



Data Science

3rd Edition

by Lillian Pierson

**for
dummies®**
A Wiley Brand

Data Science For Dummies®, 3rd Edition

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

Copyright © 2021 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: 2021944259

ISBN 978-1-119-81155-8 (pbk); ISBN 978-1-119-81166-4 (ebk); ISBN 978-1-119-81161-9 (ebk)

Contents at a Glance

Introduction	1
Part 1: Getting Started with Data Science	5
CHAPTER 1: Wrapping Your Head Around Data Science	7
CHAPTER 2: Tapping into Critical Aspects of Data Engineering	19
Part 2: Using Data Science to Extract Meaning from Your Data	37
CHAPTER 3: Machine Learning Means ... Using a Machine to Learn from Data	39
CHAPTER 4: Math, Probability, and Statistical Modeling	51
CHAPTER 5: Grouping Your Way into Accurate Predictions	77
CHAPTER 6: Coding Up Data Insights and Decision Engines	103
CHAPTER 7: Generating Insights with Software Applications	137
CHAPTER 8: Telling Powerful Stories with Data	161
Part 3: Taking Stock of Your Data Science Capabilities ...	187
CHAPTER 9: Developing Your Business Acumen	189
CHAPTER 10: Improving Operations	205
CHAPTER 11: Making Marketing Improvements	229
CHAPTER 12: Enabling Improved Decision-Making	245
CHAPTER 13: Decreasing Lending Risk and Fighting Financial Crimes	265
CHAPTER 14: Monetizing Data and Data Science Expertise	275
Part 4: Assessing Your Data Science Options	289
CHAPTER 15: Gathering Important Information about Your Company	291
CHAPTER 16: Narrowing In on the Optimal Data Science Use Case	311
CHAPTER 17: Planning for Future Data Science Project Success	327
CHAPTER 18: Blazing a Path to Data Science Career Success	341
Part 5: The Part of Tens	367
CHAPTER 19: Ten Phenomenal Resources for Open Data	369
CHAPTER 20: Ten Free or Low-Cost Data Science Tools and Applications	381
Index	397

This page intentionally left blank

Table of Contents

INTRODUCTION	1
About This Book	3
Foolish Assumptions	3
Icons Used in This Book	4
Beyond the Book	4
Where to Go from Here	4
PART 1: GETTING STARTED WITH DATA SCIENCE	5
CHAPTER 1: Wrapping Your Head Around Data Science	7
Seeing Who Can Make Use of Data Science	8
Inspecting the Pieces of the Data Science Puzzle	10
Collecting, querying, and consuming data	11
Applying mathematical modeling to data science tasks	12
Deriving insights from statistical methods	12
Coding, coding, coding — it's just part of the game	13
Applying data science to a subject area	13
Communicating data insights	14
Exploring Career Alternatives That Involve Data Science	15
The data implementer	16
The data leader	16
The data entrepreneur	17
CHAPTER 2: Tapping into Critical Aspects of Data Engineering	19
Defining Big Data and the Three Vs	19
Grappling with data volume	21
Handling data velocity	21
Dealing with data variety	22
Identifying Important Data Sources	23
Grasping the Differences among Data Approaches	24
Defining data science	25
Defining machine learning engineering	26
Defining data engineering	26
Comparing machine learning engineers, data scientists, and data engineers	27
Storing and Processing Data for Data Science	28
Storing data and doing data science directly in the cloud	28
Storing big data on-premise	32
Processing big data in real-time	35

PART 2: USING DATA SCIENCE TO EXTRACT MEANING FROM YOUR DATA..... 37

CHAPTER 3:	Machine Learning Means . . . Using a Machine to Learn from Data	39
	Defining Machine Learning and Its Processes	40
	Walking through the steps of the machine learning process	40
	Becoming familiar with machine learning terms	41
	Considering Learning Styles.....	42
	Learning with supervised algorithms	42
	Learning with unsupervised algorithms.....	43
	Learning with reinforcement.....	43
	Seeing What You Can Do	43
	Selecting algorithms based on function.....	44
	Using Spark to generate real-time big data analytics.....	48
CHAPTER 4:	Math, Probability, and Statistical Modeling	51
	Exploring Probability and Inferential Statistics	52
	Probability distributions	53
	Conditional probability with Naïve Bayes	55
	Quantifying Correlation	56
	Calculating correlation with Pearson's r.....	56
	Ranking variable-pairs using Spearman's rank correlation.....	58
	Reducing Data Dimensionality with Linear Algebra	59
	Decomposing data to reduce dimensionality	59
	Reducing dimensionality with factor analysis	63
	Decreasing dimensionality and removing outliers with PCA	64
	Modeling Decisions with Multiple Criteria Decision-Making.....	65
	Turning to traditional MCDM.....	65
	Focusing on fuzzy MCDM.....	67
	Introducing Regression Methods	67
	Linear regression.....	67
	Logistic regression.....	69
	Ordinary least squares (OLS) regression methods.....	70
	Detecting Outliers	70
	Analyzing extreme values.....	70
	Detecting outliers with univariate analysis	71
	Detecting outliers with multivariate analysis	73
	Introducing Time Series Analysis	73
	Identifying patterns in time series	74
	Modeling univariate time series data.....	75

CHAPTER 5:	Grouping Your Way into Accurate Predictions	77
	Starting with Clustering Basics	78
	Getting to know clustering algorithms	79
	Examining clustering similarity metrics	81
	Identifying Clusters in Your Data	82
	Clustering with the k-means algorithm	82
	Estimating clusters with kernel density estimation (KDE)	84
	Clustering with hierarchical algorithms	84
	Dabbling in the DBScan neighborhood	87
	Categorizing Data with Decision Tree and Random Forest Algorithms	88
	Drawing a Line between Clustering and Classification	89
	Introducing instance-based learning classifiers	90
	Getting to know classification algorithms	90
	Making Sense of Data with Nearest Neighbor Analysis	93
	Classifying Data with Average Nearest Neighbor Algorithms	94
	Classifying with K-Nearest Neighbor Algorithms	97
	Understanding how the k-nearest neighbor algorithm works	98
	Knowing when to use the k-nearest neighbor algorithm	99
	Exploring common applications of k-nearest neighbor algorithms	100
	Solving Real-World Problems with Nearest Neighbor Algorithms	100
	Seeing k-nearest neighbor algorithms in action	101
	Seeing average nearest neighbor algorithms in action	101
CHAPTER 6:	Coding Up Data Insights and Decision Engines	103
	Seeing Where Python and R Fit into Your Data Science Strategy	104
	Using Python for Data Science	104
	Sorting out the various Python data types	106
	Putting loops to good use in Python	109
	Having fun with functions	110
	Keeping cool with classes	112
	Checking out some useful Python libraries	114
	Using Open Source R for Data Science	120
	Comprehending R's basic vocabulary	121
	Delving into functions and operators	124
	Iterating in R	127
	Observing how objects work	129
	Sorting out R's popular statistical analysis packages	131
	Examining packages for visualizing, mapping, and graphing in R	133

CHAPTER 7:	Generating Insights with Software Applications	137
	Choosing the Best Tools for Your Data Science Strategy	138
	Getting a Handle on SQL and Relational Databases	139
	Investing Some Effort into Database Design	144
	Defining data types	144
	Designing constraints properly	145
	Normalizing your database	145
	Narrowing the Focus with SQL Functions	147
	Making Life Easier with Excel	151
	Using Excel to quickly get to know your data	152
	Reformatting and summarizing with PivotTables	157
	Automating Excel tasks with macros	158
CHAPTER 8:	Telling Powerful Stories with Data	161
	Data Visualizations: The Big Three	162
	Data storytelling for decision makers	162
	Data showcasing for analysts	163
	Designing data art for activists	164
	Designing to Meet the Needs of Your Target Audience	164
	Step 1: Brainstorm (All about Eve)	165
	Step 2: Define the purpose	166
	Step 3: Choose the most functional visualization type for your purpose	166
	Picking the Most Appropriate Design Style	167
	Inducing a calculating, exacting response	167
	Eliciting a strong emotional response	168
	Selecting the Appropriate Data Graphic Type	170
	Standard chart graphics	171
	Comparative graphics	173
	Statistical plots	176
	Topology structures	179
	Spatial plots and maps	180
	Testing Data Graphics	183
	Adding Context	184
	Creating context with data	184
	Creating context with annotations	185
	Creating context with graphical elements	186

PART 3: TAKING STOCK OF YOUR DATA SCIENCE CAPABILITIES 187

CHAPTER 9:	Developing Your Business Acumen	189
	Bridging the Business Gap	189
	Contrasting business acumen with subject matter expertise	190
	Defining business acumen	191

Traversing the Business Landscape	192
Seeing how data roles support the business in making money	192
Leveling up your business acumen.	195
Fortifying your leadership skills.	196
Surveying Use Cases and Case Studies	197
Documentation for data leaders.	199
Documentation for data implementers.	202
CHAPTER 10: Improving Operations	205
Establishing Essential Context for Operational Improvements Use Cases	206
Exploring Ways That Data Science Is Used to Improve Operations	207
Making major improvements to traditional manufacturing operations	208
Optimizing business operations with data science	210
An AI case study: Automated, personalized, and effective debt collection processes.	211
Gaining logistical efficiencies with better use of real-time data.	216
Another AI case study: Real-time optimized logistics routing.	217
Modernizing media and the press with data science and AI	222
Generating content with the click of a button.	222
Yet another case study: Increasing content generation rates	224
CHAPTER 11: Making Marketing Improvements	229
Exploring Popular Use Cases for Data Science in Marketing	229
Turning Web Analytics into Dollars and Sense	232
Getting acquainted with omnichannel analytics.	233
Mapping your channels	233
Building analytics around channel performance	235
Scoring your company's channels.	235
Building Data Products That Increase Sales-and-Marketing ROI	238
Increasing Profit Margins with Marketing Mix Modeling.	239
Collecting data on the four Ps	240
Implementing marketing mix modeling.	241
Increasing profitability with MMM	243
CHAPTER 12: Enabling Improved Decision-Making.	245
Improving Decision-Making	245
Barking Up the Business Intelligence Tree	247
Using Data Analytics to Support Decision-Making	249
Types of analytics	252
Common challenges in analytics.	252
Data wrangling.	253

Increasing Profit Margins with Data Science	254
Seeing which kinds of data are useful when using data science for decision support	255
Directing improved decision-making for call center agents	257
Discovering the tipping point where the old way stops working	262
CHAPTER 13: Decreasing Lending Risk and Fighting Financial Crimes	265
Decreasing Lending Risk with Clustering and Classification	266
Preventing Fraud Via Natural Language Processing (NLP)	267
CHAPTER 14: Monetizing Data and Data Science Expertise	275
Setting the Tone for Data Monetization	275
Monetizing Data Science Skills as a Service	278
Data preparation services	279
Model building services	280
Selling Data Products	282
Direct Monetization of Data Resources	283
Coupling data resources with a service and selling it	283
Making money with data partnerships	284
Pricing Out Data Privacy	285
PART 4: ASSESSING YOUR DATA SCIENCE OPTIONS	289
CHAPTER 15: Gathering Important Information about Your Company	291
Unifying Your Data Science Team Under a Single Business Vision	292
Framing Data Science around the Company's Vision, Mission, and Values	294
Taking Stock of Data Technologies	296
Inventorying Your Company's Data Resources	298
Requesting your data dictionary and inventory	298
Confirming what's officially on file	300
Unearthing data silos and data quality issues	300
People-Mapping	303
Requesting organizational charts	303
Surveying the skillsets of relevant personnel	304
Avoiding Classic Data Science Project Pitfalls	305
Staying focused on the business, not on the tech	305
Drafting best practices to protect your data science project	306
Tuning In to Your Company's Data Ethos	306
Collecting the official data privacy policy	307
Taking AI ethics into account	307
Making Information-Gathering Efficient	308

CHAPTER 16:	Narrowing In on the Optimal Data Science Use Case	311
	Reviewing the Documentation	312
	Selecting Your Quick-Win Data Science Use Cases	313
	Zeroing in on the quick win	313
	Producing a POTI model	314
	Picking between Plug-and-Play Assessments	316
	Carrying out a data skill gap analysis for your company	317
	Assessing the ethics of your company's AI projects and products	318
	Assessing data governance and data privacy policies	323
CHAPTER 17:	Planning for Future Data Science Project Success	327
	Preparing an Implementation Plan	328
	Supporting Your Data Science Project Plan	335
	Analyzing your alternatives	335
	Interviewing intended users and designing accordingly	337
	POTI modeling the future state	338
	Executing On Your Data Science Project Plan	339
CHAPTER 18:	Blazing a Path to Data Science Career Success	341
	Navigating the Data Science Career Matrix	341
	Landing Your Data Scientist Dream Job	343
	Leaning into data science implementation	345
	Acing your accreditations	346
	Making the grade with coding bootcamps and data science career accelerators	348
	Networking and building authentic relationships	349
	Developing your own thought leadership in data science	350
	Building a public data science project portfolio	351
	Leading with Data Science	354
	Starting Up in Data Science	357
	Choosing a business model for your data science business	357
	Selecting a data science start-up revenue model	359
	Taking inspiration from Kam Lee's success story	361
	Following in the footsteps of the data science entrepreneurs	364
	PART 5: THE PART OF TENS	367
CHAPTER 19:	Ten Phenomenal Resources for Open Data	369
	Digging Through data.gov	370
	Checking Out Canada Open Data	371
	Diving into data.gov.uk	372
	Checking Out US Census Bureau Data	373

Accessing NASA Data	374
Wrangling World Bank Data.....	375
Getting to Know Knoema Data	376
Queuing Up with Quandl Data	378
Exploring Exversion Data	379
Mapping OpenStreetMap Spatial Data	380
CHAPTER 20: Ten Free or Low-Cost Data Science Tools and Applications.....	381
Scraping, Collecting, and Handling Data Tools	382
Sourcing and aggregating image data with ImageQuilts.....	382
Wrangling data with DataWrangler.....	383
Data-Exploration Tools	384
Getting up to speed in Gephi.....	384
Machine learning with the WEKA suite.....	386
Designing Data Visualizations	387
Getting Shiny by RStudio	387
Mapmaking and spatial data analytics with CARTO.....	388
Talking about Tableau Public.....	390
Using RAWGraphs for web-based data visualization.....	392
Communicating with Infographics	393
Making cool infographics with Infogram	394
Making cool infographics with Piktochart	395
INDEX.....	397