

Contents

Acknowledgments	xi
Preface	xiii
1. What Is Data Science?	1
Unit 1. Data Analysis Sequence	3
Unit 2. Data Acquisition Pipeline	5
Unit 3. Report Structure	7
Your Turn	8
2. Core Python for Data Science	9
Unit 4. Understanding Basic String Functions	10
Unit 5. Choosing the Right Data Structure	13
Unit 6. Comprehending Lists Through List Comprehension	15
Unit 7. Counting with Counters	17
Unit 8. Working with Files	18
Unit 9. Reaching the Web	19
Unit 10. Pattern Matching with Regular Expressions	21
Unit 11. Globbing File Names and Other Strings	26
Unit 12. Pickling and Unpickling Data	27
Your Turn	28
3. Working with Text Data	29
Unit 13. Processing HTML Files	30
Unit 14. Handling CSV Files	34
Unit 15. Reading JSON Files	36
Unit 16. Processing Texts in Natural Languages	38
Your Turn	44
4. Working with Databases	47
Unit 17. Setting Up a MySQL Database	48

Unit 18.	Using a MySQL Database: Command Line	51
Unit 19.	Using a MySQL Database: pymysql	55
Unit 20.	Taming Document Stores: MongoDB	57
	Your Turn	61
5.	Working with Tabular Numeric Data	63
Unit 21.	Creating Arrays	64
Unit 22.	Transposing and Reshaping	67
Unit 23.	Indexing and Slicing	69
Unit 24.	Broadcasting	71
Unit 25.	Demystifying Universal Functions	73
Unit 26.	Understanding Conditional Functions	75
Unit 27.	Aggregating and Ordering Arrays	76
Unit 28.	Treating Arrays as Sets	78
Unit 29.	Saving and Reading Arrays	79
Unit 30.	Generating a Synthetic Sine Wave	80
	Your Turn	82
6.	Working with Data Series and Frames	83
Unit 31.	Getting Used to Pandas Data Structures	85
Unit 32.	Reshaping Data	92
Unit 33.	Handling Missing Data	98
Unit 34.	Combining Data	101
Unit 35.	Ordering and Describing Data	105
Unit 36.	Transforming Data	109
Unit 37.	Taming Pandas File I/O	116
	Your Turn	119
7.	Working with Network Data	121
Unit 38.	Dissecting Graphs	122
Unit 39.	Network Analysis Sequence	126
Unit 40.	Harnessing Networkx	127
	Your Turn	134
8.	Plotting	135
Unit 41.	Basic Plotting with PyPlot	136
Unit 42.	Getting to Know Other Plot Types	139
Unit 43.	Mastering Embellishments	140
Unit 44.	Plotting with Pandas	143
	Your Turn	146

9. Probability and Statistics	147
Unit 45. Reviewing Probability Distributions	148
Unit 46. Recollecting Statistical Measures	150
Unit 47. Doing Stats the Python Way	152
Your Turn	156
10. Machine Learning	157
Unit 48. Designing a Predictive Experiment	158
Unit 49. Fitting a Linear Regression	160
Unit 50. Grouping Data with K-Means Clustering	166
Unit 51. Surviving in Random Decision Forests	169
Your Turn	171
A1. Further Reading	173
A2. Solutions to Single-Star Projects	175
Bibliography	185
Index	187