Introduction to Data Science and AI – Report for Assignment 1

Big Data Technologies, Sofia University "St. Kliment Ohridski"

Stefan Dimitrov Velev, 0MI3400521

*In this assignment I had to work with data sets from **Our World in Data©** and **Python** so as to make a scatter plot of GDP per capita vs life expectancy.*

In the beginning let's analyse and process the data before making the visualisations.

The packages I've used in this assignment are: ***Pandas*** (for reading and storing data), ***NumPy*** (for making statistics) and ***Matplotlib*** (for visualisation).

Data URL: https://ourworldindata.org/grapher/life-expectancy-un-vs-gdp-per-capita-wb

Data source: UN, World Population Prospects (2024); World Bank (2023)

The data is stored in a *csv file*. The name of the separate columns can be found in the *Jupyter notebook* accompanying this report. The *csv file* contains 59 858 entries. However, most of them will not be useful for our purposes. Here comes the data cleaning part. First, I decided to remove the columns which are not required for the task – 'Code', 'Year', 'Continent'. After that, I renamed the columns left with more concise names:

'Entity' → 'Country'

'Life expectancy – Sex: all – Age: 0 – Variant: estimates' → 'Life expectancy'

'GDP per capita, PPP (constant 2017 international $)' → 'GDP per capita'

'Population (historical)' → 'Population'

The next step was to choose the year which I wanted to examine. Since I'd like up-to-date statistics, my focus is on 2022. Then I dropped the rows with missing values. After examining the data, I decided to remove the following aggregated entries with country field: 'High-income countries', 'Low-income countries', 'Lower-middle-income countries', 'Upper-middle-income countries', 'World'. The reason is that they are not necessary for the statistics about individual countries.

While inspecting the data, I noticed that there was an outlier. That is for the life expectancy of the Central African Republic for 2022 – according to the *World Bank Report* for 2022[1] it is 54.48 years, not 18.818 years. For the purposes of the task and for better visualisation, I decided to correct it. Finally, I had a *DataFrame* with 188 entries which is a satisfactory result since the total number of countries is 195 and I assume entries for some of them are probably missing or incomplete.

For the visualisation task I've made several scatter plots using different approaches. The first scatter plot I've presented is one made with the help of ***Pandas***. This diagram can only be used for an overall picture of how data is distributed. That is why the name of the countries is not shown on the scatter plot.

[1] https://data.worldbank.org/indicator/SP.DYN.LE00.IN

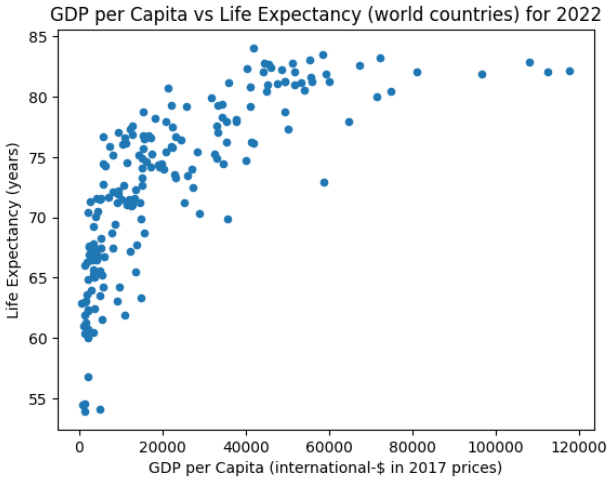

Figure 1: *GDP per Capita vs Life Expectancy (world countries) for 2022*

The most detailed scatter plot is the second one for which I've used ***Matplotlib***. For that, some considerations were taken into account:

- the scatter plot is much larger since the number of entries is 188
- the circles of the scatter plot are proportional to the population of each country for 2022
- the name of each country can be found next to its circle – for better clarity it would be good if it is visualised only on hoovering over it; however this cannot be achieved with ***Matplotlib***

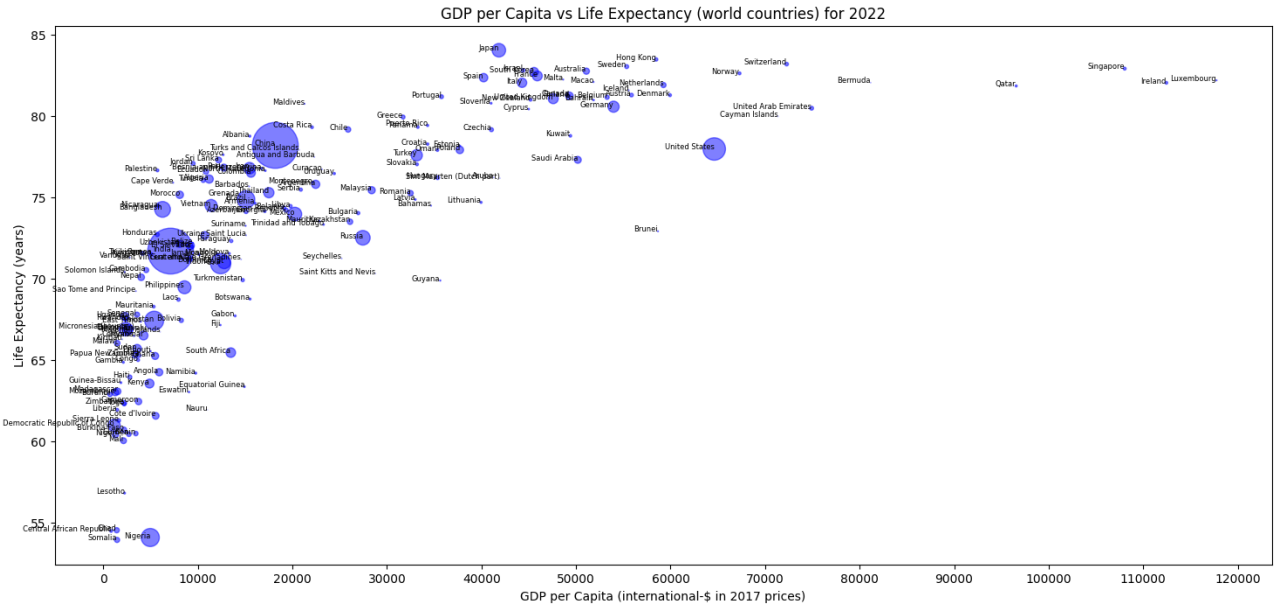

Figure 2: *GDP per Capita vs Life Expectancy (world countries) for 2022*

However large the scatter plot is, so many names of countries cannot be presented on just one diagram. That's why I've taken the decision to divide this scatter plot into two separate ones. The criteria used is 'GDP per Capita' – less than 20 000 in the first scatter plot and above it – in the second one. In that way, the names of the countries become clearer. For both scatter plots the y-limits are the same so that readers can make easier comparisons between countries from both diagrams.

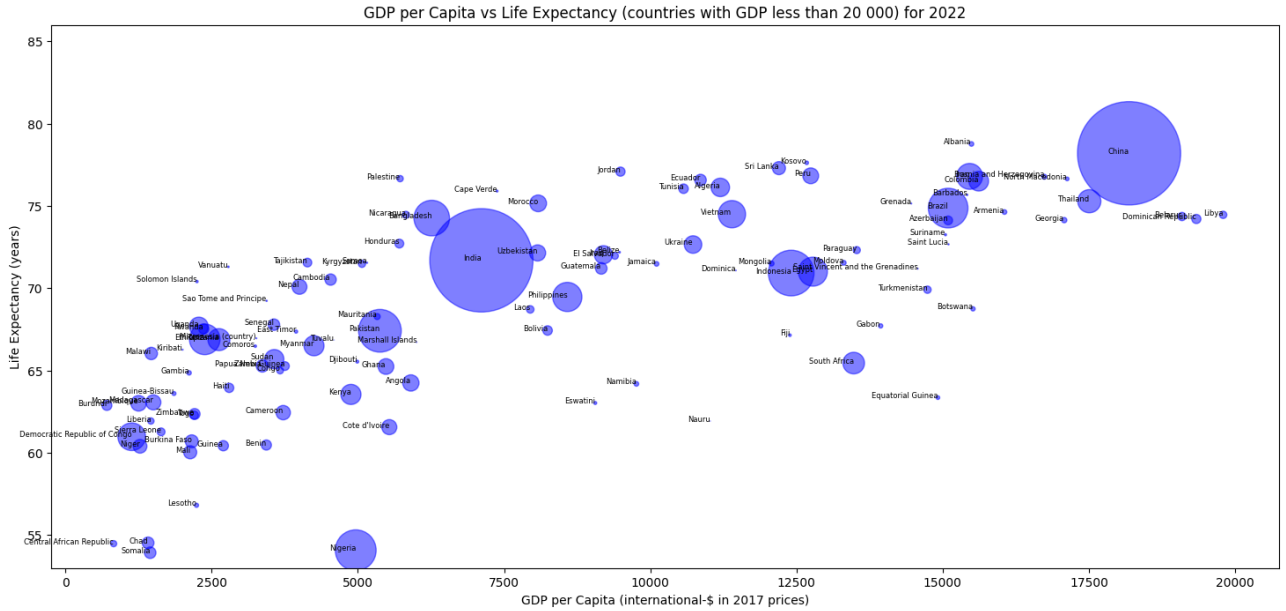

Figure 3: *GDP per Capita vs Life Expectancy (countries with GDP less than 20 000) for 2022*

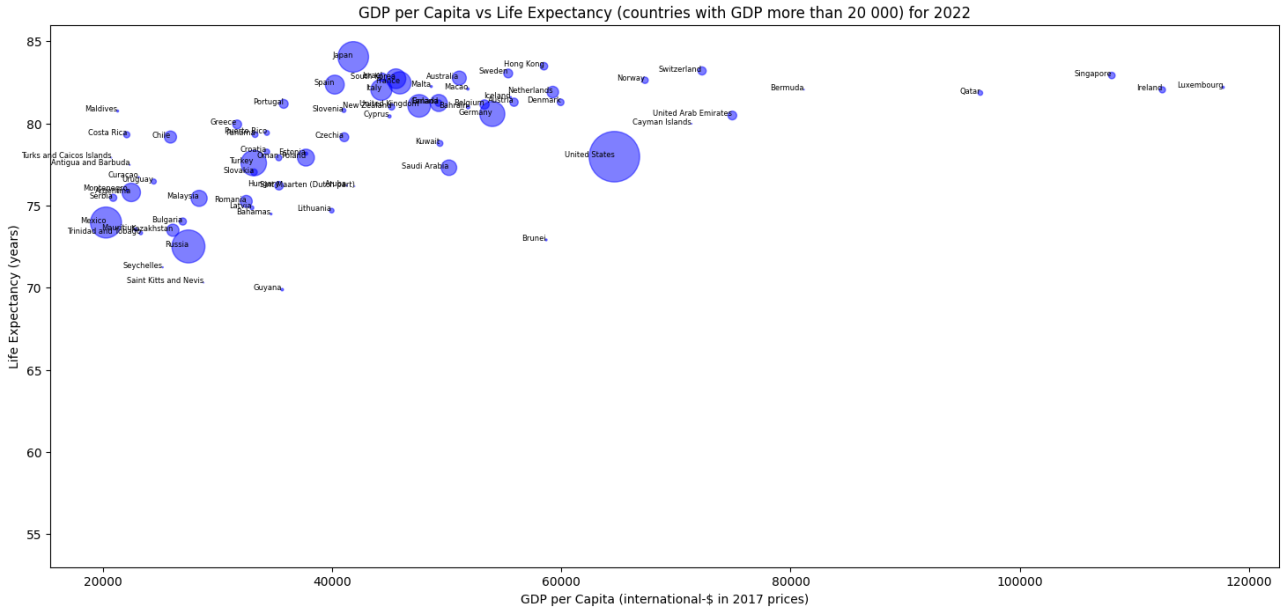

Figure 4: *GDP per Capita vs Life Expectancy (countries with GDP more than 20 000) for 2022*

Without the y-limits for the second scatter plot, the visualisation is even clearer.

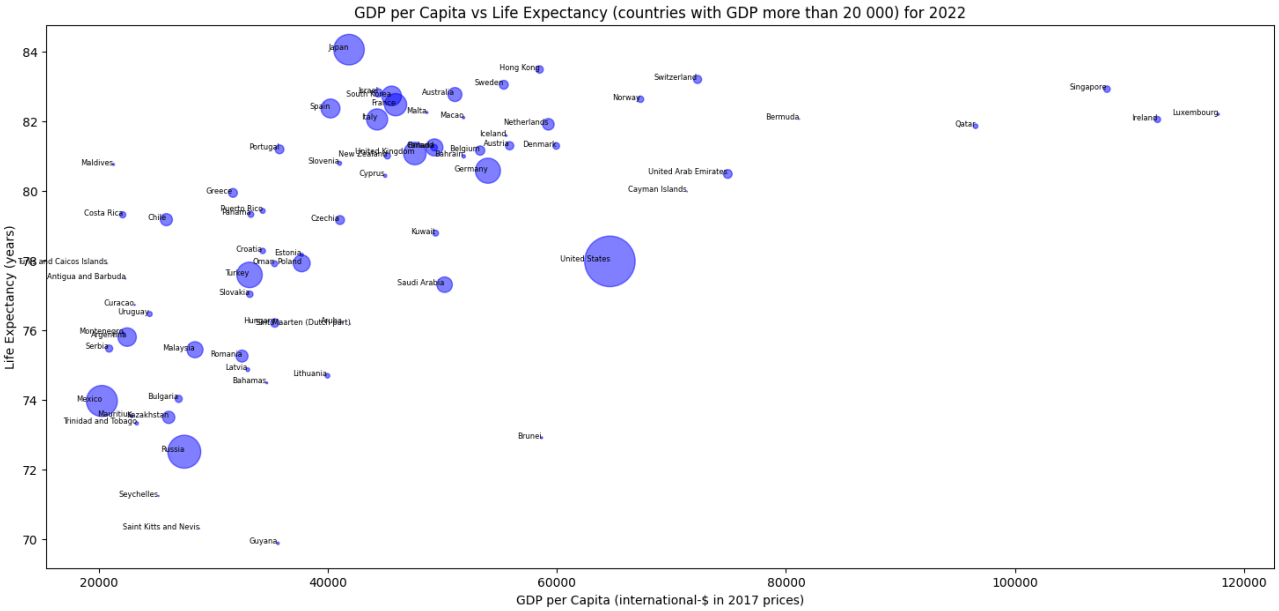

Figure 5: *GDP per Capita vs Life Expectancy (countries with GDP more than 20 000) for 2022*

Answer to the assignment questions:

b. Which countries have a life expectancy higher than one standard deviation above the mean?

With ***NumPy*** I've determined that the mean life expectancy is 72.75 and the standard deviation is 7.02. So, the lower boundary for the life expectancy of the desired countries is 79.77 (mean + standard deviation). So, using ***Pandas*** I've concluded that the number of these countries for 2022 is 36 and they are as follows:

Australia, Austria, Bahrain, Belgium, Bermuda, Canada, Cayman Islands, Cyprus, Denmark, Finland, France, Germany, Greece, Hong Kong, Iceland, Ireland, Israel, Italy, Japan, Luxembourg, Macao, Maldives, Malta, Netherlands, New Zealand, Norway, Portugal, Qatar, Singapore, Slovenia, South Korea, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom

Most of the countries are in Europe (21), 10 in Asia, 3 in North America and 2 in Australia/Oceania.

c. Which countries have high life expectancy but have low GDP? Motivate how you have chosen to define "high" and "low".

For defining "high" and "low" I've used two different measures. First, I tried to work with the means, i.e. "high" life expectancy – life expectancy above the mean life expectancy of all countries, and "low" GDP per Capita – GDP per Capita below the mean GDP per capita of all countries. With that consideration the searched number of countries is 38 and they are:

Albania, Algeria, Antigua and Barbuda, Argentina, Armenia, Azerbaijan, Bangladesh, Barbados, Belarus, Bosnia and Herzegovina, Brazil, Cape Verde, China, Colombia, Costa Rica, Dominican Republic, Ecuador, Georgia, Grenada, Iran, Jordan, Kosovo, Libya, Maldives, Mexico, Montenegro, Morocco, Nicaragua, North Macedonia, Palestine, Peru, Serbia, Sri Lanka, Suriname, Thailand, Tunisia, Turks and Caicos Islands, Vietnam

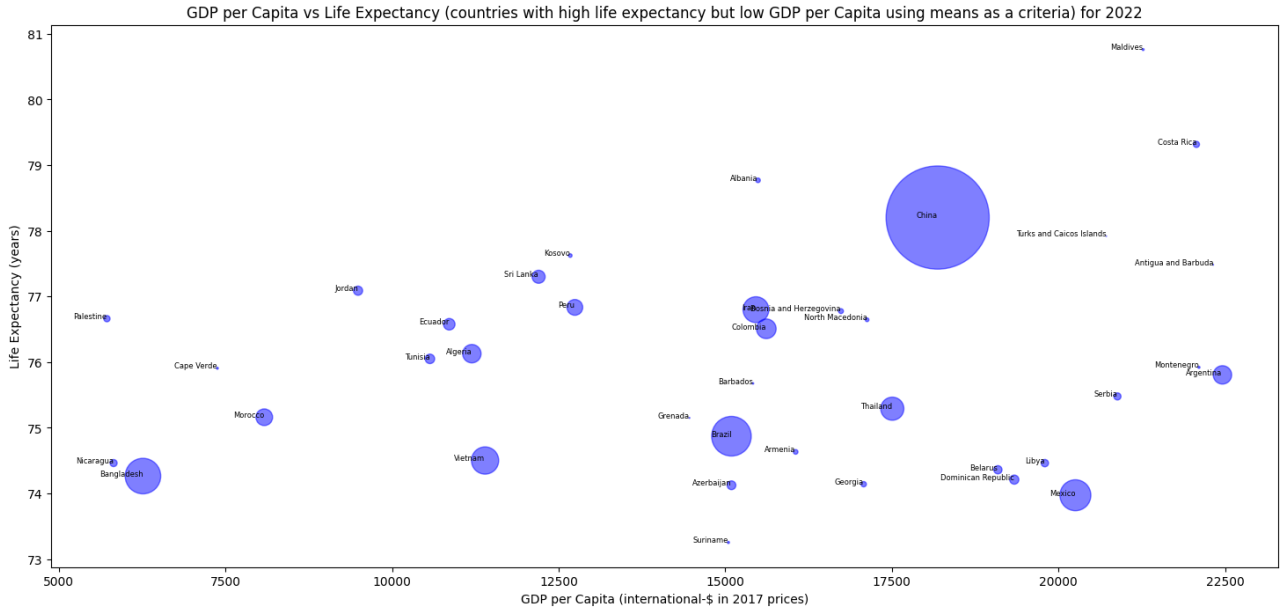

Figure 6: *GDP per Capita vs Life Expectancy (countries with high life expectancy but low GDP using means) for 2022*

Another approach is using the 75th percentile for "high" life expectancy and the 25th percentile for low GDP per Capita. In that way we're narrowing the interval of such countries and it can even be noticed that there are not such countries fulfilling that requirement for 2022. This emphasizes the connection between GDP per Capita and life expectancy.

d. Does every strong economy (normally indicated by GDP) have high life expectancy?

For defining "strong" economy I've used the 75th percentile for determining the lower boundary of high GDP per Capita countries – that is 34 831.85. The number of countries crossing that barrier is 47.

Now we have to define "high" life expectancy. I'll use the mean life expectancy and the 75th percentile – 72.75 and 77.94. The results are:

- 38 strong economies have life expectancy above the 75th percentile (> 77.94 years)

Australia, Austria, Bahrain, Belgium, Bermuda, Canada, Cayman Islands, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Hong Kong, Iceland, Ireland, Israel, Italy, Japan, Kuwait, Luxembourg, Macao, Malta, Netherlands, New Zealand, Norway, Portugal, Qatar, Singapore, Slovenia, South Korea, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, United States

- 8 strong economies have life expectancy between the mean life expectancy (>= 72.75 years) and the 75th percentile (<= 77.94 years)

Aruba, Brunei, Hungary, Lithuania, Oman, Poland, Saudi Arabia, Sint Maarten (Dutch part)

- 1 strong economy has life expectancy below the mean life expectancy (< 72.75 years)

Guyana

I've visualised these results on the last scatter plot which depicts all strong economies as well as the mean and 75th percentile life expectancy as horizontal lines outlining the intervals.

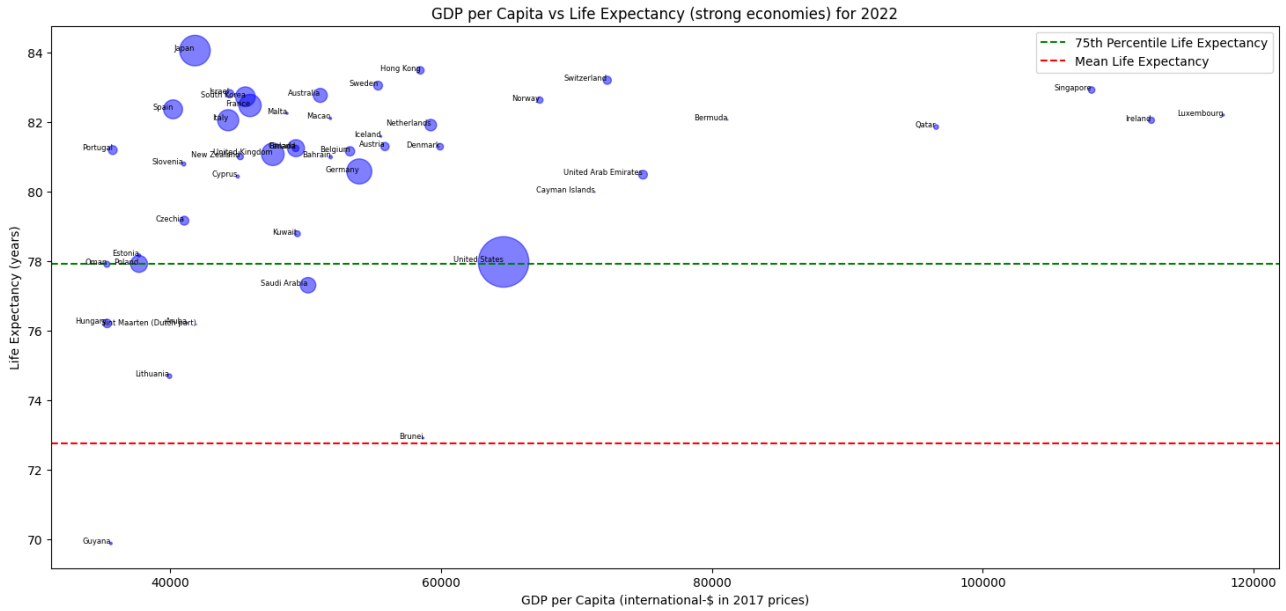

Figure 7: *GDP per Capita vs Life Expectancy (strong economies) for 2022*

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.

For this assignment I've used GDP per capita everywhere as this is the only given value in the *csv file* of the data. GDP (Gross Domestic Product) represents the total economic output of a country. GDP per capita, on the other hand, divides GDP by the population size giving an average economic output or income per person. Therefore, it's a better indicator of the standard of living and individual economic well-being in a country, regardless of its population size. If I had data for GDP, I would expect countries with large populations and economies (e.g. China, Japan, USA) to have high GDP. However, some of them might actually have low GDP per capita and as a result not so high standard of living. Countries with high GDP do not necessarily have high life expectancy. Large economies can include countries where wealth is unevenly distributed or where healthcare and living conditions vary widely. That's why GDP per capita is a better metric when being used as such an indicator.