# Python Data Analysis Essentials for Markets Analysts – 2.5 Days

### This course introduces basics of the python programming environment, including fundamental tasks such as

### Reading, manipulating and cleaning large data files

### Using the NumPy and Pandas libraries to analyze datasets

### How to visualize data using charting libraries

### Each topic combines lecture, exercise and a worked reference solution using financial market data.

The course also includes team-based **Citi Data Challenges**, a series of open-ended “mini-projects” where the analysts apply their new skills to complete real-world data analysis tasks.

### Pre-Requisites

* Understanding of financial data sets such as global market data (equities, futures, forex etc.) and reference data such as corporate actions.

### Objectives

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

* Use core python functions to analyze the range of financial data they will encounter in Citi
* Use Citi’s Dataflame environment
* Extract, fit and cleanse datasets
* Use NumPy to manage large datasets
* Using Pandas to align, rank, sort datasets
* Visualize data and create charts with matplotlib and seaborn

### Content

**Introduction - Data Science, Data Analysis & Your Role in Citi**

* Data Science Ice-Breaker
* What do we mean by data science and data analysis
* Course Overview

**[June Pilot Only] Your Feedback**

* Collating your feedback
* Daily retrospective
* Prioritizing refinements

**Guest Speaker – Citi Data Pioneer**

* Working with the volume, complexity and range of data in Citi
* Citi’s DataFlame Environment

**Module 1 – Overview of Python & Review**

* Overview of the main python language features that will be used throughout the course
* Functions, Lists, Dictionaries, Strings, Tuples
* Command Line Arguments
* The main Python Modules for numerical programming – NumPy, Pandas, matplotlib and quandl

**Lab 1a – Python Scripting**

* A simple step-by-step programming task to introduce the development environment.

**Lab 1b – Configuration and Debugging**

* Creating your own workspace with a set of pre-prepared python utilities
* Adding a new python module.

**Module 2 – Data Extraction and Consumption**

* Introduction to Regular Expressions and Python Regex Objects
* Directory navigation
* Extracting Data
* from Test Files.
* from Spreadsheets
* from Databases

**Lab 2 – Data Consumption**

* Creating a simple a Python application to consume a csv file of portfolio data containing
* Security IDs, Transaction Dates, Initial Qty, Initial Price, Currency
* Current Price, Fund Manager and Client Details
* The application should accept 3 command line arguments
* Input file
* Output file
* Format
* The format parameter will specify how to sort the output (by Portfolio Id, by Fund Manager, by Client) either ascending or descending.

**Module 3 – Data Preparation – Advanced Data preparation with Pandas**

* Dealing with Missing Data
* Pandas recipes
* NANs, Strategies,
* How to fit data
* Transforming Data
* Categorical Data
* Label Encoding
* Labels and Levels

**Lab 3 – Data consumption**

* Write a Python application to consume 2 separate datasets
* Dataset #1 – Portfolio Data
* Dataset #2 – Forex Data
* Generate a third dataset containing the value of each portfolio item in USD.
* The exchange rate file will not contain all exchange rates needed – participants will need to calculate a missing exchange rate by using 2 existing exchange rates.

**Module 4 – Data Visualization**

* Line properties
* Working with multiple axes
* Adding text to a chart
* Logarithmic and nonlinear axes
* Line graphs, block graphs, pie charts

**Lab 4 – Visualization**

Generate clearly formatted graphs using csv data containing the following data

* Dataset #1 - Salary and age – Line Chart
* Dataset #2 – Portfolio - Pie Chart of a portfolio composition

The datasets will require the participants to

* Deal with missing values
* Fit and scale data
* Deal with categorical data

### Citi Data Challenges

**Data Challenge 1 – Calculate and display the P&L of a portfolio**

This challenge will use two existing datasets from a previous exercise:

* Dataset #1 – Portfolio Data
* Dataset #2 – Forex Data

Create a Python application which

1. Consumes both datasets
2. Validates, fits and creates categorical data.
3. Produce a P&L report which includes the following data: Portfolio, Fund Manager, client
4. Split the P&L performance into 2 components

* The performance of the security, fund & fund manager
* The impact Forex rates have had on the portfolio

1. Display their results in a well-formatted chart
2. Be ready to present the solution and approach to the rest of the group.

**Data Challenge 2 – Market Data** **Correlation and Trend Analysis**

This challenge will set the participants the task of analysing two disparate datasets and identify any correlation between these 2 datasets.

The data will be in CSV files and contain market data from various sources including

* Stock Index information
* Exchange Rates
* Commodity Prices
* Asset Prices

Each team will be given two data sets and asked to write a Python application which consumes both datasets and identifies any correlation between both these data.

For example, given the following two datasets identify any overall patterns or correlation:

* Daily closing prices for S&P 500 between 2014 and 2017
* Daily closing price for Oil between 2014 and 2017

Or

* Intraday USD/EUR rates for a week in 2018
* Intraday prices for a series of Individual securities during the same week in 2018

For any of the more advance teams / individuals, the scope of the project will be expanded to have them analyse 3 data sets and look for trends and correlations