This assignment is worth 50% of the total coursework for the module.

The assignment is divided into two parts. You should submit a single Jupyter Notebook and any related scripts or SQL files as a single archive. The notebook should contain a description of your approach as well as any/all processing used to manipulate, cleanse and sanitise the data for purpose. If your dataset exceeds 10MB, then include a working sample of the data that can be used in place of the full dataset. Your project should focus on one of the following themes.

Project Themes

Theme	Scope	Example project
Premier League Football	Projects should focus on a dataset around British Premier League Football. This could include any data from 20 February 1992 up until the current season.	'How to get relegated, an analysis of poorly performing teams in the British Premier League.'
Literary Masterpieces	Projects should explore famous plays, sonnets and poems.	'What's in a name? An investigation into the names and content of the works of Shakespeare.'
HTML and Markup	Projects should focus on exploring markup from one or more of the top 50 websites (according to Alexa.)	'An analysis of the semantic features of streaming websites.'

Deliverables

The deliverable for this project is a single Jupyter Notebook and any related scripts and should be arranged as follows:

Report Format

• A report presented in a single notebook. This should include:
– introduction/context.
– brief description of data set (or output a sample), including relevant information (e.g. how it was obtained.)
– data visualisations, tables or other key descriptors.
– summarise key findings/insights.
– some form of discussion/critical analysis.
– conclusion and further work.
– references to any resources used.
• Visualisations can be presented inline in the Notebook, or in separately exported files for instance where a graph or diagram is too large (e.g. PNGs.)
• You should include a working sample of your dataset, not exceeding 10MB.
• You should include a requirements.txt file plus any additional instructions about how to replicate your approach and outcomes.
• The report must cover the following 2 parts

Part 1 (20%)

Your brief is to design a manageable data science project, and acquire the necessary dataset in a usable form. You will need to submit your notebook as well as any resources that you have used throughout the exercise.

For Part 1 you should:
• Acquire and prepare your dataset. Here, you will be expected to seek out and find your own dataset – presumably online. Be sure you are allowed to share the data with others.
• Preparation might involve collation and/or manipulation of data into a usable format.
• It may involve creating a database or a flat file format to store and manage data, for instance if you are working with very large datasets.
• It may involve writing Python which produces a dummy dataset, for instance if you are working with sensitive data. If this is your preferred option, you may need to think carefully about what you would expect the results of an analysis to look like, so you can generate the data accordingly (e.g. if generating random numbers, choose a function or method which produces a realistic distribution, and perhaps a realistic amount of noise too.)
• Explain what programming techniques you have used in the preparation of your data (including any command-line or SQL programming.)

As a topic of your choosing you should:
• Outline the idea behind your project (e.g. context.)
• Briefly detail what you intend to do with the data. You are not expected to explain the exact techniques you will use, though it is important that you identify your process as you work and any high level features of the data.
• You should carefully consider any weaknesses or potential caveats in your approach and present these too.

Part 2 (30%)

You should aim to carry out the main body of work in analysing your data and producing relevant outputs. You should summarise the project, report on your findings, and discuss any new insights/grounds for further work.

For Part 2 you should show evidence of the following:

- That an appropriate dataset has been obtained.
- There is evidence that the data has undergone some form of manipulation to get it in a suitable format.
- Some meaningful statistical values or metrics have been derived from the data (e.g. descriptives/summary statistics.)
- Visualisation(s), tables or some other suitable high level descriptions have been created programatically, and reveal aspects of the dataset.
- A report containing both parts 1 and 2 has been produced in a notebook format, a single IPYNB file:
  - The report outlines the project, and summarises key findings.
  - The report includes a descriptive and systematic rhetoric of your process and some discussion thereof e.g. including relevant sections of analysis with enough detail to understand the how/why of the process.
  - The report highlights areas of weakness/uncertainty, and suggests further work.
- The related code is written in Python and submitted inside the notebook.
- Some or all of the techniques taught have been demonstrated.
- For higher marks you must show that there has been use and some extension of the techniques taught in the module.