Introduction to Data Science and AI, Sofia, 2023

Assignment 1: Introduction to Data Science and Python

In this assignment you will work with data sets from https://ourworldindata.org/ and Python to produce thoughtful analyses and interesting visualisations. Figures should be clear (what each axis represents, units used, etc.). Consider appropriateness of different types of plot for different purposes. Motivate each step taken, and each answer given.

Download some data related to GDP per capita[1] and life expectancy[2].

a. Write a Python program that draws a scatter plot of GDP per capita vs life expectancy. State any assumptions and motivate decisions that you make when selecting data to be plotted, and in combining data.

Answer these questions:

b. Which countries have a life expectancy higher than one standard deviation above the mean?
c. Which countries have high life expectancy but have low GDP? Motivate how you have chosen to define “high” and “low”.
d. Does every strong economy (normally indicated by GDP) have high life expectancy?
e. Related to question d, what would happen if you use GDP per capita as an indicator of strong economy? Explain the results you obtained, and discuss any insights you get from comparing the results of d and e.

Self-check: please read this before submitting your report

In a data science project, it is usually not sufficient to write a program, to run it, and then to present a graph or table with results. We should also think about the data that has been used, look at the results and consider whether these seem reasonable.

Did you do any data cleaning (e.g., by removing entries that you think are not useful) for the task of drawing scatter plot(s) and the task of answering the questions above? If so, explain what kind of entries you chose to remove and why.

Check whether your results for questions b and c include just countries. Some rows of the data files might contain information aggregated per continent or on the global level, rather than data about individual countries.

Sometimes students list countries that we would consider having a high GDP among countries that “have high life expectancy but have low GDP”. This can be because an input file contains GDP figures for many years and over a century ago many countries would have had a GDP that is lower than today’s average GDP. Check whether the list of countries in your answer to question c includes countries that we would consider having a high GDP.

[1] https://ourworldindata.org/economic-growth

[2] https://ourworldindata.org/life-expectancy

Python libraries

You are encouraged to use standard Python libraries (including pandas, numpy, matplotlib) in the programming assignments in this course. In particular, we recommend using pandas to read the data files in this assignment.

What to submit

- A report (a PDF file) that includes the figures produced and the descriptions/discussions that are requested in the questions.
- All Python code written. This may take the form of (i) individual Python files (submit the Python source code **and** append the code as an appendix in your PDF report), or (ii) a Jupyter notebook (submit both the Jupyter notebook **and** a separate PDF file exported from Jupyter).

In each file that you submit, give the names of the people submitting the work. On the first page of the report state how many hours each person spent working on the assignment.

Deadline: Tuesday 24 October 2023