Fundamentals of Data Science

Coding assignment - Semester B 2023/24 (30 points, 30%)

Your datasets

Two data files with datasets are required for this project. The data files you have to use depend on **the last digit of your student ID number**. Please, use the table below (or the table in the assignment description) to find out the names of the data files you have to use.

The first dataset, the file with a filename beginning **2020exam**, contains tabulated distribution of students' grades in the module's exam in 2020. It is a three-column CSV file. The first (left) column and the second (middle) column represent the left and right edges of intervals (bins), respectively. The third column shows the number of students with the exam grade in that interval.

For instance, a line

....
60 64 16
....
means that 16 students in the module had exam marks ≥ 60 and < 64.

The second dataset, the file with a filename beginning **2024exam**, is a one-column CSV file with individual student grades (approximately 300 entries).

What needs to be done?

Write a Python code, which

- reads the data from the data files located at the same directory as your Python code. The data files (including their names) must not be changed in any way;
- creates the distributions (histograms) for the 2020 and 2024 exam grades, and plots these two distributions on a graph;
- calculates mean values and standard deviations for both distributions;
- uses the obtained distributions to calculate value V;
- prints all five values (i.e. two mean values, two standard deviation values, and V), as well as your student ID number on the graph. When rounding up, you must keep at least two significant figures.

Hence, **your plot must show two distributions and five values derived using these distributions**. It must have adequate axis labels, titles and the legend.

The code should only create one graph and save it as a PNG image with the filename $<$ $Your\,ID\,number$ $>$ .png.

Write a short report (1.5 pages maximum), which includes

- Your Student ID number;
- The description of the datasets you are given and the distributions you get;
- The formula used to calculate the mean values and SD values for both distributions;

- Discussion of your V value (equation, the value you get, its meaning);
- Discussion of the differences between the 2024 and 2020 distributions (use the mean and SD values calculated by your code);
- The plot produced by your code.

What to submit and where?

Two files must be submitted – your code and your report. Please, do not submit any other files.

Submit **your code** as a single file.

- The file must be named $< IDnumber >$.py, where $< IDnumber >$ is your student ID number. Do not include any other elements in the file name.
- This must be a single *.py file containing Python code. (A script submitted as Jupyter/Colab/other notebook file, or a code submitted as a part of another document is not considered a Python code);
- The code must read data from the file located at the same directory as the code, i.e. the command reading data from the file must not contain the full path to the file;
- The code must be executable using Spyder (https://www.anaconda.com/products/distribution) without the need for additional libraries.

You are strongly advised to test your code using Spyder before submitting it.

Submit **your report** as a single file.

- Your file must be named $< IDnumber >$.pdf, $< IDnumber >$.docx or $< IDnumber >$.odt, where $< IDnumber >$ is your student ID number. Do not include any other elements in the file name;
- This must be a single PDF, MS Word or Open Office document (PDF is preferred).

Your V value

The V value you have to calculate depends on **the last digit of your student ID number**. Please, use the list on the next page (or the table in the assignment description) to find out what input data you have to use and value you have to calculate.

If your ID number ends with ”0”:
- you have to use files **2020input0.csv** and **2024input0.csv**;
- you have to calculate value V which is the proportion of students with the grade of 50 or higher in the 2024 exam;

If your ID number ends with ”1”:
- you have to use files **2020input1.csv** and **2024input1.csv**;
- you have to calculate value V which is the proportion of students with the grade of 70 or higher in the 2024 exam;

If your ID number ends with ”2”:
- you have to use files **2020input2.csv** and **2024input2.csv**;
- you have to calculate value V which is the proportion of students with the grade below 25 in the 2024 exam;

If your ID number ends with ”3”:
- you have to use files **2020input3.csv** and **2024input3.csv**;
- you have to calculate value V such that 10% of students got the grade higher than V in the 2024 exam;

If your ID number ends with ”4”:
- you have to use files **2020input4.csv** and **2024input4.csv**;
- you have to calculate value V which is median grade in the 2024 exam;

If your ID number ends with ”5”:
- you have to use files **2020input5.csv** and **2024input5.csv**;
- you have to calculate value V which is the proportion of students with the grade of 50 or higher in the 2020 exam;

If your ID number ends with ”6”:
- you have to use files **2020input6.csv** and **2024input6.csv**;
- you have to calculate value V which is the proportion of students with the grade of 70 or higher in the 2020 exam;

If your ID number ends with ”7”:
- you have to use files **2020input7.csv** and **2024input7.csv**;
- you have to calculate value V which is the proportion of students with the grade below 25 in the 2020 exam;

If your ID number ends with ”8”:
- you have to use files **2020input8.csv** and **2024input8.csv**;
- you have to calculate value V such that 10% of students got the grade higher than V in the 2020 exam;

If your ID number ends with ”9”:
- you have to use files **2020input9.csv** and **2024input9.csv**;
- you have to calculate value V which is median grade in the 2020 exam.