
Contents

About the Authors	xi
Preface	xiii
I Part I: Introduction to Data Science	1
1 Prologue: Why data science?	3
1.1 What is data science?	4
1.2 Case study: The evolution of sabermetrics	6
1.3 Datasets	7
1.4 Further resources	8
2 Data visualization	9
2.1 The 2012 federal election cycle	9
2.2 Composing data graphics	16
2.3 Importance of data graphics: <i>Challenger</i>	24
2.4 Creating effective presentations	28
2.5 The wider world of data visualization	29
2.6 Further resources	31
2.7 Exercises	32
2.8 Supplementary exercises	33
3 A grammar for graphics	35
3.1 A grammar for data graphics	35
3.2 Canonical data graphics in R	43
3.3 Extended example: Historical baby names	53
3.4 Further resources	62
3.5 Exercises	62
3.6 Supplementary exercises	65
4 Data wrangling on one table	67
4.1 A grammar for data wrangling	67
4.2 Extended example: Ben's time with the Mets	76
4.3 Further resources	84
4.4 Exercises	84
4.5 Supplementary exercises	88
5 Data wrangling on multiple tables	89
5.1 <code>inner_join()</code>	89
5.2 <code>left_join()</code>	91
5.3 Extended example: Manny Ramirez	92
5.4 Further resources	99
5.5 Exercises	99

5.6	Supplementary exercises	101
6	Tidy data	103
6.1	Tidy data	103
6.2	Reshaping data	112
6.3	Naming conventions	120
6.4	Data intake	121
6.5	Further resources	135
6.6	Exercises	135
6.7	Supplementary exercises	138
7	Iteration	139
7.1	Vectorized operations	139
7.2	Using <code>across()</code> with dplyr functions	142
7.3	The <code>map()</code> family of functions	143
7.4	Iterating over a one-dimensional vector	144
7.5	Iteration over subgroups	146
7.6	Simulation	151
7.7	Extended example: Factors associated with BMI	153
7.8	Further resources	155
7.9	Exercises	157
7.10	Supplementary exercises	157
8	Data science ethics	159
8.1	Introduction	159
8.2	Truthful falsehoods	160
8.3	Role of data science in society	161
8.4	Some settings for professional ethics	163
8.5	Some principles to guide ethical action	167
8.6	Algorithmic bias	171
8.7	Data and disclosure	172
8.8	Reproducibility	174
8.9	Ethics, collectively	175
8.10	Professional guidelines for ethical conduct	176
8.11	Further resources	176
8.12	Exercises	177
8.13	Supplementary exercises	179
II	Part II: Statistics and Modeling	181
9	Statistical foundations	183
9.1	Samples and populations	183
9.2	Sample statistics	186
9.3	The bootstrap	190
9.4	Outliers	194
9.5	Statistical models: Explaining variation	196
9.6	Confounding and accounting for other factors	199
9.7	The perils of p-values	202
9.8	Further resources	204
9.9	Exercises	205
9.10	Supplementary exercises	206

10 Predictive modeling	207
10.1 Predictive modeling	208
10.2 Simple classification models	209
10.3 Evaluating models	216
10.4 Extended example: Who has diabetes?	223
10.5 Further resources	227
10.6 Exercises	227
10.7 Supplementary exercises	228
11 Supervised learning	229
11.1 Non-regression classifiers	229
11.2 Parameter tuning	245
11.3 Example: Evaluation of income models redux	246
11.4 Extended example: Who has diabetes this time?	250
11.5 Regularization	255
11.6 Further resources	258
11.7 Exercises	259
11.8 Supplementary exercises	261
12 Unsupervised learning	263
12.1 Clustering	263
12.2 Dimension reduction	270
12.3 Further resources	278
12.4 Exercises	278
12.5 Supplementary exercises	279
13 Simulation	281
13.1 Reasoning in reverse	281
13.2 Extended example: Grouping cancers	282
13.3 Randomizing functions	285
13.4 Simulating variability	286
13.5 Random networks	293
13.6 Key principles of simulation	293
13.7 Further resources	296
13.8 Exercises	296
13.9 Supplementary exercises	298
III Part III: Topics in Data Science	299
14 Dynamic and customized data graphics	301
14.1 Rich Web content using D3.js and <code>htmlwidgets</code>	301
14.2 Animation	306
14.3 Flexdashboard	306
14.4 Interactive web apps with Shiny	308
14.5 Customization of <code>ggplot2</code> graphics	313
14.6 Extended example: Hot dog eating	317
14.7 Further resources	322
14.8 Exercises	322
14.9 Supplementary exercises	324
15 Database querying using SQL	325
15.1 From <code>dplyr</code> to SQL	325

15.2	Flat-file databases	329
15.3	The SQL universe	331
15.4	The SQL data manipulation language	332
15.5	Extended example: FiveThirtyEight flights	352
15.6	SQL vs. R	360
15.7	Further resources	360
15.8	Exercises	360
15.9	Supplementary exercises	362
16	Database administration	363
16.1	Constructing efficient SQL databases	363
16.2	Changing SQL data	369
16.3	Extended example: Building a database	371
16.4	Scalability	375
16.5	Further resources	375
16.6	Exercises	375
16.7	Supplementary exercises	376
17	Working with geospatial data	377
17.1	Motivation: What's so great about geospatial data?	377
17.2	Spatial data structures	380
17.3	Making maps	382
17.4	Extended example: Congressional districts	391
17.5	Effective maps: How (not) to lie	399
17.6	Projecting polygons	401
17.7	Playing well with others	402
17.8	Further resources	403
17.9	Exercises	404
17.10	Supplementary exercises	405
18	Geospatial computations	407
18.1	Geospatial operations	407
18.2	Geospatial aggregation	416
18.3	Geospatial joins	418
18.4	Extended example: Trail elevations at MacLeish	419
18.5	Further resources	423
18.6	Exercises	423
18.7	Supplementary exercises	424
19	Text as data	425
19.1	Regular expressions using <i>Macbeth</i>	425
19.2	Extended example: Analyzing textual data from arXiv.org	431
19.3	Ingesting text	445
19.4	Further resources	448
19.5	Exercises	448
19.6	Supplementary exercises	450
20	Network science	451
20.1	Introduction to network science	451
20.2	Extended example: Six degrees of Kristen Stewart	456
20.3	PageRank	465
20.4	Extended example: 1996 men's college basketball	467

20.5	Further resources	474
20.6	Exercises	475
20.7	Supplementary exercises	475
21	Epilogue: Towards “big data”	477
21.1	Notions of big data	477
21.2	Tools for bigger data	479
21.3	Alternatives to R	489
21.4	Closing thoughts	489
21.5	Further resources	490
IV	Part IV: Appendices	491
A	Packages used in this book	493
A.1	The mdsr package	493
A.2	Other packages	493
A.3	Further resources	498
B	Introduction to R and RStudio	499
B.1	Installation	499
B.2	Learning R	500
B.3	Fundamental structures and objects	501
B.4	Add-ons: Packages	508
B.5	Further resources	514
B.6	Exercises	515
B.7	Supplementary exercises	517
C	Algorithmic thinking	519
C.1	Introduction	519
C.2	Simple example	519
C.3	Extended example: Law of large numbers	522
C.4	Non-standard evaluation	525
C.5	Debugging and defensive coding	527
C.6	Further resources	529
C.7	Exercises	529
C.8	Supplementary exercises	530
D	Reproducible analysis and workflow	531
D.1	Scriptable statistical computing	532
D.2	Reproducible analysis with R Markdown	532
D.3	Projects and version control	535
D.4	Further resources	537
D.5	Exercises	537
D.6	Supplementary exercises	540
E	Regression modeling	541
E.1	Simple linear regression	541
E.2	Multiple regression	546
E.3	Inference for regression	552
E.4	Assumptions underlying regression	553
E.5	Logistic regression	556
E.6	Further resources	559

E.7	Exercises	561
E.8	Supplementary exercises	562
F	Setting up a database server	563
F.1	SQLite	563
F.2	MySQL	564
F.3	PostgreSQL	567
F.4	Connecting to SQL	568
	Bibliography	573
	Indices	589
	Subject index	590
	R index	622