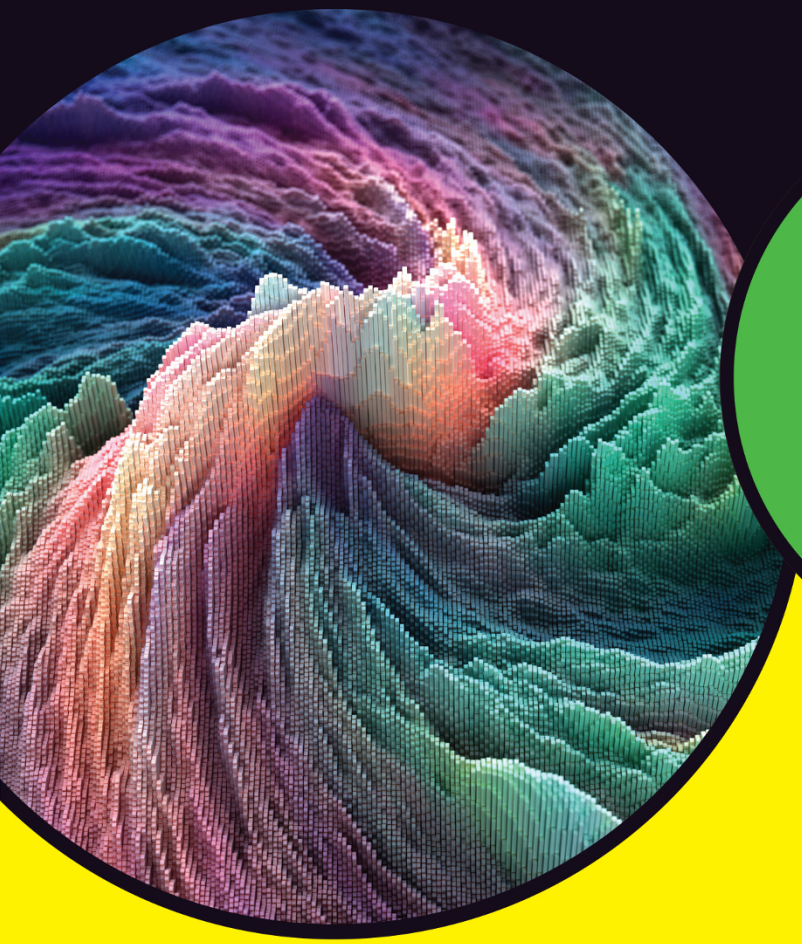
LEARNING MADE EASY

Data Science Programming

ALL-IN-ONE

for dummies®
A Wiley Brand



6 Books in one!

John Paul Mueller
Luca Massaron, GDE

Data Science Programming

ALL-IN-ONE

by John Paul Mueller and
Luca Massaron

for dummies®
A Wiley Brand

Data Science Programming All-in-One For Dummies®

Published by: **John Wiley & Sons, Inc.,** 111 River Street, Hoboken, NJ 07030-5774, www.wiley.com

Library of Congress Control Number: 2019954497

ISBN 978-1-119-62611-4; ISBN 978-1-119-62613-8 (ebk); ISBN 978-1-119-62614-5 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents at a Glance

Table of Contents

Introduction

Data science is a term that the media has chosen to minimize, obfuscate, and sometimes misuse. It involves a lot more than just data and the science of working with data. Today, the world uses data science in all sorts of ways that you might not know about, which is why you need *Data Science Programming All-in-One For Dummies*.

In the book, you start with both the data and the science of manipulating it, but then you go much further. In addition to seeing how to perform a wide range of analysis, you also delve into making recommendations, classifying real-world objects, analyzing audio, and even creating art.

However, you don't just learn about amazing new technologies and how to perform common tasks. This book also dispels myths created by people who wish data science were something different than it really is or who don't understand it at all. A great deal of misinformation swirls around the world today as the media seeks to sensationalize, anthropomorphize, and emotionalize technologies that are, in fact, quite mundane. It's hard to know what to believe. You find reports that robots are on the cusp of becoming sentient and that the giant tech companies can discover your innermost thoughts simply by reviewing your record of purchases. With this book, you can replace disinformation with solid facts, and you can use those facts to create a strategy for performing data science development tasks.

About This Book

You might find that this book starts off a little slowly because most people don't have a good grasp on getting a system prepared for data science use. Book 1 helps you configure your system. The book uses Jupyter Notebook as an Integrated Development Environment (IDE) for both Python and R. That way, if you choose to view the examples in both languages, you use the same IDE to do it. Jupyter Notebook also relies on the literate programming strategy first proposed by Donald Knuth (see `http://www.literateprogramming.com/`) to make your coding efforts significantly easier and more focused on the data. In addition, in contrast to other environments, you don't actually write entire applications before you see something; you write code and focus on the results of just that code block as part of a whole application.

After you have a development environment installed and ready to use, you can start working with data in all its myriad forms in Book 2. This book covers a great many of these forms — everything from in-memory datasets to those found on large websites. In addition, you see a number of data formats ranging from flat files to Relational Database Management Systems (RDBMSs) and Not Only SQL (NoSQL) databases.

Of course, manipulating data is worthwhile only if you can do something useful with it. Book 3 discusses common sorts of analysis, such as linear and logistic regression, Bayes' Theorem, and K-Nearest Neighbors (KNN).

Most data science books stop at this point. In this book, however, you discover AI, machine learning, and deep learning techniques to get more out of your data than you might have thought possible. This exciting part of the book, Book 4, represents the cutting edge of analysis. You use huge datasets to discover important information about large groups of people that will help you improve their health or sell them products.

Performing analysis may be interesting, but analysis is only a step along the path. Book 5 shows you how to put your analysis to use in recommender systems, to classify objects, work with nontextual data like music and video, and display the results of an analysis in a form that everyone can appreciate.

The final minibook, Book 6, offers something you won't find in many places, not even online. You discover how to detect and fix problems with your data, the logic used to interpret the data, and the code used to perform tasks such as analysis. By the time you complete Book 6, you'll know much more about how to ensure that the results you get are actually the results you need and want.

To make absorbing the concepts easy, this book uses the following conventions:

- Text that you're meant to type just as it appears in the book is in **bold**. The exception is when you're working through a step list: Because each step is bold, the text to type is not bold.
- When you see words in *italics* as part of a typing sequence, you need to replace that value with something that works for you. For example, if you see "Type ***Your Name*** and press Enter," you need to replace *Your Name* with your actual name.
- Web addresses and programming code appear in `monofont`. If you're reading a digital version of this book on a device connected to the Internet, you can click or tap the web address to visit that website, like this: `https://www.dummies.com`.
- When you need to type command sequences, you see them separated by a special arrow, like this: File ➪ New File. In this example, you go to the File menu first and then select the New File entry on that menu.

Foolish Assumptions

You might find it difficult to believe that we've assumed anything about you — after all; we haven't even met you yet! Although most assumptions are indeed foolish, we made these assumptions to provide a starting point for the book.

You need to be familiar with the platform you want to use because the book doesn't offer any guidance in this regard. (Book 1, Chapter 3 does, however, provide Anaconda installation instructions for both Python and R, and Book 1, Chapter 5 helps you install the TensorFlow and Keras frameworks used for this book.) To give you the maximum information about Python concerning how it applies to deep learning, this book doesn't discuss any platform-specific issues. You see the R version of the Python coding examples in the downloadable source, along with R-specific notes on usage and development. You really do need to know how to install applications, use applications, and generally work with your chosen platform before you begin working with this book.

You must know how to work with Python or R. You can find a wealth of Python tutorials online (see `https://www.w3schools.com/python/` and `https://www.tutorialspoint.com/python/` as examples). R, likewise, provides a wealth of online tutorials (see `https://www.tutorialspoint.com/r/index.htm`, `https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/`, and `https://www.statmethods.net/r-tutorial/index.html` as examples).

This book isn't a math primer. Yes, you see many examples of complex math, but the emphasis is on helping you use Python or R to perform data science development tasks rather than teaching math theory. We include some examples that also discuss the use of technologies such as data management (see Book 2), statistical analysis (see Book 3), AI, machine learning, deep learning (see Book 4), practical data science application (see Book 5), and troubleshooting both data and code (see Book 6). Book 1, Chapters 1 and 2 give you a better understanding of precisely what you need to know to use this book successfully. You also use a considerable number of libraries in writing code for this book. Book 1, Chapter 4 discusses library use and suggests other libraries that you might want to try.

This book also assumes that you can access items on the Internet. Sprinkled throughout are numerous references to online material that will enhance your learning experience. However, these added sources are useful only if you actually find and use them.

Icons Used in This Book

As you read this book, you see icons in the margins that indicate material of interest (or not, as the case may be). This section briefly describes each icon in this book.


TIP

Tips are nice because they help you save time or perform some task without a lot of extra work. The tips in this book are time-saving techniques or pointers to resources that you should try so that you can get the maximum benefit from Python or R, or from performing deep learning–related tasks. (Note that R developers will also find copious notes in the source code files for issues that differ significantly from Python.)


WARNING

We don't want to sound like angry parents or some kind of maniacs, but you should avoid doing anything that's marked with a Warning icon. Otherwise, you might find that your application fails to work as expected, you get incorrect answers from seemingly bulletproof algorithms, or (in the worst-case scenario) you lose data.


TECHNICAL STUFF

Whenever you see this icon, think advanced tip or technique. You might find these tidbits of useful information just too boring for words, or they could contain the solution you need to get a program running. Skip these bits of information whenever you like.


REMEMBER

If you don't get anything else out of a particular chapter or section, remember the material marked by this icon. This text usually contains an essential process or a bit of information that you must know to work with Python or R, or to perform deep learning–related tasks successfully. (Note that the R source code files contain a great deal of text that gives essential details for working with R when R differs considerably from Python.)

Beyond the Book

This book isn't the end of your Python or R data science development experience — it's really just the beginning. We provide online content to make this book more flexible and better able to meet your needs. That way, as we receive email from you, we can address questions and tell you how updates to Python, R, or their associated add-ons affect book content. In fact, you gain access to all these cool additions:

- **Cheat sheet:** You remember using crib notes in school to make a better mark on a test, don't you? You do? Well, a cheat sheet is sort of like that. It provides you with some special notes about tasks that you can do with Python and R with regard to data science development that not every other person knows. You can find the cheat sheet by going to `www.dummies.com`, searching this book's title, and scrolling down the page that appears. The cheat sheet contains really neat information, such as the most common data errors that cause people problems with working in the data science field.
- **Updates:** Sometimes changes happen. For example, we might not have seen an upcoming change when we looked into our crystal ball during the writing of this book. In the past, this possibility simply meant that the book became outdated and less useful, but you can now find updates to the book, if we have any, by searching this book's title at `www.dummies.com`.

  In addition to these updates, check out the blog posts with answers to reader questions and demonstrations of useful, book-related techniques at `http://blog.johnmuellerbooks.com/`.
- **Companion files:** Hey! Who really wants to type all the code in the book and reconstruct all those neural networks manually? Most readers prefer to spend their time actually working with data and seeing the interesting things they can do, rather than typing. Fortunately for you, the examples used in the book are available for download, so all you need to do is read the book to learn Python or R data science programming techniques. You can find these files at `www.dummies.com`. Search this book's title, and on the page that appears, scroll down to the image of the book cover and click it. Then click the More about This Book button and on the page that opens, go to the Downloads tab.

Where to Go from Here

It's time to start your Python or R for data science programming adventure! If you're completely new to Python or R and its use for data science tasks, you should start with Book 1, Chapter 1. Progressing through the book at a pace that allows you to absorb as much of the material as possible makes it feasible for you to gain insights that you might not otherwise gain if you read the chapters in a random order. However, the book is designed to allow you to read the material in any order desired.

If you're a novice who's in an absolute rush to get going with Python or R for data science programming as quickly as possible, you can skip to Book 1, Chapter 3 with the understanding that you may find some topics a bit confusing later. Skipping to Book 1, Chapter 5 is okay if you already have Anaconda (the programming product used in the book) installed with the appropriate language (Python or R as you desire), but be sure to at least skim Chapter 3 so that you know what assumptions we made when writing this book.

This book relies on a combination of TensorFlow and Keras to perform deep learning tasks. Even if you're an advanced reader who wants to perform deep learning tasks, you need to go to Book 1, Chapter 5 to discover how to configure the environment used for this book. You must configure the environment according to instructions or you're likely to experience failures when you try to run the code. However, this issue applies only to deep learning. This book has a great deal to offer in other areas, such as data manipulation and statistical analysis.

1 Defining Data Science

Contents at a Glance

IN THIS CHAPTER

- Understanding data science history and uses
- Considering the flow of data in data science
- Working with various languages in data science
- Performing data science tasks quickly

Chapter 1

Considering the History and Uses of Data Science

The burgeoning uses for data in the world today, along with the explosion of data sources, create a demand for people who have special skills to obtain, manage, and analyze information for the benefit of everyone. The data scientist develops and hones these special skills to perform such tasks on multiple levels, as described in the first two sections of this chapter.

Data needs to be funneled into acceptable forms that allow data scientists to perform their tasks. Even though the precise data flow varies, you can generalize it to a degree. The third section of the chapter gives you an overview of how data flow occurs.

As with anyone engaged in computer work today, a data scientist employs various programming languages to express the manipulation of data in a repeatable manner. The languages that a data scientist uses, however, focus on outputs expected from given inputs, rather than on low-level control or a precise procedure, as a computer scientist would use. Because a data scientist may lack a formal programming education, the languages tend to focus on declarative strategies, with the data scientist expressing a desired outcome rather than devising a specific procedure. The fourth section of the chapter discusses various languages used by data scientists, with an emphasis on Python and R.

The final section of the chapter provides a very quick overview of getting tasks done quickly. Optimization without loss of precision is an incredibly difficult task and you see it covered a number of times in this book, but this introduction is enough to get you started. The overall goal of this first chapter is to describe data science and explain how a data scientist uses algorithms, statistics, data extraction, data manipulation, and a slew of other technologies to employ it as part of an analysis.


REMEMBER

You don't have to type the source code for this chapter manually (or, actually at all, given that you use it only to obtain an understanding of the data flow process). In fact, using the downloadable source is a lot easier. The source code for this chapter appears in the `DSPD_0101_Quick_Overview.ipynb` source code file for Python. See the Introduction for details on how to find these source files.

Considering the Elements of Data Science

At one point, the world viewed anyone working with statistics as a sort of accountant or perhaps a mad scientist. Many people consider statistics and the analysis of data boring. However, data science is one of those occupations in which the more you learn, the more you want to learn. Answering one question often spawns more questions that are even more interesting than the one you just answered. However, what makes data science so sexy is that you see it everywhere, used in an almost infinite number of ways. The following sections give you more details on why data science is such an amazing field of study.

Considering the emergence of data science

Data science is a relatively new term. William S. Cleveland coined the term in 2001 as part of a paper entitled "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." It wasn't until a year later that the International Council for Science actually recognized data science and created a committee for it. Columbia University got into the act in 2003 by beginning publication of the *Journal of Data Science.*


REMEMBER

However, the mathematical basis behind data science is centuries old because data science is essentially a method of viewing and analyzing statistics and probability. The first essential use of statistics as a term comes in 1749, but statistics are certainly much older than that. People have used statistics to recognize patterns for thousands of years. For example, the historian Thucydides (in his History of the Peloponnesian War) describes how the Athenians calculated the height of the wall of Platea in fifth century BC by counting bricks in an unplastered section of