DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

Continuous Assessment Cover Sheet

Student Name:		**Student Number:**
Programme:	**Stage:**	**Complete Student Checklist:** Re-read brief ☐ References and Bibliography ☐ Proofread ☐
Module:		
Due Date:	**No. Pages:**	
Lecturer(s) Name:		
Assignment No. and/or Description/Topic:		**Mode of Submission:** Softcopy ☐ Hardcopy ☐

DECLARATION: I declare that:

- This work is entirely my own, and no part of it has been copied from any other person's words or ideas, except as specifically acknowledged through the use of inverted commas and in-text references;
- No part of this assignment has been written for me by any other person except where such collaboration has been authorised by the lecturer(s) concerned;
- I have not used generative artificial intelligence (AI) (e.g. ChatGPT) unless it has been permitted by the lecturer(s) concerned;
- I understand that I am bound by DkIT Academic Integrity Policy. I understand that I may be penalised if I have violated the policy in any way;
- This assignment has not been submitted for any other module at DkIT or any other institution, unless authorised by the relevant Lecturer(s);
- I have read and abided by all of the requirements set down for this assignment.

SIGNATURE..... **DATE**.....

Lecturer's Comments:

Provisional Mark : _____ **Lecturers Signature :** _____ **Date:** _____

CA4: Web Scraping, Data Analysis, and Flask API

Submission Date: January 4th 2026 (11:55pm)

In this continuous assessment, you will design and implement an **end-to-end data pipeline** involving web scraping, data analysis, and result presentation using a Flask-based web application.

You must devise an **automated web scraping strategy** for web-based data of your choice. Identify a suitable webpage (for example, https://www.wikipedia.org — this is only an example). You must search for and select a webpage yourself and write Python code that scrapes the data, cleans and curates it, and stores the final clean data in a structured dataset (e.g., CSV or JSON).

1. Website Selection [10 Marks]

Search for a website of **your choice** and clearly state the **web address (URL)** of the selected pages on the website.

You must explain:

- Why you selected this website
- What type of information you intend to extract
- How this information can be useful for analysis

Marks will be awarded based on the relevance of the website, clarity of explanation, and appropriateness of the selected data.

2. Data Collection (Web Scraping) [30 Marks]

Implement an **automated approach** to retrieve data programmatically using Python. You may use libraries such as *BeautifulSoup* and appropriate HTML identifiers (e.g. tags, classes, IDs) to correctly extract content from the selected webpage. During data collection, you must demonstrate an understanding of:

- The scope of the data being collected
- How this data can be transformed into meaningful variables for analysis

For example, if you retrieve dates of birth from a webpage, these may later be used to calculate ages. This is only an example; you must define your own scope and rationale for data collection.Your data collection solution must include:

- Data cleaning
- Data curation
- Structured storage of the final dataset

The final dataset must be stored in **CSV or JSON format**.

3. Exploratory Data Analysis [30 Marks]

Once the dataset has been finalised, conduct an **exploratory data analysis (EDA)**. Your analysis must include:

- Summary statistics
- Appropriate plots
- Clear interpretations of results

You must perform **at least three meaningful statistical analyses** based on the cleaned scraped data. Each analysis should lead to a **distinct insight** derived from the data.

For example, one analysis might involve counting the frequency of specific words in a text dataset. This analysis could then be extended by interpreting why certain words appear frequently and what this reveals about the underlying content. This is only an illustrative example; you must decide which analyses are appropriate for your dataset.

Marks will be awarded based on:

- Appropriateness of the chosen analyses
- Correctness of implementation
- Clarity of explanation
- Quality of result presentation

The analysis must be included in a **Jupyter Notebook**. Each analysis must be clearly explained using **Markdown cells with appropriate headings**.

4. Flask-Based API Presentation [30 Marks]

Develop a **Flask-based web application** to present your results.

The application should expose your analysis results across **multiple pages or API endpoints**. Each page must:

- Display relevant plots, tables, or summaries
- Include a clear title
- Provide appropriate explanations of the results

The Flask application should allow a user to understand your analysis without needing to inspect the Jupyter Notebook.

Submission Guidelines

You must submit the following:

1. A **Jupyter Notebook** containing:
    - Web scraping code
    - Data cleaning and curation steps
    - Exploratory data analysis and plots
2. A **Flask application**, including:
    - Source code
    - Clear instructions on how to run the application
    - All required plots and outputs
3. A **screen-cast video (up to 10 minutes)** explaining:
    - Your scraping strategy
    - Data cleaning and curation decisions
    - Analysis choices
    - Flask application structure and functionality
4. If required, a **Q&A session** may be arranged. This may involve a detailed discussion of:
    - Your web scraping strategy
    - Code originality
    - Non-functioning code or missing outputs
    - Analyses that lack clarity or justification

Note: please submitt all the files as your name. For example your_name_flask.

Academic Integrity

PLEASE PAY SPECIAL ATTENTION TO THE ISSUE OF ACADEMIC INTEGRITY. The DkIT policies are available at https://www.dkit.ie/registrars-office/academic-policies/academic-integrity-policy-procedures

In summary, all work submitted by learners for assessment purposes, or for written or oral publication, must be their own work. Where this is informed by the work of others, the source must be properly referenced using the accepted norms and formats of the appropriate academic discipline. Using generative artificial intelligence tools (e.g. ChatGPT) in this assignment unless explicitly permitted to do so and with proper acknowledgement, is a form of plagiarism.

Late Submission, a strict cut-off of January 4th for submission of projects. Any legitimate late submission must be accompanied by explanation and supporting documentation as per the policy.

https://www.dkit.ie/registrars-office/academic-policies/continuous-assessment-policy-procedures