FDSAI – Lab 1 2024/2025

We will use the NHANES (2015-2016) dataset, which comes from the National Center for Health Statistics (USA) and is part of the National Health and Nutrition Examination Survey reporting series (https://wwwn.cdc.gov/nchs/nhanes/).

The interpretations of the variables in the NHANES dataset can be obtained by accessing the following link: https://wwwn.cdc.gov/nchs/nhanes/search/default.aspx.

The variables **BMXHT** and **BMXWT**, which we will analyze further, refer to the height of the surveyed subjects (cm) and their body weight (kg), respectively. Additionally, we will also analyze their gender, recorded in the binary variable **RIAGENDR** (coded as follows: 1 = male, 2 = female).

Tasks:

1. Extract from the dataset only the information related to gender, height, and body weight. Check whether the new dataset contains missing values (NaN) and remove them if necessary.
2. Represent frequency histograms for height and body weight, both separately by gender and for the entire sample.
3. Perform descriptive statistics (sample size, mean, variance, standard deviation, minimum, maximum, median, quartiles) for height and body weight, both separately by gender and for the entire sample. Comment on the results. Also, create a scatter plot for height versus body weight, both separately by gender and for the entire sample. Calculate the correlation coefficient in each case. What do you observe?

Important note:

The NHANES reports come from a much more complex study. These data do not constitute an independent and representative sample of the target population. A correct survey data analysis should use additional information on how the data was collected. Since the complex analysis of statistical surveys (data collection) is a specialized topic beyond the scope of this course, we will ignore this aspect of the data and analyze them as if they were an independent and identically distributed sample from a statistical population.

For example, the variable **ALQ101** is a binary variable (with responses coded as 1 or 2) and refers to alcohol consumption, specifically whether a person has consumed alcohol at least 12 times in a year, where a consumption is classified as follows:

- **Low-alcoholic drinks (beer):** 330 ml
- **Moderate-alcoholic drinks (wine):** 150 ml
- **Strong alcoholic drinks (vodka, whiskey, liqueur, etc.):** 50 ml