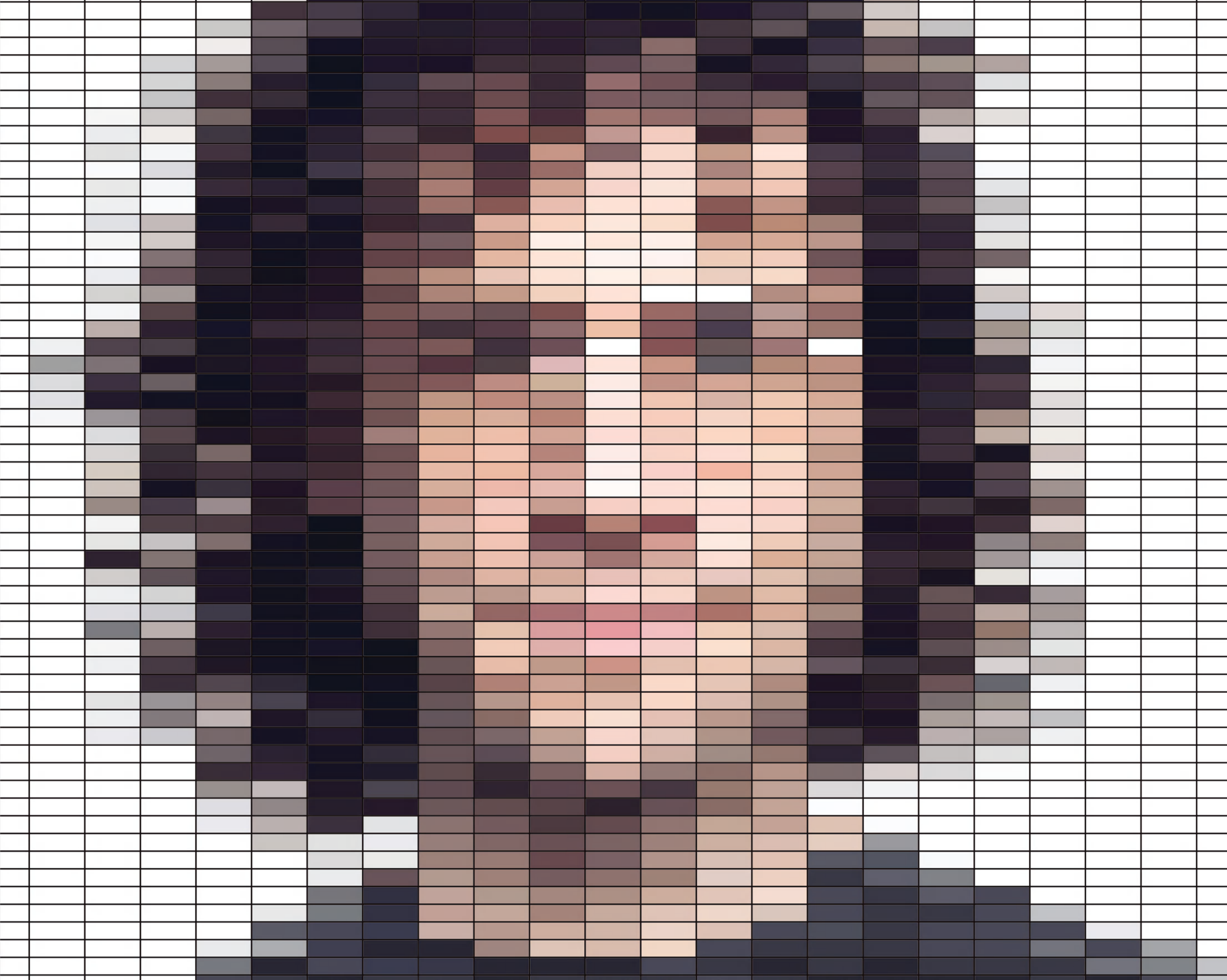

JORDAN GOLDMEIER

DATA SMART

USING DATA SCIENCE TO TRANSFORM INFORMATION INTO INSIGHT

SECOND EDITION

WILEY

Data Smart

Second Edition

Data Smart

Using Data Science to Transform Information into Insight

Second Edition

Jordan Goldmeier

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada and the United Kingdom.

ISBNs: 9781119931386 (Paperback), 9781119931485 (ePDF), 9781119931393 (ePub)

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

If you believe you've found a mistake in this book, please bring it to our attention by emailing our reader support team at wileysupport@wiley.com with the subject line "Possible Book Errata Submission."

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Control Number: 2023940016

Cover image: Courtesy of Greg Jones and Wiley
Cover design: Wiley

Dedicated to David and Terry

About the Author

Jordan Goldmeier is one of the leading global minds on data visualization and data science. His books include *Dashboards for Excel* (Apress), *Advanced Excel Essentials* (Apress), *Becoming a Data Head: How to Think, Speak, and Understand Data Science, Statistics, and Machine Learning* (Wiley). Jordan has received the prestigious Microsoft Most Valuable Professional Award many times over the years. He has consulted and provided training for Fortune 500 companies, NATO, and taught analytics for Wake Forest University. He runs multiple businesses as a digital nomad living in Lisbon, Portugal. You connect with Jordan on LinkedIn and on Instagram (`@jordangoldmeier`).

About the Technical Editors

Alex Gutman is a data scientist, corporate trainer, and Accredited Professional Statistician® who enjoys teaching a wide variety of data science topics to technical and nontechnical audiences. He's a former adjunct professor at the Air Force Institute of Technology and current adjunct at the University of Cincinnati. Alex is also the author of the book *Becoming a Data Head: How to Think, Speak, and Understand Data Science, Statistics, and Machine Learning* (Wiley). He received his BS and MS degrees in mathematics from Wright State University and his PhD in applied mathematics from the Air Force Institute of Technology.

Matthew Bernath is passionate about leveraging data to bolster economies and facilitate strategic dealmaking. Matthew has led the data analytics division of one of Africa's largest investment banks and is currently the head of data ecosystems for Africa's largest retailer. His diverse experience spans from structuring multibillion-rand project financing deals to utilizing data to uplift society, always driven by data-focused decision-making. Recognized as one of the "60 Data Changemakers to Know" by Narrative Science and a finalist for Data Analytics Leader of the Year in 2022, Matthew's achievements extend beyond his professional role. His contribution to community-building initiatives include hosting the Johannesburg Data Science and Financial Modelling meetup groups and the highly regarded *Financial Modelling Podcast*, which was awarded Financial Modelling Resource of the Year 2021. He also formerly hosted the RMB *Data Analytics* podcast. Prior to his investment banking and retail career, Matthew held leadership roles in various advisory and technology firms, bringing his data-driven approach to different industries.

Acknowledgments

Life has a weird way of coming full circle. I read the original *Data Smart* when it first came out in 2013. I had no imagination back then I would write the revised edition. Yet, here I am. If fate brought me to this place, it's because I love Excel. Therefore, it only makes sense to first acknowledge the Excel product team at Microsoft, who've managed to push Excel beyond the tool it was back in 2013.

As a Microsoft MVP, I've met some incredible folks at Microsoft over the years, who've really listened and understood the ways in which my community uses Excel. In particular, I would like to acknowledge David Gainer, Guy Lev, and Joe McDaid for continually expanding the product.

I would also like to acknowledge my peers in the Excel community who pushed the product beyond its limitations for the good of the whole. As it relates to the material in this book, I must mention George Mount, Oz du Soleil, Carlos Barboza, and Roberto Mensa for challenging the norm.

I also have to give major credit to the book's first author, John Foreman. If you weren't in the data space back in 2013, you should know it was a different world. In those days, people were enamored by the idea of "big data." Companies were rushing to implement technologies that could handle large datasets before they even had high-quality data.

But then there was John's book, which showed people how to do (or at the very least, teach) data science without big data technologies—you could just use Excel. John showed people that it wasn't about the technology, but, rather, one had to really think through the problem. And he did it without being boring. John's book served as major motivation and inspiration for my last book, *Becoming a Data Head*. It's a great honor to be working on this material.

I also have to acknowledge my technical editors, Alex Gutman and Mathew Bernath. Both are incredibly intelligent and esteemed in their fields. Alex and I wrote *Data Head* together, and it's amazing to once again have him on another project. Alex is thorough, humble, and deeply affable. He's often the smartest person in the room, but you would never know, as there's not an arrogant bone in his body. Alex's contributions are indelibly fused into the text of this book.

Mathew is perhaps the coolest data (and coffee) nerd I know. He knows his craft well and channels that knowledge into community building, bringing ideas and minds together to push the field forward. His technical advice on this book challenged many of the things I took for granted. This book is much better off for it, and I'm very grateful for his support.

I also want to acknowledge the team at Wiley. In particular, I would like to mention Jim Minatel who believed strongly in this project and really pushed to make it happen. I also want to thank John Sleeva, my development editor. John has my favorite working style—no news is good news. He's always calm, thorough, and dependable. This is my second book with Wiley and Jim and John—and I couldn't have asked for a better team.

I also have to mention Archana Pragash who worked tirelessly on proofing this book to my specifications. I often wondered when she slept. She always responded quickly—nights, weekends, etc. For a big project like this, Archana was a dependable pillar. The layout of this book is to her credit.

Finally, I would like to thank you, the reader. It's your interest that makes this book happen. I hope you enjoy it.

Contents

Introduction xix

1 Everything You Ever Needed to Know About Spreadsheets but Were Too Afraid to Ask 1

Some Sample Data 2

Accessing Quick Descriptive Statistics 3

Excel Tables. 4

Filtering and Sorting 5

Table Formatting 7

Structured References 7

Adding Table Columns 10

Lookup Formulas. 11

VLOOKUP. 11

INDEX/MATCH. 13

XLOOKUP. 15

PivotTables 16

Using Array Formulas 19

Solving Stuff with Solver 20

2 Set It and Forget It: An Introduction to Power Query 27

What Is Power Query? 27

Sample Data 28

Starting Power Query 29

Filtering Rows 32

Removing Columns. 33

Find & Replace 34

Close & Load to. . .Table. 35

3 Naïve Bayes and the Incredible Lightness of Being an Idiot 39

The World's Fastest Intro to Probability Theory 39

Totaling Conditional Probabilities 40

Joint Probability, the Chain Rule, and Independence. 40

What Happens in a Dependent Situation? 41

Bayes Rule. 42

Separating the Signal and the Noise 43

Using the Bayes Rule to Create an AI Model 44

High-Level Class Probabilities Are Often Assumed to Be Equal 45

A Couple More Odds and Ends 46

Let's Get This Excel Party Started 47

Cleaning the Data with Power Query 48

Splitting on Spaces: Giving Each Word Its Due 50

Counting Tokens and Calculating Probabilities 55

We Have a Model! Let's Use It 58

4 Cluster Analysis Part 1: Using K-Means to Segment Your Customer Base 65

Dances at Summer Camp 65

Getting Real: K-Means Clustering Subscribers in Email Marketing 70

The Initial Dataset 71

Determining What to Measure. 72

Start with Four Clusters 75

Euclidean Distance: Measuring Distances as the Crow Flies. 76

Solving for the Cluster Centers 80

Making Sense of the Results 82

Getting the Top Deals by Cluster 83

The Silhouette: A Good Way to Let Different K Values Duke It Out. 86

How About Five Clusters? 95

Solving for Five Clusters 96

Getting the Top Deals for All Five Clusters. 96

Computing the Silhouette for 5-Means Clustering 99

K-Medians Clustering and Asymmetric Distance Measurements 100

Using K-Medians Clustering. 100

Getting a More Appropriate Distance Metric 100

Putting It All in Excel 102

The Top Deals for the 5-Medians Clusters 104

5 Cluster Analysis Part II: Network Graphs and Community Detection 109

What Is a Network Graph? 110

Visualizing a Simple Graph 110

Beyond GiGraph and Adjacency Lists. 115

Building a Graph from the Wholesale Wine Data 117

Creating a Cosine Similarity Matrix 118

Producing an R-Neighborhood Graph 121

Introduction to Gephi 123

Creating a Static Adjacency Matrix 124

Bringing in Your R-Neighborhood Adjacency Matrix into Gephi 124

Node Degree 128

Touching the Graph Data 130

How Much Is an Edge Worth? Points and Penalties in Graph Modularity 132

What's a Point, and What's a Penalty? 133

Setting Up the Score Sheet 136

Let's Get Clustering! 138

Split Number 1. 138

Split 2: Electric Boogaloo 143

And. . .Split3: Split with a Vengeance 145

Encoding and Analyzing the Communities 146

There and Back Again: A Gephi Tale. 151

6 Regression: The Granddaddy of Supervised Artificial Intelligence 157

Predicting Pregnant Customers at RetailMart Using Linear Regression 158

The Feature Set. 159

Assembling the Training Data 161

Creating Dummy Variables. 163

Let's Bake Our Own Linear Regression. 165

Linear Regression Statistics: R-Squared, F-Tests, t-Tests 173

Making Predictions on Some New Data and Measuring Performance 182

Predicting Pregnant Customers at RetailMart Using Logistic Regression 192

First You Need a Link Function. 192

Hooking Up the Logistic Function and Reoptimizing. 193

Baking an Actual Logistic Regression 196

7 Ensemble Models: A Whole Lot of Bad Pizza. 203

Getting Started Using the Data from Chapter 6 203

Bagging: Randomize, Train, Repeat 204

Decision Stump is Another Name for a Weak Learner 204

Doesn't Seem So Weak to Me!. 204

You Need More Power! 207

Let's Train It 208

Evaluating the Bagged Model 220

Boosting: If You Get It Wrong, Just Boost and Try Again. 223

Training the Model—Every Feature Gets a Shot. 224

Evaluating the Boosted Model 231

8 Forecasting: Breathe Easy: You Can't Win. 235

The Sword Trade Is Hopping 236

Getting Acquainted with Time-Series Data. 236

Starting Slow with Simple Exponential Smoothing 238

Setting Up the Simple Exponential Smoothing Forecast 240

You Might Have a Trend 249

Holt's Trend-Corrected Exponential Smoothing 250

Setting Up Holt's Trend-Corrected Smoothing in a Spreadsheet 252

So Are You Done? Looking at Autocorrelations 258

Multiplicative Holt-Winters Exponential Smoothing 266

Setting the Initial Values for Level, Trend, and Seasonality 268

Getting Rolling on the Forecast 274

And. . .Optimize!. 280

Putting a Prediction Interval Around the Forecast. 283

Creating a Fan Chart for Effect 287

Forecast Sheets in Excel 289

9 Optimization Modeling: Because That "Fresh-Squeezed" Orange Juice Ain't Gonna Blend Itself 293

Wait. . .Is This Data Science? 294

Starting with a Simple Trade-Off 295

Representing the Problem as a Polytope 296

Solving by Sliding the Level Set 297

The Simplex Method: Rooting Around the Corners 298

Working in Excel 300

Fresh from the Grove to Your Glass. . .with a Pit Stop Through a Blending Model 305

Let's Start with Some Specs 307

Coming Back to Consistency 308

Putting the Data into Excel 309

Setting Up the Problem in Solver 311

Lowering Your Standards 314

Dead Squirrel Removal: the Minimax Formulation 317

If-Then and the "Big M" Constraint 320

Multiplying Variables: Cranking Up the Volume to 11,000 324

Modeling Risk 330

Normally Distributed Data 331

10 Outlier Detection: Just Because They're Odd Doesn't Mean They're Unimportant 339

Outliers Are (Bad?) People, Too 340

The Fascinating Case of *Hadlum v. Hadlum* 340

Tukey's Fences 341

Applying Tukey's Fences in a Spreadsheet 342

The Limitations of This Simple Approach 345

Terrible at Nothing, Bad at Everything 346

Preparing Data for Graphing 347

Creating a Graph 350

Getting the k-Nearest Neighbors 351

Graph Outlier Detection Method 1: Just Use the Indegree 352

Graph Outlier Detection Method 2: Getting Nuanced with k-Distance 355

Graph Outlier Detection Method 3: Local Outlier Factors Are Where It's At 358

11 Moving on From Spreadsheets 363
Getting Up and Running with R. 364
A Crash Course in R-ing 366
Show Me the Numbers! Vector Math and Factoring 367
The Best Data Type of Them All: the Dataframe. 370
How to Ask for Help in R 371
It Gets Even Better. . .Beyond Base R 372
Doing Some Actual Data Science 374
Reading Data into R 374
Spherical K-Means on Wine Data in Just a Few Lines 375
Building AI Models on the Pregnancy Data. 381
Forecasting in R. 389
Looking at Outlier Detection 393

12 Conclusion 397
Where Am I? What Just Happened?. 397
Before You Go-Go 397
Get to Know the Problem 398
We Need More Translators 398
Beware the Three-Headed Geek-Monster: Tools, Performance,
and Mathematical Perfection 399
You Are Not the Most Important Function of Your Organization 401
Get Creative and Keep in Touch! 402

Index. 403