PandasIntroduction

July 21, 2021

This week we're going to deepen our investigation to how Python can be used to manipulate, clean, and query data by looking at the Pandas data tool kit. Pandas was created by Wes McKinney in 2008, and is an open source project under a very permissive license. As an open source project it's got a strong community, with over one hundred software developers all committing code to help make it better. Before pandas existed we had only a hodge podge of tools to use, such as numpy, the python core libraries, and some python statistical tools. But pandas has quickly become the defacto library for representing relational data for data scientists.

I want to take a moment here to introduce the question answersing site Stack Overflow. Stack Overflow is used broadly within the software development community to post questions about programming, programming languages, and programming toolkits. What's special about Stack Overflow is that it's heavily curated by the community. And the Pandas community, in particular, uses it as their number one resource for helping new members. It's quite possible if you post a question to Stack Overflow, and tag it as being Pandas and Python related, that a core Pandas developer will actually respond to your question. In addition to posting questions, Stack Overflow is a great place to go to see what issues people are having and how they can be solved. You can learn a lot from browsing Stacks at Stack Overflow and with pandas, this is where the developer community is.

A second resource you might want to consider are books. In 2012 Wes McKinney wrote the definitive Pandas reference book called Python for Data Analysis and published by O'Reilly, and it's recently been update to a second edition. I consider this the go to book for understanding how Pandas works. I also appreciate the more brief book "Learning the Pandas Library" by Matt Harrison. It's not a comprehensive book on data analysis and statistics. But if you just want to learn the basics of Pandas and want to do so quickly, I think it's a well laid out volume and it can be had for a good price.

The field of data science is rapidly changing. There's new toolkits and method being created everyday. It can be tough to stay on top of it all. Marco Rodriguez and Tim Golden maintain a wonderful blog aggregator site called Planet Python. You can visit the webpage at planet-python.org, subscribe with an RSS reader, or get the latest articles from the @PlanetPython Twitter feed. There's lots of regular Python data science contributors, and I highly recommend it if you follow RSS feeds.

Here's my last plug on how to deepen your learning. Kyle Polich runs an excellent podcast called Data Skeptic. It isn't Python based per se, but it's well produced and it has a wonderful mixture of interviews with experts in the field as well as short educational lessons. Much of the word he describes is specific to machine learning methods. But if that's something you are planning to explore through this specialization this course is in, I would really encourage you to subscribe to his podcast.

That's it for a little bit of an introduction to this week of the course. Next we're going to dive right into Pandas library and talk about the series data structure.