. The book gives step-by-step how-to directions for common dirty data issues. Focused towards libraries and practicing librarians Deals with practical, real-life issues and addresses common problems that all libraries face Offers cradle-to-grave treatment for preparing and using data, including download, clean-up, management, analysis and presentation

Write maintainable, extensible, and durable software with modern C++. This book is a must for every developer, software architect, or team leader who is interested in good C++ code, and thus also wants to save development costs. If you want to teach yourself about writing clean C++, Clean C++ is exactly what you need. It is written to help C++ developers of all skill levels and shows by example how to write understandable, flexible, maintainable, and efficient C++ code. Even if you are a seasoned C++ developer, there are nuggets and data points in this book that you will find useful in your work. If you don't take care with your code, you can produce a large, messy, and unmaintainable beast in any programming language. However, C++ projects in particular are prone to be messy and tend to slip into bad habits. Lots of C++ code that is written today looks as if it was written in the 1980s. It seems that C++ developers have been forgotten by those who preach Software Craftsmanship and Clean Code principles. The Web is full of bad, but apparently very fast and highly optimized C++ code examples, with cruel syntax that completely ignores elementary principles of good design and well-written code. This book will explain how to avoid this scenario and how to get the most out of your C++ code. You'll find your coding becomes more efficient and, importantly, more fun. What You'll Learn Gain sound principles and rules for clean coding in C++ Carry out test driven development (TDD) Discover C++ design patterns and idioms Apply these design patterns Who This Book Is For Any C++ developer and software engineer with an interest in producing better code.

Many researchers jump straight from data collection to data analysis without realizing how analyses and hypothesis tests can go profoundly wrong without clean data. This book provides a clear, step-by-step process to examining and cleaning data in order to decrease error rates and increase both the power and replicability of results. Jason W. Osborne, author of Best Practices in Quantitative Methods (SAGE, 2008) provides easily-implemented suggestions that are research-based and will motivate change in practice by empirically demonstrating for each topic the benefits of following best practices and the potential consequences of not following these guidelines. If your goal is to do the best research you can do, draw conclusions that are most likely to be accurate representations of the population(s) you wish to speak about, and report results that are most likely to be replicated by other researchers, then this basic guidebook is indispensable.

Written for practitioners of data mining, data cleaning anddatabase management. Presents a technical treatment of data quality includingprocess, metrics, tools and algorithms. Focuses on developing an evolving modeling strategy through aniterative data exploration loop and incorporation of domainknowledge. Addresses methods of detecting, quantifying and correcting dataquality issues that can have a significant impact on findings anddecisions, using commercially available tools as well as newalgorithmic approaches. Uses case studies to illustrate applications in real lifescenarios. Highlights new approaches and methodologies, such as theDataSphere space partitioning and summary based analysistechniques. Exploratory Data Mining and Data Cleaning will serve as animportant reference for serious data analysts who need to analyzelarge amounts of unfamiliar data, managers of operations databases,and students in undergraduate or graduate level courses dealingwith large scale data analys is and data mining.

Development Research in Practice leads the reader through a complete empirical research project, providing links to continuously updated resources on the DIME Wiki as well as illustrative examples from the Demand for Safe Spaces study. The handbook is intended to train users of development data how to handle data effectively, efficiently, and ethically. “In the DIME Analytics Data Handbook, the DIME team has produced an extraordinary public good: a detailed, comprehensive, yet easy-to-read manual for how to manage a data-oriented research project from beginning to end. It offers everything from big-picture guidance on the determinants of high-quality empirical research, to specific practical guidance on how to implement specific workflows—and includes computer code! I think it will prove durably useful to a broad range of researchers in international development and beyond, and I learned new practices that I plan on adopting in my own research group.†? —Marshall Burke, Associate Professor, Department of Earth System Science, and Deputy Director, Center on Food Security and the Environment, Stanford University “Data are the essential ingredient in any research or evaluation project, yet there has been too little attention to standardized practices to ensure high-quality data collection, handling, documentation, and exchange. Development Research in Practice: The DIME Analytics Data Handbook seeks to fill that gap with practical guidance and tools, grounded in ethics and efficiency, for data management at every stage in a research project. This excellent resource sets a new standard for the field and is an essential reference for all empirical researchers.†? —Ruth E. Levine, PhD, CEO, IDinsight “Development Research in Practice: The DIME Analytics Data Handbook is an important resource and a must-read for all development economists, empirical social scientists, and public policy analysts. Based on decades of pioneering work at the World Bank on data collection, measurement, and analysis, the handbook provides valuable tools to allow research teams to more efficiently and transparently manage their work flows—yielding more credible analytical conclusions as a result.†? —Edward Miguel, Oxfam Professor in Environmental and Resource Economics and Faculty Director of the Center for Effective Global Action, University of California, Berkeley “The DIME Analytics Data Handbook is a must-read for any data-driven researcher looking to create credible research outcomes and policy advice. By meticulously describing detailed steps, from project planning via ethical and responsible code and data practices to the publication of research papers and associated replication packages, the DIME handbook makes the complexities of transparent and credible research easier.†? —Lars Vilhuber, Data Editor, American Economic Association, and Executive Director, Labor Dynamics Institute, Cornell University

Summary Streaming Data introduces the concepts and requirements of streaming and real-time data systems. The book is an idea-rich tutorial that teaches you to think about how to efficiently interact with fast-flowing data. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Technology As humans, we're constantly filtering and deciphering the information streaming toward us. In the same way, streaming data applications can accomplish amazing tasks like reading live location data to recommend nearby services, tracking faults with machinery in real time, and sending digital receipts before your customers leave the shop. Recent advances in streaming data technology and techniques make it possible for any developer to build these applications if they have the right mindset. This book will let you join them. About the Book Streaming Data is an idea-rich tutorial that teaches you to think about efficiently interacting with fast-flowing data. Through relevant examples and illustrated use cases, you'll explore designs for applications that read, analyze, share, and store streaming data. Along the way, you'll discover the roles of key technologies like Spark, Storm, Kafka, Flink, RabbitMQ, and more. This book offers the perfect balance between big-picture thinking and implementation details. What's Inside The right way to collect real-time data Architecting a streaming pipeline Analyzing the data Which technologies to use and when About the Reader Written for developers familiar with relational database concepts. No experience with streaming or real-time applications required. About the Author Andrew Psaltis is a software engineer focused on massively scalable real-time analytics. Table of Contents PART 1 - A NEW HOLISTIC APPROACH Introducing streaming data Getting data from clients: data ingestion Transporting the data from collection tier: decoupling the data pipeline Analyzing streaming data Algorithms for data analysis Storing the analyzed or collected data Making the data available Consumer device capabilities and limitations accessing the data PART 2 - TAKING IT REAL WORLD Analyzing Meetup RSVPs in real time

Data is the new oil, but it comes crude. To do anything meaningful - modeling, visualization, machine learning, for predictive analysis – you first need to wrestle and wrangle with data. Data Wrangling with Python teaches you the essentials that will get you up and running with data wrangling in no time.

Data in its raw state is rarely ready for productive analysis. This book not only teaches you data preparation, but also what questions you should ask of your data. It focuses on the thought processes necessary for successful data cleaning as much as on concise and precise code examples that express these thoughts.

In a conversational tone, Regression & Linear Modeling provides conceptual, user-friendly coverage of the generalized linear model (GLM). Readers will become familiar with applications of ordinary least squares (OLS) regression, binary and multinomial logistic regression, ordinal regression, Poisson regression, and loglinear models. The author returns to certain themes throughout the text, such as testing assumptions, examining data quality, and, where appropriate, nonlinear and non-additive effects modeled within different types of linear models. Available with Perusall—an eBook that makes it easier to prepare for class Perusall is an award-winning eBook platform featuring social annotation tools that allow students and instructors to collaboratively mark up and discuss their SAGE textbook. Backed by research and supported by technological innovations developed at Harvard University, this process of learning through collaborative annotation keeps your students engaged and makes teaching easier and more effective. Learn more.

Explore the mysteries of Exploratory Factor Analysis (EFA) with SAS with an applied and user-friendly approach.Exploratory Factor Analysis with SAS focuses solely on EFA, presenting a thorough and modern treatise on the different options, in accessible language targeted to the practicing statistician or researcher. This book provides real-world examples using real data, guidance for implementing best practices in the context of SAS, interpretation of results for end users, and it provides resources on the book's author page. Faculty teaching with this book can utilize these resources for their classes, and individual users can learn at their own pace, reinforcing their comprehension as they go.Exploratory Factor Analysis with SAS reviews each of the major steps in EFA: data cleaning, extraction, rotation, interpretation, and replication. The last step, replication, is discussed less frequently in the context of EFA but, as we show, the results are of considerable use. Finally, two other practices that are commonly applied in EFA, estimation of factor scores and higher-order factors, are reviewed. Best practices are highlighted throughout the chapters.A rudimentary working knowledge of SAS is required but no familiarity with EFA or with the SAS routines that are related to EFA is assumed.

Knowledge for Free... Get that job, you aspire for! Want to switch to that high paying job? Or are you already been preparing hard to give interview the next weekend? Do you know how many people get rejected in interviews by preparing only concepts but not focusing on actually which questions will be asked in the interview? Don't be that person this time. This is the most comprehensive Data Analytics interview questions book that you can ever find out. It contains: 500 most frequently asked and important Data Analytics interview questions and answers Wide range of questions which cover not only basics in Data Analytics but also most advanced and complex questions which will help freshers, experienced professionals, senior developers, testers to crack their interviews.

Access and clean up data easily using JMP®! Data acquisition and preparation commonly consume approximately 75% of the effort and time of total data analysis. JMP provides many visual, intuitive, and even innovative data-preparation capabilities that enable you to make the most of your organization's data. Preparing Data for Analysis with JMP® is organized within a framework of statistical investigations and model-building and illustrates the new data-handling features in JMP, such as the Query Builder. Useful to students and programmers with little or no JMP experience, or those looking to learn the new data-management features and techniques, it uses a practical approach to getting started with plenty of examples. Using step-by-step demonstrations and screenshots, this book walks you through the most commonly used data-management techniques that also include lots of tips on how to avoid common problems. With this book, you will learn how to: Manage database operations using the JMP Query Builder Get data into JMP from other formats, such as Excel, csv, SAS, HTML, JSON, and the web Identify and avoid problems with the help of JMP’s visual and automated data-exploration tools Consolidate data from multiple sources with Query Builder for tables Deal with common issues and repairs that include the following tasks: reshaping tables (stack/unstack) managing missing data with techniques such as imputation and Principal Components Analysis cleaning and correcting dirty data computing new variables transforming variables for modelling reconciling time and date Subset and filter your data Save data tables for exchange with other platforms

Best Practices in Exploratory Factor Analysis (EFA) is a practitioner-oriented look at this popular and often-misunderstood statistical technique. We avoid formulas and matrix algebra, instead focusing on evidence-based best practices so you can focus on getting the most from your data. Each chapter reviews important concepts, uses real-world data to provide authentic examples of analyses, and provides guidance for interpreting the results of these analysis. Not only does this book clarify often-confusing issues like various extraction techniques, what rotation is really rotating, and how to use parallel analysis and MAP criteria to decide how many factors you have, but it also introduces replication statistics and bootstrap analysis so that you can better understand how precisely your data are helping you estimate population parameters. Bootstrap analysis also informs readers of your work as to the likelihood of replication, which can give you more credibility. At the end of each chapter, the author has recommendations as to how to enhance your mastery of the material, including access to the data sets used in the chapter through his web site. Other resources include syntax and macros for easily incorporating these progressive aspects of exploratory factor analysis into your practice. The web site will also include enrichment activities, answer keys to select exercises, and other resources. The fourth "best practices" book by the author, Best Practices in Exploratory Factor Analysis continues the tradition of clearly-written, accessible guides for those just learning quantitative methods or for those who have been researching for decades. NEW in August 2014! Chapters on factor scores, higher-order factor analysis, and reliability. Chapters: 1 INTRODUCTION TO EXPLORATORY FACTOR ANALYSIS 2 EXTRACTION AND ROTATION 3 SAMPLE SIZE MATTERS 4 REPLICATION STATISTICS IN EFA 5 BOOTSTRAP APPLICATIONS IN EFA 6 DATA CLEANING AND EFA 7 ARE FACTOR SCORES A GOOD IDEA? 8 HIGHER ORDER FACTORS 9 AFTER THE EFA: INTERNAL CONSISTENCY 10 SUMMARY AND CONCLUSIONS

Data cleaning is a waste of time. If the data had been collected properly in the first place there wouldn’t be any cleaning to do, and you wouldn’t now be faced with the prospect of weeks of cleaning to get your dataset analysis-ready. Worse still, your boss won’t understand why your analysis report isn’t on his desk yet, a mere 48 hours after he’s asked for it. Bless him, he doesn’t understand – he thinks that cleaning data is just about clicking a few buttons in Excel and – ta da! – it’s all done. Even a monkey can do that, right? And – for good reason – you won’t get any help from statistics books either. Data is messy and cleaning it can be difficult, time-consuming and costly. Not to mention it’s the least sexy thing you can do with a dataset. Yet you’ve still got to do it, because, well, someone has to… But it doesn’t have to be so difficult. If you're organised and follow a few simple rules your data cleaning processes can be simple, fast and effective. Not to mention fun! Well, not fun exactly, just not quite as coma-inducing. Practical Data Cleaning (now in its 5th Edition!) explains the 19 most important tips about data cleaning with a focus on understanding your data, how to work with it, choose the right ways to analyse it, select the correct tools and how to interpret the results to get your data clean in double quick time. Best of all, there is no technical jargon – it is written in plain English and is perfect for beginners! Discover how to clean your data quickly and effectively. Get this book, TODAY!

A key task that any aspiring data-driven organization needs to learn is data wrangling, the process of converting raw data into something truly useful. This practical guide provides business analysts with an overview of various data wrangling techniques and tools, and puts the practice of data wrangling into context by asking, "What are you trying to do and why?" Wrangling data consumes roughly 50-80% of an analyst’s time before any kind of analysis is possible. Written by key executives at Trifacta, this book walks you through the wrangling process by exploring several factors—time, granularity, scope, and structure—that you need to consider as you begin to work with data. You’ll learn a shared language and a comprehensive understanding of data wrangling, with an emphasis on recent agile analytic processes used by many of today’s data-driven organizations. Appreciate the importance—and the satisfaction—of wrangling data the right way. Understand what kind of data is available Choose which data to use and at what level of detail Meaningfully combine multiple sources of data Decide how to distill the results to a size and shape that can drive downstream analysis

A practical, skill-based introduction to data analysis and literacy We are swimming in a world of data, and this handy guide will keep you afloat while you learn to make sense of it all. In Data Literacy: A User's Guide, David Herzog, a journalist with a decade of experience using data analysis to transform information into captivating storytelling, introduces students and professionals to the fundamentals of data literacy, a key skill in today’s world. Assuming the reader has no advanced knowledge of data analysis or statistics, this book shows how to create insight from publicly-available data through exercises using simple Excel functions. Extensively illustrated, step-by-step instructions within a concise, yet comprehensive, reference will help readers identify, obtain, evaluate, clean, analyze and visualize data. A concluding chapter introduces more sophisticated data analysis methods and tools including database managers such as Microsoft Access and MySQL and standalone statistical programs such as SPSS, SAS and R.

What is bad data? Some people consider it a technical phenomenon, like missing values or malformed records, but bad data includes a lot more. In this handbook, data expert Q. Ethan McCallum has gathered 19 colleagues from every corner of the data arena to reveal how they’ve recovered from nasty data problems. From cranky storage to poor representation to misguided policy, there are many paths to bad data. Bottom line? Bad data is data that gets in the way. This book explains effective ways to get around it. Among the many topics covered, you’ll discover how to: Test drive your data to see if it’s ready for analysis Work spreadsheet data into a usable form Handle encoding problems that lurk in text data Develop a successful web-scraping effort Use NLP tools to reveal the real sentiment of online reviews Address cloud computing issues that can impact your analysis effort Avoid policies that create data analysis roadblocks Take a systematic approach to data quality analysis

Jason W. Osborne’s Best Practices in Logistic Regression provides students with an accessible, applied approach that communicates logistic regression in clear and concise terms. The book effectively leverages readers’ basic intuitive understanding of simple and multiple regression to guide them into a sophisticated mastery of logistic regression. Osborne’s applied approach offers students and instructors a clear perspective, elucidated through practical and engaging tools that encourage student comprehension.

Doing data science is difficult. Projects are typically very dynamic with requirements that change as data understanding grows. The data itself arrives piecemeal, is added to, replaced, contains undiscovered flaws and comes from a variety of sources. Teams also have mixed skill sets and tooling is often limited. Despite these disruptions, a data science team must get off the ground fast and begin demonstrating value with traceable, tested work products. This is when you need Guerrilla Analytics. In this book, you will learn about: The Guerrilla Analytics Principles: simple rules of thumb for maintaining data provenance across the entire analytics life cycle from data extraction, through analysis to reporting. Reproducible, traceable analytics: how to design and implement work products that are reproducible, testable and stand up to external scrutiny. Practice tips and war stories: 90 practice tips and 16 war stories based on real-world project challenges encountered in consulting, pre-sales and research. Preparing for battle: how to set up your team's analytics environment in terms of tooling, skill sets, workflows and conventions. Data gymnastics: over a dozen analytics patterns that your team will encounter again and again in projects The Guerrilla Analytics Principles: simple rules of thumb for maintaining data provenance across the entire analytics life cycle from data extraction, through analysis to reporting Reproducible, traceable analytics: how to design and implement work products that are reproducible, testable and stand up to external scrutiny Practice tips and war stories: 90 practice tips and 16 war stories based on real-world project challenges encountered in consulting, pre-sales and research Preparing for battle: how to set up your team's analytics environment in terms of tooling, skill sets, workflows and conventions Data gymnastics: over a dozen analytics patterns that your team will encounter again and again in projects

This User’s Guide is intended to support the design, implementation, analysis, interpretation, and quality evaluation of registries created to increase understanding of patient outcomes. For the purposes of this guide, a patient registry is an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes. A registry database is a file (or files) derived from the registry. Although registries can serve many purposes, this guide focuses on registries created for one or more of the following purposes: to describe the natural history of disease, to determine clinical effectiveness or cost-effectiveness of health care products and services, to measure or monitor safety and harm, and/or to measure quality of care. Registries are classified according to how their populations are defined. For example, product registries include patients who have been exposed to biopharmaceutical products or medical devices. Health services registries consist of patients who have had a common procedure, clinical encounter, or hospitalization. Disease or condition registries are defined by patients having the same diagnosis, such as cystic fibrosis or heart failure. The User’s Guide was created by researchers affiliated with AHRQ’s Effective Health Care Program, particularly those who participated in AHRQ’s DEcIDE (Developing Evidence to Inform Decisions About Effectiveness) program. Chapters were subject to multiple internal and external independent reviews.

Feature engineering is a crucial step in the machine-learning pipeline, yet this topic is rarely examined on its own. With this practical book, you’ll learn techniques for extracting and transforming features—the numeric representations of raw data—into formats for machine-learning models. Each chapter guides you through a single data problem, such as how to represent text or image data. Together, these examples illustrate the main principles of feature engineering. Rather than simply teach these principles, authors Alice Zheng and Amanda Casari focus on practical application with exercises throughout the book. The closing chapter brings everything together by tackling a real-world, structured dataset with several feature-engineering techniques. Python packages including numpy, Pandas, Scikit-learn, and Matplotlib are used in code examples. You’ll examine: Feature engineering for numeric data: filtering, binning, scaling, log transforms, and power transforms Natural text techniques: bag-of-words, n-grams, and phrase detection Frequency-based filtering and feature scaling for eliminating uninformative features Encoding techniques of categorical variables, including feature hashing and bin-counting Model-based feature engineering with principal component analysis The concept of model stacking, using k-means as a featurization technique Image feature extraction with manual and deep-learning techniques

This highly practical handbook teaches you how to unlock the value of your existing metadata through cleaning, reconciliation, enrichment and linking and how to streamline the process of new metadata creation. Libraries, archives and museums are facing up to the challenge of providing access to fast growing collections whilst managing cuts to budgets. Key to this is the creation, linking and publishing of good quality metadata as Linked Data that will allow their collections to be discovered, accessed and disseminated in a sustainable manner. This highly practical handbook teaches you how to unlock the value of your existing metadata through cleaning, reconciliation, enrichment and linking and how to streamline the process of new metadata creation. Metadata experts Seth van Hooland and Ruben Verborgh introduce the key concepts of metadata standards and Linked Data and how they can be practically applied to existing metadata, giving readers the tools and understanding to achieve maximum results with limited resources. Readers will learn how to critically assess and use (semi-)automated methods of managing metadata through hands-on exercises within the book and on the accompanying website. Each chapter is built around a case study from institutions around the world, demonstrating how freely available tools are being successfully used in different metadata contexts. This handbook delivers the necessary conceptual and practical understanding to empower practitioners to make the right decisions when making their organisations resources accessible on the Web. Key topics include: - The value of metadata Metadata creation – architecture, data models and standards - Metadata cleaning - Metadata reconciliation - Metadata enrichment through Linked Data and named-entity recognition - Importing and exporting metadata - Ensuring a sustainable publishing model. Readership: This will be an invaluable guide for metadata practitioners and researchers within all cultural heritage contexts, from library cataloguers and archivists to museum curatorial staff. It will also be of interest to students and academics within information science and digital humanities fields. IT managers with responsibility for information systems, as well as strategy heads and budget holders, at cultural heritage organisations, will find this a valuable decision-making aid.

"This book introduces you to R, RStudio, and the tidyverse, a collection of R packages designed to work together to make data science fast, fluent, and fun. Suitable for readers with no previous programming experience"--

A charming, practical, and unsentimental approach to putting a home in order while reflecting on the tiny joys that make up a long life. In Sweden there is a kind of decluttering called döstädning, dö meaning “death” and städning meaning “cleaning.” This surprising and invigorating process of clearing out unnecessary belongings can be undertaken at any age or life stage but should be done sooner than later, before others have to do it for you. In The Gentle Art of Swedish Death Cleaning, artist Margareta Magnusson, with Scandinavian humor and wisdom, instructs readers to embrace minimalism. Her radical and joyous method for putting things in order helps families broach sensitive conversations, and makes the process uplifting rather than overwhelming. Margareta suggests which possessions you can easily get rid of (unworn clothes, unwanted presents, more plates than you’d ever use) and which you might want to keep (photographs, love letters, a few of your children’s art projects).

Digging into her late husband's tool shed, and her own secret drawer of vices, Margareta introduces an element of fun to a potentially daunting task. Along the way readers get a glimpse into her life in Sweden, and also become more comfortable with the idea of letting go.

There are awesome discoveries to be made and valuable stories to be told in datasets--and this book will help you uncover them. Whether you already work with data or just want to understand its possibilities, the techniques and advice in this practical book will help you learn how to better clean, evaluate, and analyze data to generate meaningful insights and compelling visualizations. Through foundational concepts and worked examples, author Susan McGregor provides the concepts and tools you need to evaluate and analyze all kinds of data and communicate your findings effectively. This book provides a methodical, jargon-free way for practitioners of all levels to harness the power of data. Use Python 3.8+ to read, write, and transform data from a variety of sources Understand and use programming basics in Python to wrangle data at scale Organize, document, and structure your code using best practices Complete exercises either on your own machine or on the web Collect data from structured data files, web pages, and APIs Perform basic statistical analysis to make meaning from data sets Visualize and present data in clear and compelling ways.

A beginner's guide to simplifying Extract, Transform, Load (ETL) processes with the help of hands-on tips, tricks, and best practices, in a fun and interactive way Key Features Explore data wrangling with the help of real-world examples and business use cases Study various ways to extract the most value from your data in minimal time Boost your knowledge with bonus topics, such as random data generation and data integrity checks Book Description While a huge amount of data is readily available to us, it is not useful in its raw form. For data to be meaningful, it must be curated and refined. If you're a beginner, then The Data Wrangling Workshop will help to break down the process for you. You'll start with the basics and build your knowledge, progressing from the core aspects behind data wrangling, to using the most popular tools and techniques. This book starts by showing you how to work with data structures using Python. Through examples and activities, you'll understand why you should stay away from traditional methods of data cleaning used in other languages and take advantage of the specialized pre-built routines in Python. Later, you'll learn how to use the same Python backend to extract and transform data from an array of sources, including the internet, large database vaults, and Excel financial tables. To help you prepare for more challenging scenarios, the book teaches you how to handle missing or incorrect data, and reformat it based on the requirements from your downstream analytics tool. By the end of this book, you will have developed a solid understanding of how to perform data wrangling with Python, and learned several techniques and best practices to extract, clean, transform, and format your data efficiently, from a diverse array of sources. What you will learn Get to grips with the fundamentals of data wrangling Understand how to model data with random data generation and data integrity checks Discover how to examine data with descriptive statistics and plotting techniques Explore how to search and retrieve information with regular expressions Delve into commonly-used Python data science libraries Become well-versed with how to handle and compensate for missing data Who this book is for The Data Wrangling Workshop is designed for developers, data analysts, and business analysts who are looking to pursue a career as a full-fledged data scientist or analytics expert. Although this book is for beginners who want to start data wrangling, prior working knowledge of the Python programming language is necessary to easily grasp the concepts covered here. It will also help to have a rudimentary knowledge of relational databases and SQL.

In this book, Steven Feuerstein, widely recognized as one of the world's experts on the Oracle PL/SQL language, distills his many years of programming, writing, and teaching about PL/SQL into a set of PL/SQL language "best practices"--rules for writing code that is readable, maintainable, and efficient. Too often, developers focus on simply writing programs that run without errors--and ignore the impact of poorly written code upon both system performance and their ability (and their colleagues' ability) to maintain that code over time.Oracle PL/SQL Best Practices is a concise, easy-to-use reference to Feuerstein's recommendations for excellent PL/SQL coding. It answers the kinds of questions PL/SQL developers most frequently ask about their code: How should I format my code? What naming conventions, if any, should I use? How can I write my packages so they can be more easily maintained? What is the most efficient way to query information from the database? How can I get all the developers on my team to handle errors the same way? The book contains 120 best practices, divided by topic area. It's full of advice on the program development process, coding style, writing SQL in PL/SQL, data structures, control structures, exception handling, program and package construction, and built-in packages. It also contains a handy, pull-out quick reference card. As a helpful supplement to the text, code examples demonstrating each of the best practices are available on the O'Reilly web site.Oracle PL/SQL Best Practices is intended as a companion to O'Reilly's larger Oracle PL/SQL books. It's a compact, readable reference that you'll turn to again and again--a book that no serious developer can afford to be without.

Cowritten by Ralph Kimball, the world's leading data warehousing authority, whose previous books have sold more than 150,000 copies Delivers real-world solutions for the most time- and labor-intensive portion of data warehousing-data staging, or the extract, transform, load (ETL) process Delineates best practices for extracting data from scattered sources, removing redundant and inaccurate data, transforming the remaining data into correctly formatted data structures, and then loading the end product into the data warehouse Offers proven time-saving ETL techniques, comprehensive guidance on building dimensional structures, and crucial advice on ensuring data quality

Looks at the principles and clean code, includes case studies showcasing the practices of writing clean code, and contains a list of heuristics and "smells" accumulated from the process of writing clean code.

Data quality is one of the most important problems in data management, since dirty data often leads to inaccurate data analytics results and incorrect business decisions. Poor data across businesses and the U.S. government are reported to cost trillions of dollars a year. Multiple surveys show that dirty data is the most common barrier faced by data scientists. Not surprisingly, developing effective and efficient data cleaning solutions is challenging and is rife with deep theoretical and engineering problems. This book is about data cleaning, which is used to refer to all kinds of tasks and activities to detect and repair errors in the data. Rather than focus on a particular data cleaning task, we give an overview of the end-to-end data cleaning process, describing various error detection and repair methods, and attempt to anchor these proposals with multiple taxonomies and views. Specifically, we cover four of the most common and important data cleaning tasks, namely, outlier detection, data transformation, error repair (including imputing missing values), and data deduplication. Furthermore, due to the increasing popularity and applicability of machine learning techniques, we include a chapter that specifically explores how machine learning techniques are used for data cleaning, and how data cleaning is used to improve machine learning models. This book is intended to serve as a useful reference for researchers and practitioners who are interested in the area of data quality and data cleaning. It can also be used as a textbook for a graduate course. Although we aim at covering state-of-the-art algorithms and techniques, we recognize that data cleaning is still an active field of research and therefore provide future directions of research whenever appropriate.