

# Contents

<b>Preface</b>	<b>xv</b>
<b>List of Tables</b>	<b>xix</b>
<b>List of Figures</b>	<b>xxi</b>
<b>I Introduction to Data Science</b>	<b>1</b>
<b>1 Prologue: Why data science?</b>	<b>3</b>
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
<b>2 Data visualization</b>	<b>9</b>
2.1 The 2012 federal election cycle . . . . .	9
2.1.1 Are these two groups different? . . . . .	10
2.1.2 Graphing variation . . . . .	11
2.1.3 Examining relationships among variables . . . . .	12
2.1.4 Networks . . . . .	13
2.2 Composing data graphics . . . . .	14
2.2.1 A taxonomy for data graphics . . . . .	14
2.2.2 Color . . . . .	19
2.2.3 Dissecting data graphics . . . . .	20
2.3 Importance of data graphics: <i>Challenger</i> . . . . .	23
2.4 Creating effective presentations . . . . .	27
2.5 The wider world of data visualization . . . . .	28
2.6 Further resources . . . . .	30
2.7 Exercises . . . . .	30
<b>3 A grammar for graphics</b>	<b>33</b>
3.1 A grammar for data graphics . . . . .	33
3.1.1 Aesthetics . . . . .	34
3.1.2 Scale . . . . .	37
3.1.3 Guides . . . . .	38
3.1.4 Facets . . . . .	38
3.1.5 Layers . . . . .	38
3.2 Canonical data graphics in R . . . . .	39
3.2.1 Univariate displays . . . . .	39

3.2.2	Multivariate displays . . . . .	43
3.2.3	Maps . . . . .	48
3.2.4	Networks . . . . .	48
3.3	Extended example: Historical baby names . . . . .	48
3.3.1	Percentage of people alive today . . . . .	50
3.3.2	Most common women's names . . . . .	56
3.4	Further resources . . . . .	58
3.5	Exercises . . . . .	58
<b>4</b>	<b>Data wrangling</b>	<b>63</b>
4.1	A grammar for data wrangling . . . . .	63
4.1.1	<code>select()</code> and <code>filter()</code> . . . . .	63
4.1.2	<code>mutate()</code> and <code>rename()</code> . . . . .	66
4.1.3	<code>arrange()</code> . . . . .	69
4.1.4	<code>summarize()</code> with <code>group_by()</code> . . . . .	70
4.2	Extended example: Ben's time with the Mets . . . . .	72
4.3	Combining multiple tables . . . . .	79
4.3.1	<code>inner_join()</code> . . . . .	79
4.3.2	<code>left_join()</code> . . . . .	81
4.4	Extended example: Manny Ramirez . . . . .	82
4.5	Further resources . . . . .	88
4.6	Exercises . . . . .	88
<b>5</b>	<b>Tidy data and iteration</b>	<b>91</b>
5.1	Tidy data . . . . .	91
5.1.1	Motivation . . . . .	91
5.1.2	What are tidy data? . . . . .	93
5.2	Reshaping data . . . . .	98
5.2.1	Data verbs for converting wide to narrow and <i>vice versa</i> . . . . .	100
5.2.2	Spreading . . . . .	100
5.2.3	Gathering . . . . .	101
5.2.4	Example: Gender-neutral names . . . . .	101
5.3	Naming conventions . . . . .	103
5.4	Automation and iteration . . . . .	104
5.4.1	Vectorized operations . . . . .	104
5.4.2	The <code>apply()</code> family of functions . . . . .	106
5.4.3	Iteration over subgroups with <code>dplyr::do()</code> . . . . .	110
5.4.4	Iteration with <code>mosaic::do</code> . . . . .	113
5.5	Data intake . . . . .	116
5.5.1	Data-table friendly formats . . . . .	116
5.5.2	APIs . . . . .	120
5.5.3	Cleaning data . . . . .	120
5.5.4	Example: Japanese nuclear reactors . . . . .	126
5.6	Further resources . . . . .	127
5.7	Exercises . . . . .	128
<b>6</b>	<b>Professional Ethics</b>	<b>131</b>
6.1	Introduction . . . . .	131
6.2	Truthful falsehoods . . . . .	131
6.3	Some settings for professional ethics . . . . .	134
6.3.1	The chief executive officer . . . . .	134

6.3.2	Employment discrimination . . . . .	134
6.3.3	Data scraping . . . . .	135
6.3.4	Reproducible spreadsheet analysis . . . . .	135
6.3.5	Drug dangers . . . . .	135
6.3.6	Legal negotiations . . . . .	136
6.4	Some principles to guide ethical action . . . . .	136
6.4.1	Applying the precepts . . . . .	137
6.5	Data and disclosure . . . . .	140
6.5.1	Reidentification and disclosure avoidance . . . . .	140
6.5.2	Safe data storage . . . . .	141
6.5.3	Data scraping and terms of use . . . . .	141
6.6	Reproducibility . . . . .	142
6.6.1	Example: Erroneous data merging . . . . .	142
6.7	Professional guidelines for ethical conduct . . . . .	143
6.8	Ethics, collectively . . . . .	143
6.9	Further resources . . . . .	144
6.10	Exercises . . . . .	144

## II Statistics and Modeling

147

<b>7</b>	<b>Statistical foundations</b>	<b>149</b>
7.1	Samples and populations . . . . .	149
7.2	Sample statistics . . . . .	152
7.3	The bootstrap . . . . .	155
7.4	Outliers . . . . .	157
7.5	Statistical models: Explaining variation . . . . .	159
7.6	Confounding and accounting for other factors . . . . .	162
7.7	The perils of p-values . . . . .	165
7.8	Further resources . . . . .	167
7.9	Exercises . . . . .	168
<b>8</b>	<b>Statistical learning and predictive analytics</b>	<b>171</b>
8.1	Supervised learning . . . . .	172
8.2	Classifiers . . . . .	173
8.2.1	Decision trees . . . . .	173
8.2.2	Example: High-earners in the 1994 United States Census . . . . .	174
8.2.3	Tuning parameters . . . . .	180
8.2.4	Random forests . . . . .	181
8.2.5	Nearest neighbor . . . . .	182
8.2.6	Naïve Bayes . . . . .	183
8.2.7	Artificial neural networks . . . . .	185
8.3	Ensemble methods . . . . .	186
8.4	Evaluating models . . . . .	188
8.4.1	Cross-validation . . . . .	188
8.4.2	Measuring prediction error . . . . .	189
8.4.3	Confusion matrix . . . . .	189
8.4.4	ROC curves . . . . .	189
8.4.5	Bias-variance trade-off . . . . .	192
8.4.6	Example: Evaluation of income models . . . . .	192
8.5	Extended example: Who has diabetes? . . . . .	196

8.6	Regularization . . . . .	201
8.7	Further resources . . . . .	201
8.8	Exercises . . . . .	201
<b>9</b>	<b>Unsupervised learning</b>	<b>205</b>
9.1	Clustering . . . . .	205
9.1.1	Hierarchical clustering . . . . .	206
9.1.2	$k$ -means . . . . .	210
9.2	Dimension reduction . . . . .	211
9.2.1	Intuitive approaches . . . . .	212
9.2.2	Singular value decomposition . . . . .	213
9.3	Further resources . . . . .	218
9.4	Exercises . . . . .	218
<b>10</b>	<b>Simulation</b>	<b>221</b>
10.1	Reasoning in reverse . . . . .	221
10.2	Extended example: Grouping cancers . . . . .	222
10.3	Randomizing functions . . . . .	223
10.4	Simulating variability . . . . .	225
10.4.1	The partially planned rendezvous . . . . .	225
10.4.2	The jobs report . . . . .	227
10.4.3	Restaurant health and sanitation grades . . . . .	228
10.5	Simulating a complex system . . . . .	231
10.6	Random networks . . . . .	233
10.7	Key principles of simulation . . . . .	233
10.8	Further resources . . . . .	235
10.9	Exercises . . . . .	236
<b>III</b>	<b>Topics in Data Science</b>	<b>241</b>
<b>11</b>	<b>Interactive data graphics</b>	<b>243</b>
11.1	Rich Web content using <code>D3.js</code> and <code>htmlwidgets</code> . . . . .	243
11.1.1	Leaflet . . . . .	244
11.1.2	Plot.ly . . . . .	244
11.1.3	DataTables . . . . .	244
11.1.4	dygraphs . . . . .	246
11.1.5	streamgraphs . . . . .	246
11.2	Dynamic visualization using <code>ggvis</code> . . . . .	246
11.3	Interactive Web apps with Shiny . . . . .	247
11.4	Further customization . . . . .	250
11.5	Extended example: Hot dog eating . . . . .	254
11.6	Further resources . . . . .	258
11.7	Exercises . . . . .	258
<b>12</b>	<b>Database querying using SQL</b>	<b>261</b>
12.1	From <code>dplyr</code> to SQL . . . . .	261
12.2	Flat-file databases . . . . .	265
12.3	The SQL universe . . . . .	266
12.4	The SQL data manipulation language . . . . .	267
12.4.1	<code>SELECT . . . FROM</code> . . . . .	270

12.4.2	<code>WHERE</code>	272
12.4.3	<code>GROUP BY</code>	275
12.4.4	<code>ORDER BY</code>	277
12.4.5	<code>HAVING</code>	278
12.4.6	<code>LIMIT</code>	280
12.4.7	<code>JOIN</code>	281
12.4.8	<code>UNION</code>	286
12.4.9	Subqueries	287
12.5	Extended example: FiveThirtyEight flights	289
12.6	SQL vs. R	298
12.7	Further resources	298
12.8	Exercises	298
<b>13</b>	<b>Database administration</b>	<b>301</b>
13.1	Constructing efficient SQL databases	301
13.1.1	Creating new databases	301
13.1.2	<code>CREATE TABLE</code>	302
13.1.3	Keys	303
13.1.4	Indices	304
13.1.5	<code>EXPLAIN</code>	306
13.1.6	Partitioning	308
13.2	Changing SQL data	308
13.2.1	<code>UPDATE</code>	308
13.2.2	<code>INSERT</code>	309
13.2.3	<code>LOAD DATA</code>	309
13.3	Extended example: Building a database	309
13.3.1	Extract	310
13.3.2	Transform	310
13.3.3	Load into MySQL database	310
13.4	Scalability	314
13.5	Further resources	314
13.6	Exercises	314
<b>14</b>	<b>Working with spatial data</b>	<b>317</b>
14.1	Motivation: What's so great about spatial data?	317
14.2	Spatial data structures	319
14.3	Making maps	322
14.3.1	Static maps with <code>ggmap</code>	322
14.3.2	Projections	324
14.3.3	Geocoding, routes, and distances	330
14.3.4	Dynamic maps with <code>leaflet</code>	332
14.4	Extended example: Congressional districts	333
14.4.1	Election results	334
14.4.2	Congressional districts	336
14.4.3	Putting it all together	338
14.4.4	Using <code>ggmap</code>	340
14.4.5	Using <code>leaflet</code>	343
14.5	Effective maps: How (not) to lie	343
14.6	Extended example: Historical airline route maps	345
14.6.1	Using <code>ggmap</code>	346
14.6.2	Using <code>leaflet</code>	347

14.7	Projecting polygons . . . . .	349
14.8	Playing well with others . . . . .	351
14.9	Further resources . . . . .	352
14.10	Exercises . . . . .	352
<b>15</b>	<b>Text as data</b>	<b>355</b>
15.1	Tools for working with text . . . . .	355
15.1.1	Regular expressions using <i>Macbeth</i> . . . . .	355
15.1.2	Example: Life and death in <i>Macbeth</i> . . . . .	359
15.2	Analyzing textual data . . . . .	360
15.2.1	Corpora . . . . .	364
15.2.2	Word clouds . . . . .	365
15.2.3	Document term matrices . . . . .	365
15.3	Ingesting text . . . . .	367
15.3.1	Example: Scraping the songs of the Beatles . . . . .	367
15.3.2	Scraping data from Twitter . . . . .	369
15.4	Further resources . . . . .	374
15.5	Exercises . . . . .	374
<b>16</b>	<b>Network science</b>	<b>377</b>
16.1	Introduction to network science . . . . .	377
16.1.1	Definitions . . . . .	377
16.1.2	A brief history of network science . . . . .	378
16.2	Extended example: Six degrees of Kristen Stewart . . . . .	382
16.2.1	Collecting Hollywood data . . . . .	382
16.2.2	Building the Hollywood network . . . . .	384
16.2.3	Building a Kristen Stewart oracle . . . . .	387
16.3	PageRank . . . . .	390
16.4	Extended example: 1996 men’s college basketball . . . . .	391
16.5	Further resources . . . . .	398
16.6	Exercises . . . . .	398
<b>17</b>	<b>Epilogue: Towards “big data”</b>	<b>401</b>
17.1	Notions of big data . . . . .	401
17.2	Tools for bigger data . . . . .	403
17.2.1	Data and memory structures for big data . . . . .	403
17.2.2	Compilation . . . . .	404
17.2.3	Parallel and distributed computing . . . . .	404
17.2.4	Alternatives to SQL . . . . .	411
17.3	Alternatives to R . . . . .	413
17.4	Closing thoughts . . . . .	413
17.5	Further resources . . . . .	413
<b>IV</b>	<b>Appendices</b>	<b>415</b>
<b>A</b>	<b>Packages used in this book</b>	<b>417</b>
A.1	The <code>mdsr</code> package . . . . .	417
A.2	The <code>etl</code> package suite . . . . .	417
A.3	Other packages . . . . .	418
A.4	Further resources . . . . .	420

<b>B</b>	<b>Introduction to R and RStudio</b>	<b>421</b>
B.1	Installation . . . . .	421
B.1.1	Installation under Windows . . . . .	422
B.1.2	Installation under Mac OS X . . . . .	422
B.1.3	Installation under Linux . . . . .	422
B.1.4	RStudio . . . . .	422
B.2	Running RStudio and sample session . . . . .	422
B.3	Learning R . . . . .	423
B.3.1	Getting help . . . . .	423
B.3.2	swirl . . . . .	424
B.4	Fundamental structures and objects . . . . .	425
B.4.1	Objects and vectors . . . . .	425
B.4.2	Operators . . . . .	427
B.4.3	Lists . . . . .	427
B.4.4	Matrices . . . . .	428
B.4.5	Dataframes . . . . .	428
B.4.6	Attributes and classes . . . . .	430
B.4.7	Options . . . . .	432
B.4.8	Functions . . . . .	432
B.5	Add-ons: Packages . . . . .	433
B.5.1	Introduction to packages . . . . .	433
B.5.2	CRAN task views . . . . .	434
B.5.3	Session information . . . . .	434
B.5.4	Packages and name conflicts . . . . .	435
B.5.5	Maintaining packages . . . . .	436
B.5.6	Installed libraries and packages . . . . .	436
B.6	Further resources . . . . .	436
B.7	Exercises . . . . .	440
<b>C</b>	<b>Algorithmic thinking</b>	<b>443</b>
C.1	Introduction . . . . .	443
C.2	Simple example . . . . .	443
C.3	Extended example: Law of large numbers . . . . .	446
C.4	Non-standard evaluation . . . . .	448
C.5	Debugging and defensive coding . . . . .	452
C.6	Further resources . . . . .	453
C.7	Exercises . . . . .	453
<b>D</b>	<b>Reproducible analysis and workflow</b>	<b>455</b>
D.1	Scriptable statistical computing . . . . .	456
D.2	Reproducible analysis with R Markdown . . . . .	456
D.3	Projects and version control . . . . .	459
D.4	Further resources . . . . .	459
D.5	Exercises . . . . .	461
<b>E</b>	<b>Regression modeling</b>	<b>465</b>
E.1	Simple linear regression . . . . .	465
E.1.1	Motivating example: Modeling usage of a rail trail . . . . .	466
E.1.2	Model visualization . . . . .	467
E.1.3	Measuring the strength of fit . . . . .	467
E.1.4	Categorical explanatory variables . . . . .	469

E.2	Multiple regression . . . . .	470
E.2.1	Parallel slopes: Multiple regression with a categorical variable . . . . .	470
E.2.2	Parallel planes: Multiple regression with a second quantitative variable . . . . .	471
E.2.3	Non-parallel slopes: Multiple regression with interaction . . . . .	472
E.2.4	Modelling non-linear relationships . . . . .	472
E.3	Inference for regression . . . . .	474
E.4	Assumptions underlying regression . . . . .	475
E.5	Logistic regression . . . . .	477
E.6	Further resources . . . . .	481
E.7	Exercises . . . . .	482
<b>F</b>	<b>Setting up a database server</b>	<b>487</b>
F.1	SQLite . . . . .	487
F.2	MySQL . . . . .	488
F.2.1	Installation . . . . .	488
F.2.2	Access . . . . .	488
F.2.3	Running scripts from the command line . . . . .	491
F.3	PostgreSQL . . . . .	491
F.4	Connecting to SQL . . . . .	492
F.4.1	The command line client . . . . .	492
F.4.2	GUIs . . . . .	492
F.4.3	R and RStudio . . . . .	492
F.4.4	Load into SQLite database . . . . .	497
	<b>Bibliography</b>	<b>499</b>
	<b>Indices</b>	<b>513</b>
	Subject index . . . . .	514
	R index . . . . .	543