# DAT6501: AI and Statistical Data Analysis
# Machine Learning Project

For this assignment, you will obtain a high-quality dataset, identify an opportunity for machine learning, and use Python to obtain a result. The rationale for your choice of problem and solution should be clearly laid out around your code in a Jupyter Notebook document. You may use the Jupyter Notebooks provided for our lab classes as a guide to the style expected.

- Deadline: Friday 8 November (17:00)
- Submission: A single Jupyter Notebook file (IPYNB format) to be submitted on QMPlus
- Requirements for submission:
    - Maximum length: 500 lines [of code and text, with 115 columns per line of code]
    - Detailed descriptions of analysis performed
    - Appropriate figure(s) to illustrate the data set, the model, and any key statistics, inferences, predictions, or conclusions.
    - An appended file containing your data so that the solution can be tested. (A link to an online data source could also serve this purpose.)

This assignment is worth 25% of the total marks for the module. Standard late penalties (5% of total assignment mark per day) apply. Late assignments will not be accepted after 5 working days unless there are approved extenuating circumstances (ECs).

A total of 100 marks are available:

- 20 for the choice of data and a problem of relevance to your industry
- 30 for functioning code that addresses the problem using your data set
- 30 for clear written explanation of your code
- 20 for clear and relevant figures.

More detailed marking criteria are given below. Please read these in detail, as there are specific requirements for the report and figures. It is part of these criteria that your repository, at the time of submission and without modification, should produce any results and figures referred to in your explanation.

Detailed marking criteria:

- **Choice of Data and Problem [20 marks]**: Begin your report with an Introduction section that clearly motivates the choice of data and the machine learning challenge identified. This could be a prediction capability, an estimation/measurement problem, or a hypothesis to test.
  - Data: Clearly explain the source and characteristics of the data [10 marks]. You are encouraged to use a figure to summarise the data (see below).
  - Problem: Show that this is relevant to the broad area in which you work and has commercial significance [10 marks].

- **Code to solve the problem [30 marks]**: Your code should solve the problem you have set yourself, using the data set you have obtained.
  - This includes validating and/or testing to quantify how well your model solves the problem.
  - You are encouraged to take Python code from existing sources and/or generate code using aids such as ChatGPT or Copilot. Marks will be awarded for:
    - the functioning of your code
    - the efficiency and relevance of your code, without superfluous elements
    - judicious use of brief comments in your code to clarify purpose.
  - Note that any lines of code that exceed [115] characters should be broken into multiple lines using the backslash and will be counted accordingly for the line-count of the submission.

- **Explanation of code and outputs [30 marks]:** Explain the design and the outputs of your code, going beyond any comments embedded in the code itself.
  - Show how you have solved, as far as possible, the problem you set out in the Introduction [10 marks].
  - Explain the results of your testing. Mention any limitations or caveats for your solution, and end with a Conclusions section that explains the 'take-home message' and any immediate directions for further work [10 marks].
  - Explain your use of any AI assistants in developing your code: Which assistants did you use, how was it prompted, how did you develop and integrate the code into your project? [10 marks].

- **Figures [20 marks]**
  - Present figures to illustrate the nature of your machine learning model and its key outputs.
  - One figure should summarise the data set, and one or more figures should describe the model or its outputs.
  - Between 2 and 4 figures are likely to be sufficient for a project of this length. Marks are awarded for clarity and information content.
  - Figure captions are not required where a figure is clearly introduced by the preceding text.