Install Python and Jupyter Notebook

IEOR 142

August 31, 2021

1 Overview

This note covers how to set up the coding environment needed for IEOR 142 and as a data scientist. To begin with, you need to:

- Install Python 3: `https://www.python.org/downloads/`

Next, you need to:

- Install the JupyterLab/Jupyter Notebook[1]
- Install data analytics packages (e.g., Statsmodels for statistical analysis and Matplotlib for plotting)

Below we introduce three ways to do so. You can choose the one that you prefer.

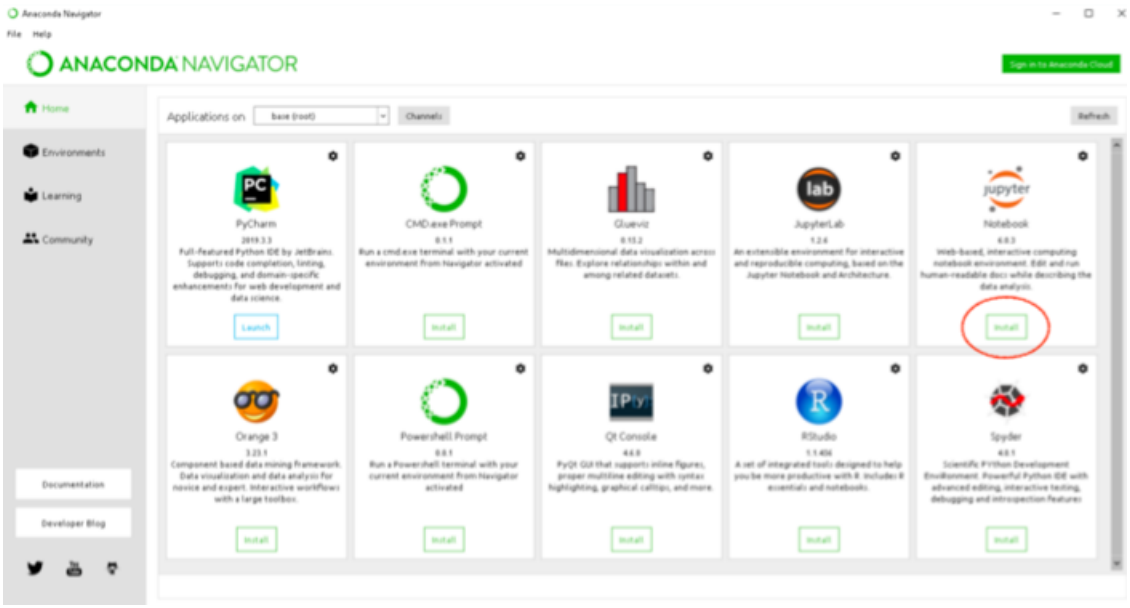
1. Use Anaconda
   - This is the easiest way to get started and perform data analytics in Python. It provides a graphical user interface and is less abstract to a beginner.
2. Use Berkeley DataHub
   - This allows you to run your program on the university's server instead of locally on your own computer.
3. Optional: Use Terminal + a package manager (pip or conda)
   - This is a slightly more sophisticated and abstract approach, but can come in handy if you want to develop skills in software development. It requires some extra work upfront, but provides more flexibility afterwards.

In sections below, we provide details on the three methods mentioned above.

[1] JupyterLab is Jupyter's Next-Generation Notebook Interface. It contains more features than the classic Jupyter Notebook. Either one would work for this course. Read more about these products at `https://jupyter.org`

1.1 Anaconda

1. Install Anaconda: `https://docs.anaconda.com/anaconda/install/`.
2. Open the Anaconda Navigator as an Application (on MacOS) or a software (On Windows) Anaconda Navigator starts automatically with a "base (root)" environment. You can either use this default environment, or create a new specific environment for this class. We recommend you do the latter by following the simple instructions in this link: Managing Environment.
3. Once you have your environment ready, check if Jupyter Notebook is installed. This should be the case if you created a new environment and ticked the Python ox. If not installed, please proceed to install it.



4. Launch Jupyter Notebook. It will open a new browser window for you.
5. Go to the 'Files' tab if not already there, select 'New': you can create a new 'Folder' in any specified path you desire. Subsequently create a new 'Python 3' notebook, click on it to open, and you are ready to code!

1.2 Berkeley DataHub

Using Berkeley DataHub, you can run your code on the university server, which might provides better computing power than your personal laptop.

1. Access Berkeley DataHub via the following link: `https://datahub.berkeley.edu/`
2. Enter your student credentials, and authorize datahub access to your account
3. Go to the 'Files' tab if not already there, select 'New': you can create a new 'Folder' in any specified path you desire. Subsequently create a new 'Python 3' notebook, click on it to open, and you are ready to code!
4. To upload data files to the server, simply drag the file from your local directory to the directory on the web page. It will be uploaded automatically.

1.3 Optional: Terminal + A Package Manager (pip or Conda)

At a high level, the learning objective of this approach is to have you set up and get familiar with the following tools in a sequential order:

- Terminal (macOS/Linux) or PowerShell (Windows) - a program that allows you to interact with your computer by entering commands
- pip or conda - a package manager
- Jupyter Notebook - a web-based interactive computational environment

1.3.1 The Terminal and Command Line Commands

You do not need a deep understanding of command line commands in the terminal, but the following techniques will come in handy when we start working with data sets:

1. Navigating through different directories (using **ls** and **cd** commands)'
2. Installing Python packages;
3. Referring to the documentation to learn how to use a tool. (Note: everything we introduce in this file has a large supporting community and very good documentations online.)

*If you are not familiar with the terminal, we recommend you to check out an introductory webpage (`https://inst.eecs.berkeley.edu/~cs61a/fa20/lab/lab00/`), created by the wonderful CS 61A team. It covers how to set up the coding environment both on MacOS and on Windows. One caveat is that we will be using the Jupyter Notebook as a text editor in IEOR 142, so you do not need to worry about installing the text editors recommended for CS 61A (such as Atom, Sublime, or Visual Studio code).

Once you get comfortable working with the terminal and have Python3 installed, you are ready to move on to the next section, which provides guidance to setup the package manager and Jupyter.

1.3.2 Package Manager - pip / Conda

There are many packages built for Python, which makes it very powerful for data science. Practitioners normally use a package manager to help manage different packages. The two most common ones are "pip" and "conda". There are similar, with minor differences.[2] For the purpose of this class, either one would do.

- Pip is the package installer for Python. It is already installed if your Python version is greater than 3.4.[3] Additional information can be found in the documentation at `https://pip.pypa.io/en/stable/`
- Conda works in a similar way, but gives you access to a broader set of packages, not limited to Python. The documentation and installation guide can be found at: `https://docs.conda.io/projects/conda/en/latest/`

Having pip or Conda installed allows you to install other packages with a single command line.

[2] `https://www.anaconda.com/blog/understanding-conda-and-pip`
[3] Installation - Pip. `https://pip.pypa.io/en/stable/installing/`

1.3.3 Jupyter Notebook

Install the classic Jupyter Notebook or JupyterLab following the steps on the official website.