The Python Data Science Ecosystem: Software Needed, Recommended Reading

6G7V0026 Principles of Data Science

Luciano Gerber

Senior Lecturer
Department of Computing and Mathematics
Faculty of Science and Engineering
Manchester Metropolitan University

Block 1, Week 1 (20 to 24-Sep-21)

Software Infra-Structure

Python Notebooks and Google Colaboratory

- Our **main data science environment** is Google Colaboratory (or Colab). It is accessible with a free Google account.
- We will implement **Python Notebooks** for programming and documenting our data science solutions.
- For an overview of what Colab is, what notebooks look like, and what you can do with them, please visit the following useful resources:
  - Introduction to Google Colab
  - Overview of Colaboratory Features
  - Markdown Guide
  - Optionally:
    - External Data: Drive, Sheets, and Cloud Storage

Google Colaboratory: Cloud-based, Online, Collaborative

- Google Colaboratory is an environment that suits our blended learning approach quite well.
  - It is cloud-based, and one only needs a browser to create and run their notebooks.
  - Can be easily shared and annotated by others.
  - Could also be run locally on one's specialist hardware.
- The notebooks created in Colab are based on **Jupyter notebooks**, which have widespread use in the Data Science community.

A Local Installation with the Python Data Science Ecosystem

- Alternatively, you could rely on a **local installation** (that is, on your own or our lab computers) of the **Python Data Science Ecosystem** .
  - Not required for the core tasks in the unit, but will be useful for you as a data scientist.
- In the University labs, the supported installation in on **Linux**. If you prefer **Windows**, that should be fine - you will need to use the `anaconda` installation.

- You will find a Python 3 installation with essential modules for data science (e.g., `pandas`, `numpy`, `matplotlib`, `scikit-learn`).
- Instead of Google Colaboratory, one can use `jupyter`, the browser-based Python notebook system.
- Remote Access to our Lab Machines can be done via Leostream.

Additional Tools for a Local Installation

- One might also like to try `ipython`, the Python REPL (interactive, exploratory) terminal-based system.
- When not using `jupyter`, a good **programming text editor** is essential (e.g., `Visual Studio Code`, `Atom`, `Brackets`, `Sublime Text`, `Notepad++`, `emacs`, `vim`). If you would like, an IDE (e.g., `PyCharm`, `Spyder`) would do a good job too.

Setting Up The Environment On Our Own Machine

- If you know what you are doing and would like to install Python/Jupyter on your machine, here are some suggestions. Please note that neither the unit team nor the specialist technicians would be able to provide support on this.
- Regardless of platform (e.g., Windows, Linux, Mac), the Anaconda distribution is probably the best way forward for installing and managing your Python installation with required modules for the Data Science ecosystem.
- In Windows, you might want to invoke Jupyter (or `ipython`/`python`) via the Anaconda shell. In Linux and Mac, similarly to in our labs, use the terminal.

Learning Python and the Associated Data Science Ecosystem

Essential and Additional Reading

- Our main text is the **Python Data Science Handbook**
  - Covers most of what we need from the Python Data Science Ecosystem, from manipulating, processing, and visualising datasets to building and evaluating machine learning models.
  - Our library has it in print, I believe, and you can access it via O'Reilly Learning using your MMU SSO.
- A Whirlwind Tour of Python (notebooks accompanying the book are available here).

Just Enough Python

- Our approach is **just enough Python** for implementing solutions for data science problems. Students will learn the basic Python needed as we go through data science scenarios, examples, and exercises.
- Little experience with programming?
  - You might want to complete the sections *Python Basics*, *Python Lists*, and *Functions and Packages* of the DataCamp's Intro to Python for Data Science
    - You would have received an invitation from DataCamp to join for free.
    - please feel free to complete *NumPy* in your independent study time.
  - a more comprehensive, introductory guide is Think Python.

- Have already some experience with programming, or would like to look further?
  - You could try A Whirlwind Tour of Python (notebooks accompanying the book are available here).
  - My suggestion is to start with the following selected sections:
    - Introduction
    - How to Run Python Code
    - A Quick Tour Of Python Language Syntax
    - Basic Python Semantics: Variables and Objects
    - Basic Python Semantics: Operators
    - Built-In Types: Simple Values
    - Built-In Data Structures
    - Control Flow
  - A more extensive Python guide (for independent study time) is Automate the Boring Stuff with Python.
- Experienced? Transitioning to Python?
  - this is a lightning-fast introduction might be useful.