

**Applied Data Science with Python**

**Course Number:** PYTH-262WA  
**Duration:** 2 days

**Overview**

This Applied Data Science with Python training course teaches attendees the fundamentals of using Python to program tasks in data science, data engineering, business analytics, and data visualization.

**Prerequisites**

Participants must have a working knowledge of Python or experience with a modern programming language and familiarity with core statistical concepts (i.e., variance, correlation, etc.).

**Materials**

All Python Data Science training attendees receive comprehensive courseware.

**Software Needed on Each Student PC**

* Anaconda Python 3.6 or later
* Spyder IDE and Jupyter notebook (Comes with Anaconda)

**Objectives**

* Use Jupyter Notebooks
* Understand Python
* Work with NumPy and pandas
* Repair and normalize data
* Work with data visualization in Python
* Perform data splitting
* Work with the Random Forest algorithm
* Understand the k-Means algorithm

**Outline**

* Introduction to Python for Data Science
  + Python Data Science-Centric Libraries
  + NumPy
  + SciPy
  + Pandas
  + Creating a pandas DataFrame
  + Fetching and Sorting Data
  + Scikit-learn
  + Matplotlib
  + Seaborn
  + Python Dev Tools and REPLs
  + IPython
  + Jupyter
  + Anaconda
* Defining Data Science
  + What is Data Science?
  + Data Science, Machine Learning, AI?
  + The Data-Related Roles
  + The Data Science Ecosystem
  + Tools of the Trade
  + Who is a Data Scientist?
  + Data Scientists at Work
  + Examples of Data Science Projects
  + An Example of a Data Product
  + Applied Data Science at Google
  + Data Science Gotchas
* Data Processing Phases
  + Typical Data Processing Pipeline
  + Data Discovery Phase
  + Data Harvesting Phase
  + Data Priming Phase
  + Exploratory Data Analysis
  + Model Planning Phase
  + Model Building Phase
  + Communicating the Results
  + Production Roll-out
  + Data Logistics and Data Governance
  + Data Processing Workflow Engines
  + Apache Airflow
  + Data Lineage and Provenance
  + Apache NiFi
* Descriptive Statistics Computing Features in Python
  + Descriptive Statistics
  + Non-uniformity of a Probability Distribution
  + Using NumPy for Calculating Descriptive Statistics Measures
  + Finding Min and Max in NumPy
  + Using pandas for Calculating Descriptive Statistics Measures
  + Correlation
  + Regression and Correlation
  + Covariance
  + Getting Pairwise Correlation and Covariance Measures
  + Finding Min and Max in pandas DataFrame
* Repairing and Normalizing Data
  + Repairing and Normalizing Data
  + Dealing with the Missing Data
  + Sample Data Set
  + Getting Info on Null Data
  + Dropping a Column
  + Interpolating Missing Data in pandas
  + Replacing the Missing Values with the Mean Value
  + Scaling (Normalizing) the Data
  + Data Preprocessing with scikit-learn
  + Scaling with the scale() Function
  + The MinMaxScaler Object
* Data Visualization in Python
  + Data Visualization
  + Data Visualization in Python
  + Matplotlib
  + Getting Started with matplotlib
  + Subplots
  + Using the matplotlib.gridspec.GridSpec Object
  + The matplotlib.pyplot.subplot() Function
  + Figures
  + Saving Figures to a File
  + Seaborn
  + Getting Started with seaborn
  + Histograms and KDE
  + Plotting Bivariate Distributions
  + Scatter plots in seaborn
  + Pair plots in seaborn
  + Heatmaps
  + ggplot
* Data Science and Machine Learning (ML) Algorithms in Scikit-Learn
  + In-Class Discussion
  + Types of Machine Learning
  + Terminology: Features and Observations
  + Representing Observations
  + Terminology: Labels
  + Terminology: Continuous and Categorical Features
  + Continuous Features
  + Categorical Features
  + Common Distance Metrics
  + The Euclidean Distance
  + What is a Model
  + Supervised vs. Unsupervised Machine Learning
  + Supervised Machine Learning Algorithms
  + Unsupervised Machine Learning Algorithms
  + Choosing the Right Algorithm
  + The scikit-learn Package
  + Scikit-learn Estimators, Models, and Predictors
  + Model Evaluation
  + The Error Rate
  + Confusion Matrix
  + The Binary Classification Confusion Matrix
  + Multi-class Classification Confusion Matrix Example
  + ROC Curve
  + Example of a ROC Curve
  + The AUC Metric
  + Feature Engineering
  + Scaling of the Features
  + Feature Blending (Creating Synthetic Features)
  + The 'One-Hot' Encoding Scheme
  + Example of 'One-Hot' Encoding Scheme
  + Bias-Variance (Underfitting vs. Overfitting) Trade-off
  + The Modeling Error Factors
  + One Way to Visualize Bias and Variance
  + Underfitting vs. Overfitting Visualization
  + Balancing Off the Bias-Variance Ratio
  + Regularization in scikit-learn
  + Regularization, Take Two
  + Dimensionality Reduction
  + PCA and isomap
  + The Advantages of Dimensionality Reduction
  + The LIBSVM format
  + Life-cycles of Machine Learning Development
  + Data Splitting into Training and Test Datasets
  + ML Model Tuning Visually
  + Data Splitting in scikit-learn
  + Cross-Validation Technique
  + Hands-on Exercise
  + Classification (Supervised ML) Examples
  + Classifying with k-Nearest Neighbors
  + k-Nearest Neighbors Algorithm
  + Regression Analysis
  + Simple Linear Regression Model
  + Linear Regression Illustration
  + Least-Squares Method (LSM)
  + Gradient Descent Optimization
  + Multiple Regression Analysis
  + Evaluating Regression Model Accuracy
  + The R2 Model Score
  + The MSE Model Score
  + Logistic Regression (Logit)
  + Interpreting Logistic Regression Results
  + Decision Trees
  + Properties of Decision Trees
  + Decision Tree Classification in the Context of Information Theory
  + The Simplified Decision Tree Algorithm
  + Using Decision Trees
  + Random Forests
  + Support Vector Machines (SVMs)
  + Naive Bayes Classifier (SL)
  + Naive Bayesian Probabilistic Model in a Nutshell
  + Bayes Formula
  + Classification of Documents with Naive Bayes
  + Unsupervised Learning Type: Clustering
  + Clustering Examples
  + k-Means Clustering (UL)
  + Global vs. Local Minimum Explained
  + XGBoost
  + Gradient Boosting
  + A Better Algorithm or More Data?
* Introduction to Python for Data Engineers
  + What is Python?
  + Additional Documentation
  + Which version of Python am I running?
  + Python Dev Tools and REPLs
  + IPython
  + Jupyter
  + Anaconda
  + Python Variables and Basic Syntax
  + Variable Scopes
  + PEP8
  + The Python Programs
  + Getting Help
  + Variable Types
  + Assigning Multiple Values to Multiple Variables
  + Null (None)
  + Strings
  + Finding the Index of a Substring
  + String Splitting
  + Triple-Delimited String Literals
  + Raw String Literals
  + String Formatting and Interpolation
  + Boolean
  + Boolean Operators
  + Numbers
  + Looking Up the Runtime Type of a Variable
  + Divisions
  + Assignment-with-Operation
  + Relational Operators
  + The if-elif-else Triad
  + Conditional Expressions (a.k.a. Ternary Operator)
  + The While-Break-Continue Triad
  + The for Loop
  + try-except-finally
  + Lists
  + Main List Methods
  + Dictionaries
  + Working with Dictionaries
  + Sets
  + Common Set Operations
  + Set Operations Examples
  + Finding Unique Elements in a List
  + Enumerate
  + Tuples
  + Unpacking Tuples
  + Functions
  + Dealing with Arbitrary Number of Parameters
  + Keyword Function Parameters
  + The range Object
  + Random Numbers
  + Python Modules
  + Creating a Runnable Application
  + List Comprehension
  + Zipping Lists
  + Working with Files
  + Reading Command-Line Parameters
  + Accessing Environment Variables
  + What is Functional Programming (FP)?
  + Terminology: Higher-Order Functions
  + Lambda Functions in Python
  + Lambdas in the Sorted Function
  + Regular Expressions
  + Python Data Science-Centric Libraries
* Conclusion