Assignment Specification

Continuous Assessment	**Class Group: TU059 / TU060 / TU256**
Worth: 60% of the overall mark for the module	**Due date:** 14/12/2025 (Sunday), 5:59pm

Penalty for late submission: 5% first day, 10% second day, 15% third day. Submissions are not accepted after 3 days.

Objective: Demonstrate your ability to import, explore, and understand data by analysing factual datasets and addressing relevant questions of interest based on your learning in the module.

Description of Tasks

Find data in publicly accessible online sources e.g. UCI, data.world, Kaggle, webpages, APIs etc.

Introduction
Explain what your data is about, where was it collected, and the meaning of all its attributes. You should have an overall goal for your data analysis. This goal will help you plan what to do in parts 1, 2 and 3.

Part 1
Import the piece/s of data and perform any cleaning and merging to produce a final dataframe. Explain the steps necessary to clean the data and justify any decisions taken. For the final dataframe, what is the meaning of each row?

(33 marks)

Part 2
Carry out a general exploration of this dataframe to develop an overall understanding of the data. You can perform single/multivariate analysis with summary statistics and relevant visualisations. Explain and discuss any values and/or visualisation reported. For example, were there any expected/unexpected results?

(33 marks)

Part 3
Focus on a particular subset of the dataframe and drill down into it extracting details to answer a series of questions that are of interest to you as an analyst. Ideally the motivation for such questions would be framed within the context of a hypothetical use case scenario.

(34 marks)

Submission

You should submit a Jupyter or RStudio notebook, using Python or R (as you prefer). Make sure you explain each code cell using markdown cells and you properly separate each part of the assignment to facilitate evaluation. Focus on what was done, flow of reasoning (motivation for steps), depth of analysis and conclusions reached. In relation to code, ensure that variable names are informative and appropriate and meaningful comments are placed.

The code should work without issue when run on the lecturer's machine. *Hence, do not use absolute paths for external files,* or the assignment might not be marked.

Plagiarism

Code sourced from elsewhere must be clearly cited (bear in mind your Python and/or R skills are also being examined). Plagiarism will result in a zero mark (0%). You should make yourself familiar with the plagiarism policy of Technological University Dublin.

General guidelines:

- Start from the beginning. Find a dataset that interests you and is complex enough to allow for an interesting analysis.
- Once a problem has been solved, test and document it before moving to the next part.
- We have gone over many notebooks so far. Review them if you feel lost or don't know how to proceed.
- An excellent assignment will likely require some self-learned concepts and the use of functions/parameters not seen in class. Still, try to complete a basic version of the assignment first and add complexities as much as possible later.

Guideline for using GenAI

- AI can be used in the assessment for brainstorming, creating structures, and generating ideas for improving work.
- AI can be used to make improvements to the clarity or quality of student-created work to improve the final output, but no new content can be created using AI.
- **AI can be used but the original work with no AI content must be provided in the appendix.**
- Any AI created content must be cited.

Grading Rubric

	Part 1	**Part 2**	**Part 3**
Challenging (70+)	Work is complex involving data in multiple formats and/or files. Exhibits elements of creativity e.g. problem solving and self-learned concepts.	Exploration is in-depth and insightful, demonstrating an excellent understanding of key characteristics and patterns in the data.	Complex manipulations of the data in pursuit of latent trends that answer a series of interesting questions.
Very good (60-69)	Work involves data from more than one format and/or file. There are some challenging aspects to the work.	Exploration covers a variety of relevant views on the data.	Some complexity in terms of manipulations with a coherent basis for operations.
Good (50-59)	Work involves data from more than one format and/or file. It employs basic aspects of data cleaning and preparation (missing data, duplicates, wrong values, etc.).	Exploration is adequate but could be improved. The analysis may feel somewhat repetitive and would benefit from more in-depth discussion and interpretation of the results.	Good manipulation and analysis, but feels a little contrived in places.
Ok (40-49)	Work based on a single piece of data. Did not require much cleaning and/or merging.	Exploration is limited/repetitive and/or misleading.	Basic manipulations of the data, work missing a clear focus and/or contains some inaccuracies.
Needs work (0-40)	Significant departures from assignment brief and/or heavily incomplete. Work does not provide clarification on the use of GenAI.	Significant departures from assignment brief and/or heavily incomplete. Work does not provide clarification on the use of GenAI.	Significant departures from assignment brief and/or heavily incomplete. Work does not provide clarification on the use of GenAI.