

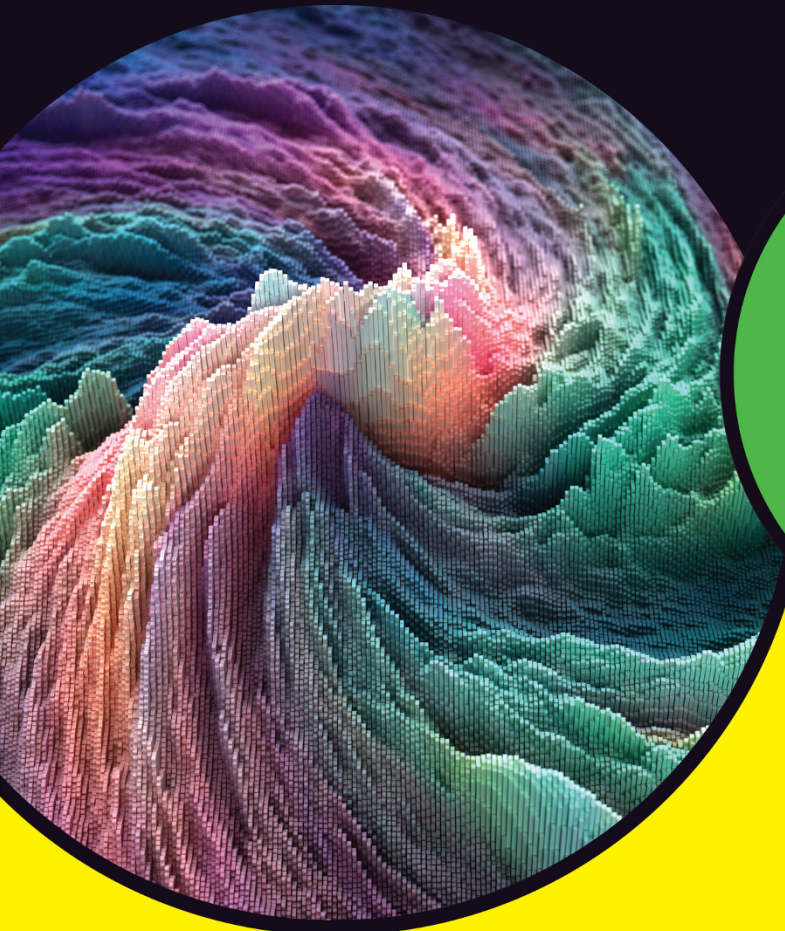
LEARNING MADE EASY



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

Copyright © 2020 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Trademarks: Wiley, For Dummies, the Dummies Man logo, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002. For technical support, please visit <https://hub.wiley.com/community/support/dummies>.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit 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

Introduction	1
Book 1: Defining Data Science	7
CHAPTER 1: Considering the History and Uses of Data Science	9
CHAPTER 2: Placing Data Science within the Realm of AI	29
CHAPTER 3: Creating a Data Science Lab of Your Own	51
CHAPTER 4: Considering Additional Packages and Libraries You Might Want	81
CHAPTER 5: Leveraging a Deep Learning Framework	95
Book 2: Interacting with Data Storage	109
CHAPTER 1: Manipulating Raw Data	111
CHAPTER 2: Using Functional Programming Techniques	131
CHAPTER 3: Working with Scalars, Vectors, and Matrices	171
CHAPTER 4: Accessing Data in Files	201
CHAPTER 5: Working with a Relational DBMS	223
CHAPTER 6: Working with a NoSQL DBMS	237
Book 3: Manipulating Data Using Basic Algorithms	253
CHAPTER 1: Working with Linear Regression	255
CHAPTER 2: Moving Forward with Logistic Regression	289
CHAPTER 3: Predicting Outcomes Using Bayes	309
CHAPTER 4: Learning with K-Nearest Neighbors	335
Book 4: Performing Advanced Data Manipulation	351
CHAPTER 1: Leveraging Ensembles of Learners	353
CHAPTER 2: Building Deep Learning Models	373
CHAPTER 3: Recognizing Images with CNNs	409
CHAPTER 4: Processing Text and Other Sequences	453
Book 5: Performing Data-Related Tasks	491
CHAPTER 1: Making Recommendations	493
CHAPTER 2: Performing Complex Classifications	509
CHAPTER 3: Identifying Objects	525
CHAPTER 4: Analyzing Music and Video	543
CHAPTER 5: Considering Other Task Types	559
CHAPTER 6: Developing Impressive Charts and Plots	579

Book 6: Diagnosing and Fixing Errors	619
CHAPTER 1: Locating Errors in Your Data.	621
CHAPTER 2: Considering Outrageous Outcomes	643
CHAPTER 3: Dealing with Model Overfitting and Underfitting	663
CHAPTER 4: Obtaining the Correct Output Presentation	689
CHAPTER 5: Developing Consistent Strategies	707
Index	721

Table of Contents

INTRODUCTION	1
About This Book	1
Foolish Assumptions	3
Icons Used in This Book	4
Beyond the Book	4
Where to Go from Here	5
BOOK 1: DEFINING DATA SCIENCE	7
CHAPTER 1: Considering the History and Uses of Data Science	9
Considering the Elements of Data Science	10
Considering the emergence of data science	10
Outlining the core competencies of a data scientist	11
Linking data science, big data, and AI	12
Understanding the role of programming	12
Defining the Role of Data in the World	13
Enticing people to buy products	13
Keeping people safer	14
Creating new technologies	15
Performing analysis for research	16
Providing art and entertainment	17
Making life more interesting in other ways	18
Creating the Data Science Pipeline	18
Preparing the data	18
Performing exploratory data analysis	18
Learning from data	19
Visualizing	19
Obtaining insights and data products	19
Comparing Different Languages Used for Data Science	20
Obtaining an overview of data science languages	20
Defining the pros and cons of using Python	22
Defining the pros and cons of using R	23
Learning to Perform Data Science Tasks Fast	25
Loading data	26
Training a model	26
Viewing a result	26

CHAPTER 2:	Placing Data Science within the Realm of AI	29
	Seeing the Data to Data Science Relationship	30
	Considering the data architecture	30
	Acquiring data from various sources	31
	Performing data analysis	32
	Archiving the data	33
	Defining the Levels of AI	33
	Beginning with AI	34
	Advancing to machine learning	39
	Getting detailed with deep learning	43
	Creating a Pipeline from Data to AI	47
	Considering the desired output	47
	Defining a data architecture	47
	Combining various data sources	47
	Checking for errors and fixing them	48
	Performing the analysis	48
	Validating the result	49
	Enhancing application performance	49
CHAPTER 3:	Creating a Data Science Lab of Your Own	51
	Considering the Analysis Platform Options	52
	Using a desktop system	53
	Working with an online IDE	53
	Considering the need for a GPU	54
	Choosing a Development Language	56
	Obtaining and Using Python	58
	Working with Python in this book	58
	Obtaining and installing Anaconda for Python	59
	Defining a Python code repository	64
	Working with Python using Google Colaboratory	69
	Defining the limits of using Azure Notebooks with Python and R	71
	Obtaining and Using R	72
	Obtaining and installing Anaconda for R	72
	Starting the R environment	73
	Defining an R code repository	75
	Presenting Frameworks	76
	Defining the differences	76
	Explaining the popularity of frameworks	77
	Choosing a particular library	79
	Accessing the Downloadable Code	80

CHAPTER 4:	Considering Additional Packages and Libraries You Might Want	81
	Considering the Uses for Third-Party Code	82
	Obtaining Useful Python Packages	83
	Accessing scientific tools using SciPy	84
	Performing fundamental scientific computing using NumPy	85
	Performing data analysis using pandas	85
	Implementing machine learning using Scikit-learn	86
	Going for deep learning with Keras and TensorFlow	86
	Plotting the data using matplotlib	87
	Creating graphs with NetworkX	88
	Parsing HTML documents using BeautifulSoup	88
	Locating Useful R Libraries	89
	Using your Python code in R with reticulate	89
	Conducting advanced training using caret	90
	Performing machine learning tasks using mlr	90
	Visualizing data using ggplot2	91
	Enhancing ggplot2 using esquisse	91
	Creating graphs with igraph	91
	Parsing HTML documents using rvest	92
	Wrangling dates using lubridate	92
	Making big data simpler using dplyr and purrr	93
CHAPTER 5:	Leveraging a Deep Learning Framework	95
	Understanding Deep Learning Framework Usage	96
	Working with Low-End Frameworks	97
	Chainer	97
	PyTorch	98
	MXNet	98
	Microsoft Cognitive Toolkit/CNTK	99
	Understanding TensorFlow	100
	Grasping why TensorFlow is so good	101
	Making TensorFlow easier by using TFLearn	102
	Using Keras as the best simplifier	102
	Getting your copy of TensorFlow and Keras	103
	Fixing the C++ build tools error in Windows	106
	Accessing your new environment in Notebook	108
	BOOK 2: INTERACTING WITH DATA STORAGE	109
CHAPTER 1:	Manipulating Raw Data	111
	Defining the Data Sources	112
	Obtaining data locally	112
	Using online data sources	117

Employing dynamic data sources	121
Considering other kinds of data sources	123
Considering the Data Forms	124
Working with pure text	124
Accessing formatted text	125
Deciphering binary data	126
Understanding the Need for Data Reliability	128
CHAPTER 2: Using Functional Programming Techniques	131
Defining Functional Programming	132
Differences with other programming paradigms	132
Understanding its goals	133
Understanding Pure and Impure Languages	134
Using the pure approach	134
Using the impure approach	134
Comparing the Functional Paradigm	135
Imperative	135
Procedural	136
Object-oriented	136
Declarative	136
Using Python for Functional Programming Needs	137
Understanding How Functional Data Works	138
Working with immutable data	139
Considering the role of state	139
Eliminating side effects	140
Passing by reference versus by value	140
Working with Lists and Strings	142
Creating lists	144
Evaluating lists	144
Performing common list manipulations	146
Understanding the Dict and Set alternatives	147
Considering the use of strings	148
Employing Pattern Matching	150
Looking for patterns in data	150
Understanding regular expressions	152
Using pattern matching in analysis	155
Working with pattern matching	156
Working with Recursion	159
Performing tasks more than once	159
Understanding recursion	161
Using recursion on lists	162
Considering advanced recursive tasks	163
Passing functions instead of variables	164

Performing Functional Data Manipulation	165
Slicing and dicing	166
Mapping your data	167
Filtering data	168
Organizing data	169
CHAPTER 3: Working with Scalars, Vectors, and Matrices.....	171
Considering the Data Forms	172
Defining Data Type through Scalars.....	173
Creating Organized Data with Vectors.....	174
Defining a vector	175
Creating vectors of a specific type	175
Performing math on vectors	176
Performing logical and comparison tasks on vectors	176
Multiplying vectors	177
Creating and Using Matrices	178
Creating a matrix.....	178
Creating matrices of a specific type	179
Using the matrix class.....	181
Performing matrix multiplication	181
Executing advanced matrix operations	183
Extending Analysis to Tensors.....	185
Using Vectorization Effectively.....	186
Selecting and Shaping Data	187
Slicing rows.....	188
Slicing columns	188
Dicing.....	189
Concatenating	189
Aggregating	194
Working with Trees	195
Understanding the basics of trees	195
Building a tree	196
Representing Relations in a Graph.....	198
Going beyond trees.....	198
Arranging graphs.....	199
CHAPTER 4: Accessing Data in Files.....	201
Understanding Flat File Data Sources	202
Working with Positional Data Files	203
Accessing Data in CSV Files	205
Working with a simple CSV file	205
Making use of header information.....	208

Moving On to XML Files	209
Working with a simple XML file	209
Parsing XML	211
Using XPath for data extraction	212
Considering Other Flat-File Data Sources	214
Working with Nontext Data	215
Downloading Online Datasets	218
Working with package datasets	218
Using public domain datasets	219
CHAPTER 5: Working with a Relational DBMS	223
Considering RDBMS Issues	224
Defining the use of tables	225
Understanding keys and indexes	226
Using local versus online databases	227
Working in read-only mode	228
Accessing the RDBMS Data	228
Using the SQL language	229
Relying on scripts	231
Relying on views	231
Relying on functions	232
Creating a Dataset	233
Combining data from multiple tables	233
Ensuring data completeness	234
Slicing and dicing the data as needed	234
Mixing RDBMS Products	234
CHAPTER 6: Working with a NoSQL DMBS	237
Considering the Ramifications of Hierarchical Data	238
Understanding hierarchical organization	238
Developing strategies for freeform data	239
Performing an analysis	240
Working around dangling data	241
Accessing the Data	243
Creating a picture of the data form	243
Employing the correct transiting strategy	244
Ordering the data	247
Interacting with Data from NoSQL Databases	248
Working with Dictionaries	249
Developing Datasets from Hierarchical Data	250
Processing Hierarchical Data into Other Forms	251

BOOK 3: MANIPULATING DATA USING BASIC ALGORITHMS	253
CHAPTER 1: Working with Linear Regression	255
Considering the History of Linear Regression	256
Combining Variables	257
Working through simple linear regression	257
Advancing to multiple linear regression	260
Considering which question to ask	262
Reducing independent variable complexity	263
Manipulating Categorical Variables	265
Creating categorical variables	266
Renaming levels	267
Combining levels	268
Using Linear Regression to Guess Numbers	269
Defining the family of linear models	270
Using more variables in a larger dataset	271
Understanding variable transformations	274
Doing variable transformations	275
Creating interactions between variables	277
Understanding limitations and problems	282
Learning One Example at a Time	283
Using Gradient Descent	283
Implementing Stochastic Gradient Descent	283
Considering the effects of regularization	287
CHAPTER 2: Moving Forward with Logistic Regression	289
Considering the History of Logistic Regression	290
Differentiating between Linear and Logistic Regression	291
Considering the model	291
Defining the logistic function	292
Understanding the problems that logistic regression solves	294
Fitting the curve	295
Considering a pass/fail example	296
Using Logistic Regression to Guess Classes	297
Applying logistic regression	297
Considering when classes are more	298
Defining logistic regression performance	300
Switching to Probabilities	301
Specifying a binary response	301
Transforming numeric estimates into probabilities	302
Working through Multiclass Regression	305
Understanding multiclass regression	305
Developing a multiclass regression implementation	306

CHAPTER 3: Predicting Outcomes Using Bayes	309
Understanding Bayes' Theorem	310
Delving into Bayes history	310
Considering the basic theorem	312
Using Naïve Bayes for Predictions	313
Finding out that Naïve Bayes isn't so naïve	314
Predicting text classifications	315
Getting an overview of Bayesian inference	318
Working with Networked Bayes	324
Considering the network types and uses	324
Understanding Directed Acyclic Graphs (DAGs)	327
Employing networked Bayes in predictions	328
Deciding between automated and guided learning	332
Considering the Use of Bayesian Linear Regression	332
Considering the Use of Bayesian Logistic Regression	333
CHAPTER 4: Learning with K-Nearest Neighbors	335
Considering the History of K-Nearest Neighbors	336
Learning Lazily with K-Nearest Neighbors	337
Understanding the basis of KNN	337
Predicting after observing neighbors	338
Choosing the k parameter wisely	341
Leveraging the Correct k Parameter	342
Understanding the k parameter	342
Experimenting with a flexible algorithm	343
Implementing KNN Regression	345
Implementing KNN Classification	347
BOOK 4: PERFORMING ADVANCED DATA MANIPULATION	351
CHAPTER 1: Leveraging Ensembles of Learners	353
Leveraging Decision Trees	354
Growing a forest of trees	356
Seeing Random Forests in action	358
Understanding the importance measures	360
Configuring your system for importance measures with Python	361
Seeing importance measures in action	361
Working with Almost Random Guesses	364
Understanding the premise	365
Bagging predictors with AdaBoost	366

Meeting Again with Gradient Descent	369
Understanding the GBM difference	369
Seeing GBM in action	371
Averaging Different Predictors	372
CHAPTER 2: Building Deep Learning Models	373
Discovering the Incredible Perceptron	374
Understanding perceptron functionality	375
Touching the nonseparability limit	376
Hitting Complexity with Neural Networks	378
Considering the neuron	379
Pushing data with feed-forward	381
Defining hidden layers	383
Executing operations	384
Considering the details of data movement through the neural network.	386
Using backpropagation to adjust learning.	387
Understanding More about Neural Networks	390
Getting an overview of the neural network process	391
Defining the basic architecture	391
Documenting the essential modules	393
Solving a simple problem	396
Looking Under the Hood of Neural Networks	399
Choosing the right activation function	399
Relying on a smart optimizer	401
Setting a working learning rate	402
Explaining Deep Learning Differences with Other Forms of AI	402
Adding more layers	403
Changing the activations	405
Adding regularization by dropout	406
Using online learning	407
Transferring learning	407
Learning end to end	408
CHAPTER 3: Recognizing Images with CNNs	409
Beginning with Simple Image Recognition	410
Considering the ramifications of sight	410
Working with a set of images	411
Extracting visual features	417
Recognizing faces using Eigenfaces	419
Classifying images	423
Understanding CNN Image Basics	427
Moving to CNNs with Character Recognition	429
Accessing the dataset	430
Reshaping the dataset	431

Encoding the categories	432
Defining the model	432
Using the model.	433
Explaining How Convolutions Work	435
Understanding convolutions	435
Simplifying the use of pooling	439
Describing the LeNet architecture	440
Detecting Edges and Shapes from Images	446
Visualizing convolutions	447
Unveiling successful architectures	449
Discussing transfer learning	450
CHAPTER 4: Processing Text and Other Sequences	453
Introducing Natural Language Processing.	454
Defining the human perspective as it relates to data science . .	454
Considering the computer perspective as it relates to data science	455
Understanding How Machines Read	456
Creating a corpus	457
Performing feature extraction.	457
Understanding the BoW.	458
Processing and enhancing text	459
Maintaining order using n-grams	461
Stemming and removing stop words	462
Scraping textual datasets from the web	465
Handling problems with raw text	470
Storing processed text data in sparse matrices	473
Understanding Semantics Using Word Embeddings	478
Using Scoring and Classification	482
Performing classification tasks	482
Analyzing reviews from e-commerce	485
BOOK 5: PERFORMING DATA-RELATED TASKS	491
CHAPTER 1: Making Recommendations	493
Realizing the Recommendation Revolution.	494
Downloading Rating Data.	495
Navigating through anonymous web data	496
Encountering the limits of rating data	499
Leveraging SVD	506
Considering the origins of SVD	506
Understanding the SVD connection	508

CHAPTER 2: Performing Complex Classifications	509
Using Image Classification Challenges	510
Delving into ImageNet and Coco	511
Learning the magic of data augmentation	513
Distinguishing Traffic Signs	516
Preparing the image data	517
Running a classification task	520
CHAPTER 3: Identifying Objects	525
Distinguishing Classification Tasks	526
Understanding the problem	526
Performing localization	527
Classifying multiple objects	528
Annotating multiple objects in images	529
Segmenting images	530
Perceiving Objects in Their Surroundings	531
Considering vision needs in self-driving cars	531
Discovering how RetinaNet works	532
Using the Keras-RetinaNet code	534
Overcoming Adversarial Attacks on Deep Learning Applications	538
Tricking pixels	539
Hacking with stickers and other artifacts	541
CHAPTER 4: Analyzing Music and Video	543
Learning to Imitate Art and Life	544
Transferring an artistic style	545
Reducing the problem to statistics	546
Understanding that deep learning doesn't create	548
Mimicking an Artist	548
Defining a new piece based on a single artist	549
Combining styles to create new art	550
Visualizing how neural networks dream	551
Using a network to compose music	551
Other creative avenues	552
Moving toward GANs	553
Finding the key in the competition	554
Considering a growing field	556
CHAPTER 5: Considering Other Task Types	559
Processing Language in Texts	560
Considering the processing methodologies	560
Defining understanding as tokenization	561
Putting all the documents into a bag	562
Using AI for sentiment analysis	566

Processing Time Series	574
Defining sequences of events	574
Performing a prediction using LSTM	575
CHAPTER 6: Developing Impressive Charts and Plots	579
Starting a Graph, Chart, or Plot	580
Understanding the differences between graphs, charts, and plots	580
Considering the graph, chart, and plot types	582
Defining the plot	583
Drawing multiple lines	584
Drawing multiple plots	584
Saving your work	586
Setting the Axis, Ticks, and Grids	587
Getting the axis	587
Formatting the ticks	590
Adding grids	590
Defining the Line Appearance	591
Working with line styles	592
Adding markers	593
Using Labels, Annotations, and Legends	594
Adding labels	595
Annotating the chart	596
Creating a legend	598
Creating Scatterplots	599
Depicting groups	599
Showing correlations	600
Plotting Time Series	603
Representing time on axes	604
Plotting trends over time	605
Plotting Geographical Data	608
Getting the toolkit	608
Drawing the map	609
Plotting the data	613
Visualizing Graphs	615
Understanding the adjacency matrix	615
Using NetworkX basics	615
BOOK 6: DIAGNOSING AND FIXING ERRORS	619
CHAPTER 1: Locating Errors in Your Data	621
Considering the Types of Data Errors	622
Obtaining the Required Data	624
Considering the data sources	624
Obtaining reliable data	625

Making human input more reliable	626
Using automated data collection	628
Validating Your Data	629
Figuring out what's in your data	629
Removing duplicates.....	631
Creating a data map and a data plan	632
Manicuring the Data	634
Dealing with missing data	634
Considering data misalignments.....	639
Separating out useful data.....	640
Dealing with Dates in Your Data	640
Formatting date and time values	641
Using the right time transformation.....	641
CHAPTER 2: Considering Outrageous Outcomes.....	643
Deciding What Outrageous Means.....	644
Considering the Five Mistruths in Data	645
Commission	645
Omission.....	646
Perspective.....	646
Bias	647
Frame-of-reference	648
Considering Detection of Outliers.....	649
Understanding outlier basics.....	649
Finding more things that can go wrong.....	651
Understanding anomalies and novel data.....	651
Examining a Simple Univariate Method.....	653
Using the pandas package	653
Leveraging the Gaussian distribution.....	655
Making assumptions and checking out	656
Developing a Multivariate Approach	657
Using principle component analysis.....	658
Using cluster analysis	659
Automating outliers detection with Isolation Forests	661
CHAPTER 3: Dealing with Model Overfitting and Underfitting.....	663
Understanding the Causes.....	664
Considering the problem	664
Looking at underfitting.....	665
Looking at overfitting	666
Plotting learning curves for insights.....	668

	Determining the Sources of Overfitting and Underfitting	670
	Understanding bias and variance	671
	Having insufficient data	671
	Being fooled by data leakage	672
	Guessing the Right Features	672
	Selecting variables like a pro	673
	Using nonlinear transformations	676
	Regularizing linear models	684
CHAPTER 4:	Obtaining the Correct Output Presentation	689
	Considering the Meaning of Correct	690
	Determining a Presentation Type	691
	Considering the audience	691
	Defining a depth of detail	692
	Ensuring that the data is consistent with audience needs	693
	Understanding timeliness	693
	Choosing the Right Graph	694
	Telling a story with your graphs	694
	Showing parts of a whole with pie charts	694
	Creating comparisons with bar charts	695
	Showing distributions using histograms	697
	Depicting groups using boxplots	699
	Defining a data flow using line graphs	700
	Seeing data patterns using scatterplots	701
	Working with External Data	702
	Embedding plots and other images	703
	Loading examples from online sites	703
	Obtaining online graphics and multimedia	704
CHAPTER 5:	Developing Consistent Strategies	707
	Standardizing Data Collection Techniques	707
	Using Reliable Sources	709
	Verifying Dynamic Data Sources	711
	Considering the problem	712
	Analyzing streams with the right recipe	714
	Looking for New Data Collection Trends	715
	Weeding Old Data	716
	Considering the Need for Randomness	717
	Considering why randomization is needed	718
	Understanding how probability works	718
	INDEX	721

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.johnmuelเลอร์books.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

CHAPTER 1:	Considering the History and Uses of Data Science	9
	Considering the Elements of Data Science	10
	Defining the Role of Data in the World	13
	Creating the Data Science Pipeline	18
	Comparing Different Languages Used for Data Science	20
	Learning to Perform Data Science Tasks Fast	25
CHAPTER 2:	Placing Data Science within the Realm of AI	29
	Seeing the Data to Data Science Relationship	30
	Defining the Levels of AI	33
	Creating a Pipeline from Data to AI	47
CHAPTER 3:	Creating a Data Science Lab of Your Own	51
	Considering the Analysis Platform Options	52
	Choosing a Development Language	56
	Obtaining and Using Python	58
	Obtaining and Using R	72
	Presenting Frameworks	76
	Accessing the Downloadable Code	80
CHAPTER 4:	Considering Additional Packages and Libraries You Might Want	81
	Considering the Uses for Third-Party Code	82
	Obtaining Useful Python Packages	83
	Locating Useful R Libraries	89
CHAPTER 5:	Leveraging a Deep Learning Framework	95
	Understanding Deep Learning Framework Usage	96
	Working with Low-End Frameworks	97
	Understanding TensorFlow	100

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 Plataea in fifth century BC by counting bricks in an unplastered section of