



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

Copyright © 2024 by John Wiley & Sons, Inc. All rights reserved.

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

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

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 Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. 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 www.wiley.com/go/permission.

Trademarks: WILEY, and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, 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: 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 I: 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

Introduction

What Am I Doing Here?

If you're reading this book, it's because on some level you understand the importance of both data and data science in your business and career.

The original *Data Smart* was written more than a decade ago. John Foreman, the first book's author, exposed a new generation of readers to the supposed magic behind the curtain of data science. John proved that data science didn't have to be so mysterious. You could both understand and do data science in something as humble as the spreadsheet.

John's words served as a prescient warning for what would come. He noted the "buzz about data science," and the pressure it created on businesses to take on data science projects and hire data scientists without even fully understanding why.

The truth is most people are going about data science all wrong. They're starting with buying the tools and hiring the consultants. They're spending all their money before they even know what they want, because a purchase order seems to pass for actual progress in many companies these days.

John's words still ring true today. Ten years after the first wave of interest in data science, the data science machine is still working in full force, churning out ideas faster than we can articulate the opportunities and challenges they present to business and society. In my last book, *Becoming a Data Head: How to Think, Speak and Understand Data Science, Statistics and Machine Learning* (Wiley, New York, NY, 2021), my coauthor and I called this the *data science industrial complex*.

To put it bluntly, despite the extensive interest in data and data science, projects still fail sometimes at alarming rates, even as data is supposed to be fact driven. In truth, as much as 87 percent data science projects won't make it into production.¹

What is and isn't a "data disaster" is perhaps up from some considerable debate. But it's fair to say the recent past is filled with examples in which technology, data, and the like were hailed as something magical before they ultimately came up short. Here are just a few examples worth considering:

¹"Why do 87% of data science projects never make it into production?" <https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes>

- An attorney used a generative AI chatbot for legal research, submitting a brief to the court with cases that did not exist, but perhaps sounded plausible.²
- The COVID-19 pandemic exposed major issues in forecasting across the board, from supply chain issues to understanding the spread of the virus.³
- When the original *Data Smart* came out, accurately predicting the outcome of the US presidential election seemed like an easy feat. In 2016, however, model after model inaccurately predicted a win for Hillary Clinton, despite increased money, time, and effort into the subject.⁴

Most data science projects and outcomes don't fail so spectacularly. Instead, data science projects die slow deaths, while management pours money and resources into chasing elusive numbers they don't entirely understand.

Yet, some of the greatest data achievements did not come from any particular technology. Rather, they came from human ingenuity. For instance, I used to lead projects for a nonprofit called DataKind, which leverages “data science and AI in service of Humanity.”

DataKind uses teams of volunteer data scientists to help mission-driven organizations design solutions to tough social problems in an ethical and socially responsible way. When I was there, we worked with major organizations like the United Nations and Habitat for Humanity.

Volunteers built all sorts of models and tools, from forecasting water demand in California to using satellite imagery to identify villages in need with machine learning. The work we did had impact, so it's not all doom and gloom. When you're done with this book, you might consider giving back in your own way.⁵ Remember: Humans solve problems not machines.

What Is Data Science?

In my last book, *Becoming a Data Head*, Alex Gutman (my coauthor) and I actually don't define data science. One reason is that the space is too hard to pin down. And we didn't want folks to get caught up in trying to justify whether or not they were data scientists. In the original *Data Smart*, John Foreman offers this working definition:

Data science is the transformation of data using mathematics and statistics into valuable insights, decisions, and products.

²“8 famous analytics and AI disasters.” www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html

³“Forecasting for COVID-19 has failed.” www.ncbi.nlm.nih.gov/pmc/articles/PMC7447267

⁴“The Real Story Of 2016.” <https://fivethirtyeight.com/features/the-real-story-of-2016>

⁵To see the impact DataKind has had, take a look at their case studies - www.datakind.org/what-we-do

John takes a broad, business-centric view. He's quick to note it's a "catchall buzzword for [everything] analytics today." Ten years later, I and the rest of the industry are still struggling to define exactly what data science is and isn't. So rather than proffer a definition as if that will get us closer to the truth, I'd rather describe what a data scientist does.

- *Data scientists identify relevant questions that can be solved with data.* This may sound obvious, but many questions can't be solved with data and technology. A good data scientist can tease out the problems in which algorithms and analyses make the most sense.
- *Data scientists extract meaningful patterns and insights from data.* Anyone can eyeball a set of numbers and draw their own conclusions. On the other hand, data scientists focus on what can be said statistically and verifiably. They separate speculation from science, focusing instead on what the data says.
- Finally, *data scientists convey results using data visualization and clear communication.* In many cases, a data scientist will have to explain how an algorithm works and what it does. Historically, this has been a challenge for many in the field. But a recent crop of books (like this one) aims at giving data scientists a way to explain how they came to their results without being too stuck into the weeds.

Incredibly, some of the techniques mentioned in the following pages are as old as World War II. They were invented at the dawn of the modern computer, long before you could easily spin up a new instance of R. The hype machine won't tell you these "new" algorithms were first developed on punch cards.

And some of the techniques in this book were invented recently, taking advantage of the wealth of data, self-service analytics, cloud computers, and new graphical processing units developed in the last 10 years.

Again, we're reminded that human ingenuity is what drives this field forward.

Age has no bearing on difficulty or usefulness. All these techniques whether or not they're currently the rage are equally useful in the right business context. It's up to you to use them correctly. That's why you need to understand how they work, how to choose the right technique for the right problem, and how to prototype with them.

Do Data Scientists Actually Use Excel?

Many (but not all) veteran data scientists will tell you they loathe spreadsheets and Excel in particular. They will say that Excel isn't the best place to create a data science model. To some extent, they're right.

But before you throw this book away, let's understand why they say this. You see, there was a time before R and before Python. It was a time when MATLAB and SPSS reigned supreme. The latter tools were expensive and often required a computer with some major

horsepower to run a model. Moreover, the files that these tools generated were not easily distributable. And, in a secure corporate or institutional environment, sending files with code in them over email would trip the unsafe-email alarms.

As a result, many in the industry began building their work in Excel. This was particularly true of models that helped support executive decision-making. Excel was the secret way around these email systems. It was a way to build a mini data application without having to get approval from the security team.

Many executive teams relied on Excel. Unfortunately, this also created a myopic view among executives who didn't really understand data science. For them, Excel was the only place to do this type of work. It was where they were most comfortable.

They knew the product. They could see what the analyst created. And the analyst could walk them through each step. In fact, that's why we're using Excel in this book.

But Excel (at the time) was limited. Limited by how much it could process at any moment. Limited by the amount of data it could store. The macro language behind Excel, Visual Basic for Applications (VBA), is still hailed by many executives as an advanced feature. But VBA is based on Visual Basic 6.0, which was deprecated in 1999. The Excel version of this language has received only the barest of updates. When today's data scientists point out that VBA can't do what R or Python can, it's hard to disagree.

On the flipside, however, Microsoft has paid attention over the last few years. The Excel product team has come to understand how data scientists use their tool. They've poured more research into some very specific use cases. For instance, we'll talk about an entirely new data wrangling tool in Excel called Power Query. Power Query can do the same data wrangling tasks as in Python and R, often more quickly. And we'll talk about new Excel functions that make data science in Excel a whole lot easier. Today, there is renewed interest in using Excel for data science problems beyond what was possible only a few years ago.

But if there's a place where Excel shines, it's in explaining and understanding data science concepts. Before getting a "yes" to your new data science project, you'll need to get buy-in from management. You can fire up an advanced algorithm in R, pull out lines of code, and explain what each function does step-by-step. Or you can walk management through the algorithm in Excel and even give them the ability to filter results and ask questions of the data.

In fact, Excel is great for prototyping. You're not running a production AI model for your online retail business out of Excel, but that doesn't mean you can't look at purchase data, experiment with features that predict product interest, and prototype a targeting model.

At the end of this book, I'll show you how to implement what we've built in Excel in R. In fact, this follows my own path in building data science tools for companies. First, I would lay out my ideas in Excel. Use the spreadsheet as a way to validate my ideas and make sure I understand exactly what the algorithms do. Then, usually, when I'm ready, I move it to R or Python.

But sometimes I don't. Because in some instances Excel just gets the job done, and the problem doesn't need more complication. As you will see, knowing how to do these techniques in Excel will give you a major advantage, whether or not you end up implementing them in something more powerful.

Conventions

To help you get the most from the text and keep track of what's happening, I've used a number of conventions throughout the book.

Frequently in this text I'll reference little snippets of Excel code like this:

```
=IF(A1 = 1, "I love Excel", "I REALLY love Excel")
```

SIDEBARS

Sidebars touch upon some side issue related to the text in detail.

WARNING

Warnings hold important, not-to-be-forgotten information that is directly relevant to the surrounding text.

NOTE

Notes cover tips, hints, tricks, or asides to the current discussion.

- We bold technical objects, when introducing them for the first time, or when it makes sense to set them off. We also use the bold font to refer to specific fields and buttons.
- We *italicize* new concepts and important words when we introduce them.
- We show filenames, URLs, and formulas within the text like so: `www.linkedin.com/in/jordangoldmeier`

Let's Get Going

A new generation of data scientists is learning how to implement work that was only theoretical when I first started. The industry is undergoing a serious reflection on what's

important. Businesses are starting to realize their most important assets aren't data, algorithms, or technology—it's people. People just like you.

As you go along your data journey, you will likely encounter more than your fair share of bad decision-making, a lack of critical thinking, ignorant management, and even some imposter syndrome. Sadly, as with many of the data successes, these are part of the legacy. But with the knowledge contained herein, you'll be set up for success. You'll understand the algorithms. You'll know how and what they do. And, you won't be fooled by buzzwords. When it comes to doing real data science work, you'll already know how to identify the data science opportunities within your own organization.

By reading this book, you're going to have a leg up on the next generation of data problems. Whether you're a veteran of the field or a student in school, by reading this book, you will become a better data scientist.

In Chapter 1, "Everything You Ever Needed to Know About Spreadsheets but Were Too Afraid to Ask," I'm going to fill in a few holes in your Excel knowledge. And, in Chapter 2, "Set & Forget it! An introduction to Power Query." I'm going to show you Power Query. After that, you'll move right into use cases. By the end of this book, you'll have experience implementing from scratch the following techniques:

- Optimization using linear and integer programming.
- Working with time-series data, detecting trends, and seasonal patterns, and forecasting with exponential smoothing.
- Using Monte Carlo simulation in optimization and forecasting scenarios to quantify and address risk.
- Applying Artificial intelligence using the general linear model, logistic link functions, ensemble methods, and naïve Bayes.
- Measuring distances between customers using cosine similarity, creating kNN graphs, calculating modularity, and clustering customers.
- Detecting outliers in a single dimension with Tukey fences or in multiple dimensions with local outlier factors.
- Using R packages to implement data science techniques quickly.

It's now time for our journey to begin. I'll see you in the next chapter!

- creating your own, 165–173
 - feature set, 159–160
 - making predictions, 182–192
 - predicting pregnant customers at
 - RetailMart using, 158–192
 - statistics for, 173–182
 - logistic
 - comparing performance of with linear regression, 198–202
 - predicting in, 192–202
 - predicting pregnant customers at
 - RetailMart using logistic, 192–202
 - statistical tests in, 197
 - removing columns, in PowerQuery, 33
 - reoptimization, 193–196
 - replace function, 377
 - resources
 - Advancing Into Analytics: From Excel to Python and R* (Mount), 364, 396
 - Applied Logistic Regression*, 3rd Edition (Hosmer, Jr., Lemeshow and Sturdivant), 197
 - Becoming a Data Head: How to Think, Speak, and Understand Data Science, Statistics, and Machine Learning* (Gutman and Goldmeier), 202, 339, 364, 399
 - Data Mining with R* (Torgo), 202
 - Effective Data Storytelling: How to Drive Change with Data, Narrative and Visuals* (Dykes), 399
 - The Elements of Statistical Learning* (Hastie, Tibshirani, and Friedman), 202
 - Excel 2019 Bible* (Alexander, Kusleika, and Walkenbach), 11
 - Excel Tables: A Complete Guide for Creating, Using and Automating Lists and Tables* (Barresse and Jones), 11
 - Hands-On Machine Learning with R* (Boehmke and Greenwell), 396
 - Present Beyond Measure: Design, Visualize, and Deliver Data Stories That Inspire Action* (Pica), 399
 - R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (Wickham and Grolemund), 396
 - The R Book*, 3rd Edition (Jones, Harden and Crawley), 396
 - Storytelling with Data: A Data Visualization Guide for Business Professionals* (Knaflic), 399
 - reweighting, 229–230
 - risk modeling, 330–331
 - r-neighborhood adjacency matrix,
 - importing into Gephi, 124–128
 - R-neighborhood graphs, creating, 121–123
 - ROC (Receiver Operating Characteristic) Curve, 190–192
 - rows
 - about, 2–3
 - filtering in Power Query, 32–33
 - R-scripting pane, 366
 - R-squared, in linear regression, 173–175
 - The R Book*, 3rd Edition (Jones, Harden and Crawley), 396
- ## S
- sample data
 - about, 2–3
 - for Power Query, 28–29
 - scale, measures of, 348
 - scale() function, 395
 - scenarios
 - creating from standard deviations, 334–335
 - setting up constraints for, 336–338

- seasonality, setting initial values
 - for, 268–274
- Select Data Source dialog box, 265
- `select()` function, 377, 379
- sensitivity, 188–190
- setting initial values for level, trend, and seasonality, 268–274
- setup
 - of cutoff values, 185–186
 - Holt’s trend-corrected exponential smoothing, 252–257
 - of linear programs, 141–143
 - problems in Solver, 311–314
 - scenario constraints, 336–338
 - of score sheets, 136–138
 - of simple exponential smoothing (SES), 240–249
- signal, separating from noise, 43
- silhouette
 - for 5-means clustering, 99
 - in K-means clustering, 86–95
- simple exponential smoothing (SES)
 - about, 238–240
 - setting up, 240–249
- simplex algorithm, 80
- simplex method, 298–300
- smoothing constants, 238, 239
- Solver
 - about, 20–26, 143
 - setting up problems in, 311–314
 - training with, 171–173
- Solver dialog box, 24
- Solver Parameters dialog box, 22, 23
- Solver Results dialog box, 24, 25
- solving
 - for cluster centers, 80–82
 - for five clusters, 96
 - level sets by sliding, 297–298
- sorting, in Excel tables, 5–7
- specifications, for optimization
 - modeling, 307–308
- specificity, 187
- spherical K-means, 375–381
- splitting
 - columns, 50–55
 - features, 206–207
- spreadsheets
 - about, 1
 - accessing quick descriptive statistics, 3
 - applying Tukey’s fences in, 342–344
 - array formulas, 19–20
 - Excel tables, 3–11
 - lookup formulas, 11–16
 - moving on from, 363–396
 - Pivot Charts, 16–19
 - PivotTables, 16–19
 - sample data, 2–3
 - Solver, 20–26
- standard deviations, creating scenarios
 - from, 334–335
- standard error, 246
- STANDARDIZE function, 349
- standardizing
 - about, 348
 - data, 74
- starting Power Query, 29–31
- static adjacency matrix, creating, 124
- statistical tests, in logistic regression, 197
- statistically insignificant, 175
- statistically significant, 175
- statistics, for linear regression, 173–182
- stop words, 52
- Storytelling with Data: A Data Visualization Guide for Business Professionals* (Knaflic), 399
- `str()` function, 371, 378, 381
- straight-lined scatter plot, 191

- structured references, in Excel tables, 7–10
 - stubs, 133, 136
 - stumps, calculating alpha value for, 228–229
 - Sturdivant, Rodney X. (author)
 - Applied Logistic Regression*, 3rd Edition, 197
 - sum of squares, 168, 174
 - sum of squares and cross products (SSCP) matrix, 178–180
 - summary() function, 368, 371, 381
 - summary statistics, in spreadsheets, 3
 - SUMPRODUCT function, 20, 262–263, 304
 - supervised machine learning, 66
- T**
- t distribution, 181
 - t() function, 369, 377, 379
 - t statistic, 181
 - tabular format, 2
 - T.DIST.2T function, 249–250
 - test sets, predicting, 231–232
 - Tibshirani, Robert (author)
 - The Elements of Statistical Learning*, 202
 - Tidyverse, 374
 - time-series data, 236–238
 - tokens
 - about, 44
 - counting, 55–57
 - tools, 399–401
 - Torgo, Luis (author)
 - Data Mining with R*, 202
 - total sum of squares, 174
 - training
 - boosting, 224–231
 - with Solver, 171–173
 - training data, assembling for linear regression, 161–163
 - translators, 398–399
 - TRANSPOSE function, 19, 20
 - trends
 - in forecasting, 249–250
 - setting initial values for, 268–274
 - trimean, 348
 - trimed/truncated mean, 348
 - triple exponential smoothing, 267
 - true negative rate, 187–188
 - true negatives, 185
 - true positive rate, 188–190
 - true positives, 162, 185
 - Tukey’s fences
 - about, 341–342
 - applying in spreadsheets, 342–344
 - limitations of, 345–346
 - two-tailed t-test, 181
 - Type I error, 162, 184–185, 188
 - Type II error, 184–185
- U**
- undirected network graph, 111
 - uniform distribution, 332
 - unsupervised machine learning, 66
- V**
- values
 - alpha
 - about, 246–247, 267, 277–280
 - calculating for stumps, 228–229
 - critical, 263
 - cutoff, 185–186
 - delta, 277–280
 - empirical, 246
 - gamma, 251–252, 267, 277–280
 - variables
 - decision, 143
 - dependent, 159
 - dummy, 163–165

- independent, 159
- multiplying, 324–330
- `varImpPlot()` function, 384
- vector math, 367–370
- visualizing graphs, 110–114
- VLOOKUP function, 11–13
- Voronoi diagram, 67

W

- Walkenbach, John (author)
 - Excel 2019 Bible*, 11
- weak learner, 204
- websites

- OpenSolver, 25
- Tidyverse, 374
- weighted adjacency matrix, 121
- weighted error, 224
- `which()` function, 368, 393
- Wickham, Hadley (author)
 - R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, 396
- Winsorized mean, 348

X

- XLOOKUP function, 15–16