Introduction to Python

Social Science Methods Workshops 2021

Lund University

Instructor: Nils Holmberg, Strategic Communication, Lund University

Contact: nils.holmberg@isk.lu.se

Dates: September 27 – October 1, at 09-12 am (schedule link)

Format: Campus-based, Eden 235

1   Introduction

Python is a programming language that has gained popularity in all types of data science. On this introductory course, we are aiming at getting acquainted with the basic syntax of Python, as well as learning how to extend this basic functionality by calling powerful modules built by members of the large Python user community. Thus, we take a "standing on the shoulders of giants" approach to Python.

Another principle that has guided the design of this course is that participants should gain an understanding of how Python can be used to accomplish fairly common research tasks within the social sciences more effectively and more transparently. Such applications of Python includes collecting and analyzing survey data, and performing content analyses on textual observations.

While we don't think any previous programming experience should be required on this course, we do think that participants will benefit from some prior knowledge of quantitative methods and data analysis, e.g. using Excel. To follow the course, you will need a personal or work laptop computer with internet access, on which you have administrative privileges to install some software.

2   Course materials

Primarily, course participants are encouraged to install a Python environment beforehand. The environment used on this course will be Anaconda3 and it consists of several modules, including support for interacting with Jupyter notebooks, which will be the main format of instruction on the course. If you plan to use a LU laptop, please check with IT support how to install Anaconda3!

It is also preferable if you are able to install the Psychopy3 2020 or 2021 software on your computer. Psychopy3 is a Python software used to design online experiments and collect data. In addition, this software can be used to undertake online surveys. Both Anaconda3 and Psychopy3 should be available through the IT support Software Center at LU, although the exact version numbers might differ slightly[1].

---

[1] Please also have a look on YouTube concerning installation and usage of Anaconda3 and Psychopy3.

3 Course contents

In this introduction to Python for social scientists you will learn how to use the Python programming language as a unified platform for handling multiple tasks and workflows connected to running research studies, collecting data, and analyzing numeric and textual data. In the table below we present an overview of the course topics covered each day, and further down we expand on each module in more detail.

Course date	Course topic	Python packages
210927	the anaconda3 python environment, read and process tabular data	numpy, scipy, pandas
210928	natural language processing and analyzing textual content	spacy
210929	experimental design, online surveys, and data collection	psychopy
210930	data analysis and data visualization using bar charts and time series	matplotlib, seaborn
211001	reproducible data analysis, sharing and collaborating on python code	jupyter

First off, we will familiarize ourselves with the Anaconda3 environment. This software consists of several modules, including the Python command prompt, a text editor for writing scripts, and an interactive development environment (IDE). We will practice writing basic flow control and calling functions, then we will analyze and process survey data from the European Social Survey (Géron, 2019; McKinney, 2012; VanderPlas, 2016).

Next up, we will focus on analyzing and processing textual data. In order to do so, we will need to extend the basic functionality of Python by adding a package called spaCy, which will allow us to leverage natural language processing (NLP) using machine learning libraries. We will practice on text preprocessing, cleaning, tokenization, and classifying movie reviews and/or news articles (Sarkar, 2016).

Another important application of Python is within experimental design, where the objective can be to present research participants with some kind of visual stimuli and collect various types of responses, ranging from self-reported questionnaires to physiological measurements. On day 3, we will practice building an online survey experiment and collect data using Psychopy3 package (Peirce & MacAskill, 2018).

On day 4, we will continue performing basic data summarization using descriptive statistics, but now we will put more emphasis on how to use quantitative data to produce compelling and intuitive visualizations. Using the matplotlib and seaborn packages, we will learn how to aggregate survey data from the European Social Survey study and present it as bar charts, pie charts, and time series (Embarak et al., 2018). If there is time, we will also use the scikit-learn package to perform multiple regression analyses in Python.

We will finish off the course by focusing on strategies for continuing to develop your Python data science skills and publishing reproducible data analyses using Jupyter notebooks. Notebooks are an excellent way for documenting, sharing and collaborating on Python code, and hopefully these practices will help you understand your own code and make your social science research more transparent and accessible (Mueller & Massaron, 2019; Toomey, 2016).

Figure 1: python logo

References

Embarak, D. O., Embarak, & Karkal. (2018). Data analysis and visualization using python. Springer.

Géron, A. (2019). Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

McKinney, W. (2012). Python for data analysis: Data wrangling with pandas, numpy, and ipython. " O'Reilly Media, Inc.".

Mueller, J. P., & Massaron, L. (2019). Python for data science for dummies. John Wiley & Sons.

Peirce, J., & MacAskill, M. (2018). Building experiments in psychopy. Sage.

Sarkar, D. (2016). Text analytics with python. Springer.

Toomey, D. (2016). Learning jupyter. Packt Publishing Ltd.

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.".