

Python for AI and Data Manipulation

Preparing reliable data with NumPy and pandas

FA-0658

Data & Analytics · Software Development

Use NumPy and pandas to load, inspect, clean, transform and validate data for analytical or machine-learning workflows, with attention to memory and leakage risks.

Introduction

This course develops practical Python data-preparation skills for learners who already know Python syntax and notebook or script workflows. It focuses on numerical arrays, DataFrames, data quality and reproducible transformations rather than AI-model construction.

Participants examine common file formats, missing and inconsistent values, joins, aggregation, reshaping and exploratory summaries. Data preparation is treated as a set of choices to validate: an outlier is not automatically an error, and correlation is not evidence of causation.

Larger-file processing and model hand-off are covered with realistic limits. pandas is primarily an in-memory tool, and chunking is suitable only for operations that can be decomposed appropriately. Training/test separation and training-fitted preprocessing help prevent misleading evaluation.

Learning outcomes

The course teaches participants to:

- Explain data collection, preparation, modelling and evaluation roles in an AI/data pipeline.
- Use NumPy array types, indexing, reshaping, broadcasting and selected numerical operations.
- Load and inspect CSV, Excel, JSON or text data with suitable dependencies.
- Clean and validate missing, inconsistent and duplicate records using justified strategies.
- Create features, aggregate groups, join datasets and reshape DataFrames.

- Summarise distributions and relationships while checking anomalies and leakage risks.
- Assess memory, chunking and scalable-tool options without assuming unlimited pandas capacity.
- Prepare reproducible feature/target datasets and documented validation checks for model hand-off.

Prerequisites

- Basic knowledge of Python syntax, variables, loops, functions, and data structures.
- Familiarity with running Python scripts and using an IDE or notebook environment.
- A general understanding of what AI and machine learning are, at a conceptual level.

Access to a prepared, supported Python environment with compatible NumPy, pandas and selected file-format dependencies. Training data should be illustrative or appropriately authorised; model-building experience is not required.

Training outline

1. 1. Python in AI and Data Workflows

Overview of AI and data pipelines

- Data collection, preparation, modeling, evaluation
- Where Python fits in end-to-end workflows
- Data-centric vs. model-centric thinking
- Common data problems encountered in real AI projects

2. 2. Numerical Computing with NumPy

Introduction to NumPy arrays

- Array creation and data types
- Memory use and performance trade-offs; measure rather than assume array operations are always faster

Array operations

- Vectorized computations
- Element-wise vs. aggregate operations
- Indexing, slicing, and reshaping
- Broadcasting rules and practical use cases
- Basic linear algebra operations for AI readiness

3. 3. Loading and Inspecting Data with pandas

Pandas core data structures

- Series and DataFrames

Loading data

- CSV, Excel, JSON, and text formats

Inspecting and understanding datasets

- Shape, schema, summaries, and basic statistics

Indexing and selection

- Label-based and position-based access
- Filtering, sorting, and conditional selection

4. 4. Cleaning, Transformation and Features

Data quality and preprocessing

Handling missing data

- Detect missing values and choose removal or imputation appropriate to the data and task

Dealing with inconsistent and dirty data

- Data type conversion
- String normalization
- Review duplicates and outliers; remove or transform only with a documented justification
- Feature scaling and normalisation; learn parameters from training data when preparing ML inputs
- Categorical encoding appropriate to the model, with consistent training and later-data handling

Feature engineering and transformations

- Creating new features from existing data

Applying functions to data

- Row-wise and column-wise transformations

Grouping and aggregation

- GroupBy patterns used in analytics and AI

Merging and joining datasets

- Inner, outer, left, and right joins

Reshaping data

- Pivoting and melting

5. 5. Exploratory Analysis and Larger Datasets

Exploratory data analysis

- Understanding distributions and trends
- Correlations and relationships, without inferring causation from association alone
- Anomalies, representativeness and data-leakage risks
- Summary statistics and descriptive analysis
- Preparing insights for downstream modeling

Memory and scale

- Performance considerations in Pandas
 - Memory optimization techniques
 - Chunked processing where the transformation permits independent or bounded-state chunks
- When to move beyond Pandas
- High-level overview of scalable tools and when an in-memory workflow is unsuitable

6. 6. Model Hand-Off and Maintainable Pipelines

Preparing data for machine learning

- Train/validation/test separation before fitting preprocessing; time/group-aware splits where appropriate
- Feature matrices and target variables
- Avoiding leakage, inconsistent transformations and unsupported cleaning assumptions
- Reproducibility and data versioning basics
- Hand-off from data manipulation to modeling workflows

Code structure and validation

- Structuring data manipulation code
- Writing readable and maintainable data pipelines
- Common pipeline pitfalls and operational limitations
- Validating data before model training

