

DAT405 Introduction to Data Science and AI

2020-2021, Reading Period 1

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 interesting visualisations.

1. Write a Python program that draws a scatter plot of GDP per capita vs life expectancy. State any assumptions and decisions that you make when selecting data to be plotted, and in combining data. Consider whether the results obtained seem reasonable.

Answer these questions:

- a. Which countries have a life expectancy higher than one standard deviation above the mean?
 - b. Which countries have high life expectancy but have low GDP?
 - c. Does every strong economy have high life expectancy?
 - d. Clean the original data by removing entries which are not useful to the previous tasks. The cleaned data should be stored in a data frame or something similar. Explain what kind of entries that you chose to remove and why.
2. Download some other data sets, e.g. related to happiness and life satisfaction, trust, corruption, etc., and make informative visualisations based on these. State any assumptions and decisions that you make when selecting data to be plotted, and in combining data. Discuss any observations that you make, or insights obtained, from the data visualisations.

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.

Deadline: Monday 7 September 2020 at 23:59.