Data Collection for Social Scientists: Leveraging Public APIs with Python

Zach Dickson (LSE)

Learning Objectives

This workshop will equip students with the skills to programmatically access data from public REST APIs, transform this data into analysis-ready datasets, and critically consider dataset design for social science research.

Specifically, by the end of this workshop, students will be able to:

- Understand the fundamentals of APIs and how they can be used to access data.
- Use Python's `requests` library to interact with REST APIs.
- Fetch, clean, and transform data from the UK House of Commons API.
- Apply data cleaning techniques to prepare datasets for analysis.
- Critically evaluate the design of datasets for social science research.

Target Audience: MSc and PhD students in Social Sciences or related fields who are interested in learning how to collect and analyze data from public APIs using Python.

Prerequisites:

- Basic to intermediate proficiency in Python (understanding data types, lists, dictionaries, loops, functions) (see below for resources).
- Familiarity with the Jupyter Notebook or a similar Python IDE (see below for resources).
- A conceptual understanding of social science research methodologies.

Github Repository:[1] https://github.com/z-dickson/my580-creating-datasets-public-APIs

[1] All content be made available prior to the workshop

Workshop Schedule

(10am-3pm with 1 hour lunch break)

The workshop is divided into two main sessions, each of which include several modules. The first session focuses on the foundational concepts of APIs and how to interact with them using Python's `requests` library. The second session introduces the UK House of Commons API, guiding students through fetching, cleaning, and transforming data for analysis.

- ***Session 1: Foundations & API Interaction*** (~2 hours)
  - **Module 1: Introduction to APIs** (20 minutes)
    - Welcome and Introduction (5-10 minutes)
    - What are APIs? (5-10 minutes)
    - Why use APIs? (5-10 minutes)
  - **Module 2: Understanding REST APIs & Python's `requests`** (45 mins)
    - URL and Endpoint Basics (10 minutes)
    - HTTP Methods (GET, POST, PUT, DELETE) (10 minutes)
    - Request Parameters and Headers (10 minutes)
    - Response Formats (JSON, XML) (10 minutes)
    - Status Codes and Error Handling (5 minutes)
  - **Module 3: `Requests` Module in Python** (55 minutes)
    - Installing the `requests` library (5 minutes)
    - Making a GET request (10 minutes)
    - Handling JSON responses (10 minutes)
    - Error handling and status codes (10 minutes)
    - Practical exercise: Fetching data from a public API (10 minutes)

Lunch Break (1 hour)

- ***Session 2: Introduction to the UK House of Commons API*** (~2 hours)
  - **Module 4: Introduction to the UK House of Commons API** (50 minutes)
    - Overview of the UK House of Commons API (5 minutes)
    - Understanding the API documentation (5 minutes)
    - Authentication (10 minutes)
    - Fetching data from the API (10 minutes)
    - *Practical exercise:* Fetching and exploring data from the UK House of Commons API (20 minutes)
  - **Module 5: Data Transformation and Cleaning** (60 minutes)
    - Introduction to data cleaning (5 minutes)
    - Using Python libraries for data manipulation (e.g., pandas) (20 minutes)
    - Structuring data for analysis (10 minutes)
    - Storing and exporting data (10 minutes)
    - *Practical exercise:* Cleaning and transforming data from the UK House of Commons API (15 minutes)

Resources

Installing Python and Jupyter Notebook

To participate in this workshop, you will need to have Python installed on your computer, along with the Jupyter Notebook environment. Python is a versatile programming language that is widely used in data science, and Jupyter Notebooks provide an interactive environment for writing and running Python code.

Python can be installed in various ways, but the most common and user-friendly method is through the Anaconda distribution, which includes Python, Jupyter Notebook, and many useful libraries for data science. If you are comfortable accessing Python through a different IDE or method, feel free to do so.

Video Tutorials for installation:

- Installing Python + Jupyter Notebook on Windows 11
- Installing Python + Jupyter Notebook on MacOS

Intro to Python:

There are many resources available to learn Python. You will not need to be an expert, but you should be comfortable with basic data types (strings, integers, floats), lists, dictionaries, loops, and functions. Below are some resources to get you started. In my experience, the best way to learn is to practice as you go, so I recommend working through at least a few of the exercises in these (or similar YouTube) tutorials:[2]

- Python for Everybody
- Python Crash Course for Beginners
- Python for Data Science
- Learning Python for Data Analysis

[2]If you are already comfortable with Python, you can skip this section.