Coding Assignment 02
Title: Analysis of Human Gut Microbiome Data from the Metalog Database

Dataset:
https://metalog.embl.de
Metalog is a repository of manually curated, open-source metadata for metagenomic samples. It includes rich contextual information such as clinical and demographic data for human subjects, as well as environmental parameters for non-human samples. Released in 2025, Metalog is actively used by researchers to gain new insights into microbiome research.

Overall Goal:
The objective of this assignment is to gain hands-on experience working with real, open-source microbiome data. You will formulate a research question and perform an initial analysis to obtain your own evidence-based insight into the human gut microbiome.

Before beginning, please choose which environment you want to analyse.
Options: Human, ocean, animal, environmental.

Task 1. Exploratory analysis of the selected dataset

Objectives
Characterize the overall structure, quality, and heterogeneity of the dataset to understand its suitability for downstream microbiome analyses.
Analysis:
Using the full metadata table, address the following questions:

- How many samples are included in the dataset in total?
- How many independent studies does the dataset contain?
- What are the minimum, maximum, and mean numbers of samples per study?
- Which metadata fields are available (for example, demographic, clinical, environmental, and technical variables)?
- For each metadata field, what proportion of samples has missing values?
- Are missing values randomly distributed, or are they concentrated in specific studies or time periods?
- Are there studies with longitudinal sampling, meaning repeated measurements from the same subject or sample source? If so, how many subjects and time points are involved?
- Is there substantial technical heterogeneity between studies (sequencing platforms, library preparation methods)?
- Are there duplicate or near-duplicate samples based on sample identifiers or identical metadata profiles?

Visualizations
Propose and generate appropriate visualizations to support your analysis. These may include, for example:

- Bar plots or histograms showing the number of samples per study
- Heatmaps or tile plots illustrating missing data patterns across metadata fields and studies
- Correlation matrices for numerical and categorical metadata variables
- Timelines or line plots showing temporal coverage and longitudinal sampling

Conclusions
Discuss the following points:

- Is the dataset dominated by a small number of large studies, or is it relatively balanced across studies?
- Are there systematic patterns of missing metadata that could bias downstream analyses?
- Are duplicate or redundant samples present, and how should they be handled?

Task 2. Formulate a question that can be answered with the data

Objective
Define clear, testable research questions that can be addressed using the available metadata and microbiome profiles in the dataset.

Candidate Research Questions

1. Does microbiome community composition differ between sampling locations, different geographic regions, different body sites?
2. Is there a relationship between environmental conditions such as temperature, salinity, or oxygen concentration and the observed microbiome structure?
3. Do samples collected in different seasons or months show systematic differences, suggesting temporal or seasonal effects on the microbiome?
4. Does microbiome community composition differ between sampling protocols?

For each selected question, clearly specify the response variables (diversity metrics), the explanatory variables of interest (location, region, etc.), and any covariates or confounders that need to be controlled for (age, sex, BMI).

Task 3. Select a subset of data for your analysis

Objective
Define a well-controlled subset of the dataset that is appropriate for addressing the chosen research question.

- Select a single study or a small group of studies that contain all metadata variables required to answer the research question. Justify the inclusion of each study based on metadata completeness and relevance.
- Aim to construct a balanced dataset to minimize confounding effects. When combining multiple studies, ensure that they have comparable structures with respect to key variables such as age, sex, sampling design, or environmental conditions relevant to the analysis.
- Define clear case and control groups based on the research question. Assess whether these groups are comparable in terms of sample size and metadata distributions. If substantial differences exist, apply matching or stratification approaches to construct balanced case–control groups where appropriate.

Task 4. Answering the Research Question

Objective
Test the selected hypothesis by comparing microbiome characteristics between defined case and control groups.

- Perform a comparative analysis of microbiome profiles between the case and control groups using the selected subset of samples. Clearly state any preprocessing or filtering steps applied prior to analysis.
- Analyze and compare alpha-diversity metrics between groups to assess within-sample diversity. Report the chosen diversity indices and apply appropriate statistical tests to evaluate whether observed differences are significant.
- Compute beta-diversity measures to quantify between-sample differences in community composition. Use ordination methods such as principal component analysis (PCA) or a suitable alternative to visualize sample separation and assess whether case and control groups differ in multivariate space.

Interpret all results in the context of the original research question, noting both statistically significant findings and potential limitations.

Final Deliverables:

- Well-documented code (comments and clear variable names required)
- A short written report summarizing results, interpretations, and limitations

IMPORTANT NOTE: The questions and visualizations proposed in this assignment are recommendations and are not compulsory.