COURSE OUTLINE

# PY-DS

# Data Science & Big Data with Python

| DURATION | LEVEL | TECHNOLOGY | DELIVERY METHOD | TRAINING CREDITS |
|---|---|---|---|---|
| 5 Days | **Advanced** | Python | Instructor Led | NA |

## INTRODUCTION

In this course we cover three main topics: Applied Data Science, Big Data Analysis and Data Science projects. We show how Jupyter Notebooks can be used with Python for various data-science applications. Aside from being an ideal "virtual playground" for data exploration, Jupyter Notebooks are equally suitable for creating reproducible data processing pipelines, visualizations, and prediction models.

We will look at various data modelling concepts using Jupyter Notebooks, and we will see the full power of Jupyter Notebooks as we work through this course.

We will also dive into Big Data Analysis, processing big data in real time is challenging due to scalability, information inconsistency, and fault tolerance. Big Data Analysis with Python teaches you how to use tools that can control this data avalanche for you. With this course, you'll learn effective techniques to aggregate data into useful dimensions for posterior analysis, extract statistical measurements, and transform datasets into features for other systems.

Data Science Projects with Python is designed to give you practical guidance on industry-standard data analysis and machine learning tools in Python, with the help of realistic data. The course will help you understand how you can use pandas and Matplotlib to critically examine a dataset with summary statistics and graphs and extract the insights you seek to derive. You will continue to build on your knowledge as you learn how to prepare data and feed it to machine learning algorithms, such as regularized logistic regression and random forest, using the scikit-learn package. You'll discover how to tune the algorithms to provide the best predictions on new and unseen data.

## AUDIENCE PROFILE

This course is designed for Python developers, data analysts, and data scientists. This course is not for beginners.

## PREREQUISITES

This course focuses on creating reproducible data analyses using Python and Jupyter and is intended for an audience with a background in Python. As such, we do not cover the basics of Python in this course. However, we will take a brief tour of the Jupyter interface.

Basic knowledge of statistical measurements and relational databases will help in understanding various concepts explained in this course.

## COURSE OBJECTIVES

After completing this course, students will be able to:

- Use Python to read and transform data into different formats
- Generate basic statistics and metrics using data on the disk
- Work with computing tasks distributed over a cluster
- Convert data from various sources into storage or querying formats
- Prepare data for statistical analysis, visualization, and machine learning
- Present data in the form of effective visuals

## COURSE CONTENT

**Lesson 1: Jupyter Fundamentals**

- Basic Functionality and Features
- Our First Analysis - The Boston Housing Dataset

**Lesson 2: Data Cleaning and Advanced Machine Learning**

- Preparing to Train a Predictive Model
- Training Classification Models

**Lesson 3: Web Scraping and Interactive Visualizations**

- Scraping Web Page Data
- Interactive Visualizations

**Lesson 4: The Python Data Science Stack**
- Python Libraries and Packages
- Using Pandas
- Data Type Conversion
- Aggregation and Grouping
- Exporting Data from Pandas
- Visualization with Pandas

**Lesson 5: Statistical Visualizations**

- Types of Graphs and When to Use Them
- Components of a Graph
- Which Tool Should Be Used?
- Types of Graphs
- Pandas DataFrames and Grouped Data
- Changing Plot Design: Modifying Graph Components
- Exporting Graphs

**Lesson 6: Working with Big Data Frameworks**
- Hadoop
- Spark
- Writing Parquet Files
- Handling Unstructured Data

**Lesson 7: Diving Deeper with Spark**
- Getting Started with Spark DataFrames
- Writing Output from Spark DataFrames
- Exploring Spark DataFrames
- Data Manipulation with Spark DataFrames
- Graphs in Spark

**Lesson 8: Handling Missing Values and Correlation Analysis**
- Setting up the Jupyter Notebook
- Missing Values
- Handling Missing Values in Spark DataFrames
- Correlation

**Lesson 9: Exploratory Data Analysis**
- Defining a Business Problem
- Translating a Business Problem into Measurable Metrics and Exploratory Data Analysis (EDA)
- Structured Approach to the Data Science Project Life Cycle

**Lesson 10: Reproducibility in Big Data Analysis**
- Reproducibility with Jupyter Notebooks
- Gathering Data in a Reproducible Way
- Code Practices and Standards
- Avoiding Repetition

**Lesson 11: Creating a Full Analysis Report**
- Reading Data in Spark from Different Data Sources
- SQL Operations on a Spark DataFrame
- Generating Statistical Measurements

**Lesson 1: Data Exploration and Cleaning**
- Python and the Anaconda Package Management System
- Different Types of Data Science Problems
- Loading the Case Study Data with Jupyter and pandas
- Data Quality Assurance and Exploration
- Exploring the Financial History Features in the Dataset
- Activity 1: Exploring Remaining Financial Features in the Dataset

**Lesson 2: Introduction to Scikit-Learn and Model Evaluation**
- Model Performance Metrics for Binary Classification
- Activity 2: Performing Logistic Regression with a New Feature and Creating a Precision-Recall Curve

**Lesson 3: Details of Logistic Regression and Feature Exploration**
- Examining the Relationships between Features and the Response
- Univariate Feature Selection: What It Does and Doesn't Do
- Building Cloud-Native Applications
- Activity 3: Fitting a Logistic Regression Model and Directly Using the Coefficients

**Lesson 4: The Bias-Variance Trade-off**
- Estimating the Coefficients and Intercepts of Logistic Regression
- Cross Validation: Choosing the Regularization Parameter and Other Hyperparameters
- Activity 4: Cross-Validation and Feature Engineering with the Case Study Data

**Lesson 5: Decision Trees and Random Forests**
- Decision trees
- Random Forests: Ensembles of Decision Trees
- Activity 5: Cross-Validation Grid Search with Random Forest

**Lesson 6: Imputation of Missing Data, Financial Analysis, and Delivery to Client**
- Review of Modelling Results
- Dealing with Missing Data: Imputation Strategies
- Activity 6: Deriving Financial Insights
- Final Thoughts on Delivering the Predictive Model to the Client

## ASSOCIATED CERTIFICATIONS & EXAM

None