Data Analysis and Decision Making

4th Year - Data Analysis - 2025

Alexandre Gazagnes

1. Objectives

The objective of this exam is to evaluate your ability to analyze a dataset, to extract insights, to make decisions and to communicate your results.

A dedicated focus will be made on the analysis, the visualisation and the implementation of a PCA / Clustering.

2. Context

- Oral presentation on April the 8th, 2025
- Group of 5 students :
  - 15 min of presentation
  - 5 min of Q&A

3. Timeline

To remember :

- AT LEAST - March 20th --> Sending your group description by email
- AT LEAST - March 24th --> Sending the dataset for validation
- AT LEAST - April 2nd --> Sending the pre-release of your notebook for validation
- AT LEAST - April 7th --> Sending the final version of your notebook for validation - powerpoint presentation

4. Modailities

4.1 About the Data

You should 1st :

- Find a dataset on the internet. Feel free to use Kaggle, UCI, etc
  - The dataset should be in CSV format
  - The dataset should contain at least 1000 rows and 10 columns
  - No same dataset for all students, first come first served !

4.2 About the mindset

Then you should :

- Find
  - a problem to solve with this dataset
  - Questions to answer
  - Analysis to perform

4.3 About the analysis

Then you should :

- Evaluate the quality of the data
- Clean the data if needed
- Prepare the data for analysis
- Perform the analysis
- Visualize the data
- Implement a PCA and/or a clustering

5. Guidances 1/3

- Powerpoint presentation :
  - From 10 to 20 slides
  - Structured approach
  - Plan :
    - 1/ Context/Intro
    - 2/ Data, preparation, cleaning etc
    - 3/ Insights, analysis, visualisation, Tests
    - Conclusion, recommendations, go further

5. Guidances 2/3

Jupyter / Google Colab notebook :

- Clean, structured, commented
- You can use Colab, plain code, Jupyter notebook etc
- 1 notebook is expected BUT feel free to add more if needed
- Feel free to use external libraries, packages, etc
- Calm down with GenAI, it's very easy to find out if the job is done by a machine or a human ;)

5. Guidances 3/3

- Your code MUST BE reproducible !
  - I should be able to run your code and get the same results as you
  - I should be able to understand your code
  - I should be able to reproduce your analysis
- If not :
  - Your work **will not be evaluated**

6. Evaluation 1/2

10 points will be given for the quality of the analysis, the insights, the visualisation, the implementation of the PCA and/or the clustering.

More precisely, 2 points will be given for each of these items :

- Data exploration
- Data preparation and cleaning
- Analysis and insights
- PCA AND/OR Clustering

6. Evaluation 2/2

5 points will be given for the quality of the presentation, the structure, the clarity, the recommendations, the go further etc

5 points will be given for the questions and the answers during the Q&A.

Appendix

Data exploration :

- Data quality / Dataset evaluation
- Data description
- Data types
- Data Structure

Data preparation and cleaning :

- Missing values
- Outliers
- Duplicates
- Feature engineering

Analysis and insights :

- Descriptive statistics
- Correlation
- Visualisation
- Global insights on the dataset

PCA AND/OR Clustering :

- Choice of the method
- Implementation
- Interpretation of the results
- Visualization of the results
- Explanation of the results