# COMP47670 Assignment 1: Data Collection & Preparation

**Deadline:** Sunday 19th March 2023

**Overview:**

The objective of this assignment is to collect a dataset from a public web API and then 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 tasks in two separate notebooks:

1) **Data Identification & Collection:**
    a) Choose one of the APIs from the list of web APIs provided at this link: API-List

    When choosing an API, if it has free and paid access alternatives, be sure that the free tier provides the data access that you need.

    If you wish to use an API not on the list, you should discuss the objectives with the module coordinator.

    b) Collect data from your chosen web API using Python. Note that, depending on the choice of API, you might need to repeat the collection process multiple times to download sufficient data for analysis.

    c) Save the collected dataset in an appropriate format for subsequent analysis.

3. **Data preparation and analysis:**
    a) Load the saved dataset from Task 1 into an appropriate data structure.

    b) Apply any preprocessing steps that might be required to clean, filter or transform the dataset before analysis.

    c) Analyse, characterise, and summarise the cleaned dataset, using tables and visualisations where appropriate.

    d) Clearly explain each step of this process and interpret the results which are produced. Markdown cells should be used for the explanations and interpretations.

    e) At the end of your notebook, summarise any insights which you gained from your analysis of the data, discuss the challenges faced in collecting data from the API, and suggest ideas for further analysis which could be performed on the data.

**Guidelines:**

- The assignment should be completed individually. All submissions will be subject to plagiarism checking. Any evidence of plagiarism will 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.
- Submit your assignment via the COMP47670 Brightspace page. Your submission should be in the form of a single ZIP file containing:
    - The two Jupyter notebooks for Task 1 and Task 2 respectively. These should be IPYNB files, not HTML.
    - The dataset you saved in Task 1. If your dataset is too large to upload to Brightspace, please include a smaller sample of the data in the ZIP file.
- In your notebooks please clearly state your student number, and the name of the web API which you have chosen for your assignment.
- Late submissions:
    - 1-5 days late: 1 grade point deduction, e.g. B to B-
    - 6-10 days late: 2 grade point deduction, e.g. B to C+
    - Assignments will not be accepted after 10 working days without Extenuating Circumstances formally approved by UCD.