Pandas Cookbook:Develop Powerful Routines for Exploring Real World Datasets

Author, Ted Petrou

A Data Science Foundation White Paper

December 2017

----

www.datascience.foundation

Pandas Cookbook - Develop Powerful Routines for Exploring Real-word Datasets

Exploring Real-World Datasets

My name is Ted Petrou , and I am the author of the newly released Pandas Cookbook. In this article, I will discuss the overall approach I took to writing Pandas Cookbook along with highlights of each chapter.

Pandas Cookbook Guiding Principles

I had three main guiding principles when writing the book:

- Use of real-world datasets
- Focus on doing data analysis
- Writing modern, idiomatic pandas

First, I wanted you, the reader, to explore real-world datasets and not randomly generated data. I tried very hard to find datasets that contained situations where an interesting or unique pandas operation could be performed. Descriptions of the main datasets used throughout the book can be found in this Jupyter notebook.

Second, I wanted to focus on doing actual data analysis by providing useful or surprising insights. I wanted to avoid a mechanical approach where pandas operations where learned in isolation or were devoid of contact with real data. In this regard, Pandas Cookbook teaches both, how to understand pandas operations, and how to generate results that would be useful for a data analysis.

Third, the pandas library has evolved quite substantially since it first started to make regular appearances in data analysis workflows in 2012. Many of the older tutorials, and especially the older answers on Stack Overflow have not been upgraded to reflect newer syntax. Pandas is confusing because there are often multiple ways to produce the same result, many of which, will be slow or ineffective. Pandas Cookbook strives to provide straightforward and efficient or 'idiomatic' pandas.

Formation of Pandas Cookbook

Pandas Cookbook was inspired by the following:

- My weeklong Data Exploration Bootcamp
- Answering 400+ questions on pandas on Stack Overflow
- Working as a data scientist at Schlumberger

- Hosting dozens of meetups for [Houston Data Science](#)

My Data Exploration Bootcamp is an intensive, weeklong class with over 700 pages of material, 250 short-answer questions and a couple projects. Much of the material for Pandas Cookbook was inspired by this class. The material was expanded and refined each time the class was taught, thanks in part to the excellent feedback from my students. Teaching showed me first hand, exactly where the greatest pain-points were.

Nothing helped me more to improve my own ability to write idiomatic pandas than answering questions on Stack Overflow. You learn an incredible amount by answering and discussing with the other top users.

As a data scientist at Schlumberger, I built scripts to clean and process data that required the use of dozens of pandas commands pieced together. Pandas Cookbook has many advanced recipes that combine operations from different parts of the library to get the required result. Also, I was given a week's worth of professional Python training, which was quite bad and sparked a desire to produce a better class. You can hear more of my story [in this podcast](#) from Undersampled Radio.

A Book must Beat the Documentation

The [official documentation](#) itself is very thorough and over 2,000 pages in total. In order for a book to be of any value, at a minimum, it must be better than the documentation. There are some major advantages of the documentation over a book. First, there is no restriction on page length, so every single aspect of the library can be covered. Second, the documentation is always up to date with the latest changes. Technical books on fast moving libraries like pandas tend to go out of date relatively fast.

How Pandas Cookbook Demolishes the Documentation

Unfortunately, the pandas documentation does not have interesting examples using real-world datasets. Nearly all of its examples are done using randomly generated or contrived data showing operations in isolation from one another. You learn how to run a single command, independent of all the other available ones. This is not at all how an analysis happens using actual data.

There is certainly lots of value for learning the mechanics of all the pandas operations and I suggest doing that in my [How to Learn Pandas](#) article. In fact, I have read through most parts of the documentation five or more times each. Pandas is a huge library and its difficult to keep all the commands in the forefront of your mind, even if you use it every single day.

Pandas Cookbook uses multiple operations one after the other in many of its recipes. This often yields a long chain of methods called from a DataFrame or Series. This is what makes Pandas Cookbook valuable — you are constantly working with real data, stringing together multiple pandas operations to complete a particular task.

Cookbook Format

It's a bit unfortunate/ludicrous that the title of the book sounds appalling for those not in the know. I suggest keeping this book in the kitchen next to your other cookbooks for some guaranteed extra laughs.

The book is composed of approximately 100 recipes, with each one containing three major sections:

- How to do it: Step-by-step code on how to complete a particular task. There are some explanations embedded into the steps themselves.
- How it works: Very detailed explanations of all the steps in the recipe. I read lots of reviews of other Packt cookbooks, and the most common complaint was the lack of explanations in this section. I took extra care to ensure that all steps and commands were fully explained.
- There's more: Extra operations, closely related to the main recipe. There are almost always tangents that you can follow when learning pandas. This section is often equivalent to an entire new recipe.

Entire Focus on Pandas

This book makes one basic assumption — that you are comfortable with the fundamentals of Python. Every single recipe (except one or two), uses pandas. Thus, the scope of the book is a bit narrower than other similar books, in that it only focuses on doing data analysis with pandas (along with matplotlib and seaborn for visualization).

Target Audience

There is no hard requirement for having any prior exposure to pandas. The recipes range from very simple to advanced, so the book is suitable for novices as well as experienced pandas users.

Getting the Most out of Pandas Cookbook

To get the most out of Pandas Cookbook, I suggest doing the following:

- Keep the official documentation open at all times
- Run the code in the Jupyter notebooks as you read the book
- Read the book sequentially, cover to cover

Pandas Cookbook strives hard to differentiate itself from the documentation. This doesn't mean it is a replacement for the documentation. Most recipes link to a specific part of the documentation, where you can get more details on a specific command. This is why I recommend keeping the documentation open as you progress through the book. Do not just read the book. Run the code as you read through each recipe. You should be doing lots of exploration and formulating questions on your own.
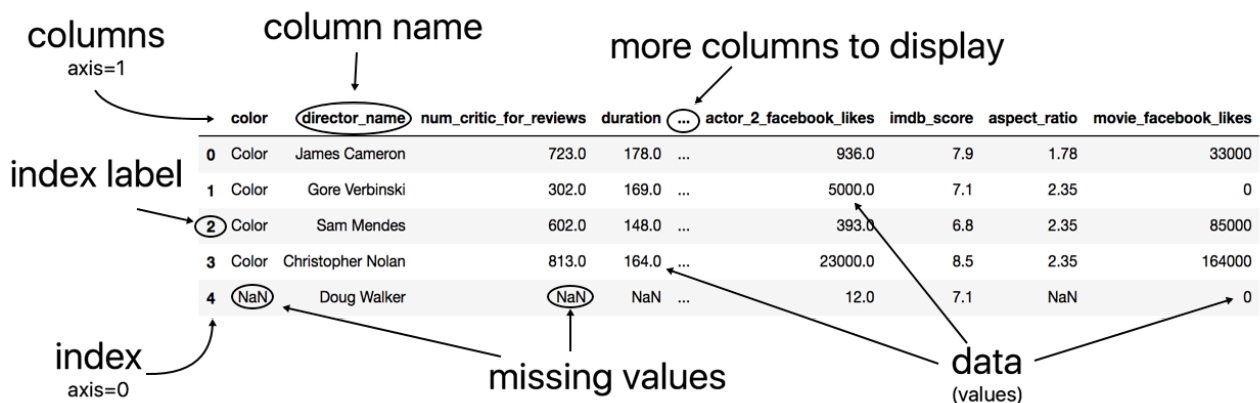
I also recommend reading this book sequentially. I recommend this whether you are a novice or an experienced pandas user. The recipes have a natural flow that progress from one to the next and tend to get more and more complex. More experienced users, of course, can skip around to recipes that appeal to them more. But, I've found that, unless you are a power user of pandas, it will still be good to drill the fundamentals, which is done by reading the book sequentially.

Chapter Highlights

Below, I discuss a few of the more important concepts and recipes of each chapter

Chapter 1: Pandas Foundations

Chapter 1 begins by dissecting the anatomy of the DataFrame and Series, the primary objects that will handle the bulk of your workload. It's vital to be aware of the DataFrame components — the index, the columns and the data (values).

The chapter continues by selecting a single column from a DataFrame as a Series. We use this Series to learn about method chaining which is an extremely common way to use pandas. The majority of the recipes in the book string together multiple methods in succession like this.

Chapter 2: Essential DataFrame Operations

Chapter 2 focuses entirely on the DataFrame. We learn how to order columns sensibly, which is a commonly overlooked task and can greatly help improve readability of the data. As a practical and fun example, we determine the persity of college campuses by using many of the concepts covered up to this point.

Chapter 3: Beginning Data Analysis

Chapter 3 covers several fairly simple but complete tasks that you might do when first starting an analysis. It can be immensely helpful to establish a routine at the beginning of a data analysis. Another recipe finds the largest/smallest value in column 'x' for every unique value in column 'y' without using a call to the groupby method. This is an example of one popular idiom that has arisen more recently.

Chapter 4: Selecting Subsets of Data

Chapter 4 selects subsets of DataFrames and Series in just about every way imaginable. Data selection is one of the most confusing aspects of the library, which is unfortunate, as it's used very frequently. Pandas

---

is partially to blame here, as indexing changed with the addition of the .loc/.iloc indexers along with the recent deprecation of .ix.

Chapter 5: Boolean Indexing

Chapter 5 covers boolean indexing, which is used to select subsets of data by the actual content of the columns and not by their label or integer location(as in chapter 4). One common theme throughout Pandas Cookbook is the comparison between different methods that produce the same results. In one recipe in this chapter, we show how boolean indexing can be replicated by placing columns into the index. For those familiar with SQL, boolean indexing is also compared to the WHERE clause.

Chapter 6: Index Alignment

All of chapter 6 is dedicated to one of the most powerful, but unexpected, feature of pandas, automatic alignment of each index. Some users can spend years using pandas without even understanding this concept. Automatic index alignment is what separates pandas from most other data analysis libraries. An absurd example is the 'Exploding Indexes' recipe, which is used to hammer-home exactly what happens when combining multiple pandas objects.

Chapter 7: Grouping for Aggregation, Filtration, and Transformation

The first 6 chapters cover the most fundamental parts of pandas in 200 pages. The remaining 5 chapters, and 300 pages, use these fundamentals in just about every recipe to do more complex and interesting analysis. The groupby method in this chapter is particularly helpful for splitting data into independent groups. One particularly fun recipe uses the transform method to calculate the results of a weight-loss bet. Also, one of the most complex recipes resides in this chapter, and finds the streaks of on-time flights for each airline.

Chapter 8: Restructuring Data into a Tidy Form

Data analysis is made easier when you have tidy data , a term popularized by Hadley Wickham. Chapter 8 transforms many different formats of messy data into tidy data with the following methods: stack, unstack, melt, and pivot. You will also be exposed to the str accessor, which is used to rip apart string data to extract new variables. Chapter 8 is probably the most unique chapter in this book, as I have not seen much discussion online on how to tidy the vast assortment of datasets as done in this chapter.

Chapter 9: Combining Pandas Objects

There are four primary methods/functions that are used to combine DataFrames/Series together: append, concat, merge and join. This chapter provides examples that are suited for each. "Comparing President Trump's and Obama's approval ratings" is one of my favorite recipes, which does intricate web-scraping, moving windows analysis and visualization all in one. This chapter also connects to a relational database with multiple tables to perform an analysis one might normally do with SQL.

Chapter 10: Time Series Analysis

Pandas has powerful time series functionality that exceeds that from the datetime and NumPy libraries. You will learn how to group simultaneously by time and another variable. Also, one of the newest additions to pandas, the merge_asof function, will be used to find the last time crime was 20% lower.

Chapter 11: Visualization with Matplotlib, Pandas, and Seaborn

One of the most infuriating and confusing things about matplotlib is its dual interface. In my opinion, all matplotlib should be written with the object-oriented interface as it's more Pythonic. Pandas Cookbook thoroughly covers how to get started with the object-oriented interface along with the Figure/Axes hierarchy which is key to understanding all of plotting in matplotlib. Pandas and seaborn both use matplotlib to make plots in completely different ways. Pandas uses wide or aggregated data while seaborn takes long or tidy data. One particularly useful recipe for data scientists involves "uncovering Simpson's paradox", which is a very common finding that gets revealed whenever you look at more granular slices of your data.

Lots More!

The chapter highlights are just a small sampling of what is contained in the book. I worked extremely hard to make Pandas Cookbook the very best book available for learning pandas while doing analysis with real-world data. I had lots of fun coming up with the recipes and hope you have fun exploring them.

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation