TREADING ON PYTHON SERIES

Learning the Pandas Library

Python Tools for Data Munging, Data Analysis, and Visualization

Matt Harrison

Learning the Pandas Library

Python Tools for Data Munging, Analysis, and Visualization

Matt Harrison

This book is for sale at http://leanpub.com/learningthepandaslibrary

This version was published on 2016-07-24

Contents

Introduction

Important Note

Though the content is here, work is still being done to fix:

- table formatting
- footnotes
- utf-8 data

(I'm working with Leanpub to solve these, sorry for the annoyances)

I have been using Python is some professional capacity since the turn of the century. One of the trends that I have seen in that time is the uptake of Python for various aspects of "data science"- gathering data, cleaning data, analysis, machine learning, and visualization. The pandas library has seen much uptake in this area.

pandas [1] is a data analysis library for Python that has exploded in popularity over the past years. The website describes it thusly:

> "pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language."
>
> -pandas.pydata.org

My description of pandas is: pandas is an in memory nosql database, that has sql-like constructs, basic statistical and analytic support, as well as graphing capability. Because it is built on top of Cython, it has less memory overhead and runs quicker. Many people are using pandas to replace Excel, perform ETL, process tabular data, load CSV or JSON files, and more. Though it grew out of the financial sector (for analysis of time series data), it is now a general purpose data manipulation library.

Because pandas has some lineage back to NumPy, it adopts some NumPy'isms that normal Python programmers may not be aware of or familiar with. Certainly, one could go out and use Cython to perform fast typed data analysis with a Python-like dialect, but with pandas, you don't need to. This work is done for you. If you are using pandas and the vectorized operations, you are getting close to C level speeds, but writing Python.

[1] pandas (http://pandas.pydata.org) refers to itself in lowercase, so this book will follow suit.

Who this book is for

This guide is intended to introduce pandas to Python programmers. It covers many (but not all) aspects, as well as some gotchas or details that may be counter-intuitive or even non-pythonic to longtime users of Python.

This book assumes basic knowledge of Python. The author has written *Treading on Python Vol 1* [2] that provides all the background necessary.

Data in this Book

Some might complain that the datasets in this book are small. That is true, and in some cases (as in plotting a histogram), that is a drawback. On the other hand, every attempt has been made to have real data that illustrates using pandas and the features found in it. As a visual learner, I appreciate seeing where data is coming and going. As such, I try to shy away from just showing tables of random numbers that have no meaning.

Hints, Tables, and Images

The hints, tables, and graphics found in this book, have been collected over almost five years of using pandas. They are derived from hangups, notes, and cheatsheets that I have developed after using pandas and teaching others how to use it. Hopefully, they are useful to you as well.

In the physical version of this book, is an index that has also been battle-tested during development. Inevitably, when I was doing analysis not related to the book, I would check that the index had the information I needed. If it didn't, I added it. Let me know if you find any omissions!

Finally, having been around the publishing block and releasing content to the world, I realize that I probably have many omissions that others might consider required knowledge. Many will enjoy the content, others might have the opposite reaction. If you have feedback, or suggestions for improvement, please reach out to me. I love to hear back from readers! Your comments will improve future versions.

[2] http://hairysun.com/books/tread/

Installation

Python 3 has been out for a while now, and people claim it is the future. As an attempt to be modern, this book will use Python 3 throughout! Do not despair, the code will run in Python 2 as well. In fact, review versions of the book neglected to list the Python version, and there was a single complaint about a superfluous `list(range(10))` call. The lone line of (Python 2) code required for compatibility is:

```
>>> from __future__ import print_function
```

Having gotten that out of the way, let's address installation of pandas. The easiest and least painful way to install pandas on most platforms is to use the Anaconda distribution [3]. Anaconda is a meta distribution of Python, that contains many additional packages that have traditionally been annoying to install unless you have toolchains to compile Fortran and C code. Anaconda allows you to skip the compile step and provides binaries for most platforms. The Anaconda distribution itself is freely available, though commercial support is available as well.

After installing the Anaconda package, you should have a `conda` executable. Running:

```
$ conda install pandas
```

Will install pandas and any dependencies. To verify that this works, simply try to import the `pandas` package:

```
$ python
>>> import pandas
>>> pandas.__version__
'0.18.0'
```

If the library successfully imports, you should be good to go.

Other Installation Options

The pandas library will install on Windows, Mac, and Linux via pip [4].

Mac and Windows users wishing to install binaries may download them from the pandas website. Most Linux distributions also have native packages pre-built and available in their repos. On Ubuntu and Debian `apt-get` will install the library:

[3] https://www.continuum.io/downloads

[4] http://pip-installer.org/

```
$ sudo apt-get install python-pandas
```

Pandas can also be installed from source. I feel the need to advise you that you might spend a bit of time going down this rabbit hole if you are not familiar with getting compiler toolchains installed on your system.

It may be necessary to prep the environment for building pandas from source by installing dependencies and the proper header files for Python. On Ubuntu this is straightforward, other environments may be different:

```
$ sudo apt-get install build-essential python-all-dev
```

Using `virtualenv` [^5] will alleviate the need for superuser access during installation. Because `virtualenv` uses pip, it can download and install newer releases of pandas if the version found on the distribution is lagging.

On Mac and Linux platforms, the following create a `virtualenv` sandbox and installs the latest pandas in it (assuming that the prerequisite files are also installed):

```
$ virtualenv pandas-env
$ source pandas-env/bin/activate
$ pip install pandas
```

After a while, pandas should be ready for use. Try to import the library and check the version:

```
$ source pandas-env/bin/activate
$ python
>>> import pandas
>>> pandas.__version__
'0.18.0'
```

scipy.stats

Some nicer plotting features require `scipy.stats`. This library is not required, but pandas will complain if the user tries to perform an action that has this dependency. `scipy.stats` has many non-Python dependencies and in practice turns out to be a little more involved to install. For Ubuntu, the following packages are required before a `pip install scipy` will work:

```
$ sudo apt-get install libatlas-base-dev gfortran
```

Installation of these dependencies is sufficiently annoying that it has lead to “complete scientific Python offerings”, such as Anaconda [^6]. These installers bundle many libraries, are available for Linux, Mac, and Windows, and have optional support contracts. They are a great way to quickly get an environment up.

Summary

Unlike “pure” Python modules, pandas is not just a pip install away unless you have an environment configured to build it. The easiest was to get going is to use the Anaconda Python distribution. Having said that, it is certainly possible to install pandas using other methods.