Course Overview

Python for Data Science is an instructor-led course taught by experienced instructors who bring anecdotes from professional experiences to solidify learning. You will learn how to use Python to build, evaluate, and implement end-to-end data analysis pipelines through a series of lectures and hands-on exercises. You will learn how to ingest data from different data streams, scrape data and extract information from the internet, explore and get insights from your data and how to articulate those insights to make data-driven decisions.

Learning Outcomes

By the end of the course, students will be able to:

- Use Python libraries such as Numpy, Pandas and Matplotlib for data analysis
- Build, evaluate, fine-tune and implement end-to-end data analysis pipelines in Python
- Explore and visualize data to identify trends and support data-driven decision making
- Articulate key findings from data analysis to recommend solutions for business problems

Module 1: Dealing with Data

Data Loading, Storage and File Formats

The first step in building a data analysis pipeline is getting raw data for pre-processing and analysis. The focus of this module will be on developing an understanding of how structured or unstructured data can be loaded into Python from a variety of file formats. We will also talk about different data structures that can be used to make organization and processing easier.

Topics: File I/O, Complex data structures in Python, Lists, Dictionaries, Pandas dataframes, Parsing JSON/XML/HTML

Data Handling Techniques

Once we have loaded our data, we'll move on to understanding Python's powerful data handling techniques. You'll be introduced to different methods that you will use to get the information you need from data stored in complex data structures.

Topics: Basics of indexing, Introduction to slicing and subsetting

Data Wrangling: Cleaning, Transforming, Merging and Reshaping

Data quality is one of the most overlooked issues in data science. We will discuss challenges and best practices in data acquisition, processing, transformation, cleaning and loading. We'll also learn how to process data in Python before it can be used for further analysis. This includes cleaning it, reshaping it, and merging the data from multiple sources/data structures, and applying transformations.

Topics: Simple data cleaning, Data transformation, Pandas merging, Pandas reshaping

Aggregation and Grouping

This module will focus on Python's powerful data aggregation and grouping techniques and how they can be used to get insights from our data.

Topics: Grouping and segmentation with pandas, Data aggregation

Data Exploration and Visualization

This module will focus on understanding how to dissect and explore our data to uncover useful insights. We will also look at different visualization techniques to help present our data
and insights findings and get acquainted with popular visualization packages in Python.

Topics: Various data visualization and exploration techniques and packages. Interpreting boxplots, histograms, density plots, scatterplots and more.

Module 2: Real World Applications

Introduction to the RESTful API

This module will be focused on the REST API, its functions and how it can be used to communicate with websites like Twitter and Reddit that have exposed endpoints to scrape data.

Topics: Introduction to the REST API, API request structure, API methods, Open data endpoints, REST API in python

Data Pipelines in Python

Once we're comfortable with the REST API, we'll look at how it can be used to get data from websites. We'll use the data we scraped along with other data loading, parsing, cleaning and exploration techniques we learned previously to build an end-to-end data pipeline.

Topics: Web scraping, building a data pipeline in Python

Python Project

The last module of the bootcamp will focus on a hands-on Python project.