Week 1 (LAB Worksheet)

Data Visualisations with Python Libraries

Environment Setup for Data Visualisation

Python-based computing environments are one of the most popular options among data scientists. Python provides standard, in-built functionalities that can be extended by importing third-party packages. The most common third-party libraries that we will use are **Matplotlib, NumPy**, and **Seaborn** etc.

1. Jupyter Notebook

We will use the Jupyter Notebook to create and share visualisations in this module. The classic Notebook interface is a document-oriented interface that allows you to create, view, and execute Python code and import various libraries into Jupyter Notebook.

You can use **Jupyter Notebook** in your web browser: https://jupyter.org/try

I recommend installing Jupyter Desktop on your local machine.

2. Anaconda Environment

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands. Follow these instructions to install Anaconda on your local machine:

https://docs.anaconda.com/free/anaconda/install/

Activity 1: Hands on with Matplotlib Figure

In this activity, you will use the **anatomy of a figure** to familiarise yourself with several matplotlib elements composing a figure, then create your own custom figure(s) templates.

Useful Tutorials

Matplotlib Tutorial

Procedural vs. Object-Oriented Plotting in Matplotlib

Matplotlib Quick start guide

Activity 2: Data Visualisation in Python

There is a wide array of libraries you can use to create Python data visualisations, including **Matplotlib, Seaborn, Plotly**, and others. These libraries help you understand data in various ways: distribution, mean, median, outlier, skewness, correlation, and spread measurements. In order to see what you can do with a Python visualisation, in this activity, we try to create visualisations combined by descriptive statistics.

Painters Dataset

The dataset of painters contains assessments of 54 classical painters on four characteristics: composition, drawing, colour, and expression. The scores are due to the eighteenth-century art critic de Piles.

Tasks:

(a) What plot would you draw for showing the distribution of all the values together? What conclusions would you draw?

(b) Draw a display to compare the distributions of the four assessments. Is it necessary to scale the variables first? What information might you lose, if you did? What comments would you make on the distributions individually and as a set?

(c) What would you expect the association between the scores for drawing and those for colour to be? Draw a scatterplot and discuss what the display shows in relation to your expectations.