COMP47670 Assignment 1
Spring 2025

Summary:

The objective of this assignment is to extract a dataset from a set of web pages and use Python to prepare, analyse, and derive insights from the collected data.

The assignment should be implemented as two Jupyter Notebooks (not script files). Your notebooks should be clearly documented, using comments and Markdown cells to explain the code and interpret the results of your analysis.

Tasks:

Complete the following three tasks in two separate notebooks:

1. **Data Collection**
    - Choose one of the four data sources listed here:
      http://mlg.ucd.ie/modules/python/assignment1
    - In your first notebook, apply web scraping in Python to collect and parse all of the data from your chosen source.
    - Save the collected data in an appropriate format for subsequent analysis.
2. **Data Preparation and Analysis**
    - In your second notebook, load the saved dataset from Task 1 into an appropriate data structure for use as an Analytics Base Table (ABT).
    - Apply any data preprocessing steps that might be required to clean, filter or transform the ABT before analysis. Use Markdown cells to explain and justify each preprocessing step.
    - Analyse, characterise, and summarise the cleaned ABT, using visualisations where appropriate. Use Markdown cells to explain each step and interpret the results.
3. **Discussion**
    a. At the end of your second notebook, discuss the following aspects of your assignment in Markdown cells:
        a. Discuss any challenges faced when scraping and cleaning the data.
        b. Summarise the key insights gained from your analysis of the data.
        c. Suggest ideas for further work which could be performed on the data (e.g. alternative analyses, integration of other sources of data).

Guidelines:

- The assignment should be completed individually. All submissions will be subject to plagiarism checking. Any evidence of plagiarism can result in a 0 grade.
- The grade awarded will depend on the complexity of the analysis and level of detail, i.e., data cleaning and preparation, analysis, interpretation etc.
- See the associated Grading Rubric document for a detailed breakdown of marks for each task.
- Submit your assignment via Brightspace. Your submission should be in the form of a single ZIP file which contains:
    1. Your two Jupyter notebooks. These should be IPYNB files, not HTML files.

2. The dataset you saved in Task 1.

- In your notebook please clearly state your student number and the data source that you selected.
- Hard deadline: Submit by end of 21st March 2025.
  Penalties will apply for late submissions:
  1. 1-5 calendar days late: 1 grade point deduction, e.g. B to B-
  2. 6-10 calendar days late: 2 grade point deduction, e.g. B to C+
  3. Assignments will not be accepted any later than 10 calendar days without Extenuating Circumstances formally approved by UCD.