Coding Assignment 01
Title: Analysis of Human Gut Microbiome Community Structure

Dataset:
https://github.com/Chartiza/KSE_microbiome/tree/main/lecture_01
You are provided with a synthetic dataset representing a cohort of human gut microbiome samples. The dataset includes three core data types commonly used in microbiome studies:

1. An abundance table with microbial species abundances per sample
2. A taxonomic annotation table linking species to higher taxonomic ranks
3. Sample metadata describing host characteristics

Overall Goal:
The goal of this assignment is to understand how different microbiome data types are structured, how they relate to each other, and what biological insights can be extracted from them. You will write code to explore the data, compute summary statistics, generate visualizations, and prepare a short analytical report describing your findings and conclusions.

---

1. Abundance Table Analysis

Objectives
Explore the microbial community structure across samples using the abundance table.
Summary Statistics
Using the abundance table, answer the following questions:

- How many samples are present in the dataset?
- How many bacterial species are detected across all samples?
- What are the top 5 most prevalent bacterial species?
- What is the prevalence of each of these top 5 species?
- What are the top 5 bacterial species with the highest mean relative abundance across samples?

Clearly state how relative abundance is calculated.
Visualizations

- Create a bar plot showing the prevalence of the most prevalent species.
- Create a scatter plot with:
  - x-axis: species prevalence
  - y-axis: mean relative abundance
    Each point should represent one species.

Make sure axes, titles, and legends are clear and readable.
Conclusions
Based on your results, discuss:

- Which species appear to dominate the gut microbiome in this dataset?
- Is there evidence for a core microbiome, defined as species present in the majority of samples?
- Do samples appear broadly similar, or is there high variability between individuals?

---

2. Sample Metadata Analysis

Objectives
Explore host characteristics and assess how representative the cohort is.
Summary Statistics
Using the metadata table, answer the following:

- How many males and females are included in the dataset?
- What is the age distribution of the cohort?
- Is the age distribution similar between males and females?
  - Use appropriate statistical tests to assess whether observed differences are statistically significant.

- Do males and females share the same most prevalent species and the same most abundant species?

Clearly state which statistical tests you use and why.

Visualizations

Propose and generate suitable visualizations to support your analysis. You may use, for example:

- Histograms or density plots for age distributions
- Box plots to compare age between sexes
- Pie charts or bar plots for sex composition
- A population pyramid to display age structure by sex

Conclusions

Discuss the following points:

- Does this dataset appear representative of a general human population?
- What limitations do you see when trying to generalize conclusions from this dataset to a broader population?
- Which metadata variables are missing or underrepresented, and how could that affect interpretation?

---

3. Taxonomic Table Analysis

Objectives

Understand how microbial species are distributed across taxonomic levels.

Summary Statistics

Using the taxonomic table:

- Aggregate species abundances at multiple taxonomic levels, such as genus, family, order, and phylum.
- Identify which phyla are most common in the gut microbiome represented in this dataset.

Visualizations

Propose and generate visualizations that show taxonomic structure, such as:

- Sankey diagrams linking taxonomic levels
- Sunburst plots showing hierarchical relationships
- Stacked bar plots at different taxonomic ranks

Choose visualizations that best communicate structure and dominance patterns.

Conclusions

Discuss:

- Whether gut microbes in this dataset are dominated by a small number of phyla or spread evenly across many phyla

---

Final Deliverables:

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

IMOPTANT NOTE: The questions and visualizations proposed in this assignment are recommendations and are not compulsory. You are encouraged to suggest and implement your own analytical questions or visualization approaches. If your proposed alternatives are well justified and provide deeper or clearer insights than the suggested ones, you will receive extra points for the assignement.