CONTENTS IN DETAIL

INTRODUCTION **xv**

Using Python for Data Science . . . xvi
Who Should Read This Book? . . . xvi
What's in the Book? . . . xvii

1
THE BASICS OF DATA **1**

Categories of Data . . . 2
  Unstructured Data . . . 2
  Structured Data . . . 2
  Semistructured Data . . . 4
  Time Series Data . . . 5
Sources of Data . . . 6
  APIs . . . 7
  Web Pages . . . 7
  Databases . . . 8
  Files . . . 9
The Data Processing Pipeline . . . 9
  Acquisition . . . 10
  Cleansing . . . 10
  Transformation . . . 11
  Analysis . . . 11
  Storage . . . 12
The Pythonic Way . . . 13
Summary . . . 13

2
PYTHON DATA STRUCTURES **15**

Lists . . . 16
  Creating a List . . . 16
  Using Common List Object Methods . . . 16
  Using Slice Notation . . . 18
  Using a List as a Queue . . . 19
  Using a List as a Stack . . . 20
  Using Lists and Stacks for Natural Language Processing . . . 21
  Making Improvements with List Comprehensions . . . 23
Tuples . . . 27
  A List of Tuples . . . 27
  Immutability . . . 28
Dictionaries . . . 28
  A List of Dictionaries . . . 29

Adding to a Dictionary with setdefault() 29
Loading JSON into a Dictionary 31
Sets 32
Removing Duplicates from Sequences 32
Performing Common Set Operations 33
Exercise #1: Improved Photo Tag Analysis 34
Summary 35

3
PYTHON DATA SCIENCE LIBRARIES 37

NumPy 37
Installing NumPy 38
Creating a NumPy Array 38
Performing Element-Wise Operations 39
Using NumPy Statistical Functions 39
Exercise #2: Using NumPy Statistical Functions 40
pandas 40
pandas Installation 41
pandas Series 41
Exercise #3: Combining Three Series 43
pandas DataFrames 43
Exercise #4: Using Different Joins 50
scikit-learn 52
Installing scikit-learn 53
Obtaining a Sample Dataset 53
Loading the Sample Dataset into a pandas DataFrame 54
Splitting the Sample Dataset into a Training Set and a Test Set 54
Transforming Text into Numerical Feature Vectors 54
Training and Evaluating the Model 55
Making Predictions on New Data 56
Summary 56

4
ACCESSING DATA FROM FILES AND APIS 57

Importing Data Using Python's open() Function 57
Text Files 58
Tabular Data Files 59
Exercise #5: Opening JSON Files 61
Binary Files 62
Exporting Data to Files 62
Accessing Remote Files and APIs 63
How HTTP Requests Work 64
The urllib3 Library 65
The Requests Library 67
Exercise #6: Accessing an API with Requests 67
Moving Data to and from a DataFrame 68
Importing Nested JSON Structures 68
Converting a DataFrame to JSON 69
Exercise #7: Manipulating Complex JSON Structures 70
Loading Online Data into a DataFrame with pandas-datareader 71
Summary 72

5
WORKING WITH DATABASES 73

Relational Databases 74
Understanding SQL Statements 75
Getting Started with MySQL 75
Defining the Database Structure 76
Inserting Data into the Database 79
Querying Database Data 80
Exercise #8: Performing a One-to-Many Join 82
Using Database Analytics Tools 82
NoSQL Databases 88
Key-Value Stores 89
Document-Oriented Databases 90
Exercise #9: Inserting and Querying Multiple Documents 92
Summary 93

6
AGGREGATING DATA 95

Data to Aggregate 96
Combining DataFrames 98
Grouping and Aggregating the Data 100
Viewing Specific Aggregations by MultiIndex 101
Slicing a Range of Aggregated Values 103
Slicing Within Aggregation Levels 103
Adding a Grand Total 104
Adding Subtotals 105
Exercise #10: Excluding Total Rows from the DataFrame 106
Selecting All Rows in a Group 106
Summary 107

7
COMBINING DATASETS 109

Combining Built-in Data Structures 110
Combining Lists and Tuples with + ... 110
Combining Dictionaries with ** 111
Combining Corresponding Rows from Two Structures 112
Implementing Different Types of Joins for Lists 114
Concatenating NumPy Arrays 116
Exercise #11: Adding New Rows/Columns to a NumPy Array 117
Combining pandas Data Structures 117
Concatenating DataFrames 118
Joining Two DataFrames 122
Summary 126

8
CREATING VISUALIZATIONS 127

Common Visualizations 128
Line Graphs 128
Bar Graphs 129

Pie Charts . . . 130
Histograms . . . 130
Plotting with Matplotlib . . . 131
Installing Matplotlib . . . 131
Using matplotlib.pyplot . . . 131
Working with Figure and Axes Objects . . . 133
Exercise #12: Combining Bins into an "Other" Slice . . . 136
Using Other Libraries with Matplotlib . . . 137
Plotting pandas Data . . . 137
Plotting Geospatial Data with Cartopy . . . 139
Exercise#13: Drawing a Map with Cartopy and Matplotlib . . . 143
Summary . . . 143

9
ANALYZING LOCATION DATA 145

Obtaining Location Data . . . 146
Turning a Human-Readable Address into Geo Coordinates . . . 146
Getting the Geo Coordinates of a Moving Object . . . 147
Spatial Data Analysis with geopy and Shapely . . . 150
Finding the Closest Object . . . 150
Finding Objects in a Certain Area . . . 152
Exercise #14: Defining Two or More Polygons . . . 154
Combining Both Approaches . . . 154
Exercise #15: Further Improving the Pick-Up Algorithm . . . 156
Combining Spatial and Nonspatial Data . . . 156
Deriving Nonspatial Attributes . . . 156
Exercise #16: Filtering Data with a List Comprehension . . . 158
Joining Spatial and Nonspatial Datasets . . . 158
Summary . . . 159

10
ANALYZING TIME SERIES DATA 161

Regular vs. Irregular Time Series . . . 161
Common Time Series Analysis Techniques . . . 163
Calculating Percentage Changes . . . 164
Rolling Window Calculations . . . 166
Calculating the Percentage Change of a Rolling Average . . . 167
Multivariate Time Series . . . 167
Processing Multivariate Time Series . . . 168
Analyzing Dependencies Between Variables . . . 169
Exercise #17: Adding More Metrics to Analyze Dependencies . . . 172
Summary . . . 174

11
GAINING INSIGHTS FROM DATA 175

Association Rules . . . 176
Support . . . 177
Confidence . . . 177
Lift . . . 178

The Apriori Algorithm 178
Creating a Transaction Dataset 179
Identifying Frequent Itemsets 180
Generating Association Rules 181
Visualizing Association Rules 182
Gaining Actionable Insights from Association Rules 186
Generating Recommendations 186
Planning Discounts Based on Association Rules 187
Exercise #18: Mining Real Transaction Data 189
Summary 192

12
MACHINE LEARNING FOR DATA ANALYSIS 193

Why Machine Learning? 194
Types of Machine Learning 194
Supervised Learning 194
Unsupervised Learning 195
How Machine Learning Works 196
Data to Learn From 196
A Statistical Model 197
Previously Unseen Data 197
A Sentiment Analysis Example: Classifying Product Reviews 198
Obtaining Product Reviews 198
Cleansing the Data 199
Splitting and Transforming the Data 201
Training the Model 203
Evaluating the Model 203
Exercise #19: Expanding the Example Set 206
Predicting Stock Trends 206
Getting Data 207
Deriving Features from Continuous Data 208
Generating the Output Variable 209
Training and Evaluating the Model 210
Exercise #20: Experimenting with Different Stocks and New Metrics 211
Summary 211

INDEX 213