

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



BIRMINGHAM - MUMBAI

Pandas 1.x Cookbook

Second Edition

Copyright © 2020 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the authors, nor Packt Publishing or its dealers and distributors, will be held liable for any damages caused or alleged to have been caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

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	vii
Chapter 1: Pandas Foundations	1
Importing pandas	1
Introduction	1
The pandas DataFrame	2
DataFrame attributes	4
Understanding data types	6
Selecting a column	10
Calling Series methods	14
Series operations	21
Chaining Series methods	27
Renaming column names	32
Creating and deleting columns	36
Chapter 2: Essential DataFrame Operations	45
Introduction	45
Selecting multiple DataFrame columns	45
Selecting columns with methods	48
Ordering column names	52
Summarizing a DataFrame	55
Chaining DataFrame methods	59
DataFrame operations	62
Comparing missing values	67
Transposing the direction of a DataFrame operation	71
Determining college campus diversity	74
Chapter 3: Creating and Persisting DataFrames	81
Introduction	81
Creating DataFrames from scratch	81

Writing CSV	84
Reading large CSV files	86
Using Excel files	95
Working with ZIP files	97
Working with databases	101
Reading JSON	102
Reading HTML tables	106
Chapter 4: Beginning Data Analysis	115
Introduction	115
Developing a data analysis routine	115
Data dictionaries	120
Reducing memory by changing data types	120
Selecting the smallest of the largest	126
Selecting the largest of each group by sorting	128
Replicating nlargest with sort_values	133
Calculating a trailing stop order price	136
Chapter 5: Exploratory Data Analysis	139
Introduction	139
Summary statistics	139
Column types	143
Categorical data	147
Continuous data	156
Comparing continuous values across categories	163
Comparing two continuous columns	169
Comparing categorical and categorical values	178
Using the pandas profiling library	185
Chapter 6: Selecting Subsets of Data	189
Introduction	189
Selecting Series data	189
Selecting DataFrame rows	196
Selecting DataFrame rows and columns simultaneously	200
Selecting data with both integers and labels	203
Slicing lexicographically	205
Chapter 7: Filtering Rows	209
Introduction	209
Calculating Boolean statistics	209
Constructing multiple Boolean conditions	213
Filtering with Boolean arrays	215
Comparing row filtering and index filtering	219

Selecting with unique and sorted indexes	222
Translating SQL WHERE clauses	225
Improving the readability of Boolean indexing with the query method	230
Preserving Series size with the .where method	232
Masking DataFrame rows	237
Selecting with Booleans, integer location, and labels	240
Chapter 8: Index Alignment	245
Introduction	245
Examining the Index object	245
Producing Cartesian products	248
Exploding indexes	251
Filling values with unequal indexes	255
Adding columns from different DataFrames	260
Highlighting the maximum value from each column	266
Replicating idxmax with method chaining	275
Finding the most common maximum of columns	282
Chapter 9: Grouping for Aggregation, Filtration, and Transformation	285
Introduction	285
Defining an aggregation	286
Grouping and aggregating with multiple columns and functions	290
Removing the MultiIndex after grouping	296
Grouping with a custom aggregation function	301
Customizing aggregating functions with *args and **kwargs	305
Examining the groupby object	309
Filtering for states with a minority majority	313
Transforming through a weight loss bet	316
Calculating weighted mean SAT scores per state with apply	325
Grouping by continuous variables	330
Counting the total number of flights between cities	334
Finding the longest streak of on-time flights	339
Chapter 10: Restructuring Data into a Tidy Form	349
Introduction	349
Tidying variable values as column names with stack	351
Tidying variable values as column names with melt	356
Stacking multiple groups of variables simultaneously	359
Inverting stacked data	362
Unstacking after a groupby aggregation	368
Replicating pivot_table with a groupby aggregation	372
Renaming axis levels for easy reshaping	376

Tidying when multiple variables are stored as column names	382
Tidying when multiple variables are stored as a single column	389
Tidying when two or more values are stored in the same cell	394
Tidying when variables are stored in column names and values	398
Chapter 11: Combining Pandas Objects	401
Introduction	401
Appending new rows to DataFrames	401
Concatenating multiple DataFrames together	408
Understanding the differences between concat, join, and merge	411
Connecting to SQL databases	421
Chapter 12: Time Series Analysis	429
Introduction	429
Understanding the difference between Python and pandas date tools	429
Slicing time series intelligently	436
Filtering columns with time data	441
Using methods that only work with a DatetimeIndex	445
Counting the number of weekly crimes	453
Aggregating weekly crime and traffic accidents separately	457
Measuring crime by weekday and year	463
Grouping with anonymous functions with a DatetimeIndex	474
Grouping by a Timestamp and another column	478
Chapter 13: Visualization with Matplotlib, Pandas, and Seaborn	485
Introduction	485
Getting started with matplotlib	486
Object-oriented guide to matplotlib	488
Visualizing data with matplotlib	499
Plotting basics with pandas	507
Visualizing the flights dataset	511
Stacking area charts to discover emerging trends	525
Understanding the differences between seaborn and pandas	530
Multivariate analysis with seaborn Grids	538
Uncovering Simpson's Paradox in the diamonds dataset with seaborn	545
Chapter 14: Debugging and Testing Pandas	553
Code to transform data	553
Apply performance	558
Improving apply performance with Dask, Pandarell, Swifter, and more	561
Inspecting code	564
Debugging in Jupyter	569
Managing data integrity with Great Expectations	573

Using pytest with pandas	582
Generating tests with Hypothesis	587
Other Books You May Enjoy	595
Index	599

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.