

WA2715 Applied Data Science with Python

This intensive training course provides theoretical and practical aspects of using Python in the realm of Data Science, Business Analytics, and Data Logistics.

2 Days

How You Will Benefit

This intensive training course provides theoretical and practical aspects of using Python in the realm of Data Science, Business Analytics, and Data Logistics. The coverage of the related core concepts,

terminology, and theory is provided as well. This training course is supplemented by a variety of hands-on labs (the list of which is provided at the bottom of this outline) that help attendees reinforce their theoretical knowledge of the learned material.

Workshop Description

Chapter 1. Python for Data Science

- In-Class Discussion
- Python Data Science-Centric Libraries
- NumPy
- NumPy Arrays
- Select NumPy Operations
- SciPy
- pandas
- Creating a pandas DataFrame
- Fetching and Sorting Data
- Scikit-learn
- Matplotlib
- Seaborn
- Python Dev Tools and REPLs
- IPython
- Jupyter
- Jupyter Operation Modes
- Jupyter Common Commands
- Anaconda
- Summary

Chapter 2. 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
- Summary

Chapter 3. Data Processing Phases

- Typical Data Processing Pipeline
- Data Discovery Phase
- Data Harvesting Phase
- Data Priming Phase
- Exploratory Data Analysis
- Model Planning Phase

Algonquin College Corporate Training specializes in industry-focused training solutions that equip employees with the knowledge, skills and expertise needed to succeed in their day-to-day work, to advance in their careers and to build organizational capacity.

For more information, contact us at (613) 727-7729 or visit algonquincollege.com/corporate

- 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
- Summary

Chapter 4. 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
- Summary

Chapter 5. 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
- Summary

Chapter 6. Data Visualization in Python

- Data Visualization
- Data Visualization in Python
- Matplotlib
- Getting Started with matplotlib

- The matplotlib.pyplot.plot() Function
- The matplotlib.pyplot.bar() Function
- The matplotlib.pyplot.pie () Function
- 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
- Summary

Chapter 7. Data Science and 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 an ROC Curve
- The AUC Metric
- Feature Engineering

Algonquin College Corporate Training specializes in industry-focused training solutions that equip employees with the knowledge, skills and expertise needed to succeed in their day-to-day work, to advance in their careers and to build organizational capacity.

For more information, contact us at (613) 727-7729 or visit algonquincollege.com/corporate

- 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
- k-Nearest Neighbors Algorithm
- Hands-on Exercise
- Regression Analysis
- Regression vs Correlation
- Regression vs Classification
- 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 Linear Logistic Regression Results
- Decision Trees
- Decision Tree Terminology
- Properties of Decision Trees
- Decision Tree Classification in the Context of Information Theory
- The Simplified Decision Tree Algorithm
- Using Decision Trees
- Random Forests

- Hands-On Exercise
- Hands-on Exercise
- 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)
- k-Means Clustering in a Nutshell
- k-Means Characteristics
- Global vs Local Minimum Explained
- Hands-On Exercise
- XGBoost
- Gradient Boosting
- Hands-On Exercise
- A Better Algorithm or More Data?
- Summary

Chapter 8. (Optional) Quick 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
- Jupyter Operation Modes
- Jupyter Common Commands
- 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 Index of a Substring
- String Splitting
- Triple-Delimited String Literals
- Raw String Literals
- String Formatting and Interpolation
- Boolean

Algonquin College Corporate Training specializes in industry-focused training solutions that equip employees with the knowledge, skills and expertise needed to succeed in their day-to-day work, to advance in their careers and to build organizational capacity.

For more information, contact us at (613) 727-7729 or visit algonquincollege.com/corporate

- Boolean Operators
- Numbers
- Looking Up the Runtime Type of a Variable
- Divisions
- Assignment-with-Operation
- Comments:
- Relational Operators
- The if-elif-else Triad
- An if-elif-else Example
- 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
- Importing Modules
- Installing Modules
- Listing Methods in a Module
- Creating Your Own Modules

- Creating a Runnable Application
- List Comprehension
- Zipping Lists
- Working with Files
- Reading and Writing Files
- Reading Command-Line Parameters
- Accessing Environment Variables
- What is Functional Programming (FP)?
- Terminology: Higher-Order Functions
- Lambda Functions in Python
- Example: Lambdas in the Sorted Function
- Other Examples of Using Lambdas
- Regular Expressions
- Using Regular Expressions Examples
- Python Data Science-Centric Libraries
- Summary

Lab Exercises

- Lab 1. Using Jupyter Notebook*
- Lab 2. Understanding Python*
- Lab 3. Understanding NumPy*
- Lab 4. Understanding pandas*
- Lab 5. Repairing and Normalizing Data*
- Lab 6. Data Visualization in Python*
- Lab 7. Data Visualization in Python Project*
- Lab 8. Data Splitting*
- Lab 9. The k-Nearest Neighbors Algorithm [OPTIONAL]*
- Lab 10. The Random Forest Algorithm*
- Lab 11. Spam Detection with Random Forest Project [OPTIONAL]*
- Lab 12. The k-Means Algorithm*
- Lab 13. Building Regression Models with XGBoost Library*
- Lab 14. A Hand-Made Classifier Project [OPTIONAL]*

Algonquin College Corporate Training specializes in industry-focused training solutions that equip employees with the knowledge, skills and expertise needed to succeed in their day-to-day work, to advance in their careers and to build organizational capacity.

For more information, contact us at (613) 727-7729 or visit algonquincollege.com/corporate

Build Your Competency In

- Applied Data Science and Business Analytics
- Common Data Science algorithms for supervised and unsupervised machine learning
- NumPy, pandas, Matplotlib, scikit-learn
- Python REPLs
- Jupyter notebooks
- Data analytics life-cycle phases
- Data repairing and normalizing
- Data aggregation and grouping
- Data visualization

Attend This Workshop If You Are A

- Business Analyst
- Developer
- IT Architect
- Technical Manager

Prerequisites

Participants should have a working knowledge of Python (or have the programming background and/or the ability to quickly pick up Python's syntax), and be familiar with core statistical concepts (variance, correlation, etc.)

Please refer to our Python 3 course for the above.

How to Register

- Call (613) 727-7729
- Email training@algonquincollege.com
- Register online:
www.algonquincollege.com/corporate
- Location for Open Enrolment Courses:
 - Algonquin College Learning Centre
 - 700 Sussex Drive, Ottawa, Ontario K1N 1K4
 - Time: 8:30 a.m. to 4:00 p.m.
 - Check-in, coffee and muffins start at 7:45 a.m. Afternoon refreshment provided.
 - Group size is limited to 20.
- Dedicated sessions are available at our site or at your location.

Algonquin College Corporate Training specializes in industry-focused training solutions that equip employees with the knowledge, skills and expertise needed to succeed in their day-to-day work, to advance in their careers and to build organizational capacity.

For more information, contact us at (613) 727-7729 or visit algonquincollege.com/corporate