Economic Indicators: Correlation between a Country's GDP per Capita and its Life Expectancy

Group 3:

Miceli, Frank

Reus, Daniela

Lopez, Erwin

Del Bando, Denise Michelle

1. Introduction / Objectives:

This report uses data from the World Health Organization (WHO) to examine the relationship between GDP per capita and life expectancy across approximately 179 countries from 2000 to 2015.. The hypothesis is that these two variables are positively correlated, meaning that as a nation's income increases, its life expectancy also tends to rise. In this report, GDP per capita is treated as the independent variable because it encompasses several macroeconomic indicators, such as a country's wealth, government spending, and health insurance availability. These factors are believed to directly influence life expectancy, which is the dependent or response variable in this analysis.

By incorporating additional factors, the report highlights underlying issues that influence global life expectancy. While the findings are not exhaustive, they provide insights that can help public officials, aid organizations, and global health policymakers allocate resources more effectively to enhance life expectancy worldwide.

Many studies have explored the link, and even potential causation, between GDP and mortality rates. Any observed correlation offers evidence that economic resources can improve a country's quality of life. However, simply increasing GDP does not provide specific guidance on how to allocate funds. A more focused investment approach, such as improving infrastructure, transportation, or education, might allow countries with lower GDPs to enhance mortality rates without needing significant and often unattainable increases in income levels for their populations.

2. Data Preparation:

The data is found on explanatory and response variables for numerous countries. We retrieved the data from an open database at Kaggle.com. Since we can't use data of all countries of the world, we will use the dataset as a sample, which includes many countries of the world (economically developing and developed countries).

The data about Population, GDP, and Life Expectancy was updated according to World Bank Data. Information about vaccinations for Measles, Hepatitis B, Polio, and Diphtheria, alcohol consumption, BMI, HIV incidents, mortality rates, and thinness were collected from World Health Organization public datasets. Information about Schooling was collected from the Our World in

Data which is a project of the University of Oxford. The data was provided in CSV format and required very little cleaning.

Table 1. Data cleaning

Metric	Value
Number of Duplicate Entries	0
Number of Missing Values	0
Number of Features	21
Number of Observations	2864

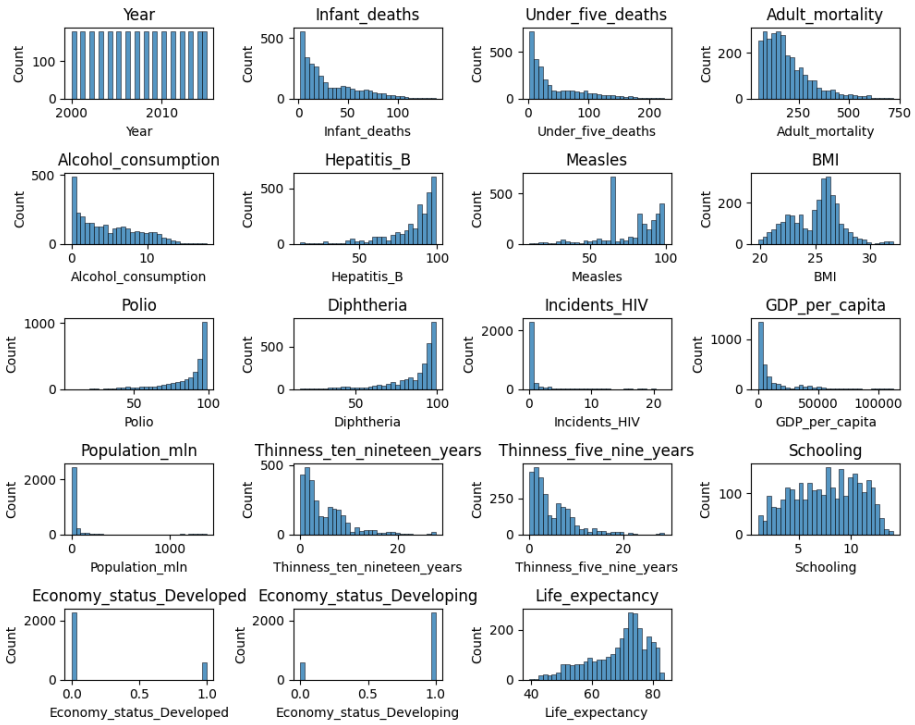
Data contains life expectancy, health, immunization, and economic and demographic information about 179 countries from 2000-2015 years. The dataset has 21 variables and 2,864 rows. This is an intriguing dataset that includes several features such as Population, GDP, Life Expectancy, Measles, Hepatitis B, Polio, and Diphtheria, alcohol consumption, BMI, HIV incidents, and mortality rates, as exemplified in **Table 2**.

Table 2. Variable types and input

Column	Variable Type
Country	object
Region	object
Year	int64
Infant_deaths	float64
Under_five_deaths	float64
Adult_mortality	float64
Alcohol_consumption	float64
Hepatitis_B	int64
Measles	int64
BMI	float64
Polio	int64
Diphtheria	int64
Incidents_HIV	float64
GDP_per_capita	int64
Population_mln	float64
Thinness_ten_nineteen_years	float64
Thinness_five_nine_years	float64
Schooling	float64
Economy_status_Developed	int64
Economy_status_Developing	int64
Life_expectancy	float64

For data cleaning, plotting and regression analysis we imported libraries like pandas, numpy, matplotlib, seaborn, scipy, sklearn, and statsmodels. Specific functions of high importance that were used are specified and described in the Analysis section. The data was analyzed and fitted using linear regression to find significant correlations between independent variables to explain and predict life expectancy.

Checking for outliers:



TEXT describing outliers check

3. Analysis

3.1. Preliminary Analysis using Descriptive Statistics

The summary output below (**Table 3**) states that life expectancy has a mean of 68.86 with a standard deviation of 9 years. In other words, the number of years one is expected to live as from the sample collected is 68.86 years. The life expectancy has a minimum and a maximum of 39.40 and 83.80, respectively. On the other hand, GDP per capita has a mean of $ 11,540.92 with a minimum and a maximum of $ 148.00 and $ 112,418.00, respectively.

Table 3. Descriptive statistics.

	Mean	Std Dev	Min	50%	Max
Year	2007.50	4.61	2000.00	2007.50	2015.00

Infant Deaths	30.36	27.54	1.80	19.60	138.10
Under-five Deaths	42.94	44.57	2.30	23.10	224.90
Adult Mortality	192.25	114.91	49.38	163.84	719.36
Alcohol Consumption	4.82	3.98	0.00	4.02	17.87
Hepatitis B	84.29	16.00	12.00	89.00	99.00
Measles	77.34	18.66	10.00	83.00	99.00
BMI	25.03	2.19	19.80	25.50	32.10
Polio	86.50	15.08	8.00	93.00	99.00
Diphtheria	86.27	15.53	16.00	93.00	99.00
HIV Incidents	0.89	2.38	0.01	0.15	21.68
GDP per Capita*	11540.92	16934.79	148.00	4217.00	112418.00
Population (M)	36.68	136.49	0.08	7.85	1379.86
Thinness (10-19)	4.87	4.44	0.10	3.30	27.70
Thinness (5-9)	4.90	4.53	0.10	3.40	28.60
Schooling*	7.63	3.17	1.10	7.80	14.10
Economy Status (Developed)	0.21	0.41	0.00	0.00	1.00
Economy Status (Developing)	0.79	0.41	0.00	1.00	1.00
Life Expectancy*	68.86	9.41	39.40	71.40	83.80

Figure 1 shows the relationship between GDP per capita and life expectancy, with data points colored by region. Key observations from the plot are that across all regions, there is a general trend where higher GDP per capita is associated with higher life expectancy. This trend is more pronounced in certain regions like the European Union. Furthermore, life expectancy at birth (e_0) seems to be higher in wealthy countries. There are distinct clusters based on regions. For example, countries in the European Union and some from the Middle East show higher GDP per capita and life expectancy, while regions such as Africa have lower values for both metrics. Swaziland has the highest life expectancy and Sierra Leon has the lowest life expectancy.

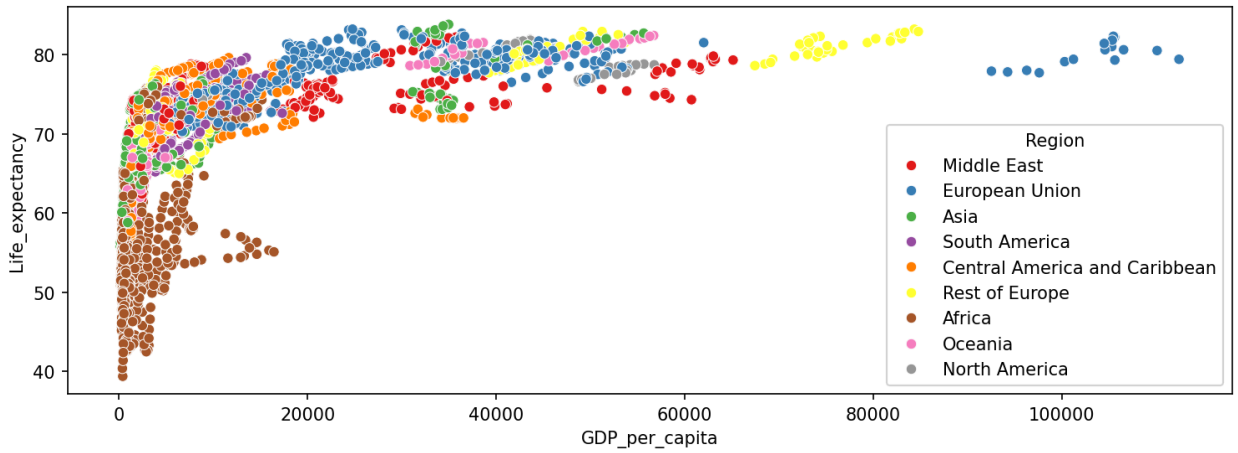


Figure 1: Correlation between GDP per capita and average life expectancy by region

But there is variation within regions as well. For instance, in the Middle East and Asia, some countries have relatively high GDP per capita but varying life expectancy levels. There are some

outliers where countries have either a very high GDP per capita but relatively lower life expectancy or vice versa.

This plot highlights the importance of considering regional contexts when analyzing the relationship between economic indicators and health outcomes. It also suggests that while GDP per capita is an important factor, other regional factors and policies likely play significant roles in determining life expectancy.

Figure 2 and **Table 4** show which variables have the highest correlation with life expectancy.

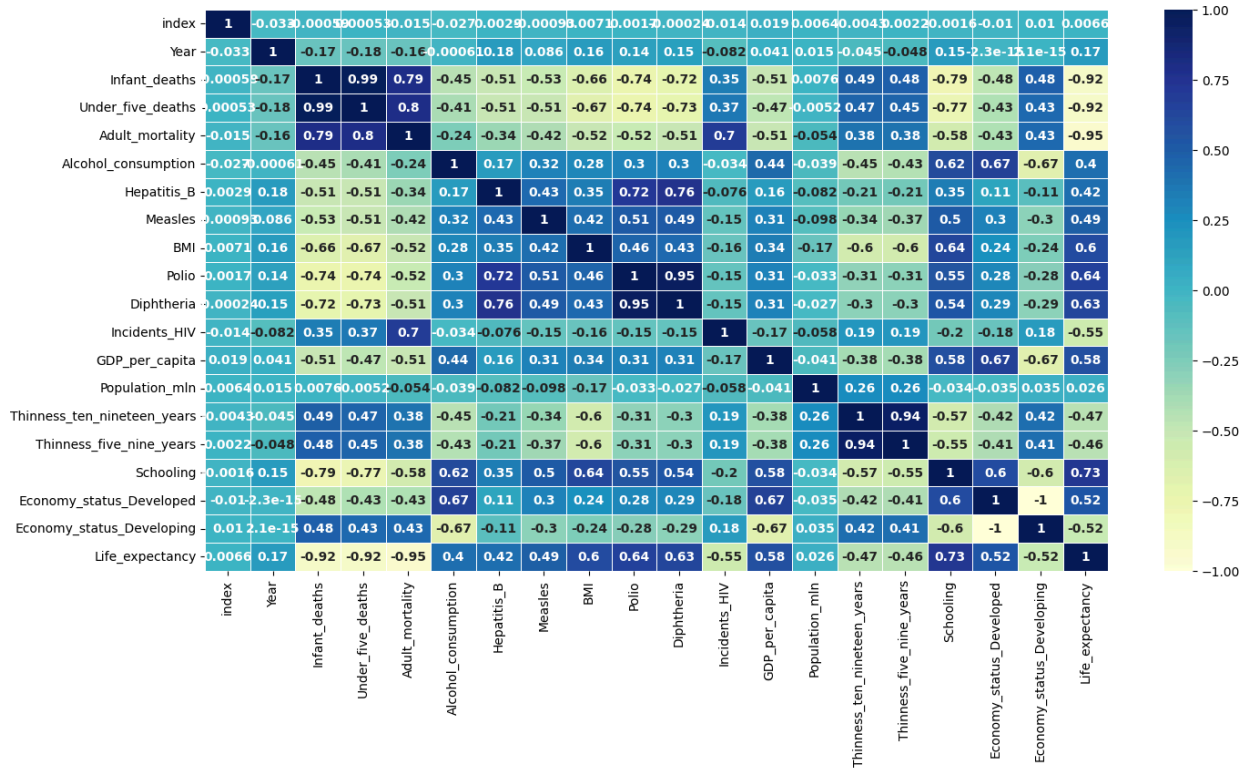


Figure 2: Correlation between life expectancy and various factors

The correlations between life expectancy and various factors, summarized in Figure 2 and Table 4, reveal several key associations. Education level (0.732) shows the strongest positive correlation, suggesting that higher schooling is linked to longer life spans. Immunization coverage for polio (0.641) and diphtheria (0.628) also strongly correlates with life expectancy, highlighting the importance of vaccinations in enhancing health outcomes. Body Mass Index (0.598) and GDP per capita (0.583) exhibit moderate positive correlations, indicating that healthier BMI ranges and greater economic wealth contribute to longer life expectancy. Developed economy status (0.524) reflects better healthcare and living standards, while measles (0.490) and hepatitis B (0.418) immunization rates reinforce the significance of vaccinations. Moderate alcohol consumption (0.399) is positively correlated, though its impact is influenced by cultural and lifestyle factors.

Table 4. Correlation of various variables with life expectancy.

Variable	Correlation with Life Expectancy
Life_expectancy	1.000000

Schooling	0.732484
Polio	0.641217
Diphtheria	0.627541
BMI	0.598423
GDP_per_capita	0.583090
Economy_status_Developed	0.523791
Measles	0.490019
Hepatitis_B	0.417804
Alcohol_consumption	0.399159
Year	0.174359
Population_mln	0.026298
Thinness_five_nine_years	-0.458166
Thinness_ten_nineteen_years	-0.467824
Economy_status_Developing	-0.523791
Incidents_HIV	-0.553027
Infant_deaths	-0.920032
Under_five_deaths	-0.920419
Adult_mortality	-0.945360

3.2. Checking for normal distribution

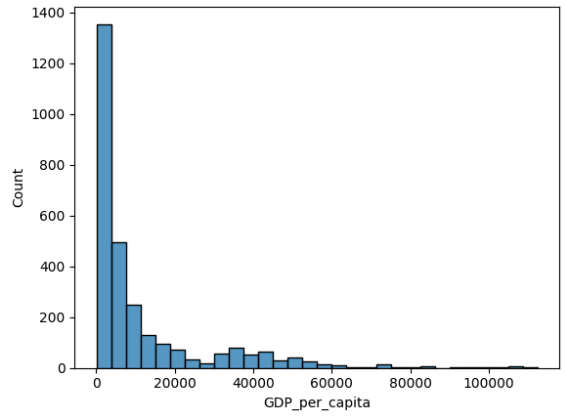


Figure 3: Distribution of GDP_per_capita

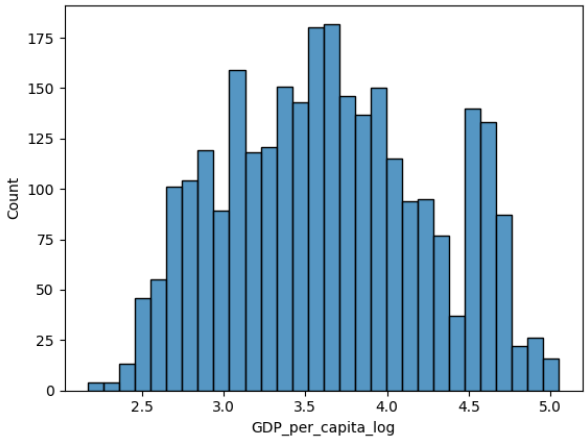


Figure 4: Distribution of GDP_per_capita after log-transformation

The histogram above shows that the observations for the GDP per capita are not normally distributed but highly skewed to the right. However, the log transformed (log 10) variable (GDP per capita) gave a histogram showing slightly normally distributed observations. Consider the graph below.

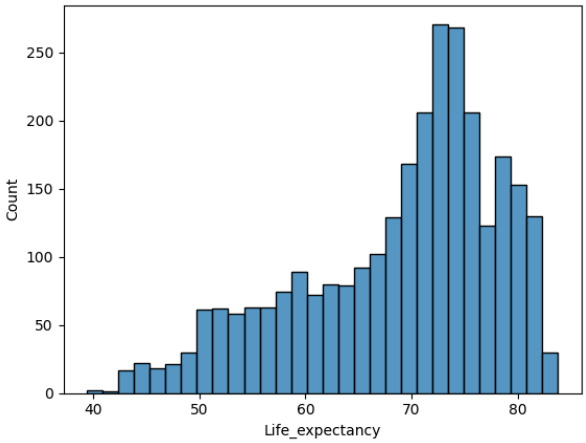


Figure 5: Distribution of Life_expectancy

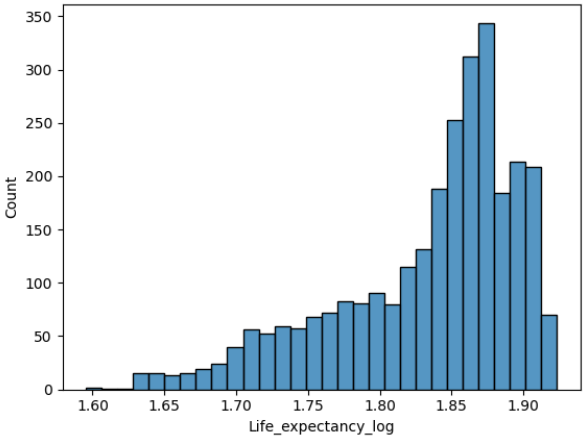


Figure 6: Distribution of Life_expectancy after log-transformation

The histogram above shows a slightly left skewed distribution. In this case, any attempt to log transform the variable (life expectancy) does not make observation to be normally distributed as it can be seen in the graph below. Figure 6 is an evidence showing that any further log transformation of the variable “Life_expectancy” does not bring about normality of the data.

3.3. T-Test

Null-hypothesis: There is no statistically significant effect of GDP per capita on life expectancy at a 5% level of significance.

Alternative hypothesis: There is a statistically significant effect of GDP per capita on life expectancy at a 5% level of significance.

```
t-test statistic = 36.265117421014146
```

```
p-value = 2.6545384236049906e-237
```

A two-tailored t-test for the means of two independent samples (no equal variances assumed) showed that the difference between GDP per capita per year with respect to life expectancy was statistically significant (t-statistic = 36.25 and p-value=0.00) with a 95% confidence interval. Thus, the null-hypothesis can be rejected.

The p-value in the two-sample t-test approach is 0,00% is equal to the paired differences test. By testing the paired differences between each individual in the population instead of the difference between the two populations, we removed the inter-subject variability. Both results reject the null hypothesis.

Before conducting a multi variante linear regression analysis Figure 6 shows whether GDP per capita as independent variable has a linear association to the dependent variable life expectancy.

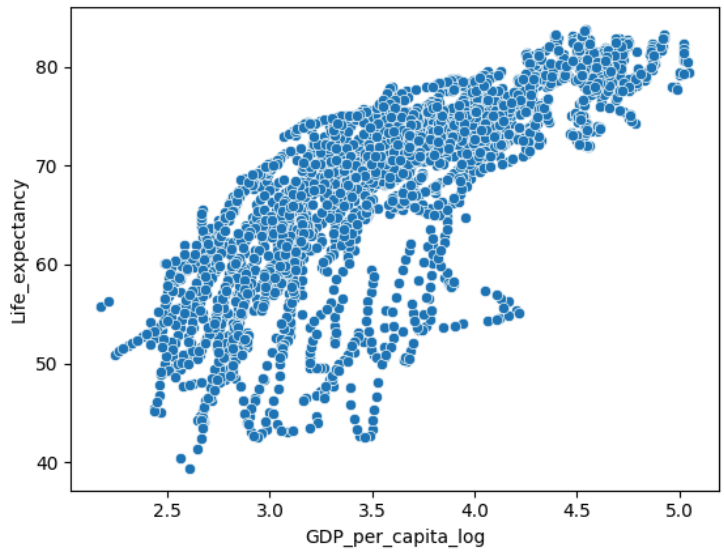
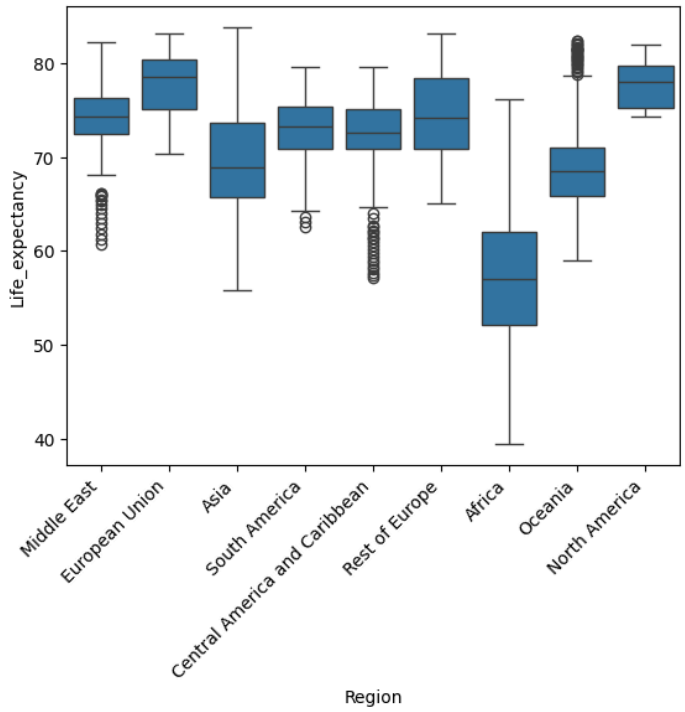


Figure 6: Linear relationship between log transformed GDP per capita and life expectancy

Figure 6 above shows a positive linear association between the log of GDP per capita and life expectancy.

3.4. ANOVA Analysis

3.4.1. Anova test for regions



The boxplot above shows that the life expectancy is comparable in Middle East, South and Central America, and Rest of Europe, and somewhat higher in the European Union as well as North

America, and lower in Africa. To conclude if this difference is statistically significant, the difference between means are evaluated and compared with the data variability.

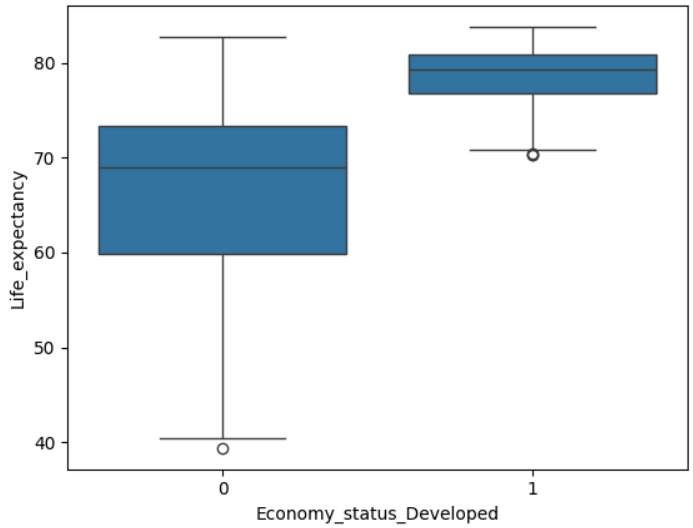
Null hypothesis: There is no difference in the life expectancy between the means of the individual countries

Alternative hypothesis: At least two country means differ from each other in regards to life expectancy

F_onewayResult(statistic=585.3498593161612, pvalue=0.0)

We have a significant test result and Null-hypothesis can be rejected since p-value is 0,0.

3.4.2. Anova Test for economical status of countries



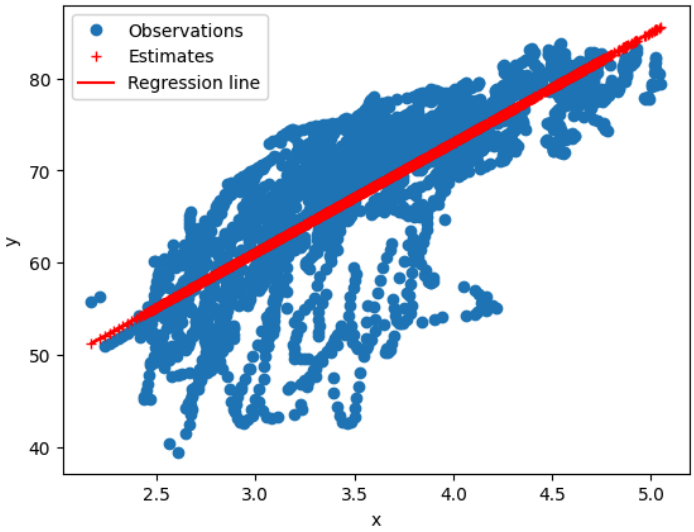
Null hypothesis: There is no difference in the life expectancy between the means of developed and undeveloped countries.

Alternative hypothesis: The means of countries with developed economies and developing economies differ from each other in regards to life expectancy.

F_onewayResult(statistic=1082.0881509799215, pvalue=1.4081397446917954e-201)

There is a significant test result and the null-hypothesis can be rejected since p-value is 0.00.

3.5. Linear Regression Model



Summary of the model shows, that approximately 63.3% of the variance in life expectancy will be explained by the model. The F-statistic is very high (4934), and the p-value is effectively zero, indicating that the model is statistically significant since we can reject null-hypothesis.

The positive coefficient for GDP_per_capita_log (11.930) suggests that for every unit increase in the log-transformed GDP per capita, life expectancy increases by approximately 11.93 years.

However, observations are slightly arching the regression line and model is skewed, which makes it necessary to normalize the data in order to get a better prediction model. Both sqrt-transformation as well as boxcox/log-transformation don't seem to improve the model.

Since the linear regression model will only explain the variance in life expectancy by 63.3% and skewness is very high (-1.178), more variables will be included in the model to make the model more precise.

3.6. Multiple Linear Regression Model

For the multiple regression analysis, a backward selection will be used in order to improve the linear regression model.

Thereby the model can be improved to 89,6%. Except for 5 variable, the p-value is effectively 0.00%, and the null-hypothesis can be rejected. By removing the variable Measles with a p-value of 0.9, Thinness_ten_nineteen_years with 0.116, Thinness_five_nine_years with 0.031 and Diphtheria with 0.010 the model can't be improved, but the dependence of the variables will be reduced. Since variable Economy_status_Developing and Economy_status_Developed are used interchangeably, one of them can be removed as well without impacting the model negatively. By removing the variables Population_mln, Hepatitis_B, BMI and Schooling the models adj. R-squared will decrease to 89,5%. 4 more reduction steps (see appendix) will be conducted to further reduce multicollinearity and dependence of variables which lead to a model

accuracy of 85,5%. Polio as variable will not be removed since this would the adj. R-squared another 6,5%.

Hence, the model with the following variables/predictors is the best model (85,5%) to describe the variance of life expectancy:

- GDP_per_capita_log
- Polio
- Incidents_HIV

4. Conclusion

The analysis showed that GDP per capita has a statistically significant influence on life expectancy, however, other variables like Polio or other diseases and schooling also influence life expectancy.

Additionally, other aspects related to GDP per capita which were not included in the dataset, like wealth, government expenditures, and health insurance availability, can be considered explanatory variables, provided they are not perfectly correlated with GDP per capita.

Furthermore, Life expectancy according to this data is measured at some point in the future, usually about 70+ years, whereas GDP is measured yearly. To identify if GDP impacts life expectancy, a comparison of the average GDP with an interval of about 70 years would be necessary. Reason for this assumption are the following observations in the dataset: There are 2 significant economic events in this dataset.

- the dot-com bubble of the early 2000's
- the financial crisis of 2007/2008

However, life expectancy remained constant during this periods of time. That led to the conclusion, that GDP per capita might whether not truly be a dependent variable of life expectancy (which was rejected by the linear regression analysis) or the life expectancy data will be impacted by this far in the future, which can’t be evaluated based on the dataset. Further analysis with more datapoints would be necessary in the future.

Appendix

Multiple Linear Regression Model: Multicollinearity and dependence of variables

Step 1: Further reducing Multicollinearity and dependence of variables

remaining variables for multi variable model		Adj. R-squared
GDP_per_capita_log + Region + Schooling + Polio + BMI + Hepatitis_B + Alcohol_consumption + Year + Population_mln + Economy_status_Developed + Incidents_HIV		0.896
	Without Incidents_HIV	0,836
	Without Economy_status_Developed	0,893
	Without Population_mln	0,895
	Without Year	0,891
	Without Alcohol_consumption	0,890
	Without Hepatitis_B	0,895
	Without BMI	0,895
	Without Polio	0,874
	Without Schooling	0,895
	Without Region	0,871
	Without GDP_per_capita_log	0,850

Step 2: Further reducing Multicollinearity and dependence of variables

remaining variables for multi variable model		Adj. R-squared

GDP_per_capita_log + Region + Polio + Alcohol_consumption + Year + Economy_status_Developed + Incidents_HIV		
	Without Incidents_HIV	0,836
	Without Economy_status_Developed	0,891
	Without Year	0,888
	Without Alcohol_consumption	0,889
	Without Polio	0,854
	Without Region	0,859
	Without GDP_per_capita_log	0,823

Step 3: Further reducing Multicollinearity and dependence of variables

remaining variables for multi variable model		Adj. R-squared
GDP_per_capita_log + Region + Polio + Alcohol_consumption + Year + Incidents_HIV		
	Without Incidents_HIV	0,831
	Without Year	0,885
	Without Alcohol_consumption	0,888
	Without Polio	0,853
	Without Region	0,858
	Without GDP_per_capita_log	0,791

Step 4: Further reducing Multicollinearity and dependence of variables

remaining variables for multi variable model		Adj. R-squared

GDP_per_capita_log + Region + Polio + Year + Incidents_HIV		
	Without Incidents_HIV	0,820
	Without Year	0,882
	Without Polio	0,849
	Without Region	0,858
	Without GDP_per_capita_log	0,788

Step 5: Further reducing Multicollinearity and dependence of variables

remaining variables for multi variable model		Adj. R-squared
GDP_per_capita_log + Region + Polio + Incidents_HIV		
	Without Incidents_HIV	0,808
	Without Polio	0,838
	Without Region	0,855
	Without GDP_per_capita_log	0,778

Step 6: Further reducing Multicollinearity and dependence of variables

remaining variables for multi variable model		Adj. R-squared
GDP_per_capita_log + Polio + Incidents_HIV		
	Without Incidents_HIV	0,708
	Without Polio	0,790
	Without GDP_per_capita_log	0,626

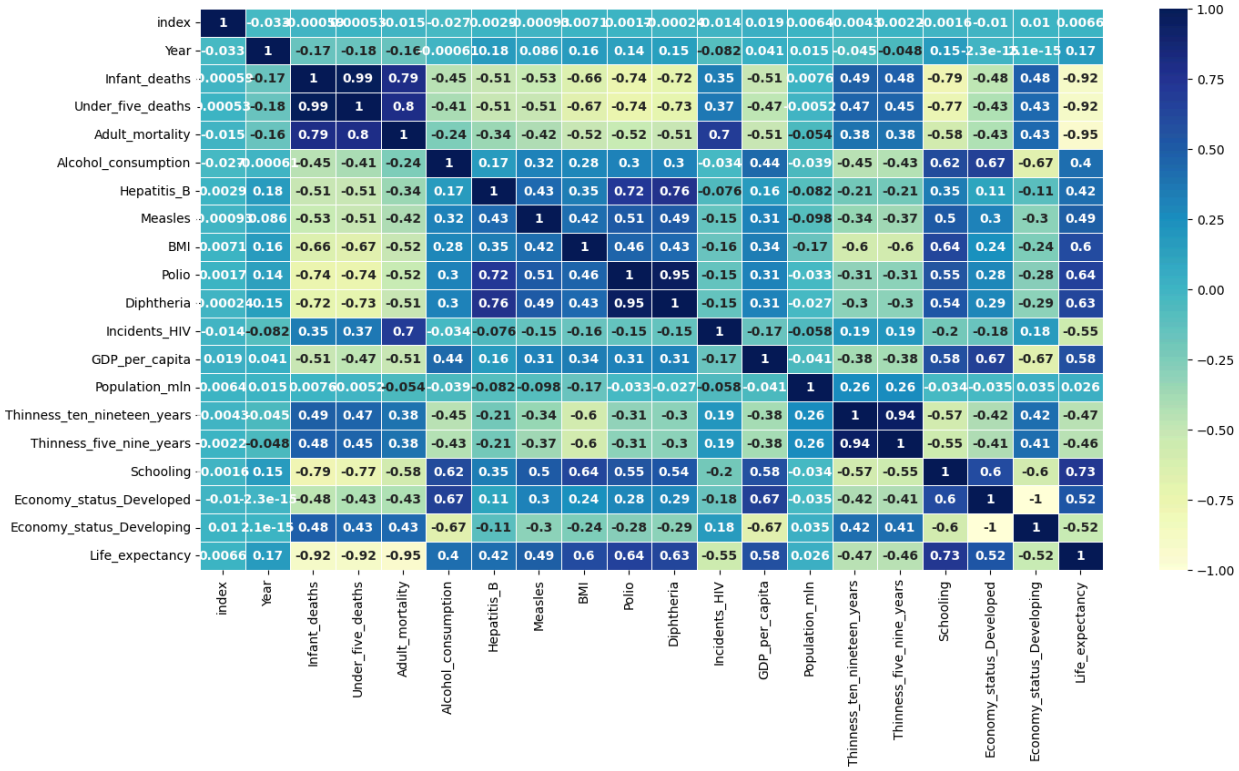
→ Last step will not be conducted since the loss of accuracy of the model is too high with 6,5%.

Appendix - A Contrarian View

Contact: Frank Miceli

There was some discussion that was not fully resolved. Below are a few thoughts that propose an alternative view showing that GDP per capita is not the best indicator of Life expectancy.

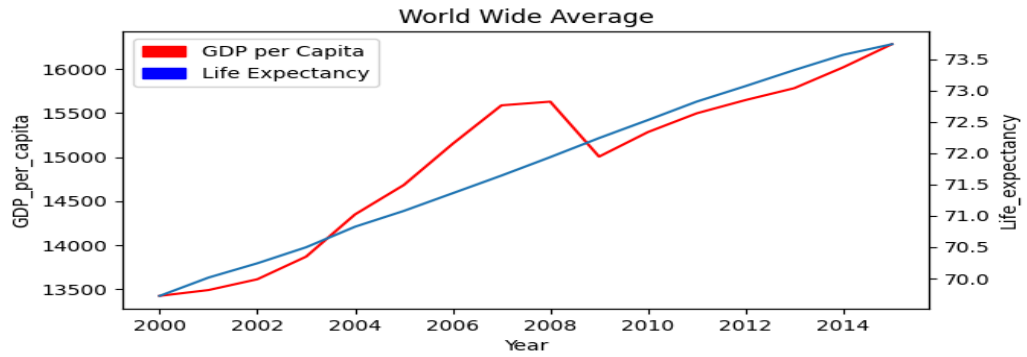
Without overloading the reader, first please review the correlation matrix of the data set.



While I am not discounting the original findings discussed above, what is not mentioned is that the highest correlation of life expectancy is a negative relationship of a combination of Infant_deaths, (-0.92), Under_five_deaths (-0.92) and Adult_mortality (-0.95).

The correlation of Life_expectancy and GDP_per_capita is much less at + 0.58.

Exploring the original conclusion, a world wide visual representation is provided as well as each region's relationship of Life_expectancy and GDP_per_capita per region from 2000 to 2015.



Definitions are key for proceeding further.

	In Kaggle DataSet	In Referenced DataSet	Estimate or Observed Value
Life_expectancy	Average life expectancy of both genders in different years from 2010 to 2015	Life Expectancy at birth (years)	Estimated
GDP_per_capita	GDP per capita in USD		Observed
Infant_deaths	deaths per 1000 population		Observed
Under_five_deaths	deaths per 1000 population		Observed
Adult_mortality	deaths per 1000 population		Observed

Quote from WorldBank.org

"The statistic "Life expectancy at birth" actually refers to the average number of years a newborn is expected to live *if mortality patterns at the time of its birth remain constant in the future"*

"**Life expectancy at birth** is the total person-years lived beyond exact age 0 divided by the number of newborns"

As I understand the process of calculating life expectancy, actuarial tables using a probability of dying and the current mortality rates result in the life expectancy estimate.

By comparison, does an observed increase or decrease of GDP_per_capita impact Life expectancy estimates?

GDP_per_capita is a country's GDP divided by its population. GDP = C + I + G + (X-M). That is, GDP = Consumer Spending + Business Investment + Government Spending + (Exports - Imports).

Governments will allocate funds to its population for health care, based on its policies. But allocation of GDP to Health Care is unknown in this dataset. It would be best to include these proportions in future analysis.

Below are a few excerpts from an OLS of various relations:

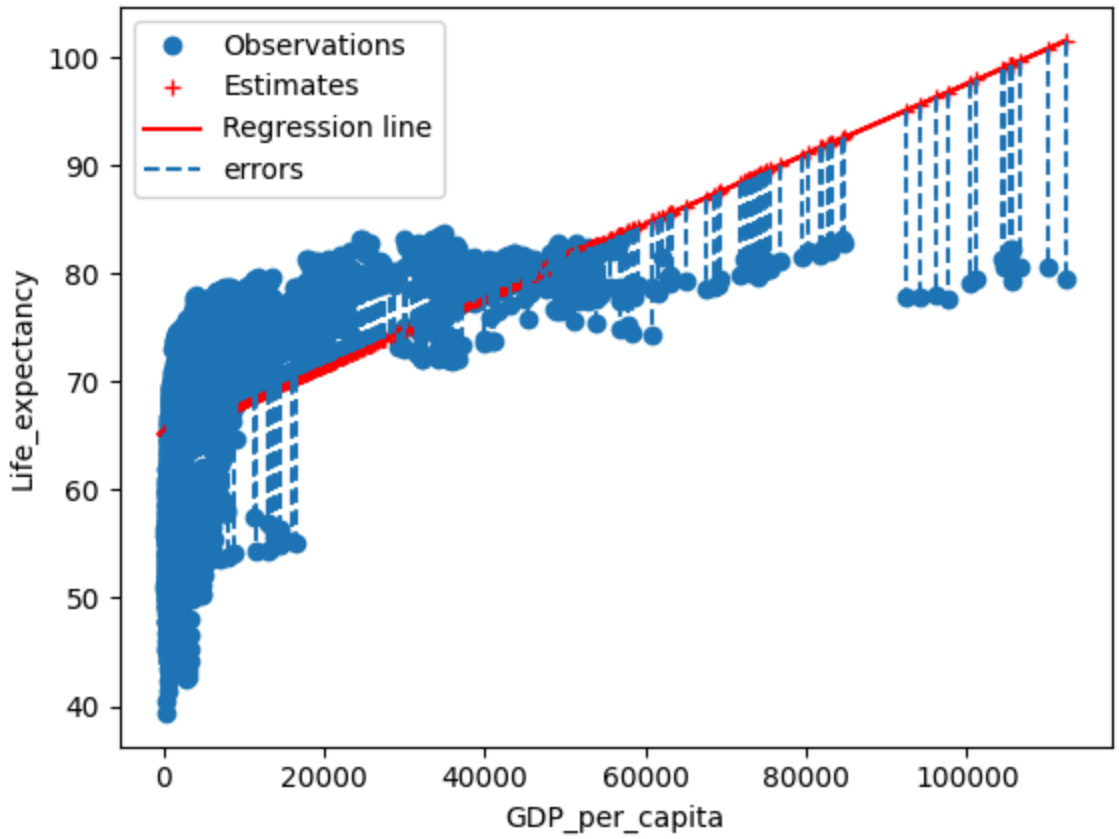
First, OLS of Life_expectancy vs GDP_per_capita

OLS Regression Results

Dep. Variable:	Life_expectancy	R-squared:	0.340
Model:	OLS	Adj. R-squared:	0.340
Method:	Least Squares	F-statistic:	1474.
Date:	Sun, 11 Aug 2024	Prob (F-statistic):	1.52e-260
Time:	15:18:59	Log-Likelihood:	-9887.4
No. Observations:	2864	AIC:	1.978e+04
Df Residuals:	2862	BIC:	1.979e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	65.1186	0.173	376.787	0.000	64.780	65.457
GDP_per_capita	0.0003	8.43e-06	38.397	0.000	0.000	0.000

Omnibus:	323.832	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	443.462
Skew:	-0.961	Prob(JB):	5.05e-97
Kurtosis:	3.158	Cond. No.	2.48e+04



Alternative approach, Life Expectancy vs Mortality

where mortality includes Infant_deaths, Under_five_deaths and Adult_mortality

OLS Regression Results

Dep. Variable:	Life_expectancy	R-squared:	0.971
Model:	OLS	Adj. R-squared:	0.971
Method:	Least Squares	F-statistic:	3.220e+04
Date:	Sun, 11 Aug 2024	Prob (F-statistic):	0.00
Time:	20:05:49	Log-Likelihood:	-5400.5
No. Observations:	2864	AIC:	1.081e+04
Df Residuals:	2860	BIC:	1.083e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	82.6150	0.066	1248.176	0.000	82.485	82.745
Infant_deaths	-0.1273	0.006	-19.829	0.000	-0.140	-0.115
Under_five_deaths	-0.0190	0.004	-4.721	0.000	-0.027	-0.011
Adult_mortality	-0.0472	0.000	-108.468	0.000	-0.048	-0.046

```
Omnibus:                        1.934   Durbin-Watson:                   2.026
Prob(Omnibus):                  0.380   Jarque-Bera (JB):                1.862
Skew:                           0.048   Prob(JB):                        0.394
Kurtosis:                       3.080   Cond. No.                         520.
==============================================================================
```

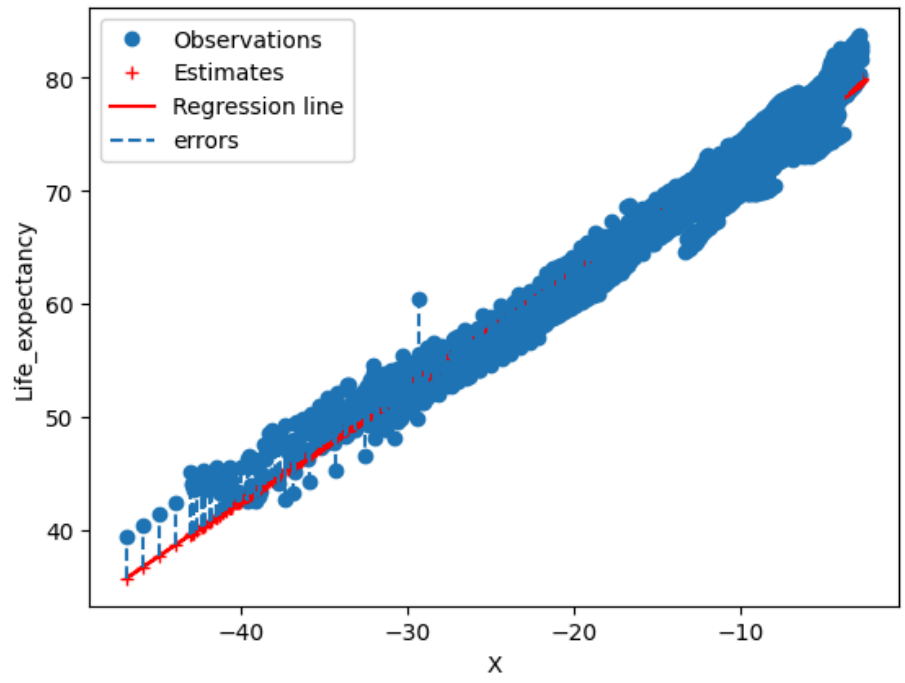
Because this regression is leveraging 3 different independent variables, a summary value called X was created to represent the parameters of the regression residuals in the plot below.

In this case,

```
dfa['X'] = dfa['Infant_deaths'] * reg1.params['Infant_deaths'] +
        dfa['Under_five_deaths']*reg1.params['Under_five_deaths'] +
        dfa['Adult_mortality']*reg1.params['Adult_mortality']
```

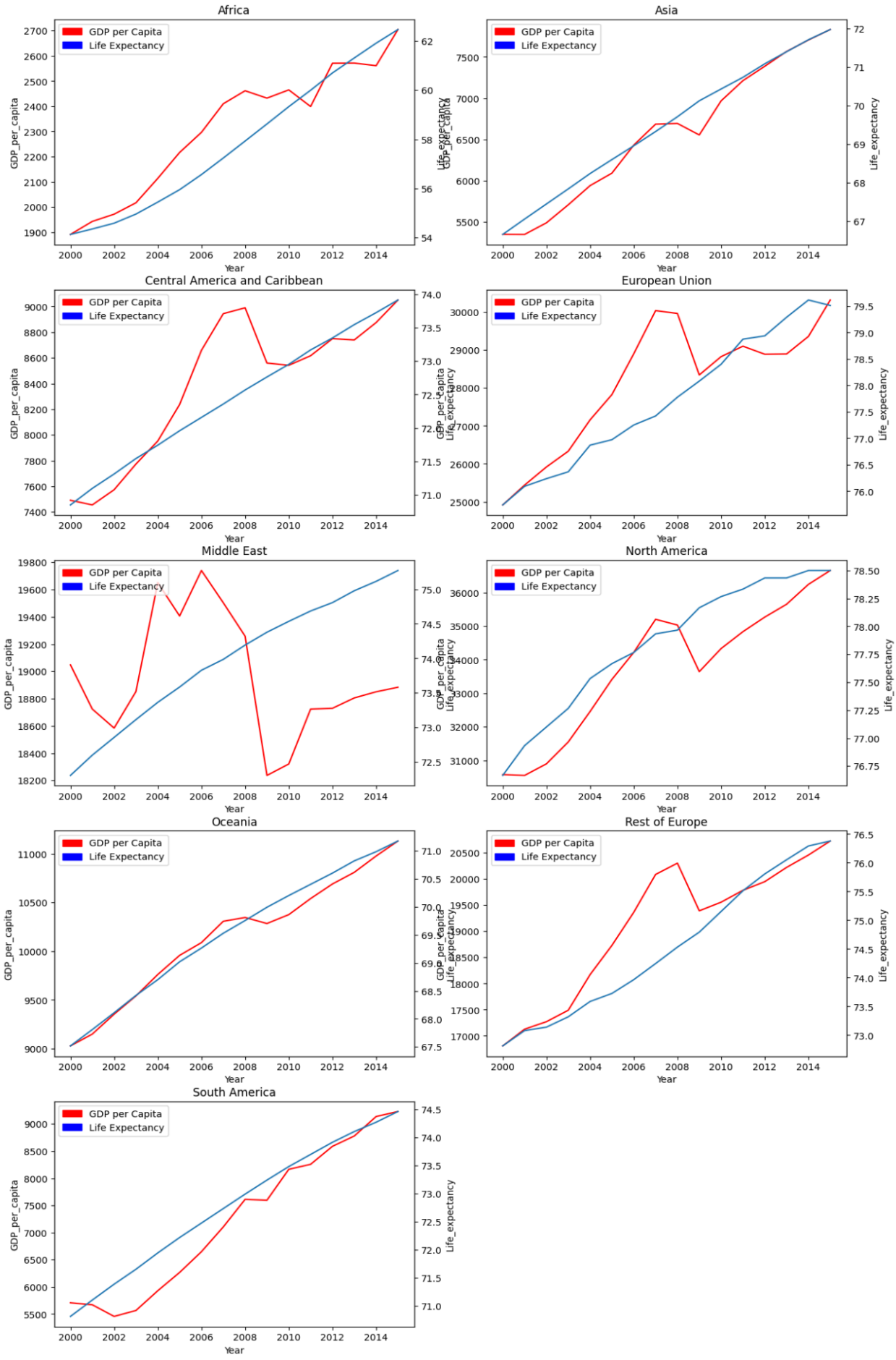
where reg1.params =

```
Intercept            82.615036
Infant_deaths        -0.127321
Under_five_deaths    -0.019050
Adult_mortality      -0.047204
dtype: float64
```



Concluding, this alternative is to say that while there is a moderate corelation between GDP and Life Expectancy, there are other factors related to how GDP directly impacts Life Expectancy Estimation such as government spending, and policies of each nation on how and where to fund health that remain unknown. The data promotes a better relationship between Life Expectancy and the mortality metrics of Infant_deaths, Under_five_deaths and Adult_mortality.

PS - I mean no disrespect to my team. There was some discussion in our forums but no consensus was reached. I felt this appendix provided a means to present an alternative conclusion.


Africa
Asia
Central America and Caribbean
European Union
Middle East
North America
Oceania
Rest of Europe
South America
GDP per Capita
Life Expectancy
GDP_per_capita
Life_expectancy
Year

Some regions appear to have a strong linear relationship between GDP and Life Expectancy.