Introduction to Python

Data Mining

Prof. Sujee Lee

Department of Systems Management Engineering

Sungkyunkwan University

Python

- **Why Python?**

  - Python is a widely used, general purpose programming language.

  - Easy to start working with.

  - Scientific computation functionality similar to Matlab and Octave.

  - Used by major machine learning frameworks such as Scikit-Learn

  - Used by major deep learning frameworks such as PyTorch and TensorFlow.

TIOBE Programming Community Index
Source: www.tiobe.com
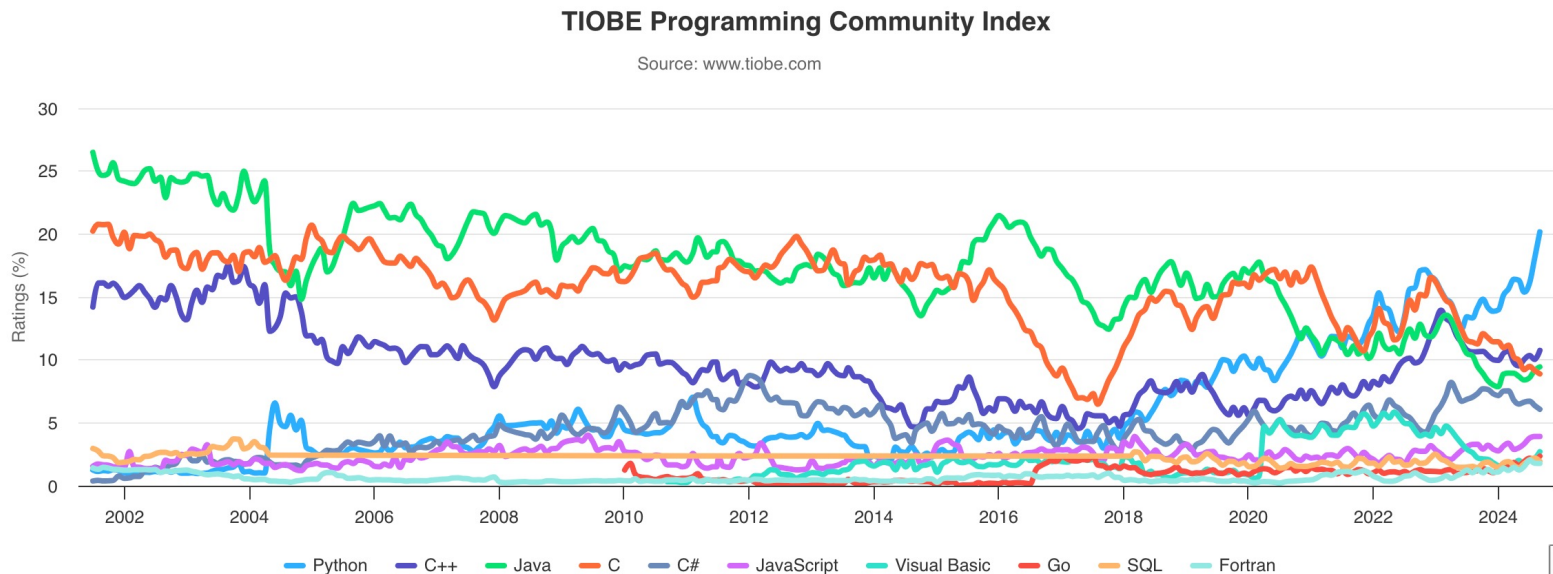
Python Environment

- **Installation**

  - As long as you can use Python and Jupyter notebook, any method can be used!

- **Recommendation**

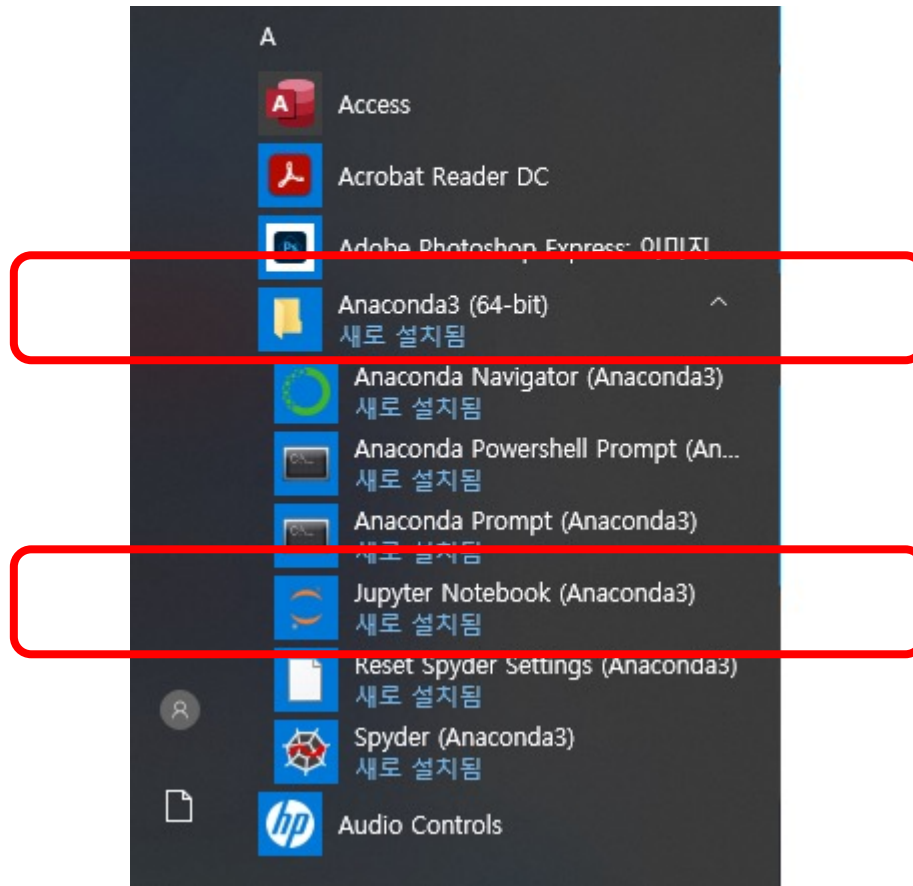 - Quick Installation : Install Anaconda3

    - https://www.anaconda.com/products/individual

  - Quick Use : Google Colab

    - https://colab.research.google.com/

    - Note : I'll use Jupyter notebook. Shortcut keys may be different between Jupyter and Colab. But, all functionalities are the same (at least for this course)
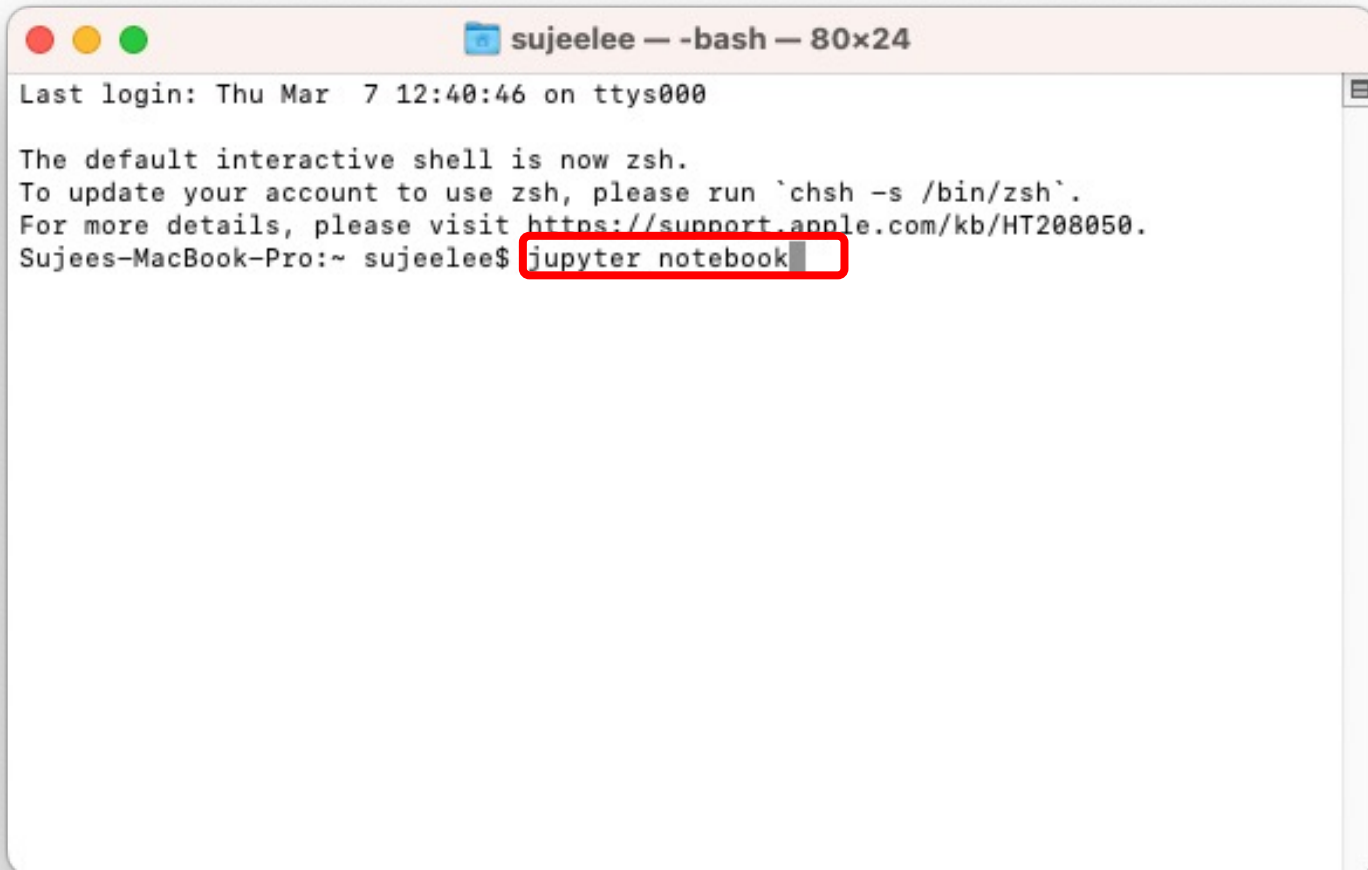
Python Environment

- After Anaconda3 is installed (Windows OS)

Python Environment

- After Anaconda3 is installed (Mac / Linux OS)

    - In Terminal,

Python Environment

- In Google Colab

Requirements

- **Python Packages for Data Analysis**

  - **Jupyter Notebook** : Browser-based interactive programming environment

  - **NumPy** : Basic array operation and manipulation

  - **SciPy** : High-level scientific computing (advanced linear algebra, optimization, statistical distributions, etc.)

  - **Pandas** : Data wrangling and analysis (Excel files, CSV files, SQL, etc.)

  - **Matplotlib & Seaborn** : Visualization

  - **scikit-learn** : Machine learning (on top of NumPy and SciPy)

Pandas

- **Pandas : An essential library for handling structured data in table or database formats.**

  - Data is represented in a table format using a data structure called DataFrame.

  - Using Pandas, you can create table-like data or load files such as CSV and Excel.

  - It is convenient for exploring and modifying data.

```
In [4]:  ## Create a simple dataset
         data = {'Name': ["John", "Anna", "Peter", "Linda"],
                 'Location' : ["New York", "Paris", "Berlin", "London"],
                 'Age' : [24, 13, 53, 33]
                 }

         df = pd.DataFrame(data)

In [5]:  df

Out[5]:
```

	Name	Location	Age
0	John	New York	24
1	Anna	Paris	13
2	Peter	Berlin	53
3	Linda	London	33