Fundamentals of Data Science

Coding assignment - January 2024 (30 points, 30%)

Data

You are given a dataset containing a statistically representative sample of salaries in some European country. The data is provided to you in a CSV file (an ASCII file that can be inspected by opening it with any basic text editor such as Notepad). Each entry in the file shows one annual salary in Euros.

What needs to be done?

Write a Python code, which

- reads the data from a datafile located at the same directory as your Python code. The data file must not be changed in any way.
- creates a probability density function representing your sample and plots this array as a histogram;
- uses the obtained probability density function to calculate $\tilde{W}$, the mean annual salary;
- uses the distribution to calculate another value, X (see section 'Your data file and required value' below to find out what value X you are required to calculate), and
- prints both values, $\tilde{W}$ and X, on the graph. If rounding up, keep at least two significant digits.

The code must create one graph only. This graph must have adequate axis labels, titles and a legend.

Write a short report (1 page maximum), addressing following points

- Describe the data you are given;
- Describe the distribution you get;
- How do you calculate the mean value? What value do you get?
- How do you calculate the required value X? What value do you get?

You should include mathematical formulae where appropriate. You can add your own comments.

What to submit and where?

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

- 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.
- The file must be named $< IDnumber >$.py, where $< IDnumber >$ is your student ID number. Do not include any other elements in the file name.

You are strongly advised to test your code using Spyder before submitting it.
Submit **your report** as a single file.

- This must be a single PDF, MS Word or Open Office document (PDF is preferred).
- It should have no more than 3000 characters (approximately, one page of Arial 12pt font).
- 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.

Your data file and required value

The data file you have to use and the value X your code should be producing (in addition to the average weight) depend on **the last digit of your student ID number**.

If it ends with '0' then you have to use file data0.csv. The value of X should be such that 25% of people have a salary above X.

If it ends with '1' then you have to use file data1.csv. The value of X should be such that 33% of people have a salary above X.

If it ends with '2' then you have to use file data2.csv. The value of X should be such that 5% of people have a salary above X.

If it ends with '3' then you have to use file data3.csv. The value of X should be such that 33% of people have a salary below X.

If it ends with '4' then you have to use file data4.csv. The value of X should be such that 25% of people have a salary below X.

If it ends with '5' then you have to use file data5.csv. The value of X should be such that 10% of people have a salary below X.

If it ends with '6' then you have to use file data6.csv. Calculate the value X, which is the fraction of population with salaries between $\tilde{W}$ and $1.25\tilde{W}$ (where $\tilde{W}$ is the mean value in your distribution).

If it ends with '7' then you have to use file data7.csv. Calculate the value X, which is the fraction of population with salaries between $0.75\tilde{W}$ and $\tilde{W}$ (where $\tilde{W}$ is the mean value in your distribution).

If it ends with '8' then you have to use file data8.csv. Calculate the value X, the standard deviation of the distribution.

If it ends with '9' then you have to use file data9.csv. Calculate the value X, which is the fraction of population with salaries between $0.8\tilde{W}$ and $1.2\tilde{W}$ (where $\tilde{W}$ is the mean value in your distribution).