

BRIEF CONTENTS

List of figures and tables	xiii
Discover this textbook's online resources!	xv
About the author	xvii
Acknowledgements	xix
Prologue	xxiii
PART I Thinking Programmatically	1
1 Introduction: Thinking of life at scale	3
2 The Series: Taming the distribution	23
3 The DataFrame: Python's tabular format	47
PART II Accessing and Converting Data	75
4 File types: Getting data in	77
5 Merging and grouping data	103
6 Accessing data on the World Wide Web using code	131
7 Accessing APIs, including Twitter and Reddit	149
PART III Interpreting Data: Expectations Versus Observations	169
8 Research questions	171
9 Visualising expectations: Comparing statistical tests and plots	185
PART IV Social Data Science in Practice: Four Approaches	219
10 Cleaning data for socially interesting features	221
11 Introducing natural language processing: Cleaning, summarising, and classifying text	249

vi | BRIEF CONTENTS

12	Introducing time-series data: Showing periods and trends	269
13	Introducing network analysis: Structuring relationships	289
14	Introducing geographic information systems: Data across space and place	315
15	Conclusion: There (to data science) and back again (to social science)	339
	References	343
	Index	353

CONTENTS

List of figures and tables	xiii
Discover this textbook's online resources!	xv
About the author	xvii
Acknowledgements	xix
Prologue	xxiii
0.1 Scaling up: Thinking about programming in the social sciences	xxiii
0.2 Who is this book for?	xxv
0.3 Why Python (and not R, Stata, Java, C, etc.)?	xxvi
0.3.1 How much Python should I already know?	xxvii
0.4 What version of Python?	xxviii
0.4.1 Part I. Thinking programmatically	xxix
0.4.2 Part II. Accessing and converting data	xxix
0.4.3 Part III. Interpreting data: Expectations versus observations	xxx
0.4.4 Part IV. Social data science in practice: Four approaches	xxxi
0.5 What about statistics?	xxxi
0.6 Writing and coding considerations	xxxii
0.6.1 My final tip before we go	xxxiii
PART I Thinking Programmatically	1
1 Introduction: Thinking of life at scale	3
1.1 From social science to what?	4
1.2 (PO)DKW: A potential theoretical framework for data science	5
1.2.1 What is data?	5
1.2.2 From data to wisdom	8
1.3 Beyond the interface	9
1.4 Fixed, variable, and marginal costs: Why not to build a barn	11
1.4.1 From economics to data science	11
1.4.2 The challenges of maximising fixed costs	13
1.5 Code should be FREE	14
1.5.1 Functioning code	14
1.5.2 Robust code	15
1.5.3 Elegant code	16
1.5.4 Efficient code	17
1.6 Pseudocode (and pseudo-pseudocode)	18
1.6.1 Attempt 1. Pseudocode as written word	18
1.6.2 Attempt 2. Pseudocode as mathematical formula	19
1.6.3 Attempt 3. Pseudocode as written code	19
1.6.4 Attempt 4. Slightly more formal pseudocode (in a Python style)	19
1.7 Summary	19
1.8 Further reading	20
1.9 Extensions and reflections	21

2	The Series: Taming the distribution	23
2.1	Introducing the Series: Python's way to store a distribution	24
2.1.1	Working from index	26
2.1.2	Working from values (and masking)	28
2.1.3	Working from distributions	30
2.1.4	Adding data to a Series	32
2.1.5	Deleting data from a Series	35
2.1.6	Working with missing data in a Series	36
2.1.7	Getting unique values in a Series	37
2.2	Changing a Series	38
2.2.1	Changing the order of items in the Series	38
2.2.2	Changing the type of the Series	39
2.2.3	Changing Series values I: Arithmetic operators	41
2.2.4	Changing Series values II: Recoding values using map	42
2.2.5	Changing Series values III: Defining your own mapping	43
2.3	Summary	45
2.4	Extensions and reflections	45
3	The DataFrame: Python's tabular format	47
3.1	From the Series to the DataFrame	48
3.2	A DataFrame with multiple columns	50
3.2.1	From a list of lists	50
3.2.2	From a dictionary	51
3.3	Getting data from a DataFrame: Querying, masking, and slicing	53
3.3.1	Getting data about the DataFrame itself	53
3.3.2	Returning a single row or column	54
3.3.3	Returning multiple columns	55
3.3.4	Returning a single element	55
3.3.5	Returning a slice of data	56
3.4	Changing data at different scales	57
3.4.1	Adding data to an existing DataFrame	57
3.4.2	Adding one DataFrame to another	60
3.4.3	Changing a column or the entire DataFrame: apply, map, and applymap	61
3.4.4	Deep versus shallow copies	65
3.5	Advanced topics: numpy and numpy arrays	67
3.5.1	Reshaping in numpy	69
3.5.2	Linear algebra and numpy	71
3.6	Summary	72
3.7	Further reading	72
3.8	Extensions and reflections	73
	PART II Accessing and Converting Data	75
4	File types: Getting data in	77
4.1	Importing data to a DataFrame	78
4.1.1	A important note on file organisation	79
4.1.2	Example data	79
4.2	Rectangular data: CSV	80
4.2.1	Using the csv library	80
4.2.2	Using the pandas CSV reader: read_csv()	82

4.3 Rectangular rich data: Excel	83
4.4 Nested data: JSON	86
4.4.1 Loading JSON	87
4.5 Nested markup languages: HTML and XML	91
4.5.1 HTML: Hypertext Markup Language	91
4.5.2 Wikipedia as a data source	92
4.5.3 Wikipedia as HTML	92
4.5.4 Using BeautifulSoup (bs4) for markup data	93
4.5.5 Data scepticism	94
4.5.6 XML	96
4.6 Serialisation	100
4.6.1 Long-term storage: Pickles and feather	101
4.7 Summary	101
4.8 Extensions and reflections	102
5 Merging and grouping data	103
5.1 Combining data across tables	104
5.2 A review of adding data to a DataFrame using concat	104
5.2.1 Adding rows	104
5.2.2 Adding columns	107
5.2.3 Multi-level indexed data	109
5.2.4 Transposing a DataFrame	110
5.3 The 'key' to merging	110
5.3.1 One-to-many versus one-to-one relationships	111
5.4 Understanding joins	114
5.4.1 A join as a kind of set logic	114
5.4.2 Inner join	116
5.4.3 Outer join	116
5.4.4 Left join	117
5.4.5 Right join	117
5.5 Grouping and aggregating data	118
5.5.1 Mean centring	120
5.6 Long versus wide data	122
5.6.1 Advanced reshaping	123
5.7 Using SQL databases	123
5.7.1 SQL basics	124
5.7.2 Using SQL for aggregation and filtering	126
5.8 Summary	128
5.9 Further reading	129
5.10 Extensions and reflections	129
6 Accessing data on the World Wide Web using code	131
6.1 Accessing data I: Remote access of webpages	132
6.1.1 What is a URL?	133
6.1.2 URL parsing	135
6.1.3 What is a web request?	136
6.2 An example web collection task using paging	138
6.3 Other web-related issues to consider	143
6.3.1 When to use your own versus someone else's program	143
6.3.2 Are there ways to simulate a browser?	143
6.4 Ethical issues to consider	143
6.4.1 What is public data and how public?	143
6.4.2 Considering data minimisation as a basic ethical principle	144

6.5 Summary	146
6.6 Further reading in ethics of data access and privacy	146
6.7 Extensions and reflections	147
7 Accessing APIs, including Twitter and Reddit	149
7.1 Accessing APIs: Abstracting from the web	150
7.1.1 Identifying yourself: Keys and tokens	150
7.1.2 Securely using credentials	152
7.2 Accessing Twitter data through the API	154
7.2.1 Troubleshooting requests	155
7.2.2 Access rights and Twitter	156
7.2.3 Strategies for navigating Twitter's API	157
7.3 Using an API wrapper to simplify data access	160
7.3.1 Collecting Reddit data using <code>praw</code>	160
7.3.2 Building a comment tree on Reddit	162
7.4 Considerations for a data collection pipeline	163
7.4.1 Version control systems and servers	163
7.4.2 Storing data remotely	164
7.4.3 Jupyter in the browser as an alternative	164
7.5 APIs and epistemology: How data access can mean knowledge access	165
7.6 Summary	167
7.7 Further reading	167
7.8 Extensions and reflections	168
PART III Interpreting data: Expectations versus Observations	169
8 Research questions	171
8.1 Introduction	172
8.1.1 What is a research question?	172
8.2 Inductive, deductive, and abductive research questions	173
8.2.1 Deductive research questions and the null hypothesis	174
8.2.2 Abductive reasoning and the educated guess	175
8.3 Avoiding description: Expectation and systematic observation in science	176
8.4 Prediction versus explanation	177
8.4.1 Prediction and resampling	178
8.5 Linking hypotheses to approaches	179
8.6 Operationalisation	180
8.7 Boundedness and research questions	181
8.8 Summary	182
8.9 Further reading	183
8.10 Extensions and reflections	183
9 Visualising expectations: Comparing statistical tests and plots	185
9.1 Introduction: Why show data?	186
9.2 Visualising distributions	188
9.2.1 Uniform distribution with histogram	190
9.3 Testing a uniform distribution using a chi-squared test	192
9.4 Testing a uniform distribution using regression	194
9.4.1 Testing against a uniform distribution: Births in the UK	198
9.4.2 Annotating a figure	201
9.4.3 Normal versus skewed distributions as being interesting	204
9.5 Comparing two distributions versus two groups	204

9.5.1 Constraining our work based on the properties of data	205
9.5.2 Two continuous distributions	207
9.5.3 PRE scores	209
9.5.4 Comparing distinct groups	213
9.5.5 Summary	215
9.6 Further reading in visualisation	216
9.7 Extensions and reflections	217
PART IV Social Data Science in Practice: Four Approaches	219
10 Cleaning data for socially interesting features	221
10.1 Data as a form of social context	223
10.2 A sustained example for cleaning: Stack Exchange	226
10.2.1 Quick summaries of the dataset	229
10.3 Setting an index	231
10.4 Handling missing data	232
10.5 Cleaning numeric data	233
10.6 Cleaning up web data	235
10.6.1 Encoding	236
10.6.2 Stripping HTML from text	236
10.6.3 Extracting links from HTML	237
10.7 Cleaning up lists of data	238
10.8 Parsing time	241
10.9 Regular expressions	242
10.9.1 Further learning for regular expressions	244
10.9.2 Regular expressions and ground truth	245
10.10 Storing our work	246
10.11 Summary	246
10.12 Further reading	247
10.13 Extensions and reflections	247
11 Introducing natural language processing: Cleaning, summarising, and classifying text	249
11.1 Reading language: Encoding text	250
11.1.1 Key definitions in text	251
11.2 From text to language	252
11.3 A sample simple NLP workflow	253
11.3.1 Preprocessing text	254
11.4 NLP approaches to analysis	258
11.4.1 Scoring documents with sentiment analysis	258
11.4.2 Extracting keywords: TF-IDF scores	261
11.4.3 Text classification	263
11.5 Summary	266
11.6 Further reading	267
11.7 Extensions and reflections	268
12 Introducing time-series data: Showing periods and trends	269
12.1 Introduction: It's about time	270
12.2 Dates and the <code>datetime</code> module	271
12.2.1 Parsing time	272
12.2.2 Timezones	274
12.2.3 Localisation and time	274

12.3 Revisiting the Movie Stack Exchange data	275
12.4 <code>pandas.datetime</code> feature extraction	276
12.5 Resampling as a way to group by time period	279
12.6 Slicing and the <code>datetime</code> index in <code>pandas</code>	281
12.7 Moving window in data	283
12.7.1 Missing data in a rolling window	284
12.8 Summary	286
12.9 Further explorations	287
12.10 Extensions and reflections	288
13 Introducing network analysis: Structuring relationships	289
13.1 Introduction: The connections that signal social structure	290
13.1.1 Doing network analysis in Python	291
13.2 Creating network graphs	291
13.2.1 Selecting a graph type	292
13.2.2 Adding nodes	293
13.2.3 Adding edges	293
13.3 Adding attributes	294
13.3.1 Working with distributions of attributes: The case of degree	295
13.4 Plotting a graph	297
13.4.1 Considering layouts for a graph	299
13.5 Subgroups and communities in a network	301
13.5.1 A goodness-of-fit metric for communities	302
13.6 Creating a network from data	303
13.6.1 Whole networks versus partial networks	305
13.6.2 Weighted networks	306
13.6.3 Bipartite networks	308
13.7 Summary	312
13.8 Further reading	313
13.9 Extensions and reflections	314
14 Introducing geographic information systems: Data across space and place	315
14.1 Introduction: From space to place	316
14.2 Kinds of spatial data	316
14.2.1 From a sphere to a rectangle	317
14.2.2 Mapping places onto spaces	319
14.2.3 Introducing the <code>geopandas</code> <code>GeoDataFrame</code>	321
14.2.4 Splitting the data into intervals using <code>mapclassify</code>	323
14.2.5 Plotting points	325
14.3 Creating your own <code>GeoDataFrame</code>	326
14.3.1 Loading your own maps	327
14.3.2 Linking maps to other data sources	329
14.4 Summary	334
14.5 Further topics and reading	335
14.6 Extensions and reflections	336
15 Conclusion: There (to data science) and back again (to social science)	339
References	343
Index	353

LIST OF FIGURES AND TABLES

FIGURES

3.1	Histogram of normal distribution ($n = 1000$)	71
6.1	Histogram showing the distribution of upvotes among recent posts in Reddit's r/aww subreddit ($n = 100$)	142
9.1	Four plots showing the distribution and trend lines of Anscombe's quartet	187
9.2	Plots of three example distributions ($n = 1000$ per distribution)	190
9.3	Plots of a uniform distribution with a horizontal line to compare the expected and observed values ($\text{bins} = 16$, $n = 8000$)	191
9.4	Scatter plot showing distribution of counts from a random uniform draw ($\text{bins} = 16$, $n = 8000$)	193
9.5	Regression plot showing trend line (with shaded confidence intervals) across distribution of counts from a random uniform draw ($\text{bins} = 16$, $n = 8000$)	195
9.6	Regression plot showing trend line across distribution of counts from a random uniform draw ($\text{bins} = 16$, $n = 8000$). Plot also includes vertical lines highlighting distance between predictions on the trend line and observations (as dots)	197
9.7	Regression plot showing trend line across average number of births in the UK for 1995-2014 (one marker per day)	200
9.8	Regression plot showing trend line across average number of births in the UK for 1995-2014 (one marker per day with annotations for Christmas and 9 months thereafter)	203
9.9	A joint plot showing two distributions as histograms and a joint distribution as a hexagonal heatmap	207
9.10	Kernel density estimate plot showing distribution of sepal widths, highlighting differences between three species of flower	210
9.11	Pairgrid showing a variety of bivariate plots concerning flower measurements, grouped by flower species	211
9.12	Heatmap of correlations between four measurements of flowers	212
9.13	Box plot of flower measurements, grouped by flower species	215
12.1	Distribution of number of posts created from 2012 to 2022 on the Movie Stack Exchange	276

12.2	Distribution of number of posts created from 2012 to 2022 on the Movie Stack Exchange, coarsened to year	277
12.3	Distribution of number of posts created from 2012 to 2022 on the Movie Stack Exchange, grouped by month	278
12.4	Distribution of number of posts created on the Movie Stack Exchange, grouped by hour of post creation	278
12.5	Comparing post frequency over time using a histogram and two coarsened distributions, one by month and one by year	281
12.6	Comparison of frequency of posts created across three different-sized rolling windows: 7, 30, and 60 days	285
12.7	Line plot showing rolling window for average number of posts, with a small tolerance for a gap in the data	286
13.1	Random layout of simple four-node graph	298
13.2	Random binomial graph ($n = 50$) with a 0.05 probability of connection between any two nodes, leading to a large component and some islands	298
13.3	The largest connected component of a random binomial graph	299
13.4	Sociogram of Zachary's Karate club data, highlighting two groups which would subsequently split	301
13.5	Sociogram of Movie Stack Exchange posters who reply to each other and have used the tag <code>dialogue</code>	307
13.6	Sociogram of the co-used tags within the Movie Stack Exchange	310
13.7	Sociogram of a subset of the co-used tags within the Movie Stack Exchange	311
14.1	Plot of the world using the default <code>geopandas</code> shapefile	320
14.2	Plot of Great Britain using the default <code>geopandas</code> shapefile, highlighting the low resolution	321
14.3	Choropleth map of the world by GDP per capita illustrating difficulties in seeing small wealthy countries on map	322
14.4	Choropleth map of the world by GDP per capita with well-contained legend showing wealth disparity	323
14.5	Choropleth map of the world by GDP per capita with well-contained legend split by quantiles	324
14.6	Map of the world, with points for capital cities	326
14.7	Map of the world, highlighting a user-defined point from a <code>GeoDataFrame</code>	327
14.8	Higher-resolution shapefile of the UK from GADM	328
14.9	Distribution of case-fatality ratio by UK subnational region	332
14.10	Case-fatality ratio distribution overlaid with multiple classifications showing differences in how algorithms split the data	333
14.11	Choropleth map of case-fatality ratio using Fisher-Jenks cutpoints at the subnational level	334

TABLES

10.1	Regular expression metacharacters and what they detect from an example phrase	244
12.1	Time-specific metacharacters for marking datetime periods	273