Coding for Economists

- Hilary Term 2025
- Lecturer: Jian Cao

Aims of Training

This training focuses on introducing Python programming concepts and applying them to data analysis, visualization, analysis, machine learning, and handling large datasets relevant to economics. The training aims to provide practical skills that participants can apply directly to their research and professional work.

Training Delivery

The training consists of 10 sessions of 1 hour each during the teaching weeks, and 5 intensive sessions of 2 hours each after the teaching weeks. The sessions use hands-on coding sessions with real economic data examples.

Learning Outcomes

Following the completion of this training, you will be able to:

- Understand the basics of Python coding—importing packages, working with built-in data types, controlling program flow with loops and conditionals, and writing user-defined functions.
- Retrieve cross-sectional and time series data from various sources (websites, public APIs, local files), store the data in Pandas objects, and transform/clean datasets (merging, reshaping, imputing) to prepare for analysis.
- Master the fundamentals of machine learning (Random Forests, K-Means, neural networks) with appropriate pre-processing (scaling, encoding), and leverage pre-trained language model (e.g., BERT) for sentiment analysis and topic modeling on textual data.
- Adopt version control (Github) and debugging to ensure efficient and replicable code. Develop skills for handling large datasets via memory management, parallel computing, and high performance computing.

Training Outline

Ten One-Hour Sessions (Foundational Skills)

Session 1: Python Fundamentals & Jupyter Notebooks

- **Anaconda/Jupyter environment**
  - Installing Anaconda
  - Navigating Jupyter Notebook
  - Markdown essentials
- **Basic Python syntax**
  - Variables, data types
  - Basic arithmetic, string operations

Session 2: Control Structures and Functions

- **Control structures**
  - `if/elif/else`
  - `for`/`while` loops
- **Functions**
  - Defining and calling functions
  - Variable scope
- **Importing and using modules**
  - Standard library modules
  - Installing and importing external modules

Session 3: File Handling Basics

- **File handling**
  - Reading/writing CSV, Excel, and Stata files
  - Best practices for file I/O
- **Introduction to data collection**
  - Basic API GET requests and JSON handling
  - Overview of web scraping concepts

Session 4: Data Manipulation with NumPy and Pandas

- **NumPy basics**
  - Arrays, shapes, indexing, vectorization
- **Pandas basics**
  - Series, DataFrames, indexing, slicing
  - Common operations (adding columns, merging DataFrames)

Session 5: Data Cleaning, Reshaping, and Handling Missing Data

- **Data cleaning**
  - Filtering, outlier handling
  - String operations in Pandas
- **Reshaping data**
  - Melting, pivoting, stacking/unstacking
- **Handling missing data**
  - Find the correct method
  - MICE, hot-deck

Session 6: Data Visualization and Exploratory Analysis

- **Data visualization**
  - Basic plots with Matplotlib (line, bar, scatter)
  - Introduction to Seaborn (e.g., `regplot`, `pairplot`, distribution plots)
- **Exploratory analysis**
  - Descriptive statistics
  - Identifying relationships and patterns (correlations, `groupby`)

Session 7: Machine Learning Concepts & Preprocessing

- **Intro to machine learning**
  - Supervised vs. unsupervised learning
  - Terminology: features, labels, train/test split
- **Preprocessing**
  - Feature scaling (min-max, standard scaler)
  - Encoding categorical variables

Session 8: Overview of Key Machine Learning Algorithms

- **Random Forests**
  - Ensemble methods, bagging
- **K-Means clustering**
  - Unsupervised learning basics, choosing k

Session 9: Applied Machine Learning & Visualization

- **Integrating learned ML algorithms with visualization**
  - Mini-project: load a dataset, clean it, apply an ML model (e.g., Random Forest), and visualize results
  - Hands-on coding practice

Session 10: Debugging, Version Control

- **Debugging & Best Practices**
  - Error handling (`try-except`), unit testing basics
- **Version control**
  - Github fundamentals
  - Reproducibility in coding projects

Five Two-Hour Sessions (Advanced Topics)

Advanced Session 1: Web Scraping & Advanced File Handling

- Deep dive into web scraping: HTML parsing with BeautifulSoup
- Advanced file handling: working with complex file types, efficient data I/O strategies

Advanced Session 2: Enhanced Machine Learning & Visualization Methods

- Exploration of additional ML algorithms and methods beyond basics
- Advanced visualization techniques with libraries like Seaborn/Plotly
- Case studies: applying advanced ML models to economic data

Advanced Session 3: Text Analysis with Pretrained Models

- Sentiment analysis using NLP: concepts, text preprocessing, using pretrained sentiment classifiers
- Topic modeling: introduction to BERTopic, Hugging Face transformers, interpreting and visualizing topics

Advanced Session 4: Large Datasets and Performance Optimization

- Strategies for working with large datasets: memory management, efficient data structures
- Parallel and distributed computing approaches in Python

Advanced Session 5: Introduction to TCD High Performance Computing (HPC)

- Architecture overview, data management
- Environment setup, job submission
- Best practices for HPC

Readings

- Wes McKinney (2022) Python for Data Analysis. O'Reilly
- Aurélien Géron (2022) Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly