Group Project Stage 1

Due: 17:00 pm on Sunday at the end of week 8 (Sep 28th)

Value: 20% of Total Mark

Note: Get started your project ASAP. Discuss with your tutors and make use of Ed to ask questions.

1 Purpose

The Stage 1 Project is a collaborative data science investigation completed in groups of 3 or 4. It assesses your ability to identify a meaningful question, prepare and clean data, summarise and analyse it using Python.

2 Group Formation

- Groups of 3–4 students.
- All group members must be enrolled in the same lab.
- The same mark is awarded to all members unless otherwise specified.
- Tutor approval is required for any group changes.

3 The Project Work for Stage 1

3.1 Define a Topic or a Question

The group should define questions or issues that are not simply a factual matter, but instead examine relationships where insights might be impactful for some stakeholder groups. We realise that you may not find data that completely resolves the issue you are targeting, but all the data should at least be helpful to provide some insights.

- Choose a topic that explores **relationships** (not just factual reporting).
- Justify the importance of your question.
- Include stakeholder relevance and real-world impact.

3.2 Select and Describe Data

Select at least three datasets within the defined topic or question.

1. Keep (and provide to us) a copy of the data as you originally obtained it.
2. State the relevant metadata about this dataset, including:
    - A data dictionary (indicating which attributes exist, and what each attribute means).
    - Provenance (giving the whole chain, from the original source of the data, through any intermediate collections, up to the place where you obtained it [and the date you obtained it]).
3. It is preferred to use publicly available data (so we can check your work if we need to). However, it is acceptable to work on privately-owned data as long as you have permission to use it, and permission to show it to the markers.

3.3 Ensure Data Quality

Members need to work with the selected datasets to ensure high-quality data that can be analysed.

1. Use Python to transform and clean each of the raw data selected. The details of this aspect vary depending on the data obtained.
    (a) The data may need to be cleaned.
    (b) If the data sources were carefully curated already, you should at least write a Python program that checks the cleanliness of the data.
2. At the end of this part of the work, you should have a dataset of high quality.

3.4 Perform Simple Analysis

- For **each dataset**, write Python code to produce at least one meaningful summary (e.g., grouped aggregate, frequency distribution, or descriptive statistics).
- Ensure that the summary for each dataset are well-labelled.
- Include all results in your report.
- After producing summaries for each dataset individually, create at least one summary or comparison that brings together information from two or more datasets. The goal is to highlight a relationship across datasets that is relevant to your research question.

3.5 Conclusion

- Write a concise summary of aimed at a non-technical audience.
- Contributions of each member.

4 Stage 1 Submission

The Front Page of the report must include:

- Course unit (DATA1002)
- Tutorial section
- Group number
- Student names, Unikeys, and student IDs

What to Upload:

- **Report (PDF)** — the written report as described above.
- **Code & Datasets (ZIP)** — a compressed archive containing your Python scripts/notebooks and both the raw and cleaned datasets used in your analysis.

Submission Links:

Report submission: [Report submission 1002]
Code & Data submission: `[Code & Data submission 1002]`

5 Marking

There are six components to be assessed. All are group-marked, and each member will receive the same score unless specific differences in contributions are explicitly noted in the report.

Component	**Weight (%)**	**Full Marks Criteria**
Define Topic or Question (max 1 page)	18%	1) Research question is clearly defined, relational (not factual). 2) Importance and relevance to stakeholders are explicitly justified. 3) Real-world impact is described.
Select and Describe Data (max 3 pages)	18%	1) At least 3 datasets with ≥300 records total are selected. 2) Each dataset is documented with schema (data dictionary), provenance (source chain + date), and noted limitations. 3) Original raw data is preserved and referenced.
Ensure Data Quality (max 5 pages)	18%	1) Python used to check and clean each dataset (missing values, formatting issues, duplicates addressed). 2) Clear explanation of data transformations (if any). 3) Final dataset is of high quality and ready for analysis.
Perform Simple Analysis (max 5 pages)	22%	1) For each dataset, at least one meaningful summary is produced using Python. 2) All summaries are clearly labelled. 3) At least one combined summary or comparison is provided that integrates information across two or more datasets to highlight a relationship relevant to the research question.
Conclusion (max 1 page)	6%	1) Concise non-technical summary of findings. 2) Contributions of each group member are explicitly listed.
Formatting & References	6%	1) All datasets and literature are cited in APA 7th style. 2) Report formatting is professional and consistent with academic standards.

Code and Data	12%	1) Provide all python code. 2) The code runs successfully within a reasonable time. 3) Code is well-structured, clearly commented, and properly documented to ensure readability and reproducibility. 4) Both the original raw data and the cleaned, processed datasets are included and appropriately organised.

6 Late Submission

As announced in the unit outline, late work (without approved special consideration or arrangements) suffers a penalty of 5% of the maximum marks, for each calendar day after the due date. That is, we subtract 0.25 marks per day from what you would otherwise get for the work. No late work will be accepted more than 10 calendar days after the due date.