

ADDISON-WESLEY DATA & ANALYTICS SERIES

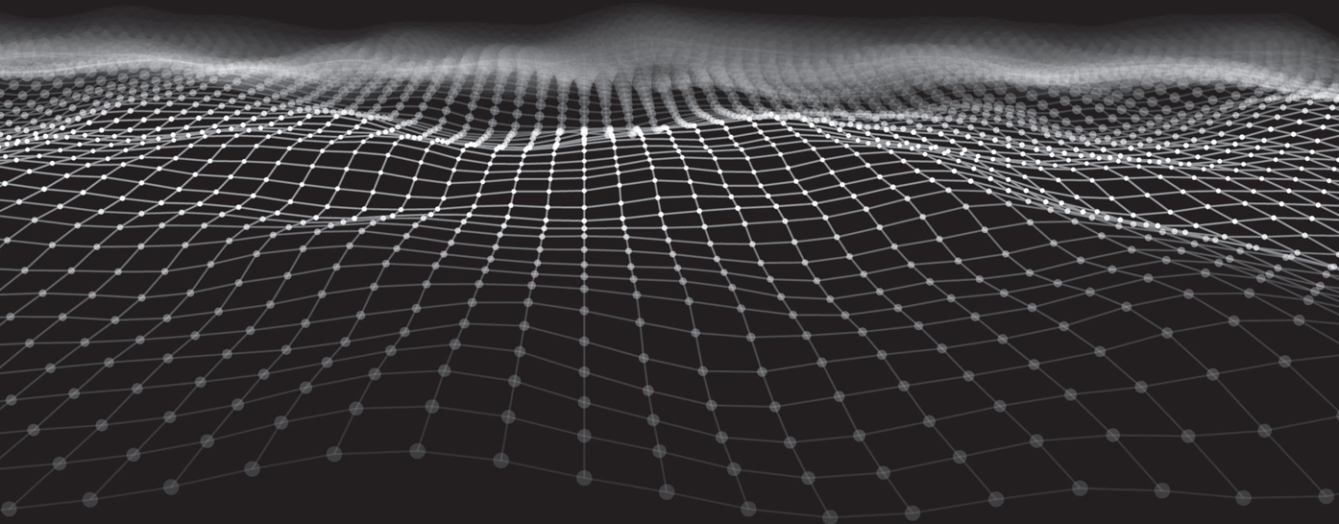


PANDAS

FOR EVERYONE

PYTHON DATA ANALYSIS

— SECOND EDITION —



DANIEL Y. CHEN

Pandas for Everyone

Pandas for Everyone: Python Data Analysis

Table of Contents

Cover

Half Title

Title Page

Copyright Page

Contents

Foreword to Second Edition

Foreword to First Edition

Preface

Breakdown of the Book

Part I

Part II

Part III

Part IV

Part V

Appendices

How to Read This Book

Newcomers

Fluent Python Programmers

Instructors

Setup

Get the Data

Set up Python

Table of Contents

Feedback, Please!

Acknowledgments

About the Author

Changes in the Second Edition

Part I: Introduction

1 Pandas DataFrame Basics

1.1 Introduction

Learning Objectives

1.2 Load Your First Data Set

1.3 Look at Columns, Rows, and Cells

1.3.1 Select and Subset Columns by Name

1.3.2 Subset Rows

1.3.3 Subset Rows by Row Number: `.iloc[]`

1.3.4 Mix It Up

1.3.5 Subsetting Rows and Columns

1.4 Grouped and Aggregated Calculations

1.4.1 Grouped Means

1.4.2 Grouped Frequency Counts

1.5 Basic Plot

Conclusion

2 Pandas Data Structures Basics

Learning Objectives

2.1 Create Your Own Data

2.1.1 Create a Series

2.1.2 Create a DataFrame

2.2 The Series

2.2.1 The Series Is ndarray-like

2.2.2 Boolean Subsetting: Series

2.2.3 Operations Are Automatically Aligned and Vectorized (Broadcasting)

2.3 The DataFrame

Table of Contents

2.3.1 Parts of a DataFrame

2.3.2 Boolean Subsetting: DataFrames

2.3.3 Operations Are Automatically Aligned and Vectorized (Broadcasting)

2.4 Making Changes to Series and DataFrames

2.4.1 Add Additional Columns

2.4.2 Directly Change a Column

2.4.3 Modifying Columns with .assign()

2.4.4 Dropping Values

2.5 Exporting and Importing Data

2.5.1 Pickle

2.5.2 Comma-Separated Values (CSV)

2.5.3 Excel

2.5.4 Feather

2.5.5 Arrow

2.5.6 Dictionary

2.5.7 JSON (JavaScript Object Notation)

2.5.8 Other DataOutput Types

Conclusion

3 Plotting Basics

Learning Objectives

3.1 Why Visualize Data?

3.2 Matplotlib Basics

3.2.1 Figure Objects and Axes Subplots

3.2.2 Anatomy of a Figure

3.3 Statistical Graphics Using matplotlib

3.3.1 Univariate (Single Variable)

3.3.2 Bivariate (Two Variables)

3.3.3 Multivariate Data

3.4 Seaborn

3.4.1 Univariate

3.4.2 Bivariate Data

3.4.3 Multivariate Data

Table of Contents

- 3.4.4 Facets
- 3.4.5 Seaborn Styles and Themes
- 3.4.6 How to GoThrough Seaborn Documentation
- 3.4.7 Next-Generation Seaborn Interface

3.5 Pandas Plotting Method

- 3.5.1 Histogram
- 3.5.2 Density Plot
- 3.5.3 Scatter Plot
- 3.5.4 Hexbin Plot
- 3.5.5 Box Plot

Conclusion

4 Tidy Data

Learning Objectives

Note About This Chapter

4.1 Columns Contain Values, Not Variables

- 4.1.1 Keep One Column Fixed
- 4.1.2 Keep Multiple Columns Fixed

4.2 Columns Contain Multiple Variables

- 4.2.1 Splitand Add Columns Individually
- 4.2.2 Split and Combine in a Single Step

4.3 Variables in Both Rows and Columns

Conclusion

5 Apply Functions

Learning Objectives

Note About This Chapter

5.1 Primer on Functions

5.2 Apply (Basics)

- 5.2.1 Apply Over a Series
- 5.2.2 Apply Over a DataFrame

5.3 Vectorized Functions

- 5.3.1 Vectorize with NumPy

Table of Contents

5.3.2 Vectorize with Numba

5.4 Lambda Functions (Anonymous Functions)

Conclusion

Part II: Data Processing

6 Data Assembly

Learning Objectives

6.1 Combine Data Sets

6.2 Concatenation

6.2.1 Review Parts of a DataFrame

6.2.2 Add Rows

6.2.3 Add Columns

6.2.4 Concatenate with Different Indices

6.3 Observational Units Across Multiple Tables

6.3.1 Load Multiple Files Using a Loop

6.3.2 Load Multiple Files Using a List Comprehension

6.4 Merge Multiple Data Sets

6.4.1 One-to-One Merge

6.4.2 Many-to-One Merge

6.4.3 Many-to-Many Merge

6.4.4 Check Your Work with Assert

Conclusion

7 Data Normalization

Learning Objectives

7.1 Multiple Observational Units in a Table (Normalization)

Conclusion

8 Group by Operations: Split-Apply-Combine

Learning Objectives

8.1 Aggregate

8.1.1 Basic One-Variable Grouped Aggregation

8.1.2 Built-In Aggregation Methods

8.1.3 Aggregation Functions

Table of Contents

8.1.4 Multiple Functions Simultaneously

8.1.5 Use a dict in.agg()/ .aggregate()

8.2 Transform

8.2.1 Z-Score Example

8.2.2 Missing Value Example

8.3 Filter

8.4 The pandas . core . groupby . DataFrameGroupBy object

8.4.1 Groups

8.4.2 Group Calculations Involving Multiple Variables

8.4.3 Selecting a Group

8.4.4 Iterating Through Groups

8.4.5 Multiple Groups

8.4.6 Flattening the Results (.reset_index())

8.5 Working With a MultiIndex

Conclusion

Part III: Data Types

9 Missing Data

Learning Objectives

9.1 What Is a NaN Value?

9.2 Where Do Missing Values Come From?

9.2.1 Load Data

9.2.2 Merged Data

9.2.3 User Input Values

9.2.4 Reindexing

9.3 Working With Missing Data

9.3.1 Find and Count Missing Data

9.3.2 Clean Missing Data

9.3.3 Calculations With Missing Data

9.4 Pandas Built-In NA Missing

Conclusion

10 Data Types

Table of Contents

Learning Objectives

10.1 Data Types

10.2 Converting Types

10.2.1 Converting to String Objects

10.2.2 Converting to Numeric Values

10.3 Categorical Data

10.3.1 Convert to Category

10.3.2 Manipulating Categorical Data

Conclusion

11 Strings and Text Data

Introduction

Learning Objectives

11.1 Strings

11.1.1 Subset and Slice Strings

11.1.2 Get the Last Character in a String

11.2 String Methods

11.3 More String Methods

11.3.1 Join

11.3.2 Splitlines

11.4 String Formatting (F-Strings)

11.4.1 Formatting Numbers

11.5 Regular Expressions (RegEx)

11.5.1 Match a Pattern

11.5.2 Remember What Your RegEx Patterns Are

11.5.3 Find a Pattern

11.5.4 Substitute a Pattern

11.5.5 Compile a Pattern

11.6 The regex Library

Conclusion

12 Dates and Times

Learning Objectives

Table of Contents

- 12.1 Python's datetime Object
- 12.2 Converting to datetime
- 12.3 Loading Data That Include Dates
- 12.4 Extracting Date Components
- 12.5 Date Calculations and Timedeltas
- 12.6 Datetime Methods
- 12.7 Getting Stock Data
- 12.8 Subsetting Data Based on Dates
 - 12.8.1 The DatetimeIndex Object
 - 12.8.2 The TimedeltaIndex Object
- 12.9 Date Ranges
 - 12.9.1 Frequencies
 - 12.9.2 Offsets
- 12.10 Shifting Values
- 12.11 Resampling
- 12.12 Time Zones
- 12.13 Arrow for Better Dates and Times
- Conclusion

Part IV: Data Modeling

- 13 Linear Regression (Continuous Outcome Variable)
 - 13.1 Simple Linear Regression
 - 13.1.1 With statsmodels
 - 13.1.2 With scikit-learn
 - 13.2 Multiple Regression
 - 13.2.1 With statsmodels
 - 13.2.2 With scikit-learn
 - 13.3 Models with Categorical Variables
 - 13.3.1 Categorical Variables in statsmodels
 - 13.3.2 Categorical Variables in scikit-learn
 - 13.4 One-Hot Encoding in scikit-learn with Transformer Pipelines
 - Conclusion

Table of Contents

14 Generalized Linear Models

About This Chapter

14.1 Logistic Regression (Binary Outcome Variable)

14.1.1 With statsmodels

14.1.2 With sklearn

14.1.3 Be Careful of scikit-learn Defaults

14.2 Poisson Regression(Count Outcome Variable)

14.2.1 With statsmodels

14.2.2 Negative Binomial Regression for Overdispersion

14.3 More Generalized Linear Models

Conclusion

15 Survival Analysis

15.1 Survival Data

15.2 Kaplan Meier Curves

15.3 Cox Proportional Hazard Model

15.3.1 Testing the Cox Model Assumptions

Conclusion

16 Model Diagnostics

16.1 Residuals

16.1.1 Q-Q Plots

16.2 Comparing Multiple Models

16.2.1 Working with Linear Models

16.2.2 Working with GLM Models

16.3 k-Fold Cross-Validation

Conclusion

17 Regularization

17.1 Why Regularize?

17.2 LASSO Regression

17.3 Ridge Regression

17.4 Elastic Net

Table of Contents

17.5 Cross-Validation

Conclusion

18 Clustering

18.1 k-Means

18.1.1 Dimension Reduction with PCA

18.2 Hierarchical Clustering

18.2.1 Complete Clustering

18.2.2 Single Clustering

18.2.3 Average Clustering

18.2.4 Centroid Clustering

18.2.5 Ward Clustering

18.2.6 Manually Setting the Threshold

Conclusion

Part V: Conclusion

19 Life Outside of Pandas

19.1 The (Scientific) Computing Stack

19.2 Performance

19.2.1 Timing Your Code

19.2.2 Profiling Your Code

19.2.3 Concurrent Futures

19.3 Dask

19.4 Siuba

19.5 Ibis

19.6 Polars

19.7 PyJanitor

19.8 Pandera

19.9 Machine Learning

19.10 Publishing

19.11 Dashboards

Conclusion

Table of Contents

20 It's Dangerous To Go Alone!

20.1 Local Meetups

20.2 Conferences

20.3 The Carpentries

20.4 Podcasts

20.5 Other Resources

Conclusion

Appendices

A Concept Maps

B Installation and Setup

B.1 Install Python

B.1.1 Anaconda

B.1.2 Miniconda

B.1.3 Uninstall Anaconda or Miniconda

B.1.4 Pyenv

B.2 Install Python Packages

B.3 Download Book Data

C Command Line

C.1 Installation

C.1.1 Windows

C.1.2 Mac

C.1.3 Linux

C.2 Basics

D Project Templates

E Using Python

E.1 Command Line and Text Editor

E.2 Python and IPython

E.3 Jupyter

E.4 Integrated Development Environments (IDEs)

Table of Contents

F Working Directories

G Environments

G.1 Conda Environments

G.2 Pyenv+Pipenv

H Install Packages

H.1 Updating Packages

I Importing Libraries

J Code Style

J.1 Line Breaks in Code

K Containers: Lists, Tuples, and Dictionaries

K.1 Lists

K.2 Tuples

K.3 Dictionaries

L Slice 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 Numpyndarray

S Classes

T SettingWithCopyWarning

T.1 Modifying a Subset of Data

T.2 Replacing a Value

Table of Contents

T.3 MoreResources

U MethodChaining

V TimingCode

W StringFormatting

W.1 C-Style

W.2 StringFormatting:.format() Method

W.3 FormattingNumbers

X Conditionals(if-elif-else)

Y NewYorkACSLogisticRegression Example

Y.0.1 Withsklearn

Z ReplicatingResultsinR

Z.1 LinearRegression

Z.2 LogisticRegression

Z.3 PoissonRegression

Z.3.1 NegativeBinomial Regressionfor Overdispersion

Index