DUNDALK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA DHÚN DEALGAN

# Continuous Assessment
# Cover Sheet

| **Student Name:** | **Student Number:** |
|---|---|

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

*Work submitted late will be subject to penalties in accordance with the DkIT Continuous Assessment Policy*

# CA1 – Data Cleaning and Preparation Using Python

## Dataset Overview

For this assignment, you will work on the Irish Weather Hourly Data dataset, available on Kaggle: https://www.kaggle.com/datasets/conorrot/irish-weather-hourly-data

This dataset contains *hourly weather observations* from multiple weather stations across Ireland. It includes variables such as temperature, rainfall, and wind speed recorded over several years.

The dataset provided to you in Moodle is a modified version, containing various data quality issues intentionally introduced for analysis and cleaning practice. Some values are missing, inconsistent, or non-sensical; there may also be outliers and rows not properly ordered by date. Your goal is to use Python and Pandas to explore, clean, and prepare this dataset for further analysis.

You must present all your work in a Jupyter Notebook (.ipynb) file, with appropriate code, markdown explanations, and clean dataset.

### 1 – Describe and Rename Columns (10 marks)

- Explore the dataset and describe what kind of data each column contains.
- Some column names are abbreviations (e.g., wdspd). Rename columns to meaningful names that describe the data clearly.
- You may research the dataset on Kaggle or other online sources to understand what each column represents.
- The goal is to make the dataset self-explanatory and easy to interpret.

### 2 – Identify Missing and Non-sense Values (10 marks)

- Investigate the dataset to find all missing, null, and non-sense values.
- Non-sense values include entries like "?", "error", "missing", "NaN", or other inconsistent symbols.
- Report which columns contain such values and how many appear in each column.
- Summarise your findings clearly using printed outputs and a markdown explanation.

### 3 – Develop and Apply a Cleaning Strategy (15 marks)

- Develop a *clear and well-structured strategy* to clean missing and non-sense data.
- A missing value can be a blank cell, while non-sense values may include symbols or strings that do not represent real data.

- Apply your cleaning strategy systematically using Pandas, and clearly justify the methods you choose (for example, why you replaced, removed, or imputed specific values).
- Explain your cleaning steps using markdown cells.

### 4 – Detect Outliers (10 marks)

- Examine the dataset for possible outliers using an *appropriate statistical method*.
- You may use descriptive statistical tests (e.g., IQR or z-score).
- Report which columns contain outliers and describe how you identified them. Explain it using markdown cells.

### 5 – Handle Outliers (10 marks)

- Choose and apply suitable methods to handle detected outliers.
- Document your reasoning and the steps you take to address them (e.g., removing, capping, or transforming values).
- Demonstrate the effect of your outlier-handling process on the dataset. For example you can compare the mean value of column having outliers before and after handling it.

### 6 – Check and Sort by Date (10 marks)

- Examine whether the dataset is properly sorted by its date or time column.
- If it is not sorted, reorder it chronologically.
- Confirm that sorting was successful.

### 7– Date-based Slicing (10 marks)

- Use Pandas slicing to extract and analyse data based on specific date ranges. Select data for the *month of your birth in a year of your choice* and select columns “rain” and “temp”
- Calculate the average rainfall and average temperature for that month.
- Present your results clearly with code and markdown explanation.

### Task 8 – Location-based Slicing (15 marks)

- Select your *favourite Irish county or weather station* and perform an analysis of weather patterns there.
- Identify which *months or seasons* are best for visiting based on temperature and rainfall data. For this task you need to compare your calculated statistics with the ideal conditions and *select days of the months in each year* for those conditions. For example, good days to visit Dublin are the ones when temperature above 20 °C and low chances of rainfall. Create similar conditional statements using suitable Pandas functions and display *suitable days range* in each year.
- Summarise your insights with calculations and a short explanation.

## Task 9 – Remove Empty or Irrelevant Columns (5 marks)

- Identify any columns that are empty or do not have any useful information.
- Remove such columns and display the shape of the final cleaned dataset.

## Task 10 – Save and Submit (5 marks)

- Save your cleaned dataset as: cleaned_hrly_Irish_weather.csv
- Ensure that all data cleaning and processing steps are completed within the same notebook.
- Submit both your cleaned dataset and your Jupyter Notebook.

## Submission Format

Submit the following two files:

1. **YourName_DataCleaning.ipynb** – Your complete Jupyter notebook with all code, outputs, and markdown explanations.
2. **cleaned_hrly_Irish_weather.csv** – The final cleaned dataset file.
3. All code must be properly commented and explained.
4. Each task should be clearly labelled in your notebook using markdown headings.
5. Your notebook must run from start to finish without errors.
6. Include short explanations for your decisions, such as why you selected a particular cleaning or filtering method.
7. The final dataset should be clean, consistent, and ready for further analysis.

## Evaluation Criteria

- The assignment will be evaluated based on quality of codes, comments, markdown explanations, and strategies taken to handle data.
- A short interview session will be conducted in lab which may include the live production of part of your presented code.