Python:
Data Analytics and Visualization

Understand, evaluate, and visualize data

A course in three modules

BIRMINGHAM - MUMBAI

Python: Data Analytics and Visualization

Published on: March 2017

Production reference: 1220317

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B32PB, UK.

ISBN: 978-1-78829-009-8

www.packtpub.com

Preface

The world generates data at an increasing pace. Consumers, sensors, or scientific experiments emit data points every day. In finance, business, administration and the natural or social sciences, working with data can make up a significant part of the job. Being able to efficiently work with small or large datasets has become a valuable skill. Python started as a general purpose language. Around ten years ago, in 2006, the first version of NumPy was released, which made Python a first class language for numerical computing and laid the foundation for a prospering development, which led to what we today call the PyData ecosystem: A growing set of high-performance libraries to be used in the sciences, finance, business or anywhere else you want to work efficiently with datasets. Python is not only about data analysis. The list of industrial-strength libraries for many general computing tasks is long, which makes working with data in Python even more compelling.

Social media and the Internet of Things have resulted in an avalanche of data. The data is powerful but not in its raw form; it needs to be processed and modeled and Python is one of the most robust tools we have out there to do so. It has an array of packages for predictive modeling and a suite of IDEs to choose from. Learning to predict who would win, lose, buy, lie, or die with Python is an indispensable skill set to have in this data age. This course is your guide to get started with Predictive Analytics using Python as the tool.

Data visualization is intended to provide information clearly and help the viewer understand them qualitatively. The well-known expression that a picture is worth a thousand words may be rephrased as “a picture tells a story as well as a large collection of words”. Visualization is, therefore, a very precious tool that helps the viewer understand a concept quickly. We are currently faced with a plethora of data containing many insights that hold the key to success in the modern day. It is important to find the data, clean it, and use the right tool to visualize it. This course explains several different ways to visualize data using Python packages, along with very useful examples in many different areas such as numerical computing, financial models, statistical and machine learning, and genetics and networks.

What this learning path covers

Module 1, Getting Started with Python Data Analysis starts with an introduction to data analysis and process, overview of libraries and its uses. Further you'll dive right into the core of the PyData ecosystem by introducing the NumPy package for high-performance computing. We will also deal with a prominent and popular data analysis library for Python called Pandas and understand the data through graphical representation. Moving further you will see how to work with time-oriented data in Pandas. You will then learn to interact with three main categories: text formats, binary formats and databases and work on some application examples. In the end you will see the working of different scikit-learn modules.

Module 2, Learning Predictive Analytics with Python, talks about aspects, scope, and applications of predictive modeling. Data cleaning takes about 80% of the modelling time and hence we will understand its importance and methods. You will see how to subset, aggregate, sample, merge, append and concatenate a dataset. Further you will get acquainted with the basic statistics needed to make sense of the model parameters resulting from the predictive models. You will also understand the mathematics behind linear and logistic regression along with clustering. You will also deal with Decision trees and related classification algorithms. In the end you will be learning about the best practices adopted in the field of predictive modelling to get the optimum results.

Module 3, Mastering Python Data Visualization, expounds that data visualization should actually be referred to as "the visualization of information for knowledge inference". You will see how to use Anaconda from Continuum Analytics and learn interactive plotting methods. You will deal with stock quotes, regression analysis, the Monte Carlo algorithm, and simulation methods with examples. Further you will get acquainted with statistical methods such as linear and nonlinear regression and clustering and classification methods using numpy, scipy, matplotlib, and scikit-learn. You will use specific libraries such as graph-tool, NetworkX, matplotlib, scipy, and numpy. In the end we will see simulation methods and examples of signal processing to show several visualization methods.

What you need for this learning path

You will need a Python programming environment installed on your system. The first module uses a recent Python 2, but many examples will work with Python 3 as well.b The versions of the libraries used in the first module are: NumPy 1.9.2, Pandas 0.16.2, matplotlib 1.4.3, tables 3.2.2, pymongo 3.0.3, redis 2.10.3, and scikit-learn 0.16.1. As these packages are all hosted on PyPI, the Python package index, they can be easily installed with pip. To install NumPy, you would write:

$ pip install numpy If you are not using them already, we suggest you take a look at virtual environments for managing isolating Python environment on your computer. For Python 2, there are two packages of interest there: virtualenv and virtualenvwrapper. Since Python 3.3, there is a tool in the standard library called pyvenv (`https://docs.python.org/3/library/venv.html`), which serves the same purpose. Most libraries will have an attribute for the version, so if you already have a library installed, you can quickly check its version:

>>> import redis

>>> redis.__version__

'2.10.3').

While all the examples in second module can be run interactively in a Python shell. We used IPython 4.0.0 with Python 2.7.10.

For the third module, you need Python 2.7.6 or a later version installed on your operating system. For the examples in this module, Mac OS X 10.10.5's Python default version (2.7.6) has been used. Install the prepackaged scientific Python distributions, such as Anaconda from Continuum or Enthought Python Distribution if possible

Who this learning path is for

This course is for Python Developers who are willing to get into data analysis and wish to visualize their analyzed data in a more efficient and insightful manner.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this course—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail `feedback@packtpub.com`, and mention the course's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt course, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files for this course from your account at `http://www.packtpub.com`. If you purchased this course elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.

You can download the code files by following these steps:

1. Log in or register to our website using your e-mail address and password.
2. Hover the mouse pointer on the **SUPPORT** tab at the top.
3. Click on **Code Downloads & Errata**.
4. Enter the name of the course in the **Search** box.
5. Select the course for which you're looking to download the code files.
6. Choose from the drop-down menu where you purchased this course from.
7. Click on **Code Download**.

You can also download the code files by clicking on the **Code Files** button on the course's webpage at the Packt Publishing website. This page can be accessed by entering the course's name in the **Search** box. Please note that you need to be logged in to your Packt account.

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 course is also hosted on GitHub at `https://github.com/PacktPublishing/Python-Data-Analytics-and-Visualization`. We also have other code bundles from our rich catalog of books, videos, and courses available at `https://github.com/PacktPublishing/`. Check them out!

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our courses—maybe a mistake in the text or the code—we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this course. If you find any errata, please report them by visiting `http://www.packtpub.com/submit-errata`, selecting your course, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to `https://www.packtpub.com/books/content/support` and enter the name of the course in the search field. The required information will appear under the **Errata** section.

Module 1: Getting Started with Python Data Analysis

Module 2: Learning Predictive Analytics with Python