MINI PROJECT 1: DATA INGESTION AND WRANGLING

Glossary

Data Ingestion is the process of obtaining, importing, and processing data for later use or storage in a database. This can be achieved manually, or automatically using a combination of software and hardware tools designed specifically for this task. IBM

Data Wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. Wikipedia

Objective

The objective of this assignment is to enable you to build and train your skills in data collection, exploration and preprocessing.

Tasks

Data comes from various origins in different formats, such as:

PDF
CSV
TXT
XLS
JSON
DOC
@
API

Before being analysed, it needs to be extracted, loaded, and transformed (ETL) into an appropriate Python data structure, such as pandas data frame or numpy numeric array.

Your tasks are:

1. Choose minimum three of the source data formats, displayed by the icons above, feel free to add others if you have any in mind.
2. Write Python functions for loading data from each of the selected external types of sources into a Python data structure.
3. Collecting data by use of your functions.
4. Explore and clean the collected data as needed.
5. Apply anonymisation if necessary.
6. Apply visualisation techniques, as many as appropriate.
7. Store the code and the visuals with brief explanation text in one Github repository, which you will be using for the rest of the course.
8. Upload a link to the repository in Moodle’s flow by 02/09/2024, 23:59.

Notes:

It is a teamwork task.

The original solutions (data loading functions) will be collected in the class repository and made available for shared use.