DATA1002-CC-L14-G1 Stage 2 Report

SID and Unikeys

SID: 520 423 035 Unikey:slin7370

SID: 520 102 934 Unikey:pwoo5799

SID: 520 246 478 Unikey:obow2075

SID: 520 463 178 Unikey:sspi3010

Section A:

Data Description:

This dataset is a collection of datasets gathered from a variety of sources merged together via Country Code. Individual Datasets include Education - obtained from Barrolee (BarroLeeDataSet, 2021) from a varying sources such as administrative data available from educational management systems, household surveys, population census etc. GDP/capita country data was obtained from the World Bank database (World Bank, 2022), sourced from World Bank national accounts data and OECD National Accounts Data. Homicide Rate data by country obtained from Our World in Data (), publishing data from Global Burden of Disease Study 2019 (GBD 2019) Results, and
Seattle, United States: Institute for Health Metrics and Evaluation (IHME, 2021). Lastly the Life expectancy data was obtained from The World Bank database (Life Expectancy at Birth, Total (Years), The World Bank), a compilation sourcing from, United Nations Population Division , Census reports from national statistical offices, Eurostat: Demographic
Statistics United Nations Statistical Division, US Census Bureau, and Secretariat of the Pacific Community.

Together these datasets were merged using an “inner” join resulting in a table with 234 entries (2 for each country, years 2005 & 2010), and 24 columns:

Country Name (String)	Country Code (String)	Life Expectancy (Years - Int)	GDP/capita (USD - Float)
Deaths by Interpersonal Violence (Rate)	2005-2010 Interpersonal Violence (Rate Ratio)	BLCode (Barrolee Country Code - unique identifier - Str)	Sex (String)
Age from (Education - Int)	Age to (Education - Int)	Percent no Schooling (Float)	Percent Primary Schooling (Float)
Percent Primary School Completion (Float)	Percent Secondary School (Float)	Percent Secondary School Completion (Float)	Percent Tertiary School (Float)
Percent Tertiary School Completion (Float)	Average Years of School Attained (Float)	Average Years Primary School Attained (Float)	Average Years Secondary School Attained (Float)
Average Years Tertiary School Attained (Float)	Population (Int)	Region Code (String)	Year (2005 or 2010 - String)

Education Rate Relation to GDP per Capita (520423035)

```
#Numerical Summaries
outfile = open("Results by Year.csv", "w")
by_year = df.groupby("Year")
print("Year,Max Years in Schooling,Min Years in Schooling,Mean Years in Schooling,Average GDP/capita (USD)", file = outfile)
for year in by_year:
    result = [year[0], round(year[1]["Average Years of Schooling Attained"].max(),2)\
        , round(year[1]["Average Years of Schooling Attained"].min(),2)\
        , round(year[1]["Average Years of Schooling Attained"].mean(), 2), round(year[1]["GDP/capita (USD)"].mean(), 2)]

    print(*result, file = outfile, sep = ",")
outfile.close()
```

Results by Year

Year	Max Years in Schooling	Min Years in Schooling	Mean Years in Schooling	Average GDP/capita (USD)
2005	11.82	1.49	7.82	12218.86
2010	11.97	2.9	8.13	15504.09

To create the group aggregate summaries, we used df.groupby() to groupby the selected attribute (in the first case it was “Year”). Then a for loop was used to access each group (different years) and inbuilt pandas functions were used to calculate numerical summaries, with the values stored in a list to allow for easy output to the outfile. The following table was the result of this process.

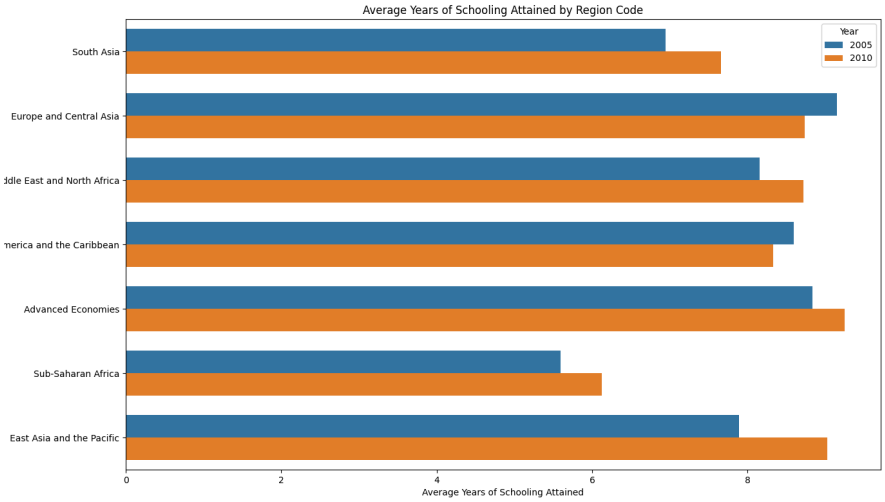
This process was repeated but this time grouping by region code and also by year. The table shows the information with respect to each different region. Here we grouped by region and year and applied the same process as explained above.

Results by Year and Region

Region	Year	Max Years in Schooling	Min Years in Schooling	Mean Years in Schooling	Average GDP/capita (USD)
Advanced Economies	2005	11.46	5.99	8.83	41266.56
Advanced Economies	2010	10.95	5.93	9.25	49233.32
East Asia and the Pacific	2005	11.57	4.8	7.89	4280.32
East Asia and the Pacific	2010	11.97	4.41	9.03	7045.81
Europe and Central Asia	2005	10.32	7.2	9.15	6619.42
Europe and Central Asia	2010	10.2	7.88	8.73	9770.42
Latin America and the Caribbean	2005	10.25	4.65	8.59	4321.37
Latin America and the Caribbean	2010	11.04	4.94	8.33	6799.53
Middle East and North Africa	2005	10.06	5.5	8.15	17843.01
Middle East and North Africa	2010	10.27	5.67	8.71	21809.28
South Asia	2005	11.82	3.39	6.94	1058.54
South Asia	2010	10.82	4.81	7.66	2019.75
Sub-Saharan Africa	2005	10.39	1.49	5.59	1497.94
Sub-Saharan Africa	2010	11.85	2.9	6.13	2133.17

```
df = pd.read_csv("integrated.csv")
outfile3 = open("Results by Year and Region.csv", "w")
by_region_and_year = df.groupby([df["region_code"],df["Year"]])
print("Region,Year,Max Years in Schooling,Min Years in Schooling,Mean Years in Schooling,Average GDP/capita (USD)", file = outfile3)
for region_year in by_region_and_year:
    result = [region_year[0][0],region_year[0][1], round(region_year[1]["Average Years of Schooling Attained"].max(),2)\
        , round(region_year[1]["Average Years of Schooling Attained"].min(),2)\
        , round(region_year[1]["Average Years of Schooling Attained"].mean(), 2), round(region_year[1]["GDP/capita (USD)"].mean(), 2)]
    print(*result, file = outfile3, sep = ",")
```

The relationship between Average Years of Schooling Attained was also displayed graphically to show the difference between 2005 and 2010 for each region.




```
import seaborn as sns


df = pd.read_csv("integrated.csv")
barplot = sns.barplot(data = df, y = "region_code", x = "Average Years of Schooling Attained", hue = "Year", orient = "h", width = 0.7, errorbar = None)
plt.title("Average Years of Schooling Attained by Region Code")
plt.show()
```

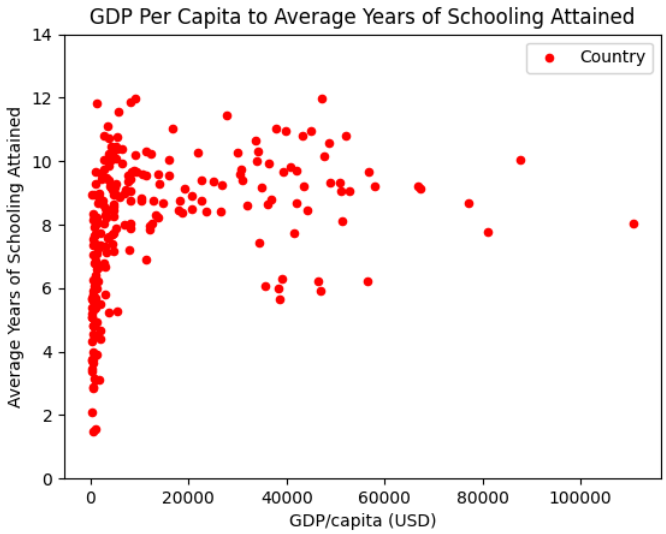
To make this bar chart, the seaborn module was used. A barplot was plotted with a horizontal orientation, meaning the qualitative data (Region) was on the y axis and the quantitative data (Average Years of Schooling Attained) on the x. The hue parameter allowed us to plot 2005 and 2010, showing the differences between the years. The errorbar was also removed. The box plot was used as we are plotting qualitative data to quantitative data and also wished to show the differences between 2005 and 2010.

From these results, it appears there may be a positive correlation between Average Years of Schooling Attained and GDP/capita (USD). To investigate this further, a scatter plot of the different countries were plotted, with GDP/capita (USD) on the x-axis and Average Years of Schooling Attained on the y-axis. A scatter plot was used as we wish to visualise the distribution of data points and we are plotting quantitative data to quantitative data. As we can see, there is rapid growth initially when GDP/capita (USD) increases but the growth eventually plateaus and this suggests that there is a

logarithmic relationship between GDP Per capita and Average Years of Schooling Attained.

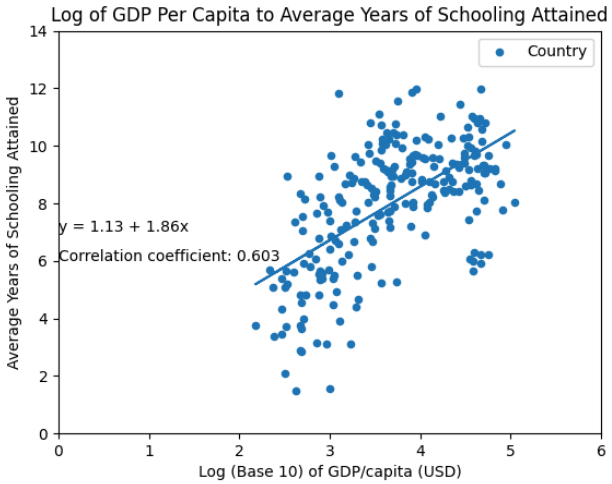
This chart was created by plotting a graph with kind = “scatter”, x axis being GDP/capita (USD) and y axis being “Average Years of Schooling Attained”. A title, axis labels were added and so was the y axis labelling to reduce visual error. The y-axis was limited from 0 to 14 as the highest average years in schooling was less than 14. The x-axis was also limited from -5000 to 120000 to clearly show the clustered set of points near 0 and also because the highest recorded GDP/capita was below 120000.

```
df = pd.read_csv("integrated.csv")

df.plot(kind='scatter',x='GDP/capita (USD)',y='Average Years of Schooling Attained',color='red')
plt.title("GDP Per Capita to Average Years of Schooling Attained")
plt.xlabel("GDP/capita (USD)")
plt.ylabel("Average Years of Schooling Attained")
plt.ylim(0,14)
plt.xlim(-5000,120000)
plt.legend(["Country"])
plt.savefig("GDP TO EDUCATION")
plt.show()
```



To investigate this further, Average Years of Schooling Attained was plotted to the logarithm of GDP per Capita. As we can see, a more strong positively correlated linear line (Correlation Coefficient of +0.603) can be observed. This provides evidence towards the logarithmic relationship between GDP per Capita and Average Years of Schooling Attained. Again, a scatter plot was used to visualise the relationship as we are looking for an overall trend. I copied the integrated data set dataframe to a new variable called data and added a new column which took the logarithm with base ten of each of the GDP/capita values. I then plotted a scatter graph with the logarithm of the GDP/capita on the x axis and Average Years of Schooling Attained on the y axis. The x-axis was limited from 0 to 6 as the highest GDP/capita found was below 1,000,000. The y-axis was limited from 0 to 14 as the highest average years in schooling was less than 14. The gradient and y-intercept of the best fit line were then calculated and plotted

on the graph. The correlation coefficient was calculated and written on the graph.




```
data = df
data['Log (Base 10) of GDP/capita (USD)'] = np.log10(data['GDP/capita (USD)'])  #Converting to logarithm
data.plot(kind='scatter',x='Log (Base 10) of GDP/capita (USD)',y='Average Years of Schooling Attained')
plt.title("Log of GDP Per Capita to Average Years of Schooling Attained")
plt.legend(["Country"])
plt.xlim(0, 6)
plt.ylim(0, 14)
#add line of best fit to plot
x = data['Log (Base 10) of GDP/capita (USD)']
y = data['Average Years of Schooling Attained']
gradient, intersect = np.polyfit(x, y, 1)   #calculating gradient and intersect
cc = round(x.corr(y), 3)    #Correlation coefficient
plt.plot(x, gradient * x + intersect)
plt.text(0, 7, f'y = {round(intersect, 2)} + {round(gradient, 2)}x', size=10) #Showing result on graph
plt.text(0, 6, f"Correlation coefficient: {cc}", size = 10)
plt.savefig("LOG OF GDP TO EDUCATION")
```

Effectiveness of Graphs:

The scatter plot is effective to show the relationship between GDP/capita and Average Years of Schooling attained as it is showing the relationship between 2 quantitative values. All the data values we have are plotted and it builds a trend that can be easily seen. However the plotted points are not weighted to population size and each point has the same weighting. Even if more data was added, it is likely that the scatter plot will still be valid as it will simply be another point on the graph. However, it may be possible that the relationship between the 2 quantitative values may be less obvious.

The bar graph is effective as we are grouping data into discrete groups (By region) and displaying them. There is also colour coding to distinguish between 2005 and 2010. If more data points were added, the bar graph should be able to adapt. Since we grouped the values into its respective region, and are taking the mean, this can be done effectively. However the graph may get more messy and unorganised if more regions were added and also if there are multiple more countries. In this case a line graph may be more applicable.

Crime Rate's Relationship to GDP (SID: 520 102 934)

Creating a summary table by grouping them according to years:

```python
import pandas as pd

df = pd.read_csv("integrated.csv")
dic = {}
grouping = df["Deaths - Interpersonal violence - Sex: Both - Age: All Ages (Rate/100000)"].groupby([df["Year"]])

dic["Mean"] = grouping.mean().round(2)
dic["Min"] = grouping.min().round(2)
dic["Max"] = grouping.max().round(2)
y = pd.DataFrame.from_dict(dic)
y.to_csv("Crime-Rate-per-100,000-People_Year_Summary.csv")

```

Death from Interpersonal Violence by Year

Year	Mean	Min	Max
2005	8.47	0.71	54.88
2010	8.07	0.59	53.13

For this section, we will mainly be looking at the year column, region code column, and crime rate column of the dataset. First, we will be summarising the crime rates of the countries in 2005 and 2010. We used pandas's "groupby" to group the dataset according to the years. Crime rates were measured by the number of deaths from interpersonal violence per 100,000 people. From comparing 2005 and 2010, we can see that all mean, minimum, and maximum crime rates decreased.

By further dividing them according to their regions:

```python

dic2 = {}
group = df["Deaths - Interpersonal violence - Sex: Both - Age: All Ages (Rate/100000)"].groupby([df["region_code"],df["Year"]])
dic2["Mean"] = group.mean().round(2)
dic2["Min"] = group.min().round(2)
dic2["Max"] = group.max().round(2)
x = pd.DataFrame.from_dict(dic2)
x.to_csv("Crime-Rate-per-100,000-People_Region_Code_Year_Summary.csv")

```

Death from Interpersonal Violence by Year and Region

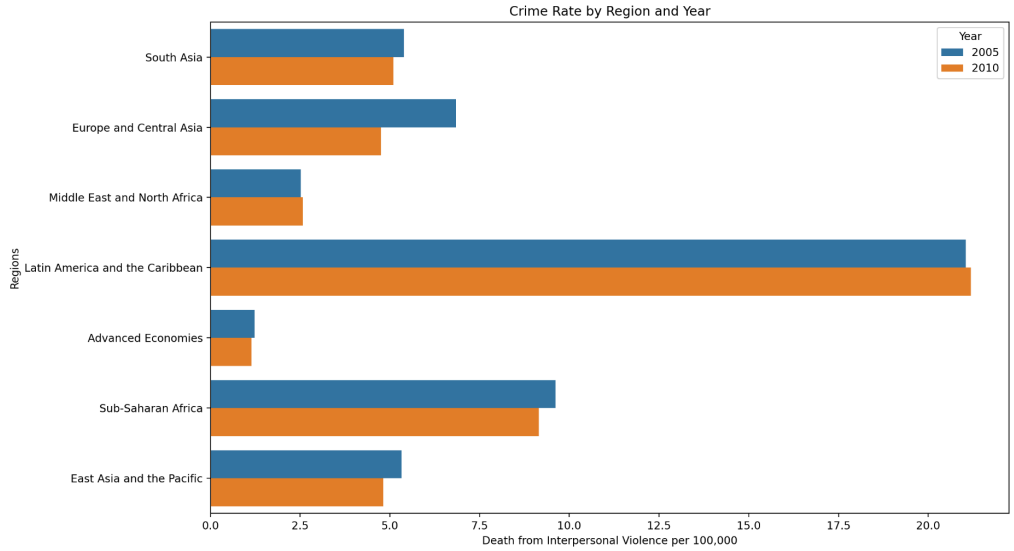
region_code	Year	Mean	Min	Max
Advanced Economies	2005	1.23	0.71	2.42
Advanced Economies	2010	1.14	0.59	2.08
East Asia and the Pacific	2005	5.32	0.79	17.54
East Asia and the Pacific	2010	4.82	0.64	15.02

Europe and Central Asia	2005	6.84	1.47	21.06
Europe and Central Asia	2010	4.76	0.82	13.89
Latin America and the Caribbean	2005	21.04	3.77	54.88
Latin America and the Caribbean	2010	21.19	3.99	53.13
Middle East and North Africa	2005	2.51	0.9	9.25
Middle East and North Africa	2010	2.58	0.87	12.87
South Asia	2005	5.39	1.74	14.41
South Asia	2010	5.09	1.85	14.43
Sub-Saharan Africa	2005	9.62	0.9	53.09
Sub-Saharan Africa	2010	9.15	0.95	42.45

According to the summary table, we can see that Latin America and the Caribbean had the highest in all mean, minimum, and maximum crime rates. The mean and minimum also increased from 2005 to 2010 in Latin America and the Caribbean. The Advanced Economies region had the lowest crime rates in all categories with a decreasing trend. The most notable change was in Europe and Central Asia with a 7.17 decrease in maximum crime rate. Most regions experienced 0.07 to 2 mean decrease.

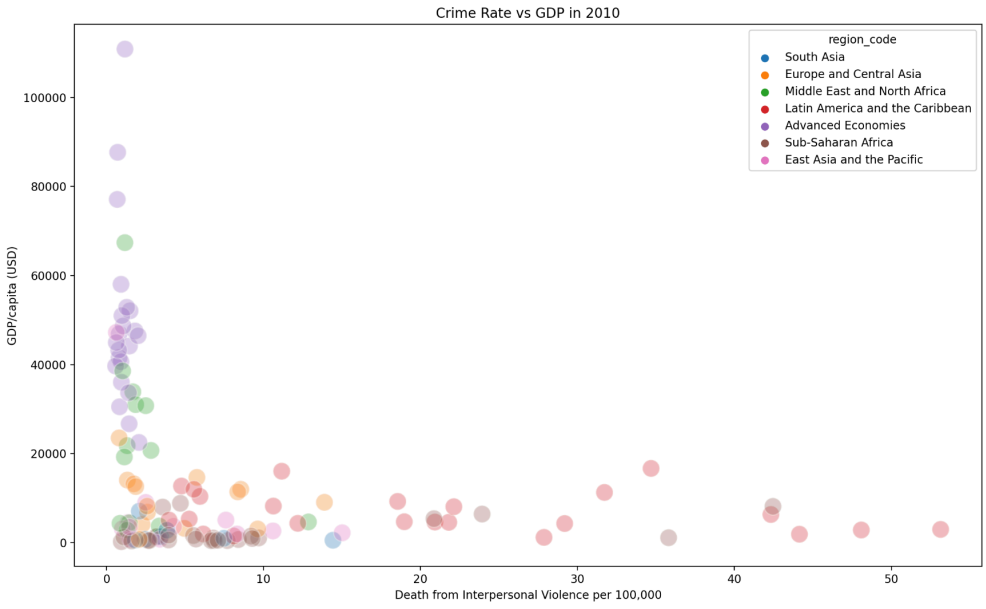
What is the relationship between Crime Rate and GDP of a Country?

```python
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns

df = pd.read_csv("integrated.csv")
df["Deaths - Interpersonal violence - Sex: Both - Age: All Ages (Rate/100000)"].groupby([df["Year"]]).describe().to_csv("Crime_Rate_summaries_year.csv")
df['Deaths - Interpersonal violence - Sex: Both - Age: All Ages (Rate/100000)'].groupby([df["region_code"],df["Year"]]).describe().to_csv("Crime_Rate_summaries_year_region.csv")

g = sns.barplot(y = "region_code", x = 'Deaths - Interpersonal violence - Sex: Both - Age: All Ages (Rate/100000)', data = df, hue = "Year",errorbar = None)
g.set_title("Crime Rate by Region and Year")
plt.ylabel("Regions")
plt.xlabel("Death from Interpersonal Violence per 100,000")
fig = g.get_figure()
# to run the first graph, uncomment line 16 and comment line 18 to 24
#plt.show()
```



Before we compare the relationship between the two, we would like to first make a plot showing the crime rates grouped by region and year. We chose a bar plot to represent our findings because we aimed to find the differences in crime rate according to their regions each year. By using a bar plot, we can clearly see which region had the highest crime rates and the increase or decrease trend between the two years in each region. Although using bar plots meant we wouldn't be able to know each country's stats individually, it was a good way to understand the summary of the data visually. It was also important to note that as the whole dataset was condensed to make a graph, there was some data removed and the bar plot originally had an errorbar. To make the plot look neater, we wrote code to remove the errorbar from the presentation.

```
fig.clf()
new_df = df[['Deaths - Interpersonal violence - Sex: Both - Age: All Ages (Rate/100000)', "region_code", "GDP/capita (USD)", "Year"]].copy()
new_df.query("Year == 2010", inplace = True)
g = sns.scatterplot(data = new_df, x = 'Deaths - Interpersonal violence - Sex: Both - Age: All Ages (Rate/100000)', y = "GDP/capita (USD)", hue = "region_code", s = 230, alpha = 0.3)
g.set_title("Crime Rate vs GDP in 2010")
plt.xlabel("Death from Interpersonal Violence per 100,000")
plt.show()
```



Now, we will be looking at the relationship between GDP per capita and crime rate according to each region. To make it clearer to look at, we chose to only put crime rates from 2010 in the graph. We also chose to create a new dataframe as we were drawing both graphs in the same file. We chose a scatter plot to present the data visually because we wanted to show how the two factors, GDP and crime rate, correlated. We also made the size of the dots in the plot bigger and more transparent as there were a lot of dots overlapping. From looking at the graph, we can see a lot of purple plots on the left side. It meant countries in Advanced Economies all had low

crime rates and less spread in crime rates data compared to other regions. The scatter plot also showed that unlike other regions that had lower GDP, Latin America and the Caribbean and Sub-Saharan Africa had higher crime rates. From looking at the plot, we could make a conclusion of a negative correlation between GDP and crime rate. However, it was important to note that it was not directly proportional as even if a country had very low GDP, it was not certain they would have a very high crime rate.

Life Expectancy's Relationship to GDP (SID: 520 246 478)

Aggregate summaries concerning the life expectancy of counties grouped by year were as follows:

```
output = pd.DataFrame(columns=["Year","Mean", "Min", "Max"])

for group in df["Life Expectancy (Years)"].groupby([df["Year"]]):
    d = {"Year": [group[0]], "Mean": [group[1].mean()],
        "Min": [group[1].min()], "Max": [group[1].max()]}
    output = pd.concat([output, pd.DataFrame(data=d)], ignore_index=True)

output.to_csv("LE_summaries_year.csv", index=False)
```

Life Expectancy Summaries by Year

Year	Mean	Min	Max
2005	69.09	42.66	81.93
2010	71.06	45.10	82.84

From these, there was an overall increase in the life expectancy of countries from years 2005 and 2010.

These can then be broken up by region as follows:

```
output = pd.DataFrame(columns=["Region","Year","Mean", "Min", "Max"])

for group in df["Life Expectancy (Years)"].groupby([df["region_code"],df["Year"]]):
    d = {"Region": [group[0][0]], "Year": [group[0][1]], "Mean": [group[1].mean()],
        "Min": [group[1].min()], "Max": [group[1].max()]}
    output = pd.concat([output, pd.DataFrame(data=d)], ignore_index=True)


output.to_csv("LE_summaries_year_region.csv", index=False)
```

Life Expectancy Summaries by Year/Region

Region	Year	Mean	Min	Max
Advanced Economies	2005	79.78	77.84	81.93
Advanced Economies	2010	80.91	79.03	82.84
East Asia and the Pacific	2005	68.46	60.53	79.99
East Asia and the Pacific	2010	69.99	62.03	81.54
Europe and Central Asia	2005	71.99	65.86	77.61
Europe and Central Asia	2010	73.77	68.45	79.42
Latin America and the Caribbean	2005	72.45	58.65	78.12
Latin America and the Caribbean	2010	73.58	60.51	78.78
Middle East and North Africa	2005	74.90	68.27	80.15
Middle East and North Africa	2010	76.04	68.57	81.60
South Asia	2005	66.87	58.29	74.70
South Asia	2010	68.83	61.03	75.91
Sub-Saharan Africa	2005	53.49	42.66	72.43
Sub-Saharan Africa	2010	57.63	45.10	72.97

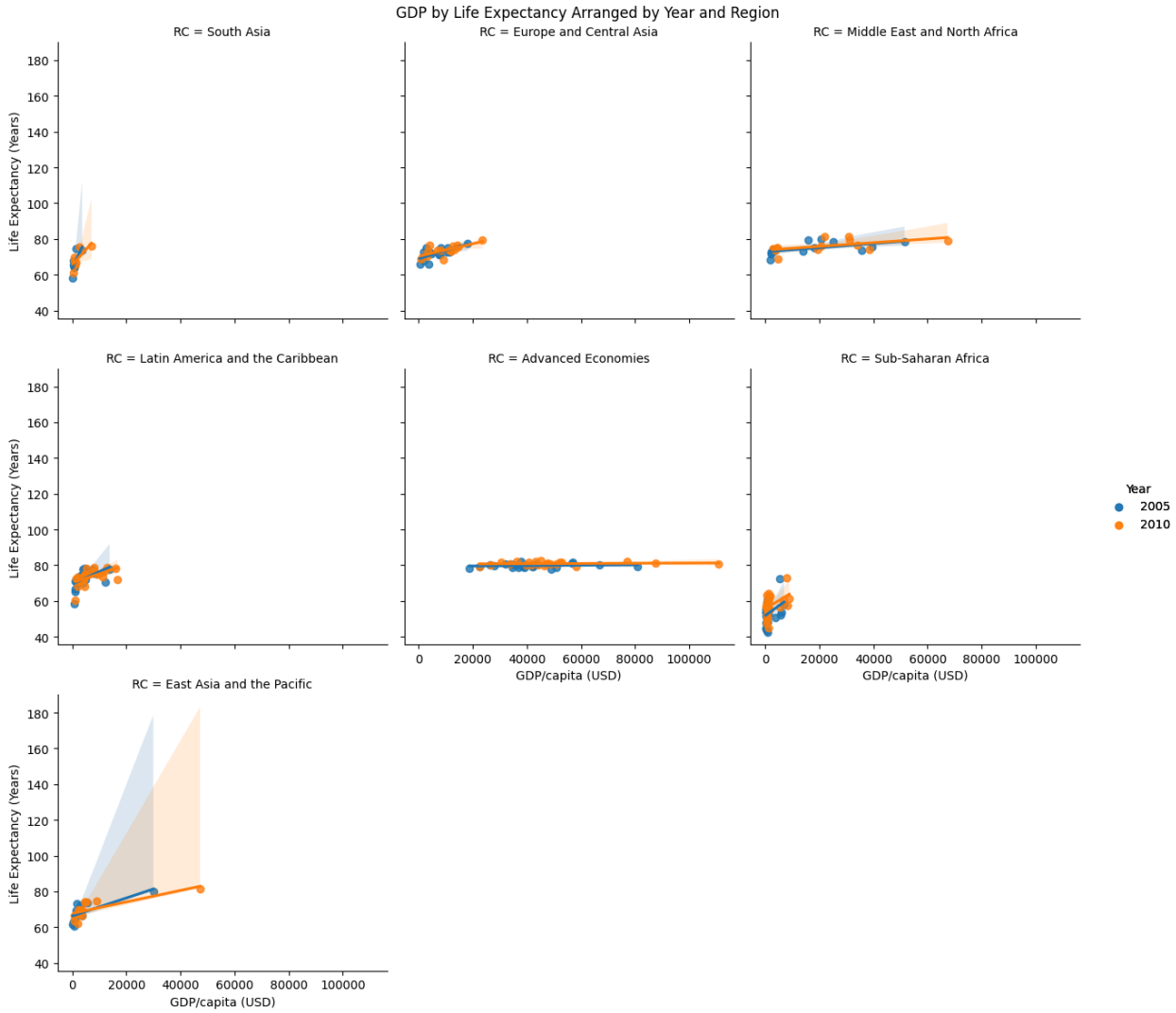
Notably within this 5 year period the biggest increases in mean life expectancy occur in Sub-Saharan Africa at a 4.1, with a mean increase of 1.81 for all countries, with most experiencing a 1-2 year mean increase.

How does GDP influence the Life Expectancy of a Country / What's the relationship between GDP and Life Expectancy?

```
g = sns.scatterplot(data = df.rename(columns ={"region_code":"RC"}),
    x = "GDP/capita (USD)",
    y = "Life Expectancy (Years)",
    hue = "Year"
    )
g.set_title("GDP by Life Expectancy")
fig = g.get_figure()
fig.savefig("GDP-LE-Year.png")
```

To answer this question we plotted the GDP of a country against their corresponding life expectancy for both years 2005 and 2010 using a scatterplot to see a trend if one is present. As shown, there is little difference in the GDP and life expectancy by year. However to ensure that a year that was picked for analysis was not an outlier we added both years. From the plot, the life expectancy appears to follow a logarithmic relationship with GDP. Overall, GDP/capita and life expectancy of a country are highly correlated, with a positive correlation between the two. Additionally, this relationship does not look linear and is more likely an exponential relationship between the GDP and life expectancy of a country. To avoid a Simpson's Paradox situation, these points were further analysed grouping by country region as follows:

```
g = sns.lmplot(data = df.rename(columns ={"region_code":"RC"}),
    x = "GDP/capita (USD)",
    y = "Life Expectancy (Years)",
    col = "RC",
    col_wrap = 3,
    height=4,
    hue = "Year"
    )
g.fig.suptitle("GDP by Life Expectancy Arranged by Year and Region")
g.add_legend()
g.tight_layout()
g.savefig("GDP-LE-Year-Region.png")
```



From this, we can establish that most regions remained relatively the same over the 5 year period, with small changes in GDP and Life Expectancy. For all regions however, the relationship seen in the overall GDP by life expectancy holds between all of the regions. However we cannot comment on the causation of these two variables as we do not have enough information to draw that conclusion.

GDP per Capita Statistics (520 463 178)

A summary containing the Minimum, Mean and Maximum grouped by the years 2005 and 2010 was created using Python coding.

```
import pandas as pd

df = pd.read_csv("intergrated.csv")
dic = {}
grouping = df["GDP"].groupby([df["Year"]])

dic["Mean"] = grouping.mean().round(2)
dic["Min"] = grouping.min().round(2)
dic["Max"] = grouping.max().round(2)
y = pd.DataFrame.from_dict(dic)
y.to_csv("GDP-per-Capita_Year_Summary.csv")
```

GDP per Capita summaries by year:

Year	Mean	Min	Max
2005	11418.41	151.68	124197.26
2010	14671.81	234.24	150737.89

As shown in the data summary above, the minimum, maximum and mean GDP per capita all increased from the years 2005 to 2010. The values were measured in US dollars, and it can be seen that the minimum GDP per capita increased in percentage the most at 54.43% (2 d.p) followed by the mean increasing by 28.49% (2 d.p) and the maximum increasing by 21.37% (2 d.p).

The following code gives the mean, minimum and maximum values by region, denoted by “region_code”.

```
12
13 dic2 = {}
14 group = df["GDP"].groupby([df["region_code"]),df["Year"]])
15 dic2["Mean"] = grouping.mean().round(2)
16 dic2["Min"] = grouping.min().round(2)
17 dic2["Max"] = grouping.max().round(2)
18 z = pd.DataFrame.from_dict(dic2)
19 z.to_csv("GDP-per-Capita_region_code_Year_Summary.csv")
```

Region	Year	Mean	Min	Max
Advanced Economies	2005	21495.35	18780.13	80988.14
Advanced Economies	2010	26794.68	22520.64	110885.99
East Asia and the Pacific	2005	1056.89	216.31	29961.26
East Asia and the Pacific	2010	2105.38	746.95	47236.96
Europe and Central Asia	2005	5907.80	340.58	18098.91
Europe and Central Asia	2010	8462.63	749.55	23532.48
Latin America and the Caribbean	2005	2208.70	781.28	13822.74
Latin America and the Caribbean	2010	3579.18	1191.97	16056.12
Middle East and North Africa	2005	17843.01	1855.52	51455.95
Middle East and North Africa	2010	21809.28	2839.93	67403.09
South Asia	2005	1058.44	242.03	3640.01

South Asia	2010	2019.75	543.30	7076.74
Sub-Saharan Africa	2005	478.91	151.68	6891.36
Sub-Saharan Africa	2010	856.18	234.24	8849.32

It can be discerned that there is an overall upwards shift in GDP per capita in every region from the years 2005 to 2010. However, poorer regions such as the sub-saharan Africa region had a larger percentage increase in GDP per capita from 2005-2010, with a 78.8% increase compared to the 24.7% in Gdp per capita for advanced economies.

Section B:

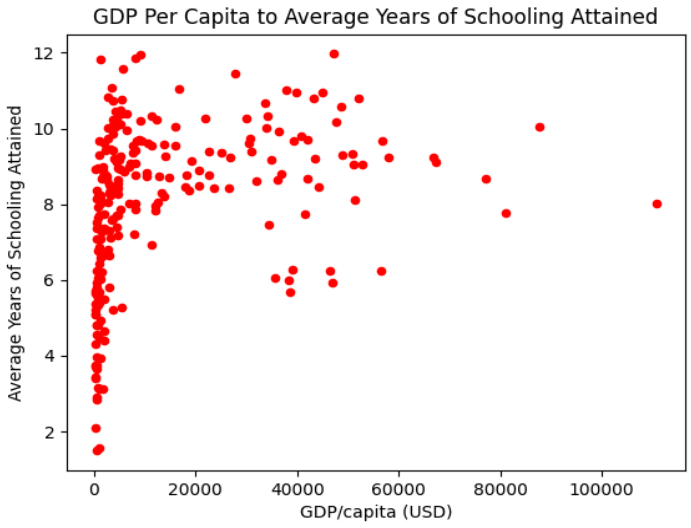
TOPIC: How does a country's GDP affect their education rate, crime rate and life expectancy?

Results by Year

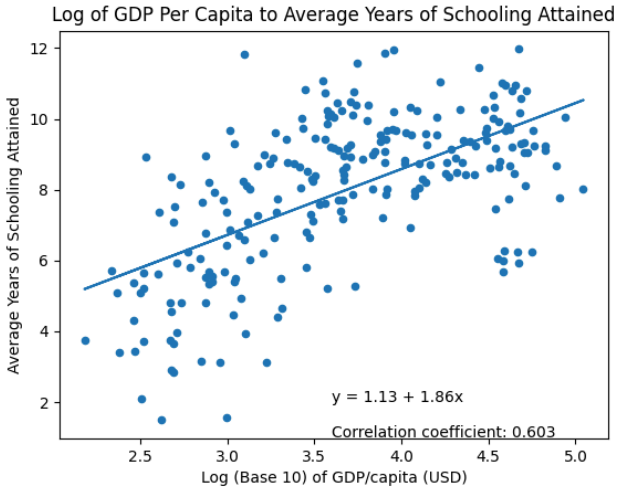
Year	Max Years in Schooling	Min Years in Schooling	Mean Years in Schooling	Average GDP/capita (USD)
2005	11.82	1.49	7.82	12218.86
2010	11.97	2.9	8.13	15504.09

From 2005 to 2010, all figures increased (Max, Min, Mean Years in Schooling and Average GDP/capita (USD)). GDP/capita increased the most at around 25% whereas Mean Years in Schooling only increased by about 4%. This suggests that there may be a positive relationship between GDP Per Capita, and Average Years of Schooling Attained and that it may not be a linear relationship.

To determine the relationship between GDP Per Capita (GDPC) and Average Years of Schooling Attained (AYS), a scatter plot was graphed to visualise the relationship. It appears that GDPC and AYS follow a logarithmic trend. Initially, an increase in GDPC appears to cause a large increase in AYS, from 0 GDPC to about 10000 GDPC (shown by the vertical trend in the data points). However, as GDPC increases beyond roughly 10000 GDPC, the growth plateaus and a horizontal relationship can be observed.



To reinforce that there is indeed a logarithmic relationship, AYS was plotted against the logarithm of GDPC. This gave us a linear trend with a correlation coefficient of 0.603, which suggests there is a strong positive correlation between the logarithm of GDPC and AYS.



This logarithmic relationship seems plausible as a typical developed education system consists of primary and secondary education on average from ages 5-18 (13 years) and tertiary education on average from ages 18-21 (3 years). Therefore it can be expected that AYS eventually plateaus even if there is a substantial increase in GDPC. In developed countries such as Australia, the government imposes mandatory school attendance from ages 6 to 15. Education beyond these 9 years is optional so it is unlikely that AYS will increase beyond 9-16 years.

It is also important to note that, despite there being a strong positive relationship between AYS and and the logarithm of GDPC, we are unable to prove whether there is a causal relationship between GDPC and AYS so an increase in GDPC may not cause and increase in AYS and an increase in AYS may not necessarily cause an increase in GDPC. This may

When analysing the life expectancy data by drawing summaries we found that between the 5 year period of 2005 to 2010 there was an overall increase in life expectancy for min max and average, with the largest summary increase being the minimum life expectancy:

Life Expectancy Summaries by Year

Year	**Mean**	**Min**	**Max**
2005	69.09	42.66	81.93
2010	71.06	45.10	82.84

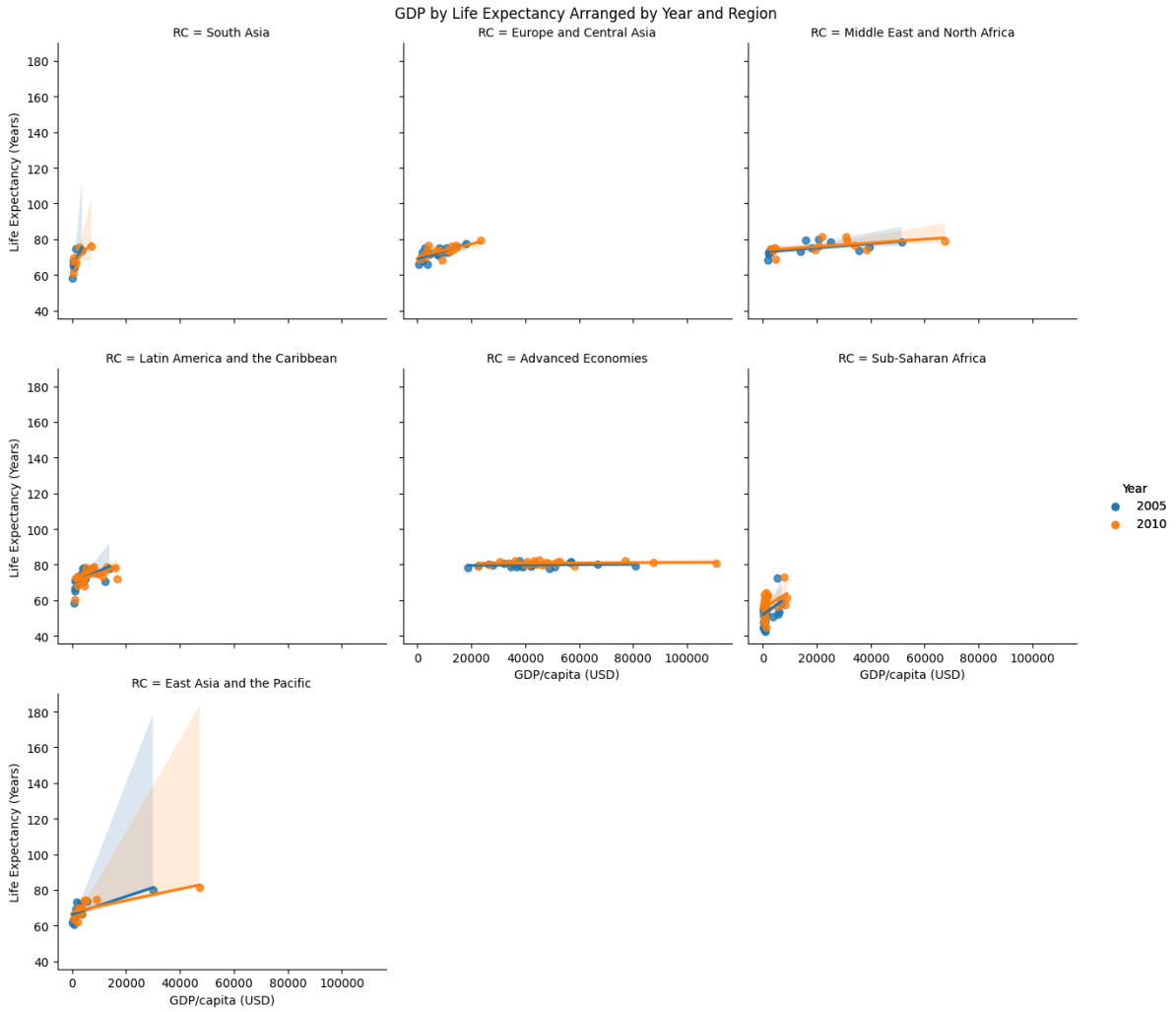
These were then broken down into respective regions to aid in our understanding of which areas of the world have lower or higher life expectancy summaries:

Life Expectancy Summaries by Year/Region

Region	**Year**	**Mean**	**Min**	**Max**
Advanced Economies	2005	79.78	77.84	81.93
Advanced Economies	2010	80.91	79.03	82.84
East Asia and the Pacific	2005	68.46	60.53	79.99
East Asia and the Pacific	2010	69.99	62.03	81.54
Europe and Central Asia	2005	71.99	65.86	77.61
Europe and Central Asia	2010	73.77	68.45	79.42
Latin America and the Caribbean	2005	72.45	58.65	78.12
Latin America and the Caribbean	2010	73.58	60.51	78.78
Middle East and North Africa	2005	74.90	68.27	80.15
Middle East and North Africa	2010	76.04	68.57	81.60
South Asia	2005	66.87	58.29	74.70
South Asia	2010	68.83	61.03	75.91
Sub-Saharan Africa	2005	53.49	42.66	72.43
Sub-Saharan Africa	2010	57.63	45.10	72.97

In observing a relationship between life expectancy and GDPC we can see in a plot that they are at least correlated:

From the plot, the relationship between GDPC and life expectancy appears to follow a logarithmic relationship (or exponential), where a low GDPC implies a lower life expectancy and a higher GDPC implies a high life expectancy. To check that this positive trend between GDPC and life expectancy remained between regions the chart was sectioned as follows:



From this graph we can see that this trend remains positive and can therefore say that the region has no impact on this trend.

In conclusion, while we cannot say life expectancy or GDPC have a direct impact on one another, it is evident that one another are highly correlated and have a positive relationship, that an increase in one does not necessarily entail an increase in the other, but it can be implied that an increase in one can cause an increase in the other.

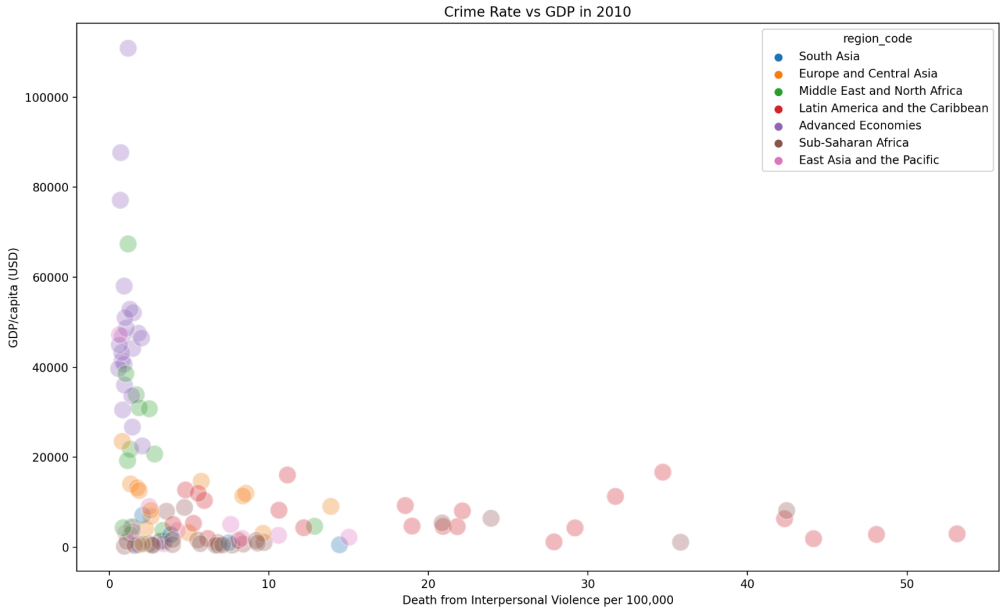
To determine the relationship between GDPC and crime rates, we first created a table to show the overall trend of the world. The numbers presented the number of deaths from interpersonal violence per 100,000 people. Even though there could be more factors that would add when looking at overall crime rate, we will only focus on

Year	Mean	Min	Max
2005	8.47	0.71	54.88
2010	8.07	0.59	53.13

the number of deaths for this report. According to the table, we can see the numbers decreased for all three categories: mean, minimum, and maximum number of deaths.

region_code	Year	Mean	Min	Max
Advanced Economies	2005	1.23	0.71	2.42
Advanced Economies	2010	1.14	0.59	2.08
East Asia and the Pacific	2005	5.32	0.79	17.54
East Asia and the Pacific	2010	4.82	0.64	15.02
Europe and Central Asia	2005	6.84	1.47	21.06
Europe and Central Asia	2010	4.76	0.82	13.89
Latin America and the Caribbean	2005	21.04	3.77	54.88
Latin America and the Caribbean	2010	21.19	3.99	53.13
Middle East and North Africa	2005	2.51	0.9	9.25
Middle East and North Africa	2010	2.58	0.87	12.87
South Asia	2005	5.39	1.74	14.41
South Asia	2010	5.09	1.85	14.43
Sub-Saharan Africa	2005	9.62	0.9	53.09
Sub-Saharan Africa	2010	9.15	0.95	42.45

We further grouped the dataset into regions and years to get more details about the trends between regions. We can see that Latin America and the Caribbean and Sub-Saharan Africa had the highest rate of crime out of all the regions while Advanced Economies had the lowest rate of crime.



This scatter plot shows some ideas about how crime rate could be related to GDPC in 2010 specifically. We can see there are a lot of purple dots on the left side of the plot. This means many countries from Advanced Economies have lower crime rates. As the name suggests, countries from Advanced Economies have higher GDPC. Meanwhile, countries from Latin America and the Caribbean and Sub-Saharan Africa

tend to have higher crime rates while having lower GDPC. There are a lot of countries with low GDPC with low crime rates.

In conclusion we can say there is a negative correlation between GDPC and crime rate, meaning the higher the GDPC of a country, the more likely they have a low crime rate. However, it is important to note that GDPC and crime rate are not directly proportional. As we can see, more countries with low GDPC also have less extreme crime rate instead of high crime rate as our conclusion suggests. Also this data is only looking at one year time hence less reliability. Moreover, we cannot know the exact relationship of the two factors as it is hard to find if GDPC determines the crime rate or vice versa.

Limitations

One limitation of our dataset is that it is outdated. Our dataset is made up of values from the years 2005 and 2010 which may not be applicable to the present. For example, COVID19 could have caused unforeseen effects that make this comparison not applicable. Social norms are also constantly changing which may also have an effect on the reliability of data from 2005 and 2010. For example, there may be more, or less, promotion and fixation on attaining education which can affect its relationship with GDPC.

To note, the years used, particularly 2010, follows 2 years after the 2008 Global Financial Crisis. This may have an effect on particular parts of the data, for example, over between these years there was an observed drop in many statistical summaries compared to that of 2005. While this may not necessarily be the cause, it is important to take into consideration.

References

Life expectancy at birth, total (years), The World Bank, Retrieved September 9, 2022, from: https://data.worldbank.org/indicator/SP.DYN.LE00.IN

Roser, M., & Ritchie, H. (2019). Deaths - Interpersonal Violence - Sex: Both - Age: All Ages
(Rate) [Data Set]. Our World in Data. https://ourworldindata.org/homicides

Robert J. Barro, Jong-Wha Lee. (2021). Barro-Lee Education Attainment Dataset. BarroLeeDataSet. Retrieved from: http://www.barrolee.com

Appendix

Regions:
Advanced Economies: Australia, Austria, Belgium, Canada, Switzerland, Germany, Denmark, Spain, Finland, France, United Kingdom, Greece, Ireland, Iceland, Italy, Japan, Luxembourg, Netherlands, Norway, New Zealand, Portugal, Sweden

East Asia and the Pacific: China, Fiji, Indonesia, Cambodia, Myanmar, Mongolia, Malaysia, Philippines, Papua New Guinea, Singapore, Thailand, Tonga

Europe and Central Asia: Albania, Armenia, Bulgaria, Estonia, Croatia, Hungary, Kazakhstan, Lithuania, Latvia, Poland, Romania, Slovenia, Tajikistan, Ukraine

Latin America and the Caribbean: Argentina, Belize, Bolivia, Brazil, Barbados, Chile, Colombia, Costa Rica, Cuba, Ecuador, Guatemala, Guyana, Honduras, Haiti, Jamaica, Mexico, Nicaragua, Panama, Peru, Paraguay, El Salvador, Trinidad and Tobago, Uruguay

Middle East and North Africa: United Arab Emirates, Bahrain, Cyprus, Algeria, Iraq, Israel, Jordan, Kuwait, Morocco, Malta, Qatar, Saudi Arabia, Tunisia
South Asia: Afghanistan, Bangladesh, India, Sri Lanka, Maldives, Nepal, Pakistan

Sub-Saharan Africa: Burundi, Benin, Botswana, Central African Republic, Cameroon, Gabon, Ghana, Kenya, Liberia, Lesotho, Mali, Mozambique, Mauritania, Mauritius, Malawi, Namibia, Niger, Rwanda, Sudan, Senegal, Sierra Leone, Togo, Uganda, South Africa, Zambia, Zimbabwe