ProTech protechtraining.com *"Charting the Course ... ... to Your Success!"*

Python For Data Analysis

Course Summary

Description

This course takes beginning or intermediate Python programmers into the world of Python data analysis. From importing to searching to exporting, many facets of handling large datasets are covered.

This is a hands-on programming class. All concepts are reinforced by informal practice during the lecture followed by graduated lab exercises.

Python For Data Analysis is a practical introduction to a working programming language, not an academic overview of syntax and grammar. Students will immediately be able to use Python to complete tasks in the real world.

THIS COURSE MAY BE CUSTOMIZED

Objectives

At the end of this course, students will be able to:

- Extract data from binary files or other binary data streams
- Create data structures using classes and named tuples
- Search and replace text with regular expressions
- Read and write CSV and other data formats
- Serialize data to pickle files, JSON, and XML
- Consume and process data from the Web
- Deal with missing data
- Share data with Excel spreadsheets
- Analyze data with scipy/numpy

Topics

- File I/O
- Classes
- Generators and Other Iterables
- Data structures
- Serializing Data
- Consuming Data from the Web
- Excel Spreadsheets
- Dates and Times
- Regular Expressions
- Working with Binary Data
- Analyzing Datasets
- Bigger Data – Working with PyTables

Audience

This course is appropriate for Python developers who work with medium to large datasets.

Prerequisites

Students should be comfortable with basic Python programming.

Duration

Three days

Python For Data Analysis

Course Outline

I. File I/O

A.Opening a file
B.Iterating over lines
C. Reading characters or bytes
D. Reading all lines
E.Formatted output
F. Using fileinput

This section covers reading and writing text files, with some exposure to binary data. It introduces the with statement. Labs involve reading data from files, as well as writing data to files.
LENGTH: 60 minutes

II. Classes

A.Defining classes
B.Constructors
C. Instance methods and data
D. Class/static methods and data

This chapter is an intro to object-oriented programming in Python. Defining classes with constructors, instance methods, and properties is shown, along with some advanced features such as inheritance and class data and methods. The lab for this chapter is using a class to represent a record from a data file.
LENGTH: 70 minutes

III. Generators and Other Iterables

A.Iterables
B.Saving memory with generators
C. Generator expressions
D. Generator functions
E.Generator classes
F. Stacking generators

This chapter presents Python's powerful concepts of iterables. It describes the difference between "real" iterables (lists, maps, etc) and "virtual" iterables (generators). Labs including creating generator expressions and functions.
LENGTH: 60 minutes

IV. Data Structures

A.How to store data
B.The basics: lists and tuples
C. Named access with dictionaries
D. Named tuples: best of both worlds
E.Using classes as data structures

This chapter covers all of Python's import builtin data structures, including the under- utilized namedtuple. Students learn which data structures are best for which situation. Labs involve creating and using each type of structure.
LENGTH: 60 minutes

V. Serializing Data

A.Pickle
B.JSON
C. CSV
D. XML

In this chapter, students learn tools provided by Python for data serialization. In addition to reading and writing data with XML and JSON, we cover Pickle, CSV, YAML, and other formats. Labs for this chapter involve reading data from XML and JSON, as well as creating new files in those formats.
LENGTH: 60 minutes

VI. Consuming Data from the Web

A.Web data sources
B.Data via URL
C. RESTful data
D. Screen-scraping

This chapter teaches students how to use the requests module to execute HTTP requests an d extract data from the HTTP response. It covers RESTful services as well as normal HTML and other documents provided by web servers. Labs include downloading CSV data and processing it.

LENGTH: 60 minutes

VII. Excel Spreadsheets

A.The xlrd, xlwr, and xlutil modules
B.Reading an existing spreadsheet
C. Creating a spreadsheet from scratch
D. Modifying an existing spreadsheet

This chapter covers reading and writing data to and from Excel spreadsheets. Students will learn to open a workbook and select individual worksheets, read data from any cell, and update worksheets by adding or changing data. They will also learn to create formulas and change the style of cells.
LENGTH: 60 minutes

Python For Data Analysis

Course Outline (cont'd)

VIII. Dates and Times

A.Python date and time objects
B.The time module
C. Using calendars
D. Converting between formats
E.Parsing and printing
F. Time zones

This chapter covers how to work with dates and times. It covers Python's builtin time- related classes, as well as Unix-style times, and conversion among all types.
Labs including parsing and manipulating user-provided time and date strings. LENGTH: 60 minutes

IX. Regular Expressions

A.RE syntax overview
B.Basic patterns
C. RE Objects
D. Searching and matching
E.Compilation flags
F. Grouping
G. Replacing text
H. Splitting a string

Students will learn how to make sense of often-confusing regular expression syntax, then learn how to work with text using Python's re module. Labs including parsing data from text files and displaying it from a data structure.
LENGTH: 60 minutes

X. Analyzing Datasets

A.Sorting data
B.Filtering values
C. Basic Statistics
D. Leveraging scipy/numpy
E.Using PANDAS

This chapter explores general data analysis using NumPy and SciPy. Many techniques are discussed, as well as other modules for data analysis, including SciKit-Learn. Labs for this chapter are a little free-form, depending on student interests.
LENGTH: 60-90 minutes (depending on audience)

XI. Using Pandas

A.Pandas basics
B.Creating dataframes
C. Fetching data
D. Saving to file
E.Selecting data
F. Plotting from Pandas

This chapter starts with a discussion of what Pandas is, and how it relates to the R language. Then we go into how to create and manipulate dataframes. After discussing the many ways to index and manipulate data, we present the powerful I/O capabilities of Pandas, and demonstrate reading in data files. Useful features such as time series, dropping invalid data, and matrix match are also covered. Labs involve using Pandas to read in a dataset and perform calculations on the data.
LENGTH: 60 minutes

XII. Testing tools

A.The unittest module
B.Skipping tests
C. Test runners
D. Test discovery
E.Mocking
F. Other Python testing tools

This chapter is a comprehensive look at using the unittest module to create and run unit tests for Python code. Labs include writing unit tests for scripts written in previous labs.
LENGTH: 60 minutes