Assignment 1

Descriptive Statistics & Exploratory Data Analysis with Python
Course: Statistics and Data Science 1
Topics covered:
Topic 1: Python & Pandas refresher
Topic 2: Descriptive statistics, data types
Weight: 40 points of 1st attestation
Deadline: *24.12.2025*
Format: Individual work

Objective

The goal of this assignment is to practice the **basic statistical workflow in Python**:

- loading and inspecting a real dataset,
- identifying data types,
- computing descriptive statistics,
- visualizing distributions,
- and interpreting numerical results in words.

This assignment focuses on understanding the data, not on prediction or machine learning.

Dataset choice (choose ONE)

Students must choose **one dataset** from the list below.

Option A - Titanic Passengers
Source: Kaggle
Link: https://www.kaggle.com/c/titanic/data
Description:
Passenger information from the Titanic disaster (age, gender, class, fare, survival).

Option B - Coffee Quality Data
Source: Coffee Quality Institute
Link: https://github.com/jldbc/coffee-quality-database
Description:
Sensory evaluation scores for coffee beans.

Option C - IMDb Top Movies Metadata
Source: Kaggle
Link: https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset/data
Description:
Ratings, years, runtimes, genres for top movies.

Option D - Weather in Szeged (Historical)
Source: Kaggle
Link: https://www.kaggle.com/datasets/budincsevity/szeged-weather
Description:
Daily weather measurements (temperature, humidity, pressure).

Option E - Human Height & Weight
Source: Open dataset
Link: https://www.kaggle.com/datasets/mustafaali96/weight-height

Description:
Height and weight measurements with gender.

Tasks

Task 1 - Data loading & inspection (10 points)

Using **Pandas**:

1. Load the dataset.
2. Print:
   first 5 rows,
   last 5 rows.
3. Display:
   column names,
   data types.
4. Identify:
   number of rows and columns,
   missing values (if any).

Task 2 - Data types & variables (5 points)

For **at least 5 variables**, classify them as:

- categorical / numerical,
- discrete / continuous (if numerical).

Briefly justify your choices.

Task 3 - Descriptive statistics (15 points)

For **at least 3 numerical variables**, compute:

- mean
- median
- standard deviation
- variance
- minimum & maximum

Use **Pandas / NumPy**, not manual formulas.

Comment on:

- spread of the data,
- presence of possible outliers,
- whether mean and median differ significantly.

Task 4 - Visualization (5 points)

Create **at least 3 plots**:

- 1 histogram,
- 1 boxplot,
- 1 additional plot of your choice (scatter plot, line graph, etc.).

All plots must:

- have titles,
- have axis labels,
- be readable.

Task 5 - Interpretation (5 points)

Write a **project report** answering:

What does this dataset describe?
What did you learn from the statistics?

Which variables show high variability?
Any surprising patterns or observations?
Description of code snippets
Explanation of the results

Submission requirements
Students must submit ONE ZIP or folder containing:

Jupyter Notebook (.ipynb)

- all code,
- plots,
- brief comments.

Report (PDF or DOCX)

- answers to Task 1 and Task 5,
- screenshots of key plots, description of plots.

Academic integrity note
You may use ChatGPT or similar tools for clarification, but all analysis and interpretation must be your own.
Copy-pasting notebook code or text from others will result in 0 points.