

Contents

Foreword xix

Preface xxi

Acknowledgments xxvii

About the Author xxxi

I Introduction 1

1 Pandas DataFrame Basics 3

- 1.1 Introduction 3
- 1.2 Loading Your First Data Set 4
- 1.3 Looking at Columns, Rows, and Cells 7
 - 1.3.1 Subsetting Columns 7
 - 1.3.2 Subsetting Rows 8
 - 1.3.3 Mixing It Up 12
- 1.4 Grouped and Aggregated Calculations 18
 - 1.4.1 Grouped Means 19
 - 1.4.2 Grouped Frequency Counts 23
- 1.5 Basic Plot 23
- 1.6 Conclusion 24

2 Pandas Data Structures 25

- 2.1 Introduction 25
- 2.2 Creating Your Own Data 26
 - 2.2.1 Creating a Series 26
 - 2.2.2 Creating a DataFrame 27
- 2.3 The Series 28
 - 2.3.1 The Series Is ndarray-like 30
 - 2.3.2 Boolean Subsetting: Series 30
 - 2.3.3 Operations Are Automatically Aligned and Vectorized (Broadcasting) 33

2.4	The DataFrame	36
2.4.1	Boolean Subsetting: DataFrames	36
2.4.2	Operations Are Automatically Aligned and Vectorized (Broadcasting)	37
2.5	Making Changes to Series and DataFrames	38
2.5.1	Add Additional Columns	38
2.5.2	Directly Change a Column	39
2.5.3	Dropping Values	43
2.6	Exporting and Importing Data	43
2.6.1	pickle	43
2.6.2	CSV	45
2.6.3	Excel	46
2.6.4	Feather Format to Interface With R	47
2.6.5	Other Data Output Types	47
2.7	Conclusion	47

3 Introduction to Plotting 49

3.1	Introduction	49
3.2	Matplotlib	51
3.3	Statistical Graphics Using matplotlib	56
3.3.1	Univariate	57
3.3.2	Bivariate	58
3.3.3	Multivariate Data	59
3.4	Seaborn	61
3.4.1	Univariate	62
3.4.2	Bivariate Data	65
3.4.3	Multivariate Data	73
3.5	Pandas Objects	83
3.5.1	Histograms	84
3.5.2	Density Plot	85
3.5.3	Scatterplot	85
3.5.4	Hexbin Plot	86
3.5.5	Boxplot	86
3.6	Seaborn Themes and Styles	86
3.7	Conclusion	90

II Data Manipulation 91

4 Data Assembly 93

- 4.1 Introduction 93
- 4.2 Tidy Data 93
 - 4.2.1 Combining Data Sets 94
- 4.3 Concatenation 94
 - 4.3.1 Adding Rows 94
 - 4.3.2 Adding Columns 98
 - 4.3.3 Concatenation With Different Indices 99
- 4.4 Merging Multiple Data Sets 102
 - 4.4.1 One-to-One Merge 104
 - 4.4.2 Many-to-One Merge 105
 - 4.4.3 Many-to-Many Merge 105
- 4.5 Conclusion 107

5 Missing Data 109

- 5.1 Introduction 109
- 5.2 What Is a NaN Value? 109
- 5.3 Where Do Missing Values Come From? 111
 - 5.3.1 Load Data 111
 - 5.3.2 Merged Data 112
 - 5.3.3 User Input Values 114
 - 5.3.4 Re-indexing 114
- 5.4 Working With Missing Data 116
 - 5.4.1 Find and Count missing Data 116
 - 5.4.2 Cleaning Missing Data 118
 - 5.4.3 Calculations With Missing Data 120
- 5.5 Conclusion 121

6 Tidy Data 123

- 6.1 Introduction 123
- 6.2 Columns Contain Values, Not Variables 124

6.2.1	Keep One Column Fixed	124
6.2.2	Keep Multiple Columns Fixed	126
6.3	Columns Contain Multiple Variables	128
6.3.1	Split and Add Columns Individually (Simple Method)	129
6.3.2	Split and Combine in a Single Step (Simple Method)	131
6.3.3	Split and Combine in a Single Step (More Complicated Method)	132
6.4	Variables in Both Rows and Columns	133
6.5	Multiple Observational Units in a Table (Normalization)	134
6.6	Observational Units Across Multiple Tables	137
6.6.1	Load Multiple Files Using a Loop	139
6.6.2	Load Multiple Files Using a List Comprehension	140
6.7	Conclusion	141

III Data Munging 143

7 Data Types 145

7.1	Introduction	145
7.2	Data Types	145
7.3	Converting Types	146
7.3.1	Converting to String Objects	146
7.3.2	Converting to Numeric Values	147
7.4	Categorical Data	152
7.4.1	Convert to Category	152
7.4.2	Manipulating Categorical Data	153
7.5	Conclusion	153

8 Strings and Text Data 155

- 8.1 Introduction 155
- 8.2 Strings 155
 - 8.2.1 Subsetting and Slicing Strings 155
 - 8.2.2 Getting the Last Character in a String 157
- 8.3 String Methods 158
- 8.4 More String Methods 160
 - 8.4.1 Join 160
 - 8.4.2 Splitlines 160
- 8.5 String Formatting 161
 - 8.5.1 Custom String Formatting 161
 - 8.5.2 Formatting Character Strings 162
 - 8.5.3 Formatting Numbers 162
 - 8.5.4 C printf Style Formatting 163
 - 8.5.5 Formatted Literal Strings in Python 3.6+ 163
- 8.6 Regular Expressions (RegEx) 164
 - 8.6.1 Match a Pattern 164
 - 8.6.2 Find a Pattern 168
 - 8.6.3 Substituting a Pattern 168
 - 8.6.4 Compiling a Pattern 169
- 8.7 The regex Library 170
- 8.8 Conclusion 170

9 Apply 171

- 9.1 Introduction 171
- 9.2 Functions 171
- 9.3 Apply (Basics) 172
 - 9.3.1 Apply Over a Series 173
 - 9.3.2 Apply Over a DataFrame 174
- 9.4 Apply (More Advanced) 177
 - 9.4.1 Column-wise Operations 178
 - 9.4.2 Row-wise Operations 180

- 9.5 Vectorized Functions 182
 - 9.5.1 Using numpy 184
 - 9.5.2 Using numba 185
- 9.6 Lambda Functions 185
- 9.7 Conclusion 187

10 Groupby Operations: Split–Apply–Combine 189

- 10.1 Introduction 189
- 10.2 Aggregate 190
 - 10.2.1 Basic One-Variable Grouped Aggregation 190
 - 10.2.2 Built-in Aggregation Methods 191
 - 10.2.3 Aggregation Functions 192
 - 10.2.4 Multiple Functions Simultaneously 195
 - 10.2.5 Using a dict in agg/aggregate 195
- 10.3 Transform 197
 - 10.3.1 z-Score Example 197
- 10.4 Filter 201
- 10.5 The pandas.core.groupby.DataFrameGroupBy Object 202
 - 10.5.1 Groups 202
 - 10.5.2 Group Calculations Involving Multiple Variables 203
 - 10.5.3 Selecting a Group 204
 - 10.5.4 Iterating Through Groups 204
 - 10.5.5 Multiple Groups 206
 - 10.5.6 Flattening the Results 206
- 10.6 Working With a MultiIndex 207
- 10.7 Conclusion 211

11 The datetime Data Type 213

- 11.1 Introduction 213
- 11.2 Python's datetime Object 213
- 11.3 Converting to datetime 214
- 11.4 Loading Data That Include Dates 217
- 11.5 Extracting Date Components 217

- 11.6 Date Calculations and Timedeltas 220
- 11.7 Datetime Methods 221
- 11.8 Getting Stock Data 224
- 11.9 Subsetting Data Based on Dates 225
 - 11.9.1 The DatetimeIndex Object 225
 - 11.9.2 The TimedeltaIndex Object 226
- 11.10 Date Ranges 227
 - 11.10.1 Frequencies 228
 - 11.10.2 Offsets 229
- 11.11 Shifting Values 230
- 11.12 Resampling 237
- 11.13 Time Zones 238
- 11.14 Conclusion 240

IV Data Modeling 241

12 Linear Models 243

- 12.1 Introduction 243
- 12.2 Simple Linear Regression 243
 - 12.2.1 Using statsmodels 243
 - 12.2.2 Using sklearn 245
- 12.3 Multiple Regression 247
 - 12.3.1 Using statsmodels 247
 - 12.3.2 Using statsmodels With Categorical Variables 248
 - 12.3.3 Using sklearn 249
 - 12.3.4 Using sklearn With Categorical Variables 250
- 12.4 Keeping Index Labels From sklearn 251
- 12.5 Conclusion 252

13 Generalized Linear Models 253

- 13.1 Introduction 253
- 13.2 Logistic Regression 253
 - 13.2.1 Using Statsmodels 255
 - 13.2.2 Using Sklearn 256

- 13.3 Poisson Regression 257
 - 13.3.1 Using Statsmodels 258
 - 13.3.2 Negative Binomial Regression for Overdispersion 259
- 13.4 More Generalized Linear Models 260
- 13.5 Survival Analysis 260
 - 13.5.1 Testing the Cox Model Assumptions 263
- 13.6 Conclusion 264

14 Model Diagnostics 265

- 14.1 Introduction 265
- 14.2 Residuals 265
 - 14.2.1 Q-Q Plots 268
- 14.3 Comparing Multiple Models 270
 - 14.3.1 Working With Linear Models 270
 - 14.3.2 Working With GLM Models 273
- 14.4 k -Fold Cross-Validation 275
- 14.5 Conclusion 278

15 Regularization 279

- 15.1 Introduction 279
- 15.2 Why Regularize? 279
- 15.3 LASSO Regression 281
- 15.4 Ridge Regression 283
- 15.5 Elastic Net 285
- 15.6 Cross-Validation 287
- 15.7 Conclusion 289

16 Clustering 291

- 16.1 Introduction 291
- 16.2 k -Means 291
 - 16.2.1 Dimension Reduction With PCA 294
- 16.3 Hierarchical Clustering 297
 - 16.3.1 Complete Clustering 298
 - 16.3.2 Single Clustering 298
 - 16.3.3 Average Clustering 299

- 16.3.4 Centroid Clustering 299
- 16.3.5 Manually Setting the Threshold 299
- 16.4 Conclusion 301

V Conclusion 303

17 Life Outside of Pandas 305

- 17.1 The (Scientific) Computing Stack 305
- 17.2 Performance 306
 - 17.2.1 Timing Your Code 306
 - 17.2.2 Profiling Your Code 307
- 17.3 Going Bigger and Faster 307

18 Toward a Self-Directed Learner 309

- 18.1 It's Dangerous to Go Alone! 309
- 18.2 Local Meetups 309
- 18.3 Conferences 309
- 18.4 The Internet 310
- 18.5 Podcasts 310
- 18.6 Conclusion 311

VI Appendixes 313

A Installation 315

- A.1 Installing Anaconda 315
 - A.1.1 Windows 315
 - A.1.2 Mac 316
 - A.1.3 Linux 316
- A.2 Uninstall Anaconda 316

B Command Line 317

- B.1 Installation 317
 - B.1.1 Windows 317
 - B.1.2 Mac 317
 - B.1.3 Linux 318
- B.2 Basics 318

C Project Templates 319**D Using Python 321**

- D.1 Command Line and Text Editor 321
- D.2 Python and IPython 322
- D.3 Jupyter 322
- D.4 Integrated Development Environments (IDEs) 322

E Working Directories 325**F Environments 327****G Install Packages 329**

- G.1 Updating Packages 330

H Importing Libraries 331**I Lists 333****J Tuples 335****K Dictionaries 337****L Slicing Values 339****M Loops 341****N Comprehensions 343****O Functions 345**

- O.1 Default Parameters 347
- O.2 Arbitrary Parameters 347
 - O.2.1 *args 347
 - O.2.2 **kwargs 348

P Ranges and Generators 349

Q Multiple Assignment 351**R numpy ndarray 353****S Classes 355****T Odo: The Shapeshifter 357****Index 359**