Table of Contents

Preface 1

Section 1: Getting Started with Pandas

Chapter 1: Introduction to Data Analysis 9
- Chapter materials 10
- Fundamentals of data analysis 12
  - Data collection 13
  - Data wrangling 14
  - Exploratory data analysis 15
  - Drawing conclusions 16
- Statistical foundations 17
  - Sampling 18
  - Descriptive statistics 19
    - Measures of central tendency 19
      - Mean 19
      - Median 20
      - Mode 20
    - Measures of spread 21
      - Range 21
      - Variance 21
      - Standard deviation 22
      - Coefficient of variation 23
      - Interquartile range 23
      - Quartile coefficient of dispersion 24
    - Summarizing data 24
    - Common distributions 28
    - Scaling data 30
    - Quantifying relationships between variables 30
    - Pitfalls of summary statistics 32
  - Prediction and forecasting 34
  - Inferential statistics 38
- Setting up a virtual environment 40
  - Virtual environments 40
    - venv 41
      - Windows 41
      - Linux/macOS 42
    - Anaconda 43
  - Installing the required Python packages 44
  - Why pandas? 44
  - Jupyter Notebooks 45
    - Launching JupyterLab 45
    - Validating the virtual environment 47

Reordering, reindexing, and sorting data 147
Restructuring the data 158
Pivoting DataFrames 161
Melting DataFrames 166
Handling duplicate, missing, or invalid data 170
Finding the problematic data 170
Mitigating the issues 177
Summary 185
Exercises 185
Further reading 186
Chapter 4: Aggregating Pandas DataFrames 187
Chapter materials 188
Database-style operations on DataFrames 189
Querying DataFrames 190
Merging DataFrames 192
DataFrame operations 203
Arithmetic and statistics 203
Binning and thresholds 206
Applying functions 211
Window calculations 213
Pipes 216
Aggregations with pandas and numpy 220
Summarizing DataFrames 221
Using groupby 222
Pivot tables and crosstabs 228
Time series 233
Time-based selection and filtering 233
Shifting for lagged data 238
Differenced data 239
Resampling 240
Merging 244
Summary 247
Exercises 247
Further reading 249
Chapter 5: Visualizing Data with Pandas and Matplotlib 251
Chapter materials 252
An introduction to matplotlib 252
The basics 253
Plot components 259
Additional options 263
Plotting with pandas 265
Evolution over time 267
Relationships between variables 270

Closing JupyterLab 48
Summary 49
Exercises 49
Further reading 51
Chapter 2: Working with Pandas DataFrames 53
Chapter materials 54
Pandas data structures 55
Series 59
Index 60
DataFrame 61
Bringing data into a pandas DataFrame 64
From a Python object 65
From a file 69
From a database 73
From an API 75
Inspecting a DataFrame object 78
Examining the data 78
Describing and summarizing the data 82
Grabbing subsets of the data 85
Selection 86
Slicing 88
Indexing 90
Filtering 93
Adding and removing data 99
Creating new data 100
Deleting unwanted data 108
Summary 112
Exercises 112
Further reading 113

Section 2: Using Pandas for Data Analysis

Chapter 3: Data Wrangling with Pandas 117
Chapter materials 118
What is data wrangling? 120
Data cleaning 120
Data transformation 121
The wide data format 122
The long data format 124
Data enrichment 128
Collecting temperature data 128
Cleaning up the data 138
Renaming columns 139
Type conversion 140

Exploratory data analysis 376
The Visualizer class family 380
Visualizing a stock 393
Visualizing multiple assets 402
Technical analysis of financial instruments 409
The StockAnalyzer class 410
The AssetGroupAnalyzer class 417
Comparing assets 419
Modeling performance 422
The StockModeler class 422
Time series decomposition 428
ARIMA 429
Linear regression with statsmodels 431
Comparing models 433
Summary 436
Exercises 436
Further reading 437
Chapter 8: Rule-Based Anomaly Detection 439
Chapter materials 440
Simulating login attempts 440
Assumptions 440
The login_attempt_simulator package 442
Helper functions 442
The LoginAttemptSimulator class 444
Simulating from the command line 454
Exploratory data analysis 459
Rule-based anomaly detection 471
Percent difference 472
Tukey fence 477
Z-score 478
Evaluating performance 480
Summary 486
Exercises 486
Further reading 487

Section 4: Introduction to Machine Learning with Scikit-Learn

Chapter 9: Getting Started with Machine Learning in Python 491
Chapter materials 492
Learning the lingo 493
Exploratory data analysis 496
Red wine quality data 496
White and red wine chemical properties data 499

Distributions 276
Counts and frequencies 283
The pandas.plotting subpackage 291
Scatter matrices 292
Lag plots 294
Autocorrelation plots 296
Bootstrap plots 297
Summary 298
Exercises 299
Further reading 300
Chapter 6: Plotting with Seaborn and Customization Techniques 301
Chapter materials 302
Utilizing seaborn for advanced plotting 303
Categorical data 303
Correlations and heatmaps 306
Regression plots 315
Distributions 319
Faceting 320
Formatting 322
Titles and labels 322
Legends 325
Formatting axes 326
Customizing visualizations 331
Adding reference lines 331
Shading regions 337
Annotations 339
Colors 341
Summary 354
Exercises 354
Further reading 355

Section 3: Applications - Real-World Analyses Using Pandas

Chapter 7: Financial Analysis - Bitcoin and the Stock Market 359
Chapter materials 360
Building a Python package 362
Package structure 362
Overview of the stock_analysis package 364
Data extraction with pandas 366
The StockReader class 366
Bitcoin historical data from HTML 373
S&P 500 historical data from Yahoo! Finance 374
FAANG historical data from IEX 375

Ensemble methods 596
Random forest 597
Gradient boosting 598
Voting 600
Inspecting classification prediction confidence 602
Addressing class imbalance 605
Under-sampling 608
Over-sampling 609
Regularization 610
Summary 613
Exercises 614
Further reading 617
Chapter 11: Machine Learning Anomaly Detection 619
Chapter materials 620
Exploring the data 622
Unsupervised methods 629
Isolation forest 630
Local outlier factor 633
Comparing models 634
Supervised methods 638
Baselining 640
Dummy classifier 640
Naive Bayes 642
Logistic regression 646
Online learning 648
Creating the PartialFitPipeline subclass 649
Stochastic gradient descent classifier 650
Building our initial model 651
Evaluating the model 652
Updating the model 658
Presenting our results 659
Further improvements 660
Summary 661
Exercises 662
Further reading 663

Section 5: Additional Resources

Chapter 12: The Road Ahead 667
Data resources 668
Python packages 668
Seaborn 668
Scikit-learn 669
Searching for data 669
APIs 669

Planets and exoplanets data 502
Preprocessing data 509
Training and testing sets 509
Scaling and centering data 512
Encoding data 513
Imputing 517
Additional transformers 518
Pipelines 520
Clustering 521
k-means 522
Grouping planets by orbit characteristics 523
Elbow point method for determining k 526
Interpreting centroids and visualizing the cluster space 528
Evaluating clustering results 531
Regression 533
Linear regression 533
Predicting the length of a year on a planet 534
Interpreting the linear regression equation 535
Making predictions 536
Evaluating regression results 537
Analyzing residuals 538
Metrics 539
Classification 543
Logistic regression 544
Predicting red wine quality 545
Determining wine type by chemical properties 546
Evaluating classification results 547
Confusion matrix 547
Classification metrics 550
Accuracy and error rate 550
Precision and recall 551
F score 553
Sensitivity and specificity 554
ROC curve 555
Precision-recall curve 559
Summary 563
Exercises 564
Further reading 566
Chapter 10: Making Better Predictions - Optimizing Models 569
Chapter materials 570
Hyperparameter tuning with grid search 572
Feature engineering 580
Interaction terms and polynomial features 581
Dimensionality reduction 583
Feature unions 592
Feature importances 593

Websites 670
Finance 670
Government data 670
Health and economy 671
Social networks 671
Sports 671
Miscellaneous 672
Practicing working with data 672
Python practice 673
Summary 674
Exercises 675
Further reading 675
Solutions 681
Appendix A: Appendix 683
Data analysis workflow 683
Choosing the appropriate visualization 684
Machine learning workflow 685
Other Books You May Enjoy 687
Index 691