Contents

Introduction ................................................................. xiii

1 Everything You Ever Needed to Know about Spreadsheets but Were Too Afraid to Ask ................................................. 1
   Some Sample Data ..................................................... 2
   Moving Quickly with the Control Button ......................... 2
   Copying Formulas and Data Quickly ................................. 4
   Formatting Cells ....................................................... 5
   Paste Special Values .................................................. 7
   Inserting Charts ....................................................... 8
   Locating the Find and Replace Menus ............................... 9
   Formulas for Locating and Pulling Values ......................... 10
   Using VLOOKUP to Merge Data ...................................... 12
   Filtering and Sorting ................................................ 13
   Using PivotTables .................................................... 16
   Using Array Formulas ............................................... 19
   Solving Stuff with Solver ............................................ 20
   OpenSolver: I Wish We Didn't Need This, but We Do .......... 26
   Wrapping Up .......................................................... 27

2 Cluster Analysis Part I: Using K-Means to Segment Your Customer Base .................................................. 29
   Girls Dance with Girls, Boys Scratch Their Elbows ............ 30
   Getting Real: K-Means Clustering Subscribers in E-mail Marketing .................................................. 35
       Joey Bag O’ Donuts Wholesale Wine Emporium ................. 36
       The Initial Dataset ................................................. 36
       Determining What to Measure .................................. 38
       Start with Four Clusters .......................................... 41
       Euclidean Distance: Measuring Distances as the Crow Flies 41
       Distances and Cluster Assignments for Everybody! .......... 44
       Solving for the Cluster Centers ................................ 46
       Making Sense of the Results ...................................... 49
Getting the Top Deals by Cluster .............................................. 50

The Silhouette: A Good Way to Let Different K Values
Duke It Out ............................................................ 53

How about Five Clusters? ............................................. 60
Solving for Five Clusters ............................................. 60

Getting the Top Deals for All Five Clusters ......................... 61
Computing the Silhouette for 5-Means Clustering ............. 64

K-Medians Clustering and Asymmetric Distance Measurements ........................................... 66
Using K-Medians Clustering ..................................... 66
Getting a More Appropriate Distance Metric .................. 67
Putting It All in Excel ............................................... 69
The Top Deals for the 5-Medians Clusters ..................... 70
Wrapping Up ........................................................... 75

3 Naïve Bayes and the Incredible Lightness of Being an Idiot ............ 77
When You Name a Product Mandrill, You’re Going to Get Some Signal and
Some Noise ................................................................ 77
The World’s Fastest Intro to Probability Theory ............... 79
Totaling Conditional Probabilities .................................. 80
Joint Probability, the Chain Rule, and Independence .......... 80
What Happens in a Dependent Situation? ...................... 81
Bayes Rule ................................................................. 82
Using Bayes Rule to Create an AI Model ....................... 83
High-Level Class Probabilities Are Often Assumed to Be Equal 84
A Couple More Odds and Ends ..................................... 85
Let’s Get This Excel Party Started ............................... 87
Removing Extraneous Punctuation ................................ 87
Splitting on Spaces .................................................... 88
Counting Tokens and Calculating Probabilities ............... 92
And We Have a Model! Let’s Use It ............................. 94
Wrapping Up ........................................................... 98

4 Optimization Modeling: Because That “Fresh Squeezed” Orange Juice
Ain’t Gonna Blend Itself ............................................. 101
Why Should Data Scientists Know Optimization? .......... 102
1 Starting with a Simple Trade-Off .................................................. 103
   Representing the Problem as a Polytope .................................. 103
   Solving by Sliding the Level Set ............................................. 105
   The Simplex Method: Rooting around the Corners ...................... 106
   Working in Excel ............................................................... 108
   There’s a Monster at the End of This Chapter ............................ 117

5 Cluster Analysis Part II: Network Graphs and Community Detection .... 155
   What Is a Network Graph? ..................................................... 156
   Visualizing a Simple Graph .................................................. 157
   Brief Introduction to Gephi .................................................. 159
      Gephi Installation and File Preparation ................................. 160
      Laying Out the Graph ..................................................... 162
      Node Degree ............................................................... 165
      Pretty Printing ............................................................ 166
      Touching the Graph Data .................................................. 168
   Building a Graph from the Wholesale Wine Data ....................... 170
      Creating a Cosine Similarity Matrix .................................... 172
      Producing an r-Neighborhood Graph .................................. 174
   How Much Is an Edge Worth? Points and Penalties in Graph Modularity 178
### Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Let’s Train It</td>
<td>258</td>
</tr>
<tr>
<td></td>
<td>Evaluating the Bagged Model</td>
<td>267</td>
</tr>
<tr>
<td>8</td>
<td>Forecasting: Breathe Easy; You Can’t Win</td>
<td>285</td>
</tr>
<tr>
<td></td>
<td>The Sword Trade Is Hopping</td>
<td>286</td>
</tr>
<tr>
<td></td>
<td>Getting Acquainted with Time Series Data</td>
<td>286</td>
</tr>
<tr>
<td></td>
<td>Starting Slow with Simple Exponential Smoothing</td>
<td>288</td>
</tr>
<tr>
<td></td>
<td>Setting Up the Simple Exponential Smoothing Forecast</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>You Might Have a Trend</td>
<td>296</td>
</tr>
<tr>
<td></td>
<td>Holt’s Trend-Corrected Exponential Smoothing</td>
<td>299</td>
</tr>
<tr>
<td></td>
<td>Setting Up Holt’s Trend-Corrected Smoothing in a Spreadsheet</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>So Are You Done? Looking at Autocorrelations</td>
<td>306</td>
</tr>
<tr>
<td></td>
<td>Multiplicative Holt-Winters Exponential Smoothing</td>
<td>313</td>
</tr>
<tr>
<td></td>
<td>Setting the Initial Values for Level, Trend, and Seasonality</td>
<td>315</td>
</tr>
<tr>
<td></td>
<td>Getting Rolling on the Forecast</td>
<td>319</td>
</tr>
<tr>
<td></td>
<td>And...Optimize!</td>
<td>324</td>
</tr>
<tr>
<td></td>
<td>Please Tell Me We’re Done Now!!!</td>
<td>326</td>
</tr>
<tr>
<td></td>
<td>Putting a Prediction Interval around the Forecast</td>
<td>327</td>
</tr>
<tr>
<td></td>
<td>Creating a Fan Chart for Effect</td>
<td>331</td>
</tr>
<tr>
<td></td>
<td>Wrapping Up</td>
<td>333</td>
</tr>
<tr>
<td>9</td>
<td>Outlier Detection: Just Because They’re Odd Doesn’t Mean They’re</td>
<td>335</td>
</tr>
<tr>
<td></td>
<td>Unimportant</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Outliers Are (Bad?) People, Too</td>
<td>335</td>
</tr>
<tr>
<td></td>
<td>The Fascinating Case of Hadlum v. Hadlum</td>
<td>336</td>
</tr>
<tr>
<td></td>
<td>Tukey Fences</td>
<td>337</td>
</tr>
<tr>
<td></td>
<td>Applying Tukey Fences in a Spreadsheet</td>
<td>338</td>
</tr>
<tr>
<td></td>
<td>The Limitations of This Simple Approach</td>
<td>340</td>
</tr>
<tr>
<td></td>
<td>Terrible at Nothing, Bad at Everything</td>
<td>341</td>
</tr>
<tr>
<td></td>
<td>Preparing Data for Graphing</td>
<td>342</td>
</tr>
</tbody>
</table>
## Contents

Creating a Graph ................................................................. 345
Getting the k Nearest Neighbors ............................................. 347
Graph Outlier Detection Method 1: Just Use the Indegree .............. 348
Graph Outlier Detection Method 2: Getting Nuanced with k-Distance  .... 351
Graph Outlier Detection Method 3: Local Outlier Factors Are Where It’s At 353
Wrapping Up ........................................................................ 358

### 10 Moving from Spreadsheets into R. ........................................ 361
Getting Up and Running with R ................................................. 362
Some Simple Hand-Jamming ..................................................... 363
Reading Data into R ................................................................. 370
Doing Some Actual Data Science ................................................. 372
  Spherical K-Means on Wine Data in Just a Few Lines .................. 372
  Building AI Models on the Pregnancy Data ................................. 378
Forecasting in R ..................................................................... 385
Looking at Outlier Detection ..................................................... 389
Wrapping Up ........................................................................ 394

Conclusion ............................................................................. 395
Where Am I? What Just Happened? ........................................... 395
Before You Go-Go ................................................................... 395
Get to Know the Problem ........................................................ 396
We Need More Translators ......................................................... 397
Beware the Three-Headed Geek-Monster: Tools, Performance, and Mathematical Perfection .............................................. 397
You Are Not the Most Important Function of Your Organization ........ 400
Get Creative and Keep in Touch! ................................................. 400

Index ....................................................................................... 401