Assignment 1: Introduction to Data Science and Python

Due September 4, 2023 at 13:00

1 Instructions

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

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.

1.1 Tasks

Download some data related to GDP per capita[1] and life expectancy[2]. Then, 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, as well as any choices made in combining data.

Then, answer the following questions:

1. Which countries have a life expectancy higher than one standard deviation above the mean?
2. Which countries have high life expectancy but have low GDP (note the difference between *GDP* and *GDP per capita*)? Motivate how you have chosen to define "high" and "low."
3. Does every strong economy (normally indicated by GDP) have high life expectancy?
4. Related to the above question (question 3), what happens if you use GDP *per capita* as an indicator of a strong economy as opposed to GDP alone? Explain the results you obtain through this analysis, and discuss any insights you get from comparing these results to question 3.

1.2 What To Submit

- A PDF report 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).
- **Do not submit "zip" files**.

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.

Ensure that all group members have joined the Assignment Group in Canvas before submitting your solution.

2 Self-Check: *Please read this before submitting your report*

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

For this assignment in particular, it is important to check for yourself the following points:

- ☐ Did you do any data cleaning (e.g., by removing entries that you believe are not useful) for the task of drawing the scatter plot(s) and answering the questions above? If so, be sure to explain what kind of entries you chose to remove, and why.
- ☐ Check whether your results for questions 1 & 2 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 to have 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 2 includes countries that we would consider to have a high GDP.

Additionally, for all assignments, **always perform the following self-checks** before submission:

- ☐ Have you answered all questions to the best of your ability?
- ☐ Is all the required information on the front page (e.g., you and your partner's names, the hours each partner worked, the correct file name, etc)
- ☐ Anything else you can easily check? (e.g., clearly labeled axes in figures, clearly labeled units, clear terminology and arguments, clearly stated answers, etc)

Do not submit an incomplete assignment! We teachers are available to help you, and you can receive a short extension if you contact us.

References

[1] Roser, M., Arriagada, P., Hasell, J., Ritchie, H., & Ortiz-Ospina, E. (2019). *Economic Growth*. Retrieved August 26, 2023, from https://ourworldindata.org/economic-growth

[2] Roser, M., Ortiz-Ospina, E., & Ritchie, H. (2019). *Life Expectancy*. Retrieved August 26 2023, from https://ourworldindata.org/life-expectancy