

DAT405/DIT405 Introduction to Data Science and AI,

Assignment 1: Introduction to Data Science and Python

In this assignment you will work with several data sets from <https://ourworldindata.org/> and Python to produce thoughtful analyses and interesting visualisations. Figures (what each axis represents, unit used, etc.) should be clear/not ambiguous. Consider appropriateness of different types of plot for different purposes. Motivate each step taken, and each answer given.

1. Download some data related to GDP per capita and life expectancy.
 - 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. [1p]
 - b. Consider whether the results obtained seem reasonable and discuss what might be the explanation for the results you obtained. [1p]

Answer these questions:

- c. 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 d, e, f, and g? If so, explain what kind of entries that you chose to remove and why. If not, explain why you did not need to. [0.5p]
 - d. Which countries have a life expectancy higher than one standard deviation above the mean? [0.5p]
 - e. Which countries have high life expectancy but have low GDP (per capita)? [0.5p]
 - f. Does every strong economy (normally indicated by GDP) have high life expectancy? [1p]
 - g. Related to question f, 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 g and f. [1p]
2. Download some other data sets, e.g. related to happiness and life satisfaction, trust, corruption, etc.
 - a. Think of several meaningful questions that can be answered with these data, make several informative visualisations to answer those questions. State any assumptions and motivate decisions that you make when selecting data to be plotted, and in combining data. [2.5p]
 - b. Discuss any observations that you make, or insights obtained, from the data visualisations. [2p]

What to submit

- All Python code written.
- A report that includes the figures produced and the descriptions/discussions that are requested in the questions.

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.

