UNIVERSITY OF TARTU

Andmetehnika mitteinformaatikutele (LTAT.02.026)

GDP, Life Expectancy and Literacy analysis

Final project

Margus Varjak
Fred Väärtnõu
Eva Meinson

Tartu 2023

Project's objective

The primary objective of this project was to identify correlations among the GDP, life expectancy, and literacy rates across various countries and regions.

Data sources and ETL process

We derived data from The World Bank publicly available repository:

- https://data.worldbank.org/

We downloaded three different datasets, for data fetching, cleaning, combining, and visualization.

- GDP per capita: https://api.worldbank.org/v2/en/indicator/NY.GDP.PCAP.PP.CD?downloadformat=csv
- Life expectancy: https://api.worldbank.org/v2/en/indicator/SP.DYN.LE00.IN?downloadformat=csv
- Literacy: https://api.worldbank.org/v2/en/indicator/SE.ADT.LITR.ZS?downloadformat=csv

Our aim was to download *.csv files, however, .zip files were provided by the website. Thus, *.zip files were downloaded and unpacked, a separate folder was made for each indicator. Following, zip files were unpacked into the corresponding '*unpacked' folder, the file starting with 'API*' denotes the *.csv file of interest, other two are for meta information.

We also required another file, which contains a list of countries and their corresponding regions and sub-regions. That information is needed for grouping as the analysis progresses. This file was downloaded from:

- https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/raw/master/all/all.csv

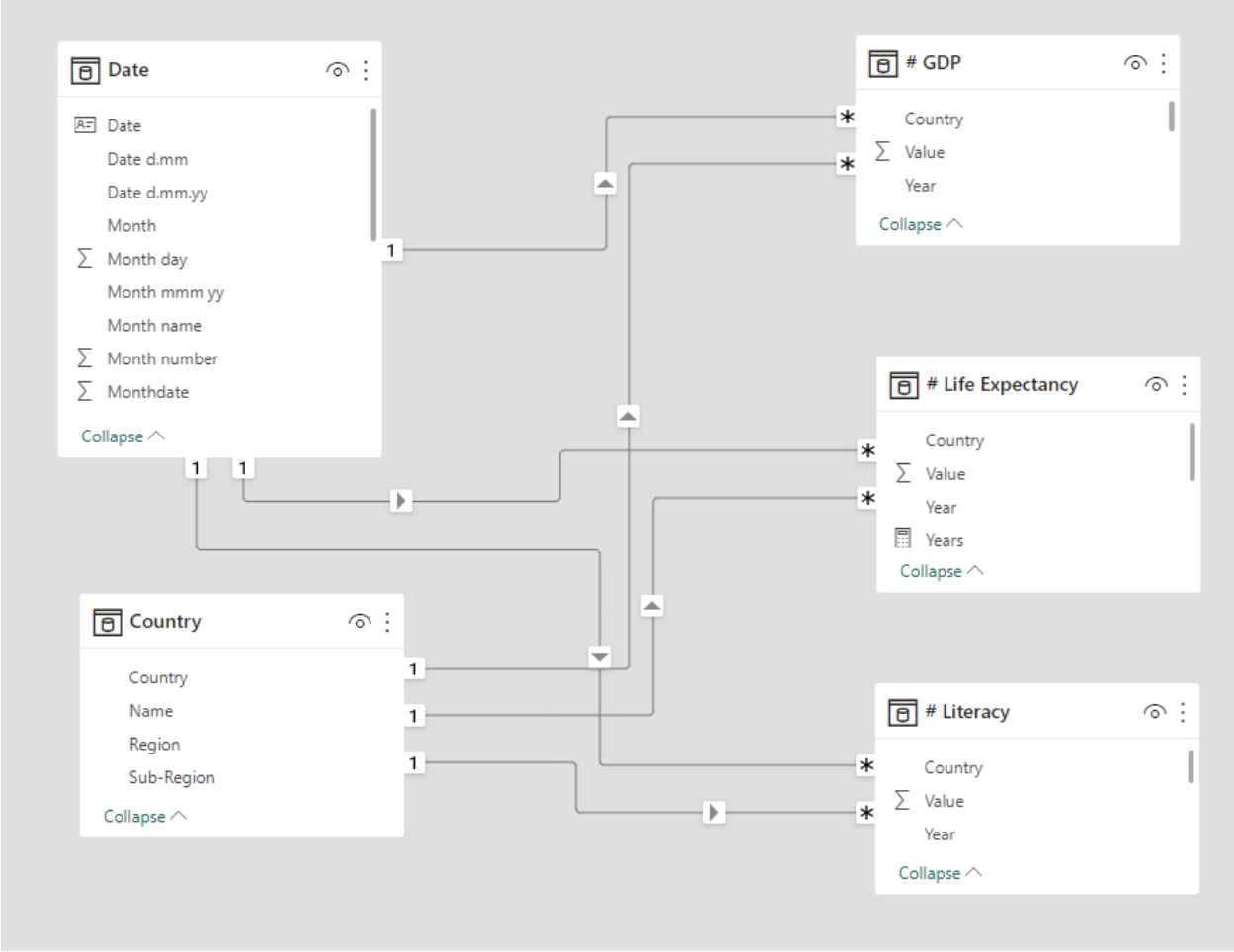
In that file, for each country, there is its name, country codes; also regions, sub-regions, and their codes.

Important notes regarding processing: due to the nature of datasets, there are many NaN values, essentially data earlier than the 1990s is lacking, thus we can look back 30 years only. Further, literacy data is relatively incomplete, meaning that not for every year literacy values are given. However, to estimate the literacy levels in 2021, we extrapolated the values from an earlier year, when measurements had been conducted.

Visualization

PowerBI Dashboard

First was created the PowerBI data model including Country, GDP, Life expectancy and Literacy. The data model gets data directly from the .csv files in Github repository (created by the ETL process). Following is the picture of the final data model in Power BI:



The main visualization was done using Microsoft Power BI Desktop. The report was uploaded to powerbi.com account and published to the Web. The report is publicly accessible for all on following link:

- https://app.powerbi.com/view?r=eyJrIjoiMjk3N2YyMDktYmUwYS00MjdlLWFlODEtMzk0NGVkMGRiYjQwIiwidCI6IjM2YTc0ZjA4LTYwMDUtNDA5OS05ZTFjLTg1NDU5ZGFjMjEzZiIsImMiOjh9

The PowerBI dashboard consists of three different sheets, each having focus on the specific data area. It is possible to navigate between sheets using navigation buttons on the dashboard header. The whole dashboard can be filtered using filters on the right side of the dashboard. In the left upper corner of the dashboard are displayed the dynamic fact boxes based on user selection.

The main visual on the dashboard compares the average value in the world to user selected value. This enables comparison of specific region or country to the world average and see how numbers have changed over time.

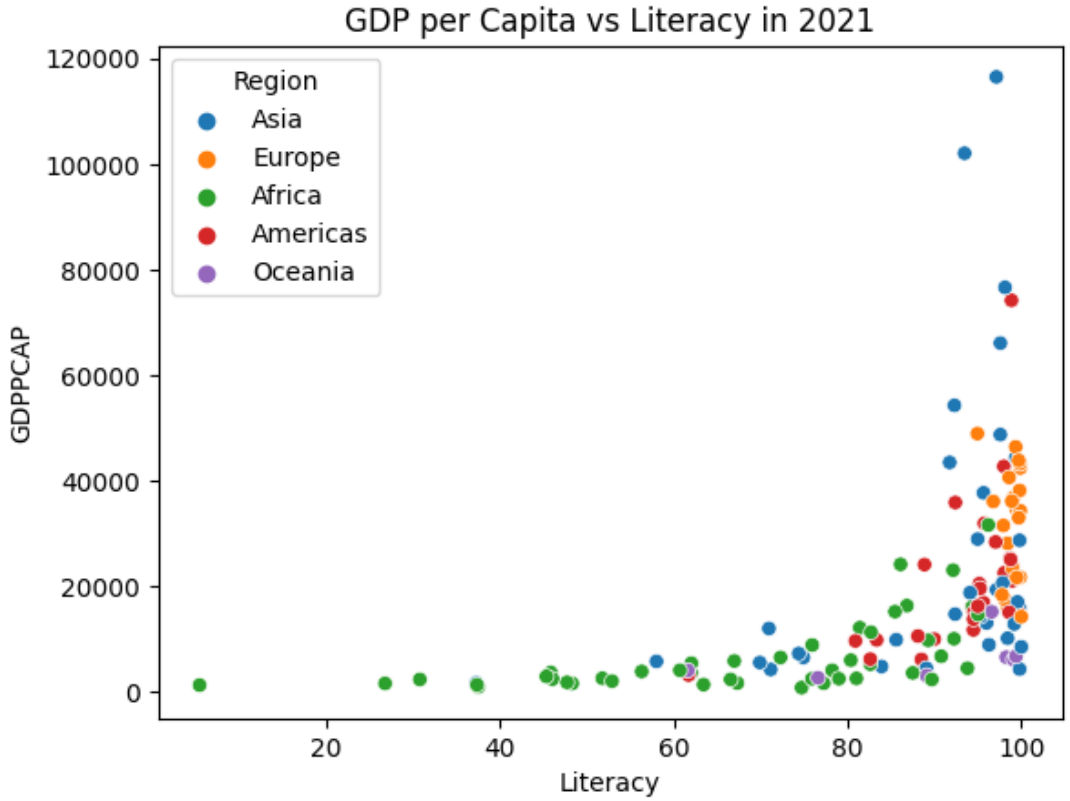
The scatter plot visual on bottom of the dashboard shows the correlation of the selected metric to the GDP. The tooltip on the datapoint shows information about the country and it's location on the map.

Visualization in Python

The additional visualizations were done using Python Matplotlib and Seaborn libraries.

Figure 1:

GDP per capita vs Literacy rate in 2021, coloring done based on Regions.

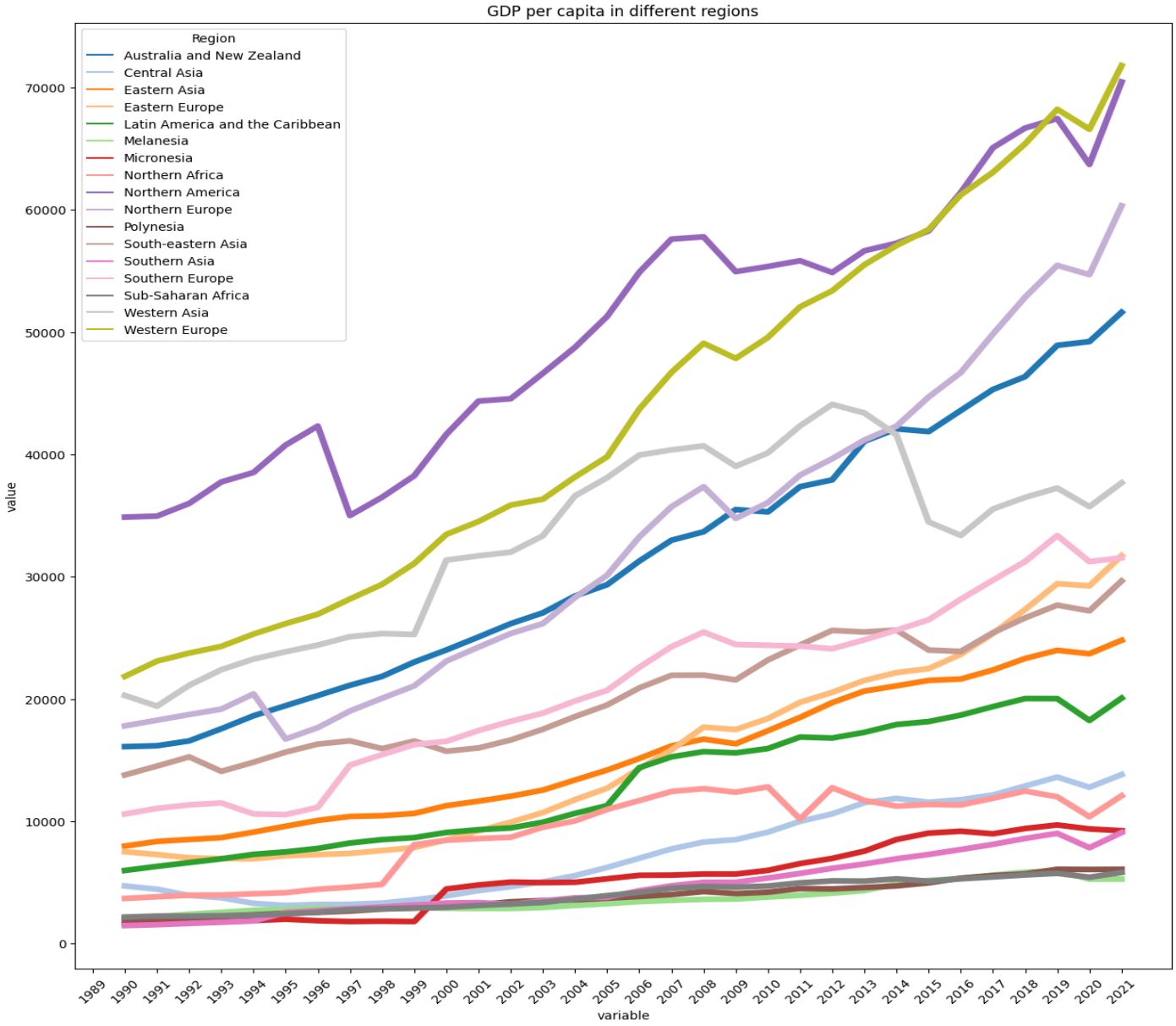


- Strong correlation between country's richness and literacy

- However, not all countries with high literacy are rich
- In Africa many countries have low literacy rates and these are poorer as well
- Europe has a very high literacy rate, Americas and Asia are mixed

Figure 2:

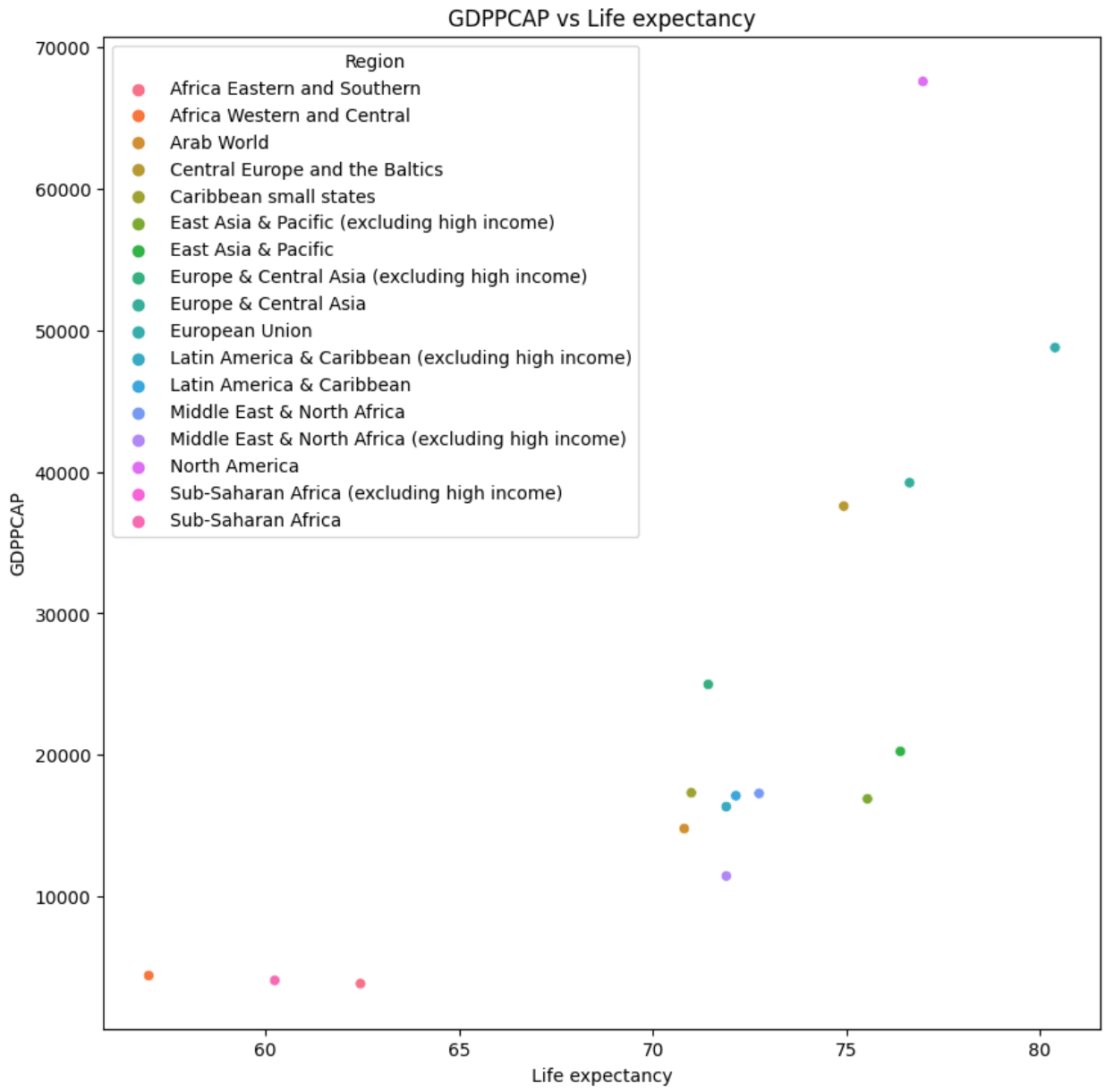
Line graph, progression of GDP per capita over last 30 years, in different sub-regions.



- Based on that North America and Western Europe are the richest (no surprises)
- Eastern Asia has gone through rapid progress
- Eastern Europe has still catching up to do (caught up with Southern Europe)
- African countries do not show much progress

Figure 3:

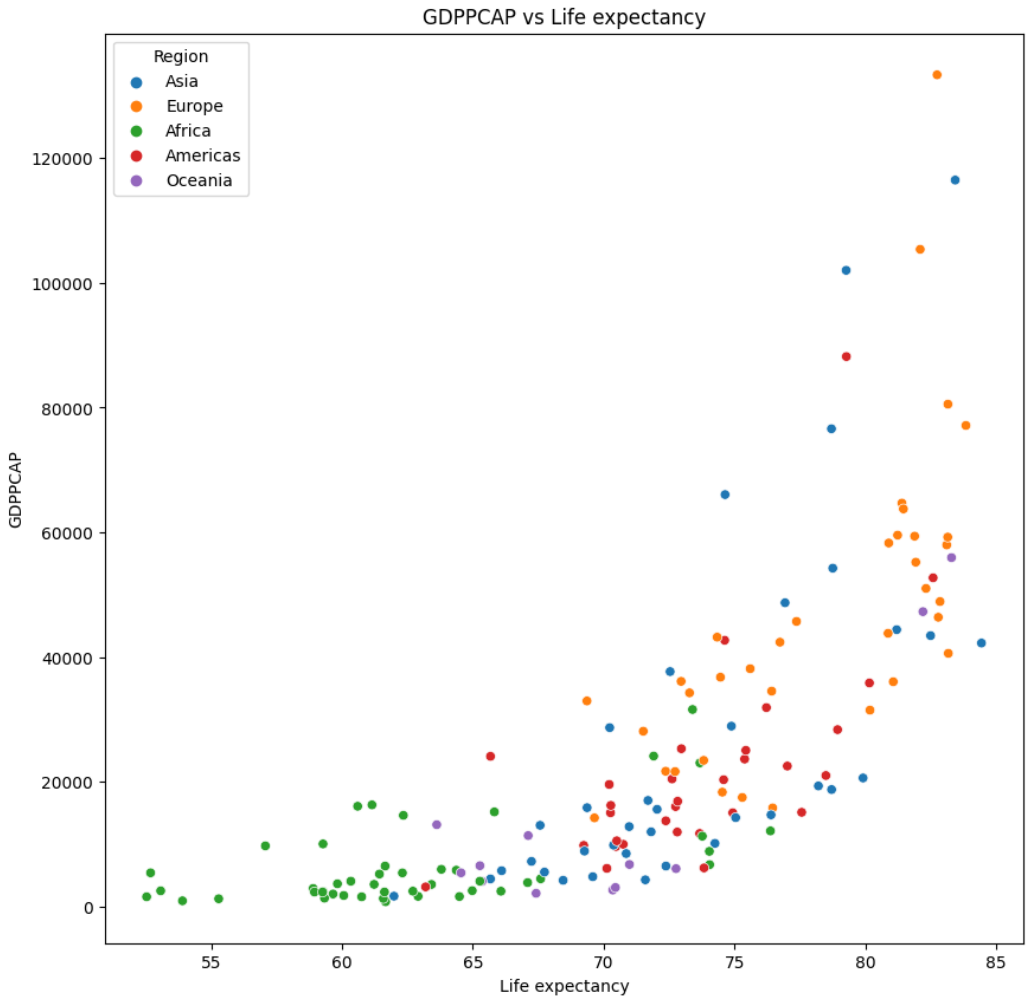
Scatterplot, GDP per capita vs Life expectancy in 2021 in different sub-regions



➢ Strong correlation between country's richness and life expectancy

Figure 4:

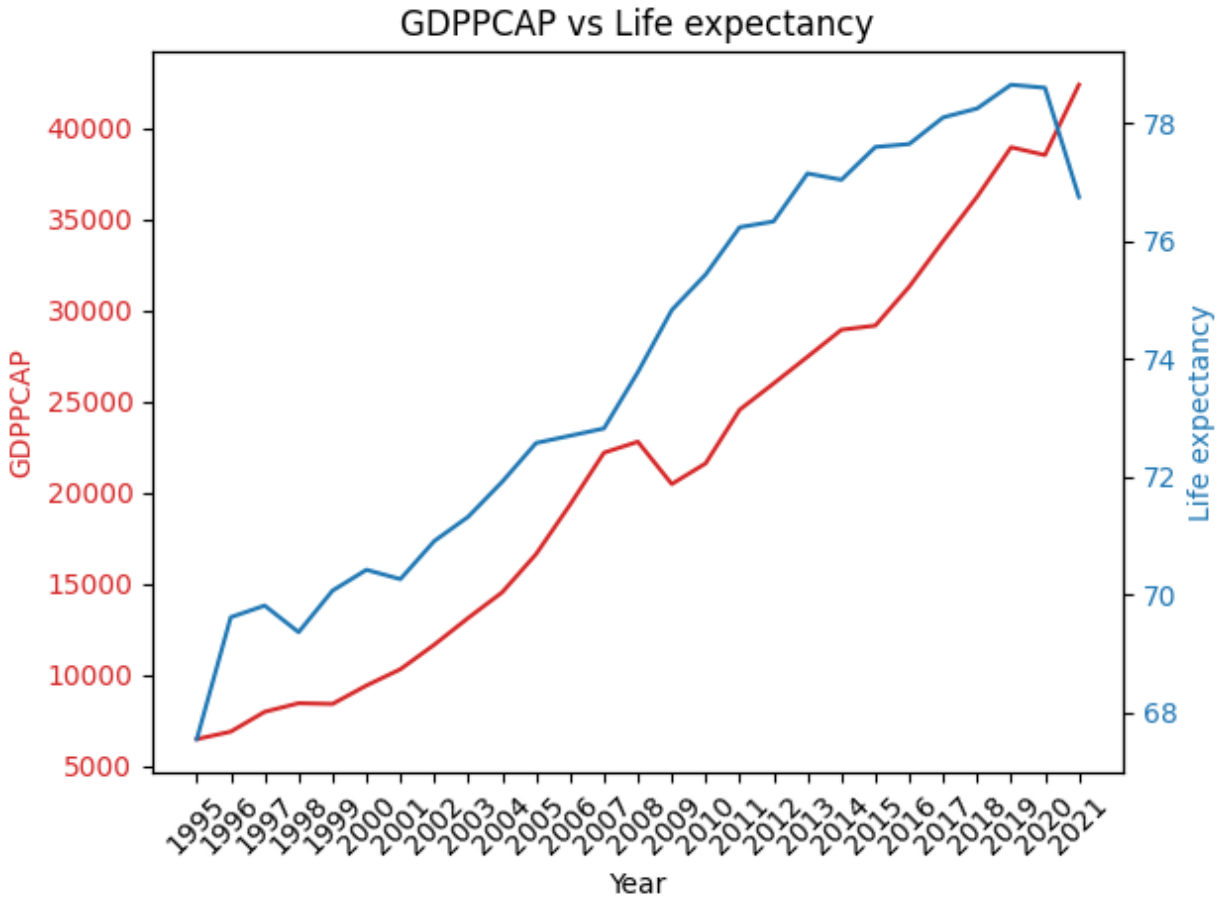
GDP per capita vs Life expectancy in 2021 in different countries, coloring based on regions



➢ Again, very strong correlation

Figure 5:

Line graph, Estonia specific, GDP per capita progression and life expectancy, over the last 30 years.



➢ From the graph, we can see that people are living longer, as the country gets richer