[I0P16A] Data Analysis Project Instructions 2024-2025

Prof Jan Aerts

IMPORTANT: You *are* allowed to collaborate with another student on the coding part, but check the collaboration policy below.

Objective

Analyse a dataset to demonstrate your ability to perform multivariate data analysis, including clustering, dimensionality reduction, and topological data analysis. Extend the methodologies from lectures to a higher level in terms of methodology, graphics, and presentation of results.

The report should be 5 pages long, not including the appendix with the code that you write.

Datasets

There are several options for a dataset:

- the Spotify Tracks dataset, available at https://huggingface.co/datasets/maharshipandya/spotify-tracks-dataset/blob/main/dataset.csv
- Data from data.world https://data.world, e.g. the US air pollution dataset at https://data.world/data-society/us-air-pollution-data
- Biodiversity data from Naturalis https://www.naturalis.nl/en/collection/netherlands-biodiversity-data-services
- Data for climate action http://dataforclimateaction.org/
- Dataset list at https://blog.journeyofanalytics.com/50-free-datasets-for-data-science-projects/
- Harvard Dataverse https://blog.journeyofanalytics.com/50-free-datasets-for-data-science-projects/
- BYOD (Bring Your Own Data, e.g. related to your thesis).

You can also search Google Dataset Search at https://datasetsearch.research.google.com/

Requirements for your dataset:

- Number of observations should be between 500 and 50,000 if you want to do clustering
- Number of features should be >20 if you want to do dimensionality reduction or topological data analysis
- Before you start, fill in the assignment on Toledo "Bring Your Own Data for project"

Report sections

1. Define your research question

(1 paragraph)

- Formulate a unique research question or objective related to the dataset
- Explain the significance of your question and what you aim to discover through your analysis

2. Data selection and preprocessing

(max 0.5 pages)

- Did you subset the data (i.e. chosen a subset of the data relevant to your research question, e.g. specific genres, years, popularity ranges)? Why?
- Which preprocessing steps did you take? Why?
    - Handle missing values, if any
    - Encode categorical variables
    - Scale or normalise features as appropriate

3. Multivariate analysis strategy

(max 1 page)
Describe the following (actual code should be in appendix):

- Which methods did you use? Choose suitable clustering algorithms, dimensionality reduction techniques, and/or topological data analysis
- Why did you select these? Explain why these methods are appropriate for your analysis.
- Did you explore the effect of different parameters? Why? How?

4. Analysis and results

(max 3 pages)
This section will show actual results after you applied your chosen algorithms.

- Provide visualisations of the results (max 25% of contents of this section = 0.5 pages total)
- Interpret the results in the context of your research question

5. Interpretation and insights

(max 0.75 pages)

- Summarise the key findings for your analysis
- Discuss the implications of your results in relation to your research question
- Acknowledge any limitations in your analysis and suggest potential improvements

6. Reflection

(max 0.25 pages)

- Reflect on what you learned during the assignment
- Discuss any challenges you faced and how you addressed them
- Provide thoughts on how this analysis could be extended or applied in real-world contexts

7. Collaboration (if a team)

If you collaborated on your projects with someone else for coding assistance (pair programming), mention in this section who you collaborated with.

Assessment criteria

Grading will be based on:

- **Understanding**: your demonstrated ability to select and apply appropriate methods
- **Analysis depth**: The thoroughness and depth of analysis, including parameter exploration (if relevant)
- **Interpretation**: quality of insights and conclusions drawn from the data
- **Communication**: clarity, organisation and professionalism in the report
- **Originality**: creativity and originality in research question, approach and analysis

Additional instructions

- **Be creative**: use this opportunity to explore aspects in the data that interest you personally
- **Do a deep dive**: go beyond surface-level analysis to uncover meaningful patterns and insights

Collaboration policy

Each student is required to work on their *own individual project*, including defining their research question, selecting methods, conducting analysis, interpreting results, and writing their report. However, collaboration is permitted under the following conditions:

- **Programming assistance:** You may collaborate with **one other student** specifically to help each other with programming challenges, such as debugging code or understanding how to implement methods.
- **Different datasets:** You are not allowed to analyse the exact same dataset. The *source* of the data may be the same (e.g. both working on data from Naturalis), but then you will have to use different subsets of that data.
- **Independent work:** All other aspects of the project, such as the analysis plan, choice of methods, interpretation of results, and report writing, must be completed independently and reflect your own individual effort.
- **Unique deliverables:** Your research question, analyses, visualisations, and report must be entirely your own and should not duplicate or closely mirror that of the student you collaborated with.

Academic integrity and responsibility

This assignment must represent your own work. You are expected to engage with the dataset, perform analyses, and present findings that are uniquely your own.

Plagiarism will be checked. What constitutes plagiarism?

- **Copying from others**: submitting work that is not your own, including copying analysis, code, or written content from classmates, previous students, or any unauthorised sources. While collaboration is permitted for **programming assistance**, sharing entire solutions, code implementations, or results is strictly prohibited.
- **Unattributed sources**: using ideas, code snippets, or text from books, articles, website, or other resources without proper citation and acknowledgment. This extends to generative AI tools like ChatGPT or CodePilot. You are allowed to use these, but need to acknowledge.
- **Violations of the collaboration policy:** Analysing the same dataset subset, duplicating research questions, or producing reports that closely mirror another student's work will also be considered plagiarism.

Format of the reports

- The report should be uploaded to Toledo as a PDF (not a Word document).
- Length: max 5 pages (excluding code appendix)
- Structure:
    - Title and author information (name, student number, master programme)
    - The sections mentioned above
    - Code appendix

- Code for data preprocessing
- Code for data analysis
- Code for visualisation

Clearly structure the code appendix with titles and subtitles. It is not necessary to add many comments, but it will be important to find certain parts of the code back easily.