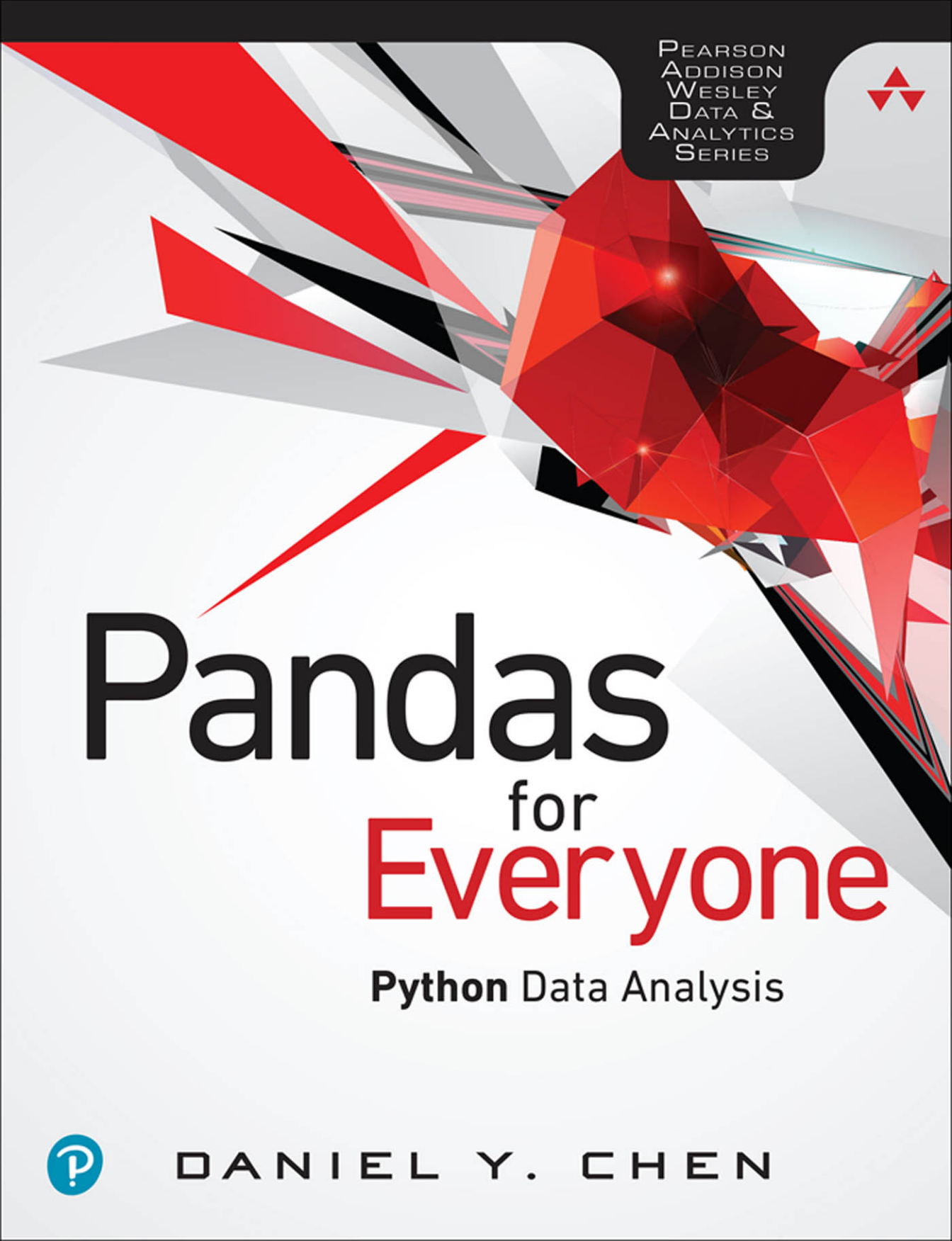




PEARSON
ADDISON
WESLEY
DATA &
ANALYTICS
SERIES



Pandas for Everyone

Python Data Analysis



DANIEL Y. CHEN

Pandas for Everyone

Pandas for Everyone: Python Data Analysis

Table of Contents

Cover

Title Page

Copyright Page

Contents

Foreword

Preface

Acknowledgments

About the Author

I: Introduction

1 Pandas DataFrame Basics

1.1 Introduction

1.2 Loading Your First Data Set

1.3 Looking at Columns, Rows, and Cells

1.3.1 Subsetting Columns

1.3.2 Subsetting Rows

1.3.3 Mixing It Up

1.4 Grouped and Aggregated Calculations

1.4.1 Grouped Means

1.4.2 Grouped Frequency Counts

1.5 Basic Plot

1.6 Conclusion

2 Pandas Data Structures

Table of Contents

2.1 Introduction

2.2 Creating Your Own Data

2.2.1 Creating a Series

2.2.2 Creating a DataFrame

2.3 The Series

2.3.1 The Series Is ndarray-like

2.3.2 Boolean Subsetting: Series

2.3.3 Operations Are Automatically Aligned and Vectorized (Broadcasting)

2.4 The DataFrame

2.4.1 Boolean Subsetting: DataFrames

2.4.2 Operations Are Automatically Aligned and Vectorized (Broadcasting)

2.5 Making Changes to Series and DataFrames

2.5.1 Add Additional Columns

2.5.2 Directly Change a Column

2.5.3 Dropping Values

2.6 Exporting and Importing Data

2.6.1 pickle

2.6.2 CSV

2.6.3 Excel

2.6.4 Feather Format to Interface With R

2.6.5 Other Data Output Types

2.7 Conclusion

3 Introduction to Plotting

3.1 Introduction

3.2 Matplotlib

3.3 Statistical Graphics Using matplotlib

3.3.1 Univariate

3.3.2 Bivariate

3.3.3 Multivariate Data

3.4 Seaborn

3.4.1 Univariate

3.4.2 Bivariate Data

Table of Contents

3.4.3 Multivariate Data

3.5 Pandas Objects

3.5.1 Histograms

3.5.2 Density Plot

3.5.3 Scatterplot

3.5.4 Hexbin Plot

3.5.5 Boxplot

3.6 Seaborn Themes and Styles

3.7 Conclusion

II: Data Manipulation

4 Data Assembly

4.1 Introduction

4.2 Tidy Data

4.2.1 Combining Data Sets

4.3 Concatenation

4.3.1 Adding Rows

4.3.2 Adding Columns

4.3.3 Concatenation With Different Indices

4.4 Merging Multiple Data Sets

4.4.1 One-to-One Merge

4.4.2 Many-to-One Merge

4.4.3 Many-to-Many Merge

4.5 Conclusion

5 Missing Data

5.1 Introduction

5.2 What Is a NaN Value?

5.3 Where Do Missing Values Come From?

5.3.1 Load Data

5.3.2 Merged Data

5.3.3 User Input Values

5.3.4 Re-indexing

Table of Contents

5.4 Working With Missing Data

5.4.1 Find and Count missing Data

5.4.2 Cleaning Missing Data

5.4.3 Calculations With Missing Data

5.5 Conclusion

6 Tidy Data

6.1 Introduction

6.2 Columns Contain Values, Not Variables

6.2.1 Keep One Column Fixed

6.2.2 Keep Multiple Columns Fixed

6.3 Columns Contain Multiple Variables

6.3.1 Split and Add Columns Individually (Simple Method)

6.3.2 Split and Combine in a Single Step (Simple Method)

6.3.3 Split and Combine in a Single Step (More Complicated Method)

6.4 Variables in Both Rows and Columns

6.5 Multiple Observational Units in a Table (Normalization)

6.6 Observational Units Across Multiple Tables

6.6.1 Load Multiple Files Using a Loop

6.6.2 Load Multiple Files Using a List Comprehension

6.7 Conclusion

III: Data Munging

7 Data Types

7.1 Introduction

7.2 Data Types

7.3 Converting Types

7.3.1 Converting to String Objects

7.3.2 Converting to Numeric Values

7.4 Categorical Data

7.4.1 Convert to Category

7.4.2 Manipulating Categorical Data

7.5 Conclusion

Table of Contents

8 Strings and Text Data

8.1 Introduction

8.2 Strings

8.2.1 Subsetting and Slicing Strings

8.2.2 Getting the Last Character in a String

8.3 String Methods

8.4 More String Methods

8.4.1 Join

8.4.2 Splitlines

8.5 String Formatting

8.5.1 Custom String Formatting

8.5.2 Formatting Character Strings

8.5.3 Formatting Numbers

8.5.4 C printf Style Formatting

8.5.5 Formatted Literal Strings in Python 3.6+

8.6 Regular Expressions (RegEx)

8.6.1 Match a Pattern

8.6.2 Find a Pattern

8.6.3 Substituting a Pattern

8.6.4 Compiling a Pattern

8.7 The regex Library

8.8 Conclusion

9 Apply

9.1 Introduction

9.2 Functions

9.3 Apply (Basics)

9.3.1 Apply Over a Series

9.3.2 Apply Over a DataFrame

9.4 Apply (More Advanced)

9.4.1 Column-wise Operations

9.4.2 Row-wise Operations

Table of Contents

9.5 Vectorized Functions

9.5.1 Using numpy

9.5.2 Using numba

9.6 Lambda Functions

9.7 Conclusion

10 Groupby Operations: SplitApplyCombine

10.1 Introduction

10.2 Aggregate

10.2.1 Basic One-Variable Grouped Aggregation

10.2.2 Built-in Aggregation Methods

10.2.3 Aggregation Functions

10.2.4 Multiple Functions Simultaneously

10.2.5 Using a dict in agg/aggregate

10.3 Transform

10.3.1 z-Score Example

10.4 Filter

10.5 The pandas.core.groupby .DataFrameGroupBy Object

10.5.1 Groups

10.5.2 Group Calculations Involving Multiple Variables

10.5.3 Selecting a Group

10.5.4 Iterating Through Groups

10.5.5 Multiple Groups

10.5.6 Flattening the Results

10.6 Working With a MultiIndex

10.7 Conclusion

11 The datetime Data Type

11.1 Introduction

11.2 Python's datetime Object

11.3 Converting to datetime

11.4 Loading Data That Include Dates

11.5 Extracting Date Components

Table of Contents

11.6 Date Calculations and Timedeltas

11.7 Datetime Methods

11.8 Getting Stock Data

11.9 Subsetting Data Based on Dates

11.9.1 The DatetimeIndex Object

11.9.2 The TimedeltaIndex Object

11.10 Date Ranges

11.10.1 Frequencies

11.10.2 Offsets

11.11 Shifting Values

11.12 Resampling

11.13 Time Zones

11.14 Conclusion

IV: Data Modeling

12 Linear Models

12.1 Introduction

12.2 Simple Linear Regression

12.2.1 Using statsmodels

12.2.2 Using sklearn

12.3 Multiple Regression

12.3.1 Using statsmodels

12.3.2 Using statsmodels With Categorical Variables

12.3.3 Using sklearn

12.3.4 Using sklearn With Categorical Variables

12.4 Keeping Index Labels From sklearn

12.5 Conclusion

13 Generalized Linear Models

13.1 Introduction

13.2 Logistic Regression

13.2.1 Using Statsmodels

13.2.2 Using Sklearn

Table of Contents

13.3 Poisson Regression

13.3.1 Using Statsmodels

13.3.2 Negative Binomial Regression for Overdispersion

13.4 More Generalized Linear Models

13.5 Survival Analysis

13.5.1 Testing the Cox Model Assumptions

13.6 Conclusion

14 Model Diagnostics

14.1 Introduction

14.2 Residuals

14.2.1 Q-Q Plots

14.3 Comparing Multiple Models

14.3.1 Working With Linear Models

14.3.2 Working With GLM Models

14.4 k-Fold Cross-Validation

14.5 Conclusion

15 Regularization

15.1 Introduction

15.2 Why Regularize?

15.3 LASSO Regression

15.4 Ridge Regression

15.5 Elastic Net

15.6 Cross-Validation

15.7 Conclusion

16 Clustering

16.1 Introduction

16.2 k-Means

16.2.1 Dimension Reduction With PCA

16.3 Hierarchical Clustering

16.3.1 Complete Clustering

Table of Contents

- 16.3.2 Single Clustering
- 16.3.3 Average Clustering
- 16.3.4 Centroid Clustering
- 16.3.5 Manually Setting the Threshold

16.4 Conclusion

V: Conclusion

17 Life Outside of Pandas

17.1 The (Scientific) Computing Stack

17.2 Performance

17.2.1 Timing Your Code

17.2.2 Profiling Your Code

17.3 Going Bigger and Faster

18 Toward a Self-Directed Learner

18.1 Its Dangerous to Go Alone!

18.2 Local Meetups

18.3 Conferences

18.4 The Internet

18.5 Podcasts

18.6 Conclusion

VI: Appendixes

A Installation

A.1 Installing Anaconda

A.1.1 Windows

A.1.2 Mac

A.1.3 Linux

A.2 Uninstall Anaconda

B Command Line

B.1 Installation

B.1.1 Windows

B.1.2 Mac

Table of Contents

B.1.3 Linux

B.2 Basics

C Project Templates

D Using Python

D.1 Command Line and Text Editor

D.2 Python and IPython

D.3 Jupyter

D.4 Integrated Development Environments (IDEs)

E Working Directories

F Environments

G Install Packages

G.1 Updating Packages

H Importing Libraries

I Lists

J Tuples

K Dictionaries

L Slicing Values

M Loops

N Comprehensions

O Functions

O.1 Default Parameters

O.2 Arbitrary Parameters

O.2.1 *args

O.2.2 **kwargs

P Ranges and Generators

Q Multiple Assignment

R numpy ndarray

S Classes

Table of Contents

T Odo: The Shapeshifter

Index