Investigation on the Impact of Tertiary Education on Life Expectancy

Grandi, Maccaferri, Zatti, Erling

Big Data For Social Sciences Project

University of Bologna

Professor Nicola Barban

A.Y. 2024/2025

Authored by:

- Francesco Grandi - 0001071804 - francesco.grandi6@studio.unibo.it
- Tommaso Maccaferri - 0001071630 - tommaso.maccaferri@studio.unibo.it
- Cosimo Zatti - 001068010 - cosimo.zatti@studio.unibo.it
- Johannes Erling - 0001073471 - johannes.erling@studio.unibo.it

The work in this project was equally divided.

Research Question

Despite rapid technological advancements and increasing global awareness, significant health and living standard disparities persist across nations. A good indicator of different health standards between countries is the comparison of their life expectancy. While numerous factors determine how long a person lives on average, we were specifically curious about the role education plays in shaping these outcomes. We hypothesize that higher tertiary education enrollment positively influences life expectancy, while other socioeconomic factors like GDP per capita, health expenditure, and fertility rates also contribute significantly to health outcomes. By identifying these relationships, we aim to uncover actionable insights for potentially enhancing global health outcomes.

Dataset Source and Description

The data used for this research comes from reliable online source Kaggle, and contains detailed information on various economic, health, and demographic indicators of different countries.

- **Primary dataset**: "Global Country Information Dataset 2023: A Comprehensive Dataset Empowering In-Depth Analysis and Cross-Country Insights".

  The original dataset, downloaded from Kaggle, featured 35 variables. For the purpose of our investigation, we reduced the dataset's variables, ending up with a dataset composed by the following:

Variable	Type of Variable	Description of Variable
Country	Nominal	Name of the country
Fertility Rate	Continuous	Average number of children born to a woman during her lifetime
GDP	Continuous	Gross Domestic Product, the total value of goods and services produced in the country
Gross Tertiary Education Enrollment (%)	Continuous	Gross enrollment ratio for tertiary education. Chosen as a proxy for education access and quality, hypothesized to enhance health literacy and outcomes.
Life Expectancy	Continuous	Average number of years a newborn is expected to live
Out-of-Pocket Health Expenditure (%)	Continuous	Percentage of total health expenditure paid out-of-pocket by individuals. Included to capture financial barriers to healthcare access, which might inversely affect life expectancy.
Population	Continuous	Total population of the country

Dataset Preparation

After importing the data, it was reviewed to identify any missing values. Imputation methods were applied to fill in the gaps and ensure dataset completeness. Subsequently, transformations were performed to merge the variables of interest and facilitate comparative analyses.

Then we calculate GDP per capita by dividing the countries' GDP by their population. This makes wealth data comparable across countries, which is essential for examining its relationship with education and life expectancy.

Analysis

Correlations and Relationships

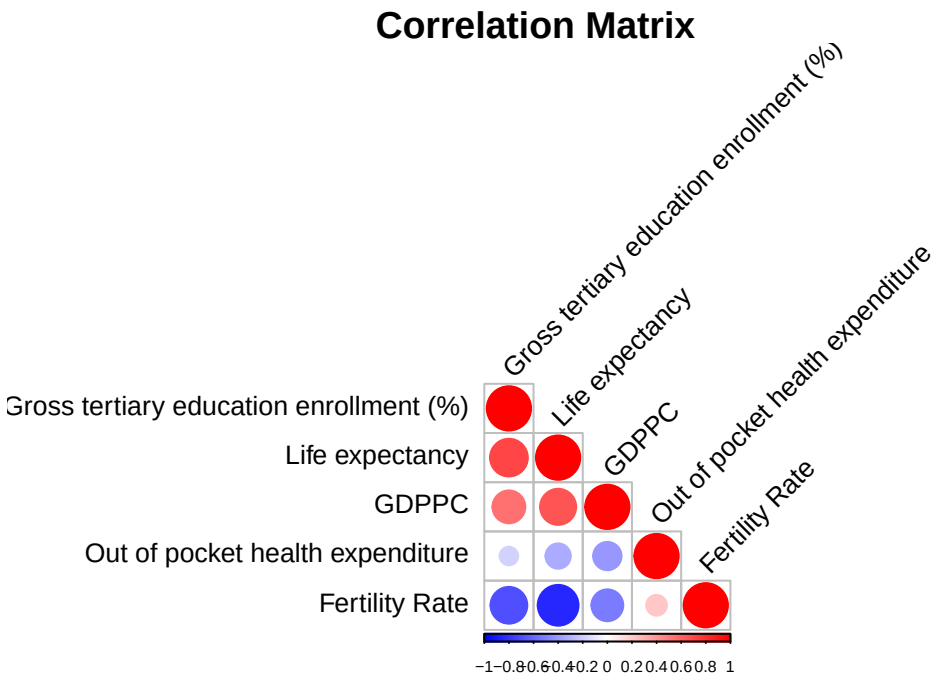
The average life expectancy was analyzed in relation to the gross tertiary education enrollment (%); in particular a correlation analysis was conducted to examine the association between tertiary education and other socioeconomic variables.

We can observe a high positive correlation between education enrollment and life expectancy with a correlation coefficient of 0.73. This supports the hypothesis that education is a significant driver of health outcomes.

There is a strong negative relationship between education enrollment and the fertility rate. (-0.69) An even stronger negative correlation (-0.85) can be witnessed between life expectancy and the fertility rate. These findings might reflect a demographic transition. Countries with better education and health outcomes tend to have lower fertility rates.

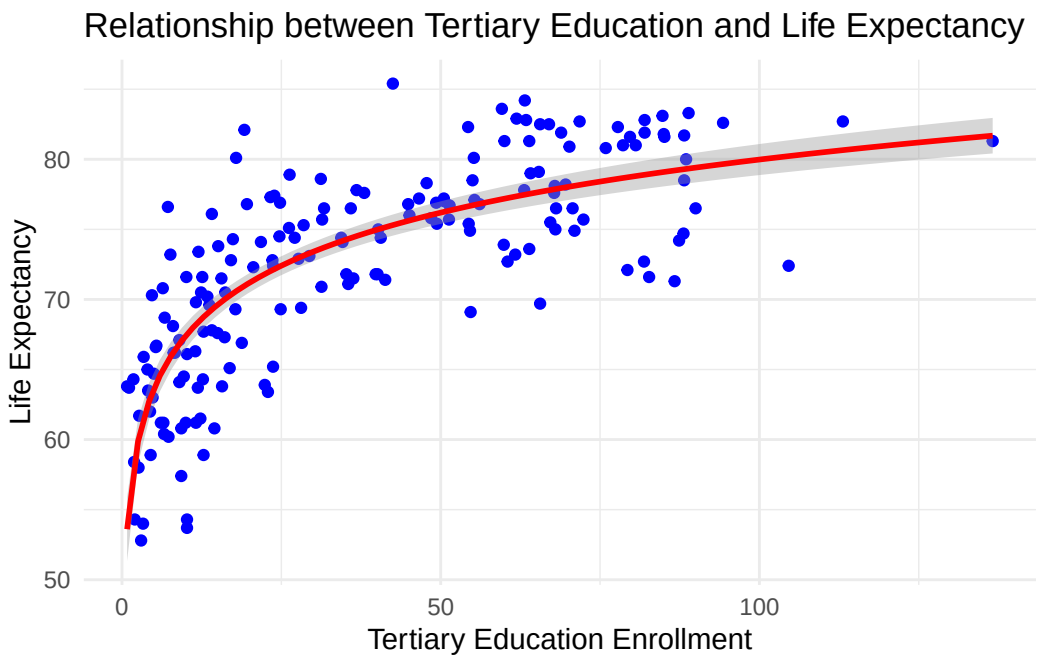
Correlation Matrix Data

Variable	Tertiary Edu Enroll	Life Expectancy	GDPPC	Health Expenditure	Fertility Rate
Gross tertiary education enrollment (%)	1.0000000	0.7260693	0.5506759	-0.1759809	-0.6885968
Life expectancy	0.7260693	1.0000000	0.6638918	-0.3260383	-0.8473891
GDPPC	0.5506759	0.6638918	1.0000000	-0.4099383	-0.5162118
Out of pocket health expenditure	-0.1759809	-0.3260383	-0.4099383	1.0000000	0.2194340
Fertility Rate	-0.6885968	-0.8473891	-0.5162118	0.2194340	1.0000000



Visualizations

Scatter plot were created to illustrate the relationship between tertiary education enrollment and life expectancy. The logarithmic regression line highlights the diminishing marginal returns of the effect of education on life expectancy. While life expectancy increases with education, the impact diminishes at higher levels. This suggests additional factors might play a more significant role in already highly educated countries.



PCA and Clustering

To reduce the dimensionality of the dataset and identify complex patterns, a principal component analysis (PCA) was performed. The first principal component (PC1) explains 63.1% of the total variance, while the second one (PC2) explains 18.8%. These components are linear combinations of the original variables and represent two new orthogonal dimensions. PC1 reflects what could be considered overall "development," summarizing variables like education, life expectancy (positive contributions), and fertility rate (negative contribution). PC2 captures economic spending patterns, distinguishing countries by GDP per capita and out-of-pocket health expenditure. So the results showed that education significantly contributes to the variance explained in the principal components.

```
Importance of components:
                          PC1    PC2     PC3     PC4    PC5
Standard deviation     1.776 0.9705 0.66916 0.57235 0.3571
Proportion of Variance 0.631 0.1884 0.08955 0.06552 0.0255
Cumulative Proportion  0.631 0.8194 0.90898 0.97450 1.0000
```

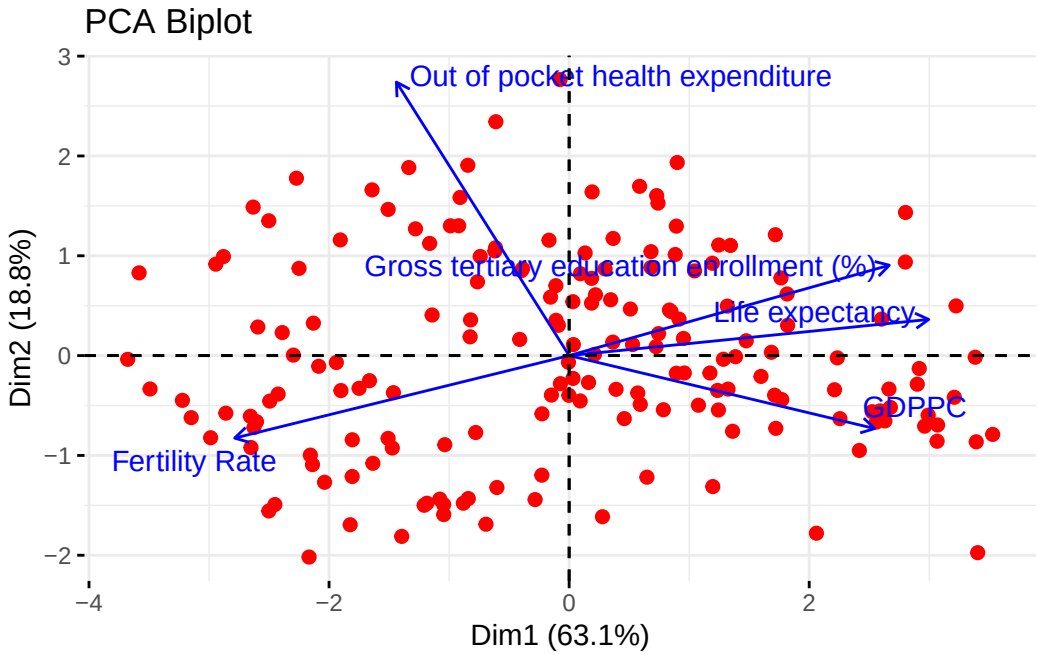
```
# A tibble: 25 x 3
   column                                    PC   value
   <chr>                                  <dbl>   <dbl>
 1 Fertility Rate                             1 -0.489
```

```
 2 Fertility Rate                              2 -0.266
 3 Fertility Rate                              3  0.427
 4 Fertility Rate                              4  0.395
 5 Fertility Rate                              5  0.593
 6 Gross tertiary education enrollment (%)     1  0.469
 7 Gross tertiary education enrollment (%)     2  0.291
 8 Gross tertiary education enrollment (%)     3  0.0225
 9 Gross tertiary education enrollment (%)     4  0.832
10 Gross tertiary education enrollment (%)     5 -0.0531
# i 15 more rows
```

```
# A tibble: 5 x 4
     PC std.dev percent cumulative
  <dbl>   <dbl>   <dbl>      <dbl>
1     1   1.78   0.631       0.631
2     2   0.971  0.188       0.819
3     3   0.669  0.0896      0.909
4     4   0.572  0.0655      0.974
5     5   0.357  0.0255      1
```

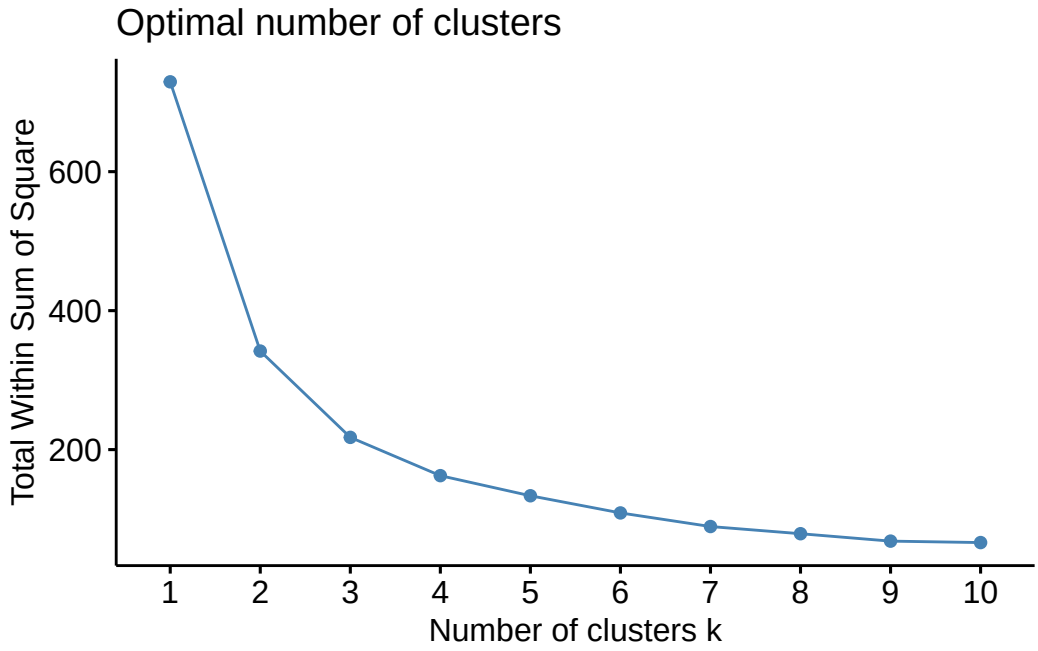
PCA Biplot

The Biplot shows again that wealthier countries, with higher health spending, also tend to have higher education and life expectancy. This suggests that economic factors complement education in achieving better health outcomes. The separation of countries in the biplot reflects different development levels, with those closer to the education and life expectancy arrows likely representing more developed nations.



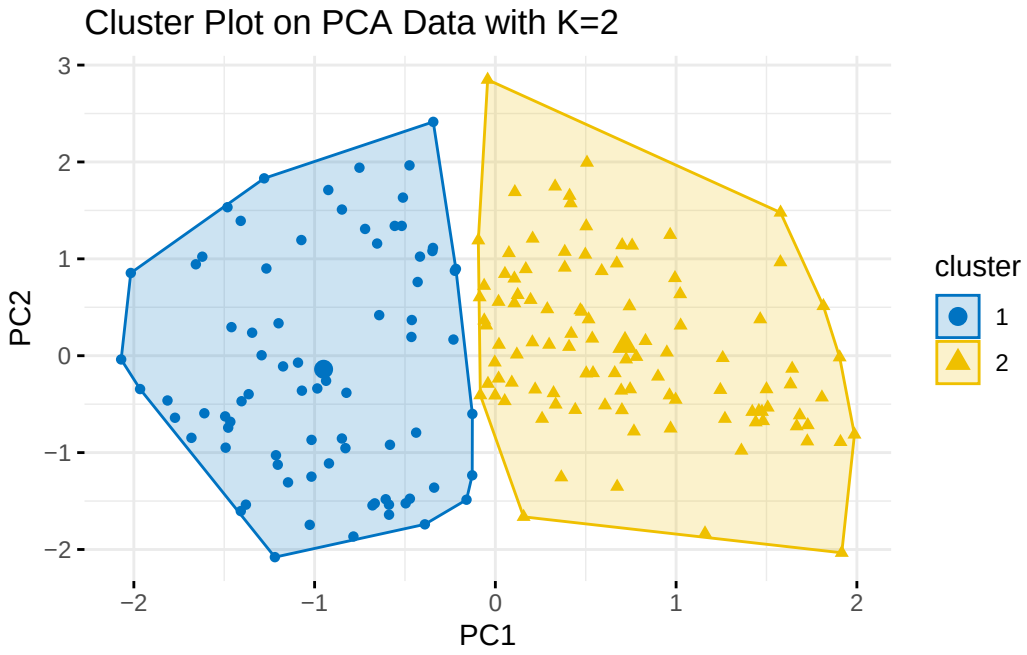
Optimal number of clusters

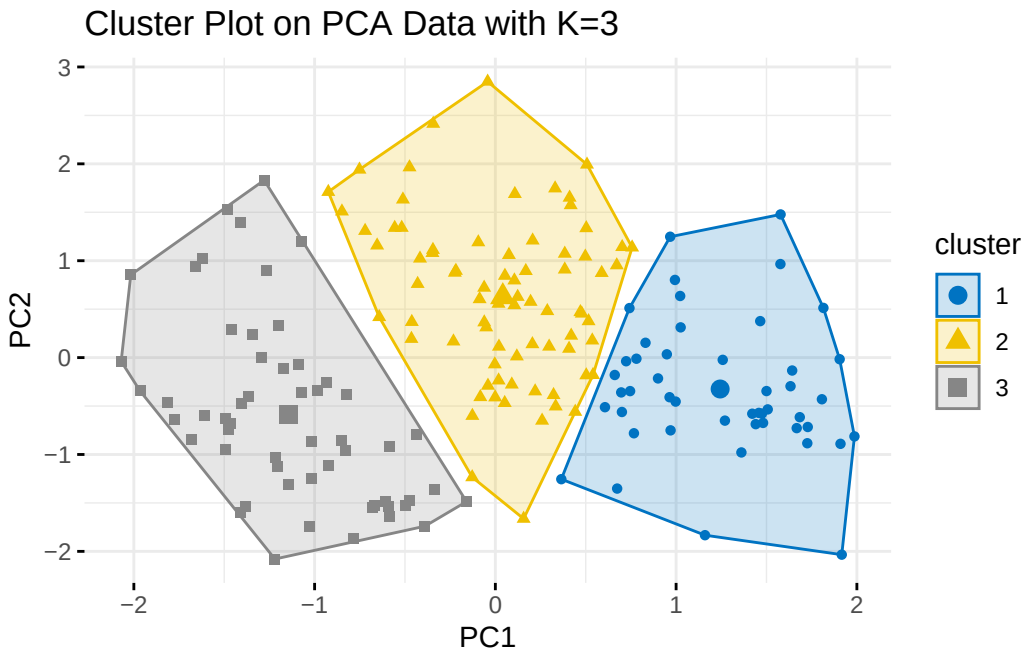
By using the "elbow rule" we determine the ideal number of clusters which here lies between two and three. Therefore, we do two cluster plots, one with K=2 and one with K=3.

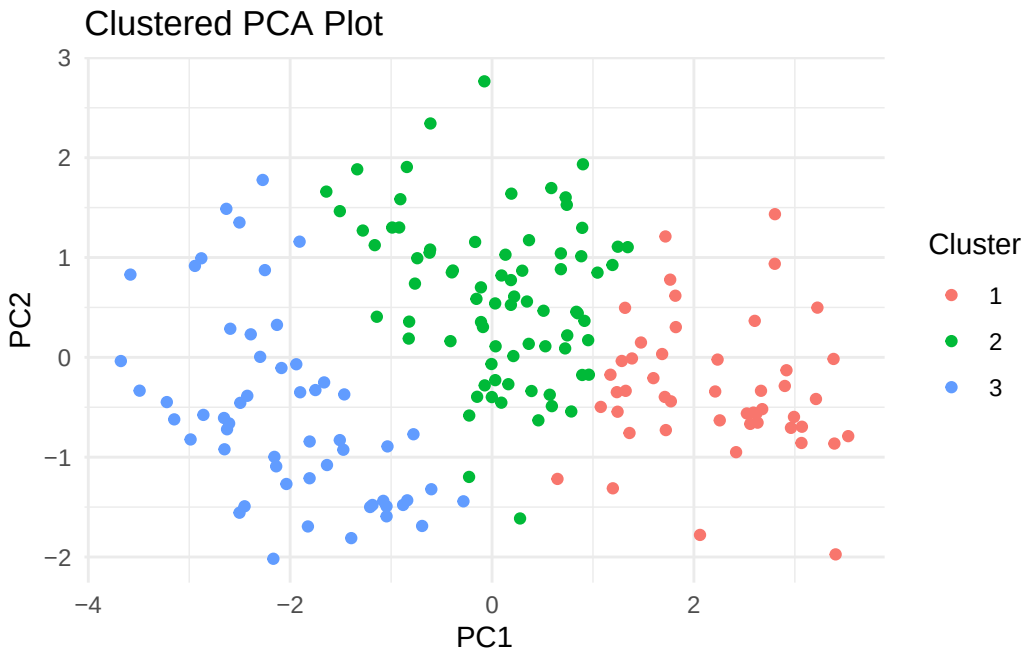


Cluster Plots

In the plot with three clusters, cluster number three likely resembles low-income countries with low education and life expectancy, and high fertility rates. The first cluster represents populations with high-income, high education and high life expectancy, while having low fertility rates. Cluster number two is made up of middle-income countries with moderate metrics. In the cluster plot with K=2, these countries with more moderate metrics are split up between the other two.





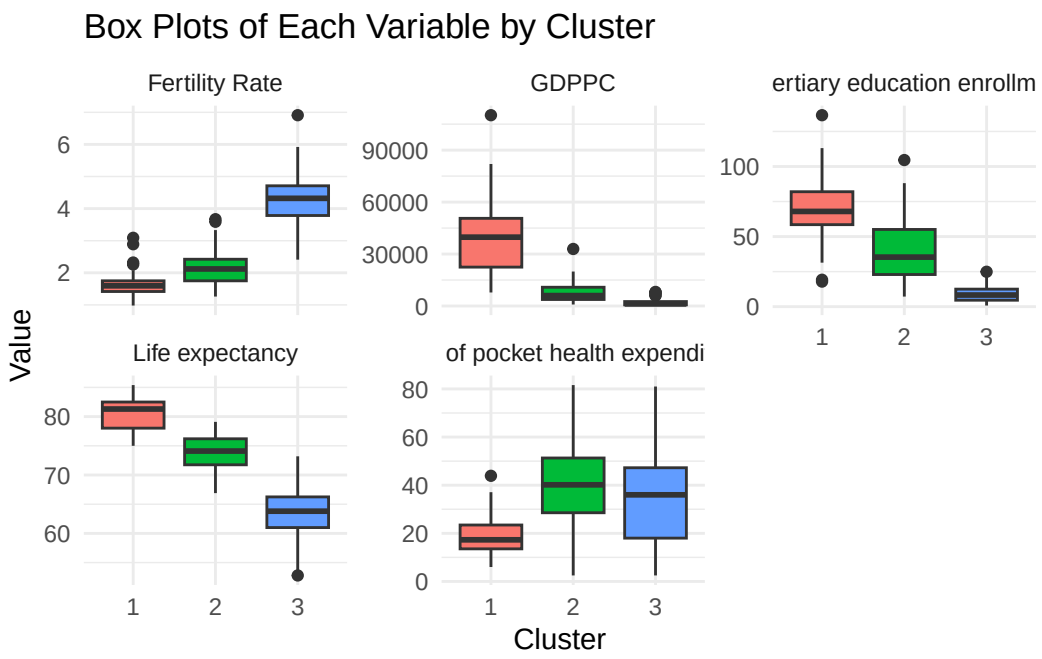


Box Plots

The boxplots underline the clear differences between clusters, showing that countries with higher education enrollment tend to have higher life expectancy and lower fertility rates, aligning with the research question. GDP per capita and health expenditure are also higher in

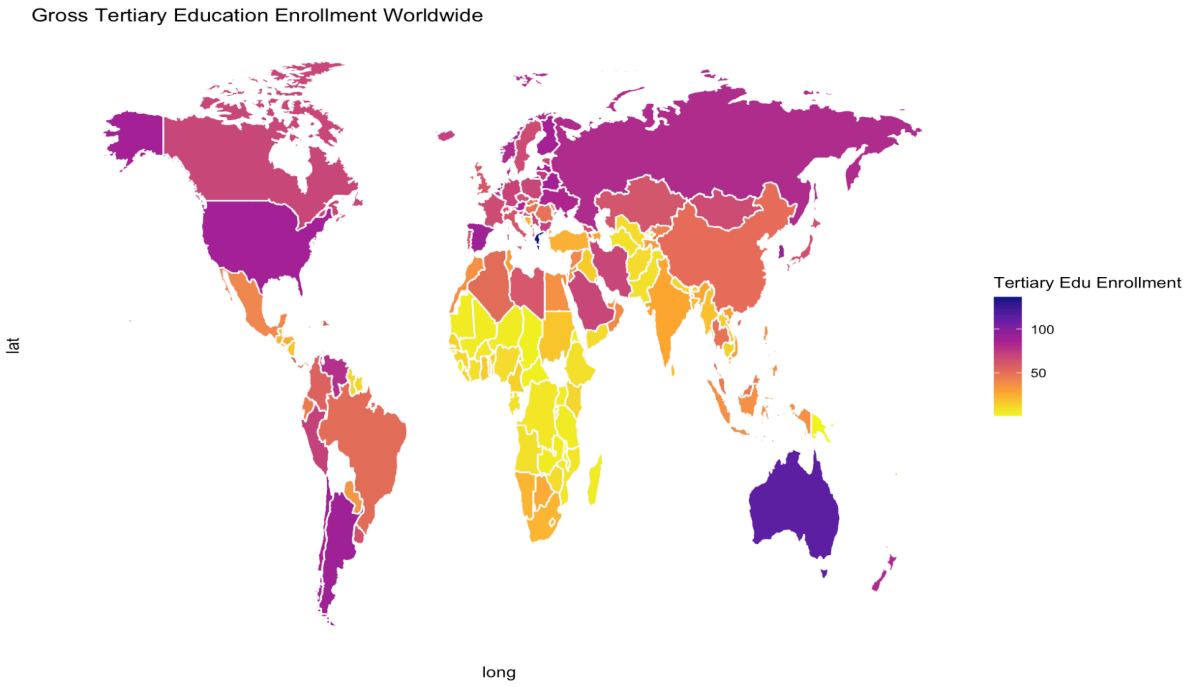
clusters with better education and health outcomes, suggesting that economic factors further enhance these relationships. Conversely, the cluster with the lowest education enrollment demonstrates significantly lower life expectancy and higher fertility, highlighting the strong link between education and demographic trends.

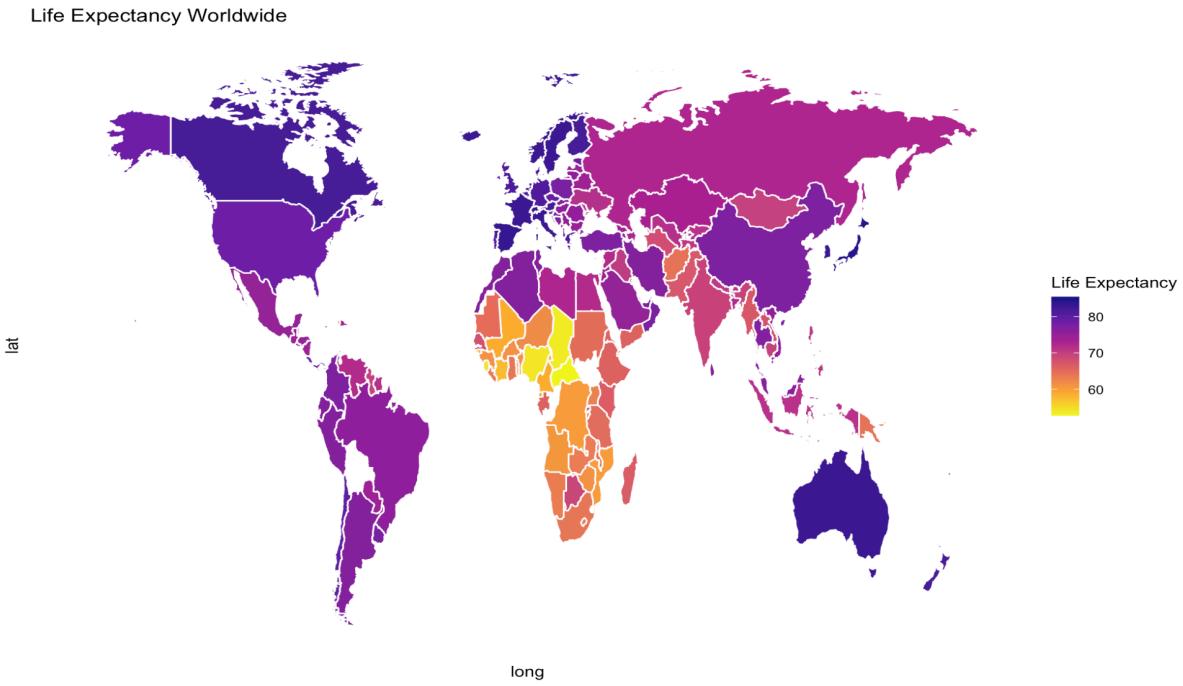
```
# A tibble: 1,790 x 4
   Country     cluster Variable                                 Value
   <chr>       <fct>   <chr>                                    <dbl>
 1 Afghanistan 3       Fertility Rate                            4.47
 2 Afghanistan 3       Gross tertiary education enrollment (%)   9.7
 3 Afghanistan 3       Life expectancy                          64.5
 4 Afghanistan 3       Out of pocket health expenditure         78.4
 5 Afghanistan 3       GDPPC                                   502.
 6 Afghanistan 3       .fittedPC1                               -2.63
 7 Afghanistan 3       .fittedPC2                                1.49
 8 Afghanistan 3       .fittedPC3                                1.01
 9 Afghanistan 3       .fittedPC4                               -0.209
10 Afghanistan 3       .fittedPC5                                0.303
# i 1,780 more rows
```

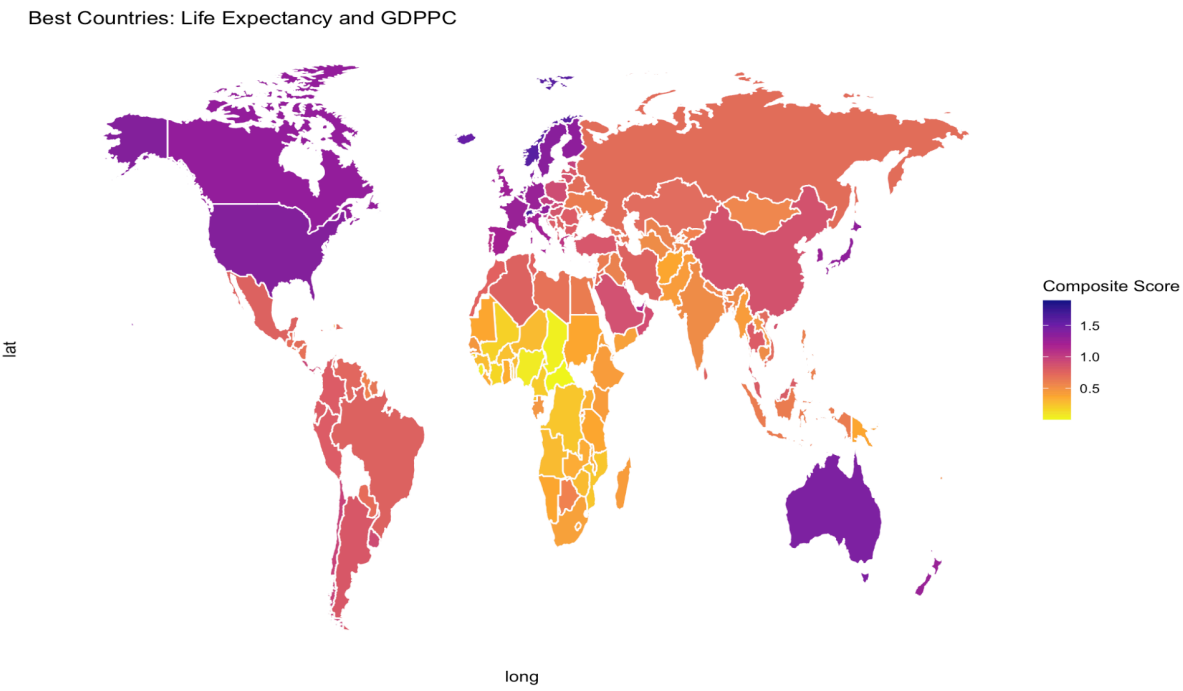


Mapping

The first map presents gross tertiary education enrollment rates worldwide, with darker colors signifying higher participation in higher education. This metric serves as a proxy for the strength of a country's educational infrastructure, economic development, and investment in human capital, while lighter regions indicate areas where limited access to tertiary education may hinder socio-economic progress. The second map illustrates global life expectancy, providing a visual representation of disparities in health outcomes across countries. Darker regions indicate higher life expectancy, often associated with advanced healthcare systems, better living conditions, and effective public health policies, whereas lighter regions highlight areas with lower life expectancy, reflecting socio-economic challenges and limited healthcare access. The third map combines normalized data on life expectancy and GDP per capita to produce a composite score, highlighting countries that achieve a balance between health and economic prosperity. The visualization identifies the best countries that combine high life expectancy, which often reflects a strong healthcare system and quality of life, with high GDP per capita, which suggests access to commodities such as advanced public transportation, commercial activities, and technological progress. Several South American and North African countries exhibit life expectancies nearly comparable to those of North America and Europe; however, when GDP per capita is considered, these countries fall behind in economic wealth and technological advancement, underscoring global inequalities in these dimensions.




Life Expectancy Worldwide
lat
long
Life Expectancy
80
70
60


Best Countries: Life Expectancy and GDPPC
lat
long
Composite Score
1.5
1.0
0.5

Conclusions

The main part of our initial research question aimed to get a deeper understanding of the interplay between the average education level and the quality of life in different countries. As these factors are both highly connected with other decisive variables, we decided to include also life expectancy, the countries respective fertility rates and their out-of-pocket health expenditure. This turned out to be a fruitful measure as we found strong positive correlations not only between GDPPC and the Gross tertiary education enrolment, but also between these two variables and life expectancy. All of these were negatively correlated with the out-of-pocket health expenditure and with the fertility rates, whereby the correlations with the latter were especially strong.

We established these patterns, doing a correlation analysis of which we illustrated the results, using a correlation matrix. To highlight the strong positive correlation between our two main variables of interest, GDPPC and tertiary education enrolment, we used a scatterplot.

In order to reduce the complexity of the dataset and to be able to see more complex patterns we performed a Principal Component Analysis. We focused on PC1 and PC2 which explained respectively 63.1% and 18.84% of the variance in the data set. These orthogonal variables represent overall development of countries (PC1) and their economic spending patterns (PC2). Since the elbow method suggested a value between two and three clusters, we tried both options. This resulted in different clusters that represent combinations of countries grouped according to what could be generally referred to as their level of wealth and development.

All of these instruments helped us to confirm our initial assumption that there is a strong link between a country's education standard and its wealth. Since also our other supposition, the interconnectedness with other factors, could be confirmed, there is proof of the high complexity of effects that these variables have on one another. Although we should be prudent in drawing causal conclusions from this study, the evidence highlights the critical importance of education as a driver of development. Ultimately, this research serves as a reminder that effective allocation of aid to struggling countries must be informed by comprehensive statistical analysis. Without this, the intricate relationships among factors such as education, health, and economic resources risk being overlooked, potentially leading to misguided interventions.

Limitations

Despite its strengths, our analysis has several limitations. First, the cross-sectional nature of the dataset does not allow us to infer causality, meaning we can only identify correlations rather than determine the direction or strength of causal relationships. Additionally, some important variables, such as healthcare quality, political stability, and cultural influences, were not included in our dataset but likely play a significant role in shaping the observed outcomes. Furthermore, the imputation of missing data, while necessary for completeness, may introduce biases that could affect the robustness of our results. Finally, the global averages

we examined may obscure within-country disparities, which would require a more granular analysis to uncover.

In conclusion, while this study highlights the significance of education and its interplay with other development indicators, it also underscores the need for future research that incorporates more comprehensive datasets and longitudinal approaches to better understand the causal mechanisms at play.