Introduction to Python for Social Data Science

Maksim Zubok

DPIR, Oxford University

Hillary Term 2025

Course Overview

This course introduces Python programming with a focus on applications in data analysis. Participants will learn to write efficient, clean, and readable code while adhering to widely recognised Python coding conventions. The curriculum combines conceptual instruction and hands-on practice, split evenly between the two formats.

Seminar Room: Room B
Time: Tuesdays, 4:00 PM - 6:00 PM
Python Clinic: Tuesdays, 6:00 PM - 7:00 PM (Optional, for personalised assistance and project consultations.)

This course is a prerequisite for the "Large Language Models for Social Sciences" course, offered in Trinity Term. The suggested readings are available on SOLO. Although optional, reading before sessions is highly recommended to enhance your understanding of the material.

Course Policies and Prerequisites

This course requires no prior experience with Python or programming. Familiarity with R or other programming languages may make the course easier but is not necessary.
We will use Google Colab as the primary coding environment, enabling you to write and run Python code directly in your browser without requiring local setup. You would need

to register with your google account or create one to use Colab.

This course is optional and there are no graded assignments. However, weekly tasks will be provided to support your learning. You are encouraged to complete these tasks and bring questions to the seminar or clinic.

Accessibility and inclusion: If you need disability-related accommodations or special arrangements, please contact me as soon as possible or discuss your needs after first class. For formal accommodations, a Student Support Plan must be obtained from DAS and sent directly to me. Nursing parents are welcome to bring their babies to class. Snacks and drinks are allowed, but please reserve full meals for breaks.

The Python Clinic provides additional support for debugging code, clarifying concepts, and receiving feedback on individual projects. They are for you and you are welcome to attend, regardless of previous experience or level of confidence.

Course Outline

Week 1: Introduction and Basic Concepts

In this session, we will cover the basic syntax and logic of Python, introduce the concept of object-oriented programming (OOP), Python's data types, and essential programming concepts such as variables, loops, conditionals, and functions.
Suggested Readings:

- McKinney, Wes. (2022). Python for Data Analysis, Chapters 2 & 3.
- Python Enhancement Proposals (PEP8).
- Overview of Google Colab features.

Week 2: Numerical Python (NumPy)

NumPy is an essential Python library that enables efficient computations across arrays of data. For that reason, NumPy constitutes the backbone of many key libraries, such as

Pandas or PyTorch, which we will explore later. In this session, we will learn how NumPy handles arrays, including their creation, manipulation, and computations. In addition, we will implement a linear regression algorithm from scratch for demonstration purposes.
Suggested Readings:

- McKinney, Wes. (2022). Python for Data Analysis, Chapter 4.
- James, Gareth & Witten, Daniela & Hastie, Trevor & Tibshirani, Robert & Taylor, Jonathan. (2023). Introduction to Statistical Learning, Chapters 3.1 & 3.2.
- VanderPlas, Jake. (2022). Python Data Science Handbook, Chapters 4-12.
- NumPy: the absolute basics for beginners.

Week 3: Pandas

Pandas is a powerful Python library for data manipulation and analysis. This session will cover creating, transforming, and analysing datasets using Pandas. We will also explore its integration with other libraries for seamless data workflows.
Suggested Readings:

- McKinney, Wes. (2022). Python for Data Analysis, Chapters 5 & 6.
- VanderPlas, Jake. (2022). Python Data Science Handbook, Chapters 13-24.
- Pandas Documentation and User Guides.
- Pandas Comparison with R.

Week 4: Visualization with Matplotlib and Seaborn

Data visualization is crucial for understanding trends and patterns in data. This session focuses on creating informative and aesthetically pleasing plots using Matplotlib and Seaborn. We will cover line plots, scatter plots, bar charts, and customisation options.
Suggested Readings:

- McKinney, Wes. (2022). Python for Data Analysis, Chapter 9.
- VanderPlas, Jake. (2022). Python Data Science Handbook, Chapters 25-36.

- Matplotlib Documentation and User Guide.
- Seaborn User Guide and Tutorial.

Week 5: Python Classes

We will dive deeper into OOP by introducing Python classes, methods, and inheritance. The session will demonstrate how OOP principles can help structure complex code, making it more readable and reusable.

Suggested Readings:

- Martelli, Alex & Martelli Ravenscroft, Anna & Holden, Steve & McGuire. (2023). Python in a Nutshell, Chapter 4.
- Object-Oriented Programming (OOP) in Python.
- Python Documentation and Tutorial on Classes.

Week 6: Machine Learning with SciKit Learn

This session will introduce machine learning workflows using SciKit Learn. We will cover data preprocessing, training simple models like regression and classification, evaluating model performance, and cross-validation techniques.

Suggested Readings:

- VanderPlas, Jake. (2022). Python Data Science Handbook, Chapters 37-40.
- James, Gareth & Witten, Daniela & Hastie, Trevor & Tibshirani, Robert & Taylor, Jonathan. (2023). Introduction to Statistical Learning, Chapters 2.1 & 2.2 & 4.1-4.3 & 5.1.
- Scikit Learn User Guide.

Week 7: Word Embeddings, Gensim, NLTK and SpaCy

In this session, we will dive into Natural Language Processing and explore Python libraries designed to process and analyse text. We will learn about word embeddings and topic modeling, with applications relevant to political science research.

Suggested Readings:

- NLP Course For You: Word Embeddings.

- Natural Language Toolkit (NLTK) Documentation.

- spaCy Documentation.

- Gensim Documentation and Tutorials.

Week 8: Deep Learning with PyTorch

This session will provide an introduction to deep learning using PyTorch. We will cover tensor operations, building neural networks, and implementing a simple deep learning model to demonstrate its application.
Suggested Readings:

- Patterson, Josh & Gibson, Adam. (2017). Deep Learning with Python, Chapters 1-3.

- PyTorch Tutorials.

- James, Gareth & Witten, Daniela & Hastie, Trevor & Tibshirani, Robert & Taylor, Jonathan. (2023). Introduction to Statistical Learning, Chapters 10.1-10.2 & 10.7.

- Goodfellow, Ian & Bengio, Yoshua & Courville, Aaron. (2016). Deep Learning, Chapter 6.