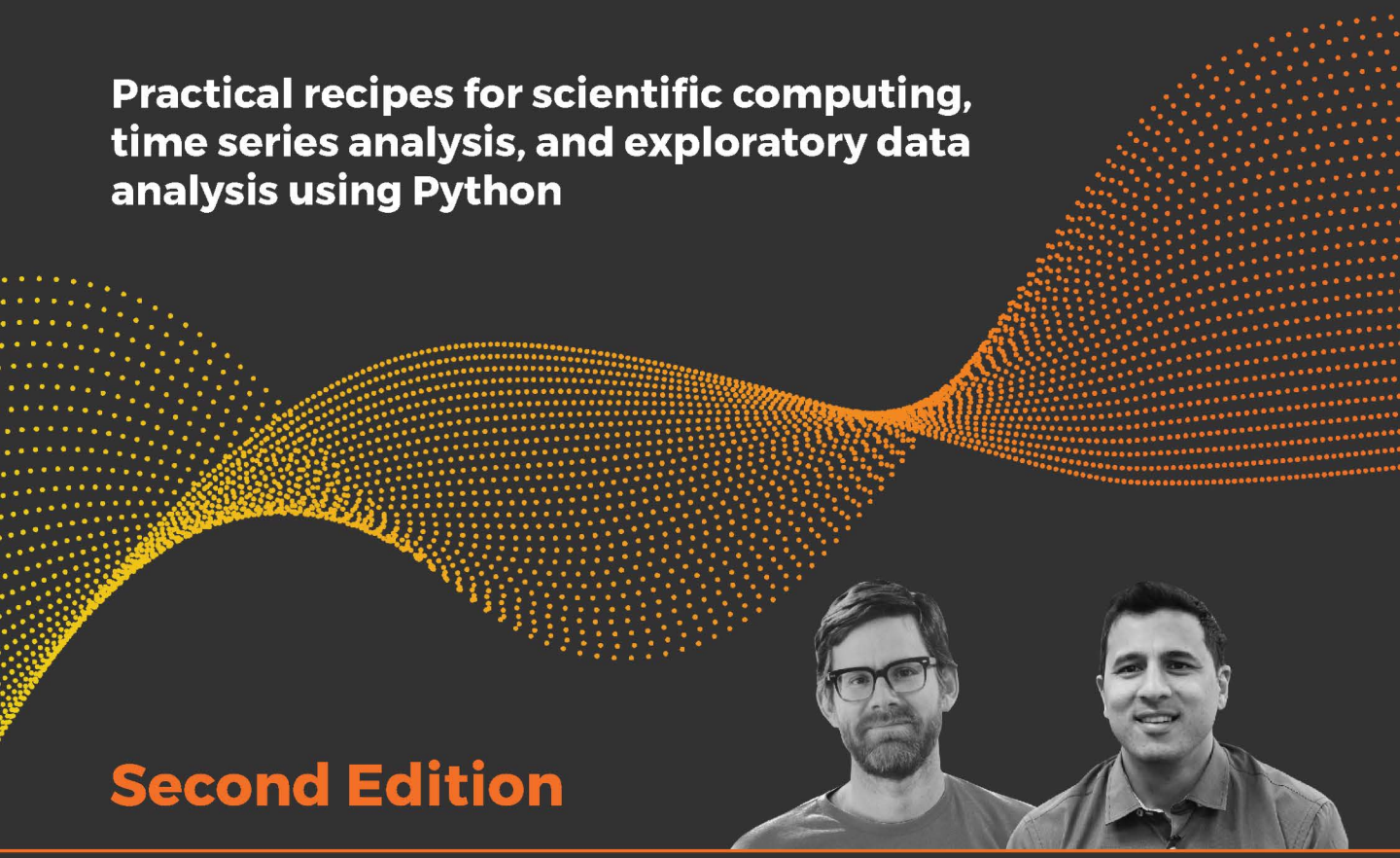
EXPERT INSIGHT

Pandas 1.x Cookbook

Practical recipes for scientific computing, time series analysis, and exploratory data analysis using Python



Second Edition

Matt Harrison
Theodore Petrou

Packt

Pandas 1.x Cookbook

Second Edition

Practical recipes for scientific computing, time series analysis, and exploratory data analysis using Python

Matt Harrison
Theodore Petrou

Packt

BIRMINGHAM - MUMBAI

Pandas 1.x Cookbook

Second Edition

Producer: Tushar Gupta
Acquisition Editor – Peer Reviews: Suresh Jain
Content Development Editor: Kate Blackham
Technical Editor: Gaurav Gavas
Project Editor: Kishor Rit
Proofreader: Safis Editing
Indexer: Pratik Shirodkar
Presentation Designer: Sandip Tadge

First published: October 2017
Second edition: February 2020

Production reference: 1260220

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-83921-310-6

www.packt.com

Packt.com

Subscribe to our online digital library for full access to over 7,000 books and videos, as well as industry leading tools to help you plan your personal development and advance your career. For more information, please visit our website.

Why subscribe?

- Spend less time learning and more time coding with practical eBooks and Videos from over 4,000 industry professionals
- Learn better with Skill Plans built especially for you
- Get a free eBook or video every month
- Fully searchable for easy access to vital information
- Copy and paste, print, and bookmark content

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.Packt.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at customercare@packtpub.com for more details.

At www.Packt.com, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on Packt books and eBooks.

Contributors

About the authors

Matt Harrison has been using Python since 2000. He runs MetaSnake, which provides corporate training for Python and Data Science.

He is the author of *Machine Learning Pocket Reference*, the best-selling *Illustrated Guide to Python 3*, and *Learning the Pandas Library*, as well as other books.

Theodore Petrou is a data scientist and the founder of Dunder Data, a professional educational company focusing on exploratory data analysis. He is also the head of Houston Data Science, a meetup group with more than 2,000 members that has the primary goal of getting local data enthusiasts together in the same room to practice data science. Before founding Dunder Data, Ted was a data scientist at Schlumberger, a large oil services company, where he spent the vast majority of his time exploring data.

Some of his projects included using targeted sentiment analysis to discover the root cause of part failure from engineer text, developing customized client/server dashboarding applications, and real-time web services to avoid the mispricing of sales items. Ted received his masters degree in statistics from Rice University, and used his analytical skills to play poker professionally and teach math before becoming a data scientist. Ted is a strong supporter of learning through practice and can often be found answering questions about pandas on Stack Overflow.

About the reviewer

Simon Hawkins holds a master's degree in aeronautical engineering from Imperial College London. During the early part of his career, he worked exclusively in the defense and nuclear sectors as a technology analyst focusing on various modelling capabilities and simulation techniques for high-integrity equipment. He then transitioned into the world of e-commerce and the focus shifted toward data analysis. Today, he is interested in all things data science and is a member of the pandas core development team.

Table of Contents

Preface

pandas is a library for creating and manipulating structured data with Python. What do I mean by structured? I mean tabular data in rows and columns like what you would find in a spreadsheet or database. Data scientists, analysts, programmers, engineers, and more are leveraging it to mold their data.

pandas is limited to "small data" (data that can fit in memory on a single machine). However, the syntax and operations have been adopted or inspired other projects: PySpark, Dask, Modin, cuDF, Baloo, Dexplo, Tabel, StaticFrame, among others. These projects have different goals, but some of them will scale out to big data. So there is a value in understanding how pandas works as the features are becoming the defacto API for interacting with structured data.

I, Matt Harrison, run a company, MetaSnake, that does corporate training. My bread and butter is training large companies that want to level up on Python and data skills. As such, I've taught thousands of Python and pandas users over the years. My goal in producing the second version of this book is to highlight and help with the aspects that many find confusing when coming to pandas. For all of its benefits, there are some rough edges or confusing aspects of pandas. I intend to navigate you to these and then guide you through them, so you will be able to deal with them in the real world.

If your company is interested in such live training, feel free to reach out (`matt@metasnake.com`).

Who this book is for

This book contains nearly 100 recipes, ranging from very simple to advanced. All recipes strive to be written in clear, concise, and modern idiomatic pandas code. The *How it works...* sections contain extremely detailed descriptions of the intricacies of each step of the recipe. Often, in the *There's more...* section, you will get what may seem like an entirely new recipe. This book is densely packed with an extraordinary amount of pandas code.

As a generalization, the recipes in the first seven chapters tend to be simpler and more focused on the fundamental and essential operations of pandas than the later chapters, which focus on more advanced operations and are more project-driven. Due to the wide range of complexity, this book can be useful to both novice and everyday users alike. It has been my experience that even those who use pandas regularly will not master it without being exposed to idiomatic pandas code. This is somewhat fostered by the breadth that pandas offers. There are almost always multiple ways of completing the same operation, which can have users get the result they want but in a very inefficient manner. It is not uncommon to see an order of magnitude or more in performance difference between two sets of pandas solutions to the same problem.

The only real prerequisite for this book is a fundamental knowledge of Python. It is assumed that the reader is familiar with all the common built-in data containers in Python, such as lists, sets, dictionaries, and tuples.

What this book covers

Chapter 1, *Pandas Foundations*, covers the anatomy and vocabulary used to identify the components of the two main pandas data structures, the Series and the DataFrame. Each column must have exactly one type of data, and each of these data types is covered. You will learn how to unleash the power of the Series and the DataFrame by calling and chaining together their methods.

Chapter 2, *Essential DataFrame Operations*, focuses on the most crucial and typical operations that you will perform during data analysis.

Chapter 3, *Creating and Persisting DataFrames*, discusses the various ways to ingest data and create DataFrames.

Chapter 4, *Beginning Data Analysis*, helps you develop a routine to get started after reading in your data.

Chapter 5, *Exploratory Data Analysis*, covers basic analysis techniques for comparing numeric and categorical data. This chapter will also demonstrate common visualization techniques.

Chapter 6, *Selecting Subsets of Data*, covers the many varied and potentially confusing ways of selecting different subsets of data.

Chapter 7, *Filtering Rows*, covers the process of querying your data to select subsets of it based on Boolean conditions.

Chapter 8, *Index Alignment*, targets the very important and often misunderstood index object. Misuse of the Index is responsible for lots of erroneous results, and these recipes show you how to use it correctly to deliver powerful results.

Chapter 9, *Grouping for Aggregation, Filtration, and Transformation*, covers the powerful grouping capabilities that are almost always necessary during data analysis. You will build customized functions to apply to your groups.

Chapter 10, *Restructuring Data into a Tidy Form*, explains what tidy data is and why it's so important, and then it shows you how to transform many different forms of messy datasets into tidy ones.

Chapter 11, *Combining Pandas Objects*, covers the many available methods to combine DataFrames and Series vertically or horizontally. We will also do some web-scraping and connect to a SQL relational database.

Chapter 12, *Time Series Analysis*, covers advanced and powerful time series capabilities to dissect by any dimension of time possible.

Chapter 13, *Visualization with Matplotlib, Pandas, and Seaborn*, introduces the matplotlib library, which is responsible for all of the plotting in pandas. We will then shift focus to the pandas plot method and, finally, to the seaborn library, which is capable of producing aesthetically pleasing visualizations not directly available in pandas.

Chapter 14, *Debugging and Testing Pandas*, explores mechanisms of testing our DataFrames and pandas code. If you are planning on deploying pandas in production, this chapter will help you have confidence in your code.

To get the most out of this book

There are a couple of things you can do to get the most out of this book. First, and most importantly, you should download all the code, which is stored in Jupyter Notebooks. While reading through each recipe, run each step of code in the notebook. Make sure you explore on your own as you run through the code. Second, have the pandas official documentation open (`http://pandas.pydata.org/pandas-docs/stable/`) in one of your browser tabs. The pandas documentation is an excellent resource containing over 1,000 pages of material. There are examples for most of the pandas operations in the documentation, and they will often be directly linked from the *See also* section. While it covers the basics of most operations, it does so with trivial examples and fake data that don't reflect situations that you are likely to encounter when analyzing datasets from the real world.

What you need for this book

pandas is a third-party package for the Python programming language and, as of the printing of this book, is on version 1.0.1. Currently, Python is at version 3.8. The examples in this book should work fine in versions 3.6 and above.

There are a wide variety of ways in which you can install pandas and the rest of the libraries mentioned on your computer, but an easy method is to install the Anaconda distribution. Created by Anaconda, it packages together all the popular libraries for scientific computing in a single downloadable file available on Windows, macOS, and Linux. Visit the download page to get the Anaconda distribution (`https://www.anaconda.com/distribution`).

In addition to all the scientific computing libraries, the Anaconda distribution comes with Jupyter Notebook, which is a browser-based program for developing in Python, among many other languages. All of the recipes for this book were developed inside of a Jupyter Notebook and all of the individual notebooks for each chapter will be available for you to use.

It is possible to install all the necessary libraries for this book without the use of the Anaconda distribution. For those that are interested, visit the pandas installation page (`http://pandas.pydata.org/pandas-docs/stable/install.html`).

Download the example code files

You can download the example code files for this book from your account at `www.packt.com`. If you purchased this book elsewhere, you can visit `www.packtpub.com/support/errata` and register to have the files emailed directly to you.

You can download the code files by following these steps:

1. Log in or register at `www.packt.com`.
2. Select the **Support** tab.
3. Click on **Code Downloads**.
4. Enter the name of the book in the **Search** box and follow the on-screen instructions.

Once the file is downloaded, please make sure that you unzip or extract the folder using the latest version of:

- WinRAR / 7-Zip for Windows
- Zipeg / iZip / UnRarX for Mac
- 7-Zip / PeaZip for Linux

The code bundle for the book is also hosted on GitHub at `https://github.com/PacktPublishing/Pandas-Cookbook-Second-Edition`. In case there's an update to the code, it will be updated on the existing GitHub repository.

We also have other code bundles from our rich catalog of books and videos available at `https://github.com/PacktPublishing/`. Check them out!

Running a Jupyter Notebook

The suggested method to work through the content of this book is to have a Jupyter Notebook up and running so that you can run the code while reading through the recipes. Following along on your computer allows you to go off exploring on your own and gain a deeper understanding than by just reading the book alone.

Assuming that you have installed the Anaconda distribution on your machine, you have two options available to start the Jupyter Notebook, from the Anaconda GUI or the command line. I highly encourage you to use the command line. If you are going to be doing much with Python, you will need to feel comfortable from there.

After installing Anaconda, open a command prompt (type `cmd` at the search bar on Windows, or open a Terminal on Mac or Linux) and type:

```
$ jupyter-notebook
```

It is not necessary to run this command from your home directory. You can run it from any location, and the contents in the browser will reflect that location.

Although we have now started the Jupyter Notebook program, we haven't actually launched a single individual notebook where we can start developing in Python. To do so, you can click on the New button on the right-hand side of the page, which will drop down a list of all the possible kernels available for you to use. If you just downloaded Anaconda, then you will only have a single kernel available to you (Python 3). After selecting the Python 3 kernel, a new tab will open in the browser, where you can start writing Python code.

You can, of course, open previously created notebooks instead of beginning a new one. To do so, navigate through the filesystem provided in the Jupyter Notebook browser home page and select the notebook you want to open. All Jupyter Notebook files end in `.ipynb`.

Alternatively, you may use cloud providers for a notebook environment. Both Google and Microsoft provide free notebook environments that come preloaded with pandas.

Download the color images

We also provide a PDF file that has color images of the screenshots/diagrams used in this book. You can download it here: `https://static.packt-cdn.com/downloads/9781839213106_ColorImages.pdf`.

Conventions

There are a number of text conventions used throughout this book.

`CodeInText`: Indicates code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles. Here is an example: "You may need to install `xlwt` or `openpyxl` to write XLS or XLSX files respectively."

A block of code is set as follows:

```
import pandas as pd
import numpy as np
movies = pd.read_csv("data/movie.csv")
movies
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
import pandas as pd
import numpy as np
movies = pd.read_csv("data/movie.csv")
movies
```

Any command-line input or output is written as follows:

```
>>> employee = pd.read_csv('data/employee.csv')
>>> max_dept_salary = employee.groupby('DEPARTMENT')['BASE_SALARY'].max()
```

Bold: Indicates a new term, an important word, or words that you see on the screen, for example, in menus or dialog boxes, also appear in the text like this. Here is an example: "Select **System info** from the **Administration** panel."

Warnings or important notes appear like this.

Tips and tricks appear like this.

Assumptions for every recipe

It should be assumed that at the beginning of each recipe pandas, NumPy, and matplotlib are imported into the namespace. For plots to be embedded directly within the notebook, you must also run the magic command `%matplotlib inline`. Also, all data is stored in the `data` directory and is most commonly stored as a CSV file, which can be read directly with the `read_csv` function:

```
>>> %matplotlib inline
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>> import pandas as pd
>>> my_dataframe = pd.read_csv('data/dataset_name.csv')
```

Dataset descriptions

There are about two dozen datasets that are used throughout this book. It can be very helpful to have background information on each dataset as you complete the steps in the recipes. A detailed description of each dataset may be found in the `dataset_descriptions` Jupyter Notebook found at `https://github.com/PacktPublishing/Pandas-Cookbook-Second-Edition`. For each dataset, there will be a list of the columns, information about each column and notes on how the data was procured.

Sections

In this book, you will find several headings that appear frequently.

To give clear instructions on how to complete a recipe, we use these sections as follows:

How to do it...

This section contains the steps required to follow the recipe.

How it works...

This section usually consists of a detailed explanation of what happened in the previous section.