

**REPORT OF ASSIGNMENT 2**

**Carnegie  
Mellon  
University  
Africa**

**Submitted By: Peace Ekundayo Bakare**

**Course: Data, Inference, and Applied Machine Learning**

## **Libraries Used:**

*Pip Installed quandl, pandas, matplotlib, numpy*

1. *Matplotlib.pyplot as peaceplt* for visualizations and graph plotting
2. *Pandas as peacepd* for exploring and manipulating data in data sets
3. *Google.colab.drive* for file storage and retrieval
4. *Seaborn as peacesns* for visualization and statistical graphics
5. *Numpy as peacenp* for calculations
6. *Tabulate as tabulate* for tabulation
7. *Display as display* for display of tables

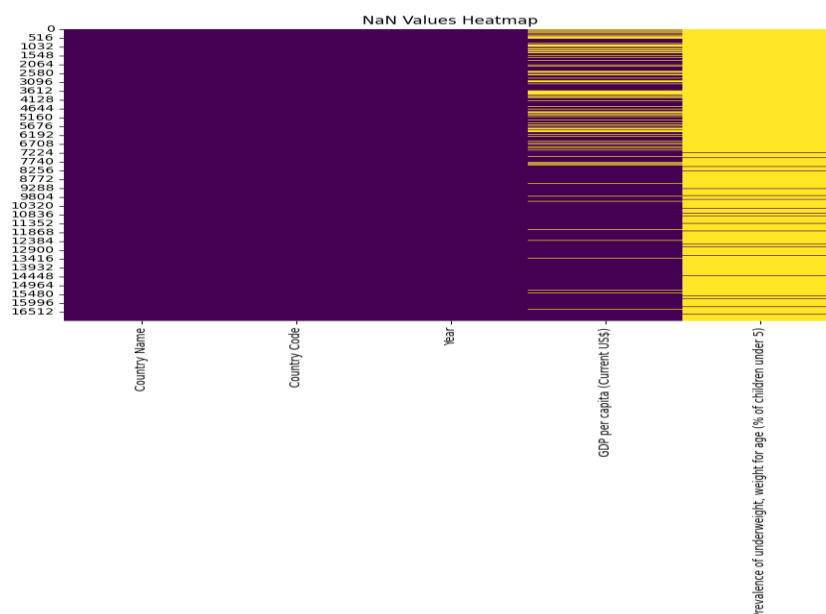
## QUESTION 1

**What kind of relationship do you expect?** I expected to see countries with low GDP per capita to have more prevalence of underweight and countries with high GDP per capita to have less prevalence of underweight.

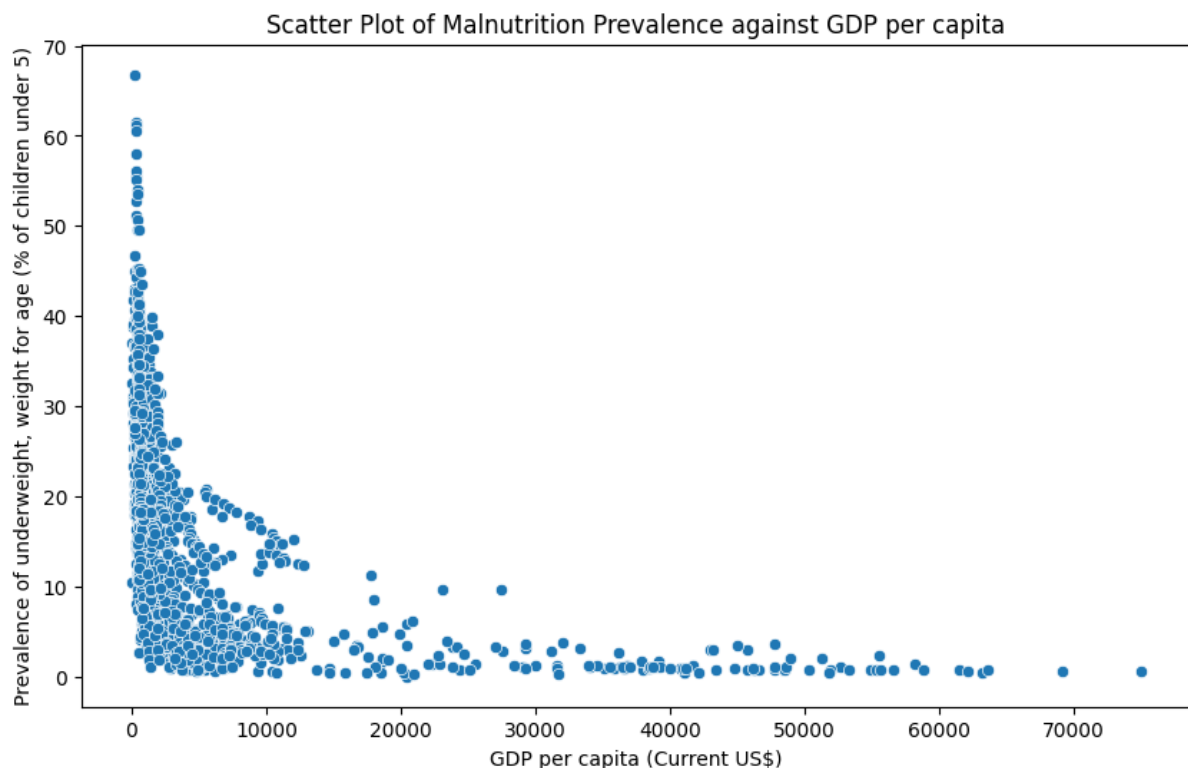
My approach to solving this question is first by cleaning the data and making it ready to plot the required graphs. I used the .csv files.

These are the steps to be precise:

1. **Reading the datasets.** I read the GDP dataset as *df\_GDP* and the prevalence of underweight as *df\_weight*. The first 4 rows in the .csv file are not useful, so I used the *skiprows* parameter to escape those lines.
2. **Melting and Merging the dataframes.** I used the *melt()* method on the dataframes to transform the dataframe from wide to long format i.e. turning the columns into rows, making the data easier to analyze or to merge. I performed the melting function on the 2 dataframes and stored the melted dataframes in *df\_GDP\_melted* and *df\_weight\_melted* dataframes, for the GDP and Prevalence of underweight respectively. After melting, I merged the dataframes together into another dataframe, *merged\_data* on the 'Country Name' and 'Year' columns, using inner join.
3. **An Add-on.** I plotted an heatmap to show the NaN values frequency in the *merged\_data* dataframe. It can be deduced that the Prevalence of Underweight, weight for age (% of children under 5) has more NaN values, followed by GDP per capita (Current US\$) column.

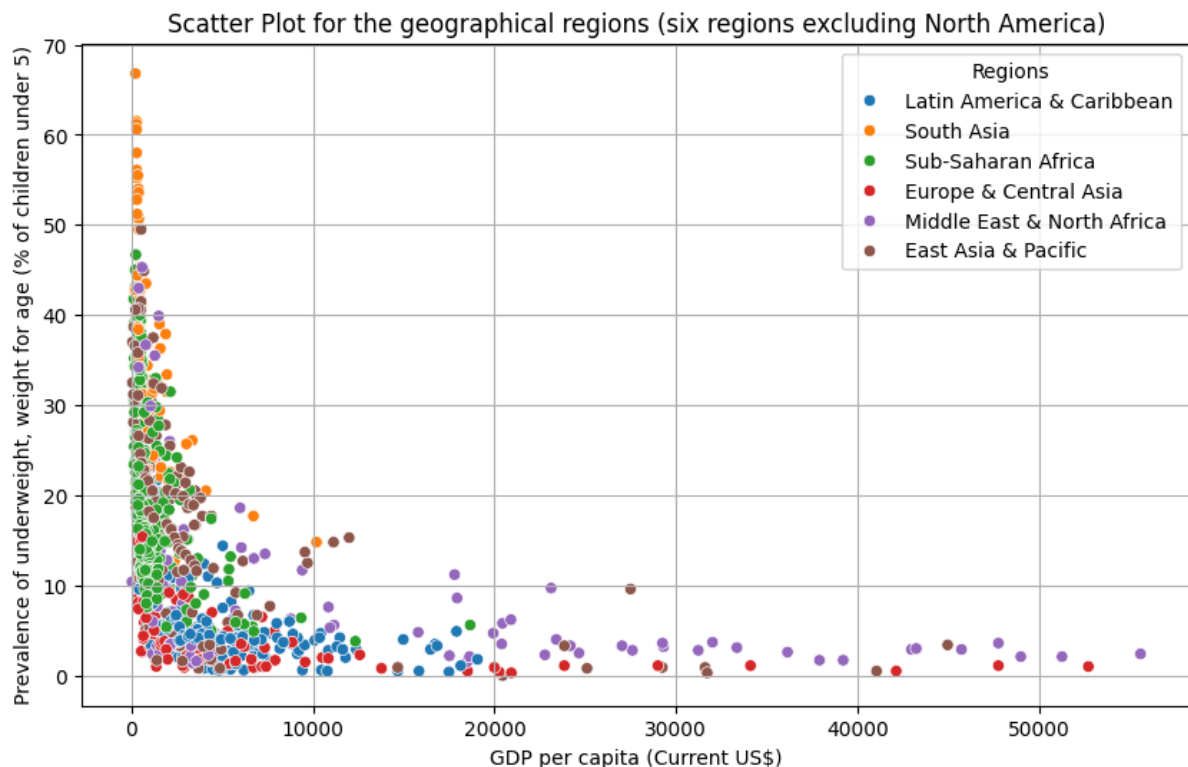


4. **Plotted the Scatter Plot of Malnutrition Prevalence against GDP per capita** using the *matplotlib.pyplot* library as *peaceplt* and the dataframe, *merged\_data*.



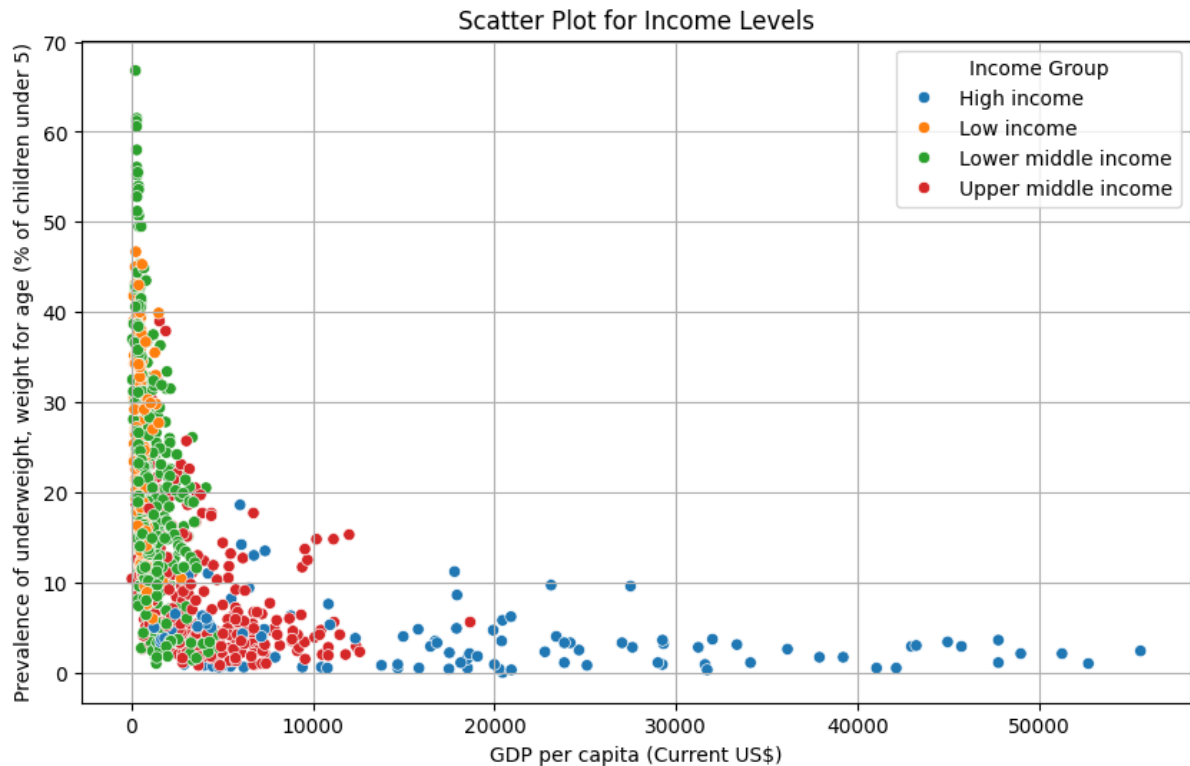
**What kind of relationship do you see?** I see that as the GDP per capita of a country increases from 0 to \$70,000, the prevalence of underweight is closer to 0%.

5. **Read the GDP Metadata.** I read the GDP metadata into the dataframe, *df\_GDP\_geo\_regions*, and merged it with the *merged\_data* on 'Country Code', to get the country regions and the Income levels to be used for the subsequent scatter plots. The new dataframe, *merged\_data\_with\_metadata* is formed from this merger.
6. **Plotted the Scatter Plot for the geographical regions.** To plot the graph excluding North America, I created another dataframe, *merged\_data\_without\_america*, and excluded the North America region. The scatter plot of the six regions of the prevalence of underweight versus GDP per capita is plotted with the legend indicating the Regions.



**Explanation of the Graph:** Most Sub-Saharan African countries have GDP per capita between \$0 and \$10,000, and would be rated as the region with the second highest prevalence of underweight. The South Asia region has the highest record of prevalence of underweight in children under 5 years. This can be explained seeing that most countries in South Asia have very low GDP per capita between \$0 to \$4,000. One of the countries in Middle East & North Africa region stood out, having the highest GDP per capita, although not with the lowest prevalence of underweight. Considering a perfect blend of the prevalence of underweight and GDP per capita i.e. the country and region with the lowest prevalence of underweight and considerable high GDP per capita, one country in Europe & Central Asia stood out, having a GDP per capita of \$42,000 and prevalence of underweight value of 0.

7. **Plotted the Scatter Plot for Income Levels.** To plot the graph for Income levels excluding North America, I used the R dataframe, *merged\_data\_without\_america*, that excluded the North America region. The scatter plot of the four Income groups, of the prevalence of underweight versus GDP per capita is plotted with the legend indicating the Income Groups.



**Insight:** The scatter plots show that richer countries tend to have fewer malnourished children, while poorer countries face a tougher battle against child malnutrition. However, some countries did not follow the trend, and this could be due to differences in how wealth is distributed or the effectiveness of policies aimed at improving child health.

In the battle against child malnutrition, the plots have improved our understanding of the most problematic areas and the areas that most require assistance, particularly in areas with high rates of malnutrition and low GDP. It also emphasizes how crucial fair access to resources, social safety nets, and economic growth are to lowering the rate of malnutrition.

## Question 2.

### Approach when using Quandl API:

**Installation, Importation & Understanding Dataset:** Using the quandl API, I installed quandl using pip and then imported quandl. I used my API Key after signing up on Nasdaq and created three dataframes, *test\_wheat\_data*, *test\_crude\_oil\_data*, *test\_gold\_prices\_data*, to store the data for wheat, crude oil and gold prices respectively. These dataframes are to be used to understand the dataset in order to synchronize the dates of the three items, wheat, crude oil and gold, before merging them to plot the time series graph.

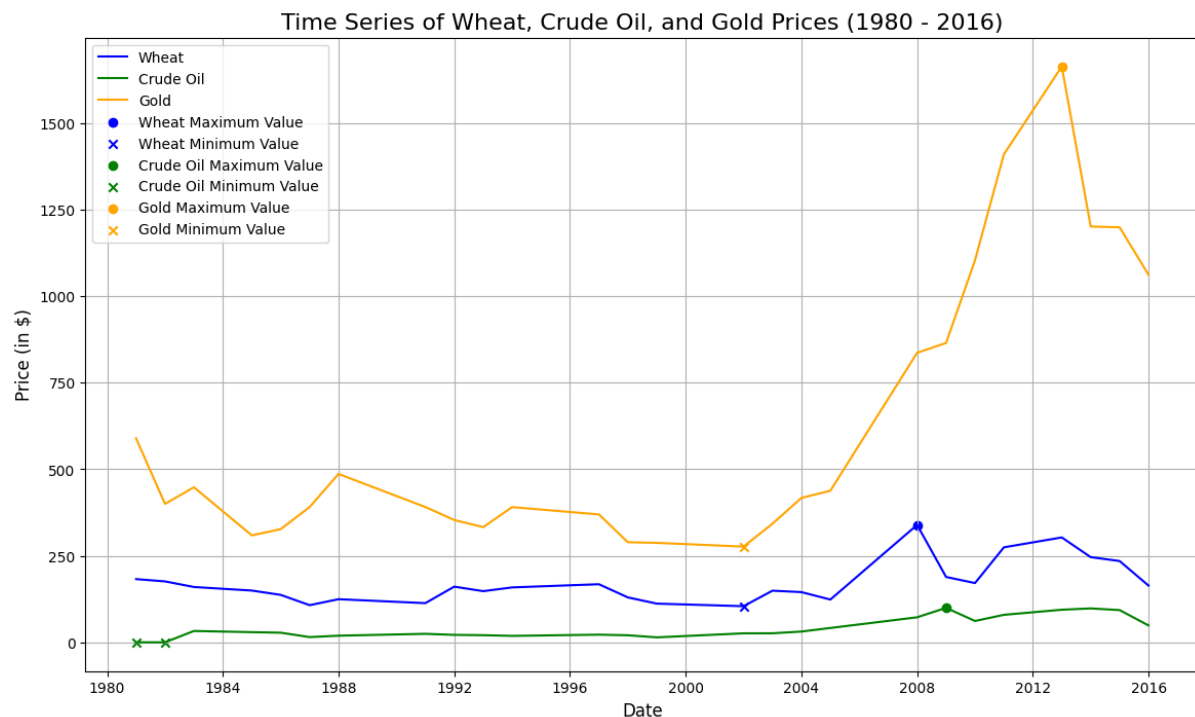
**Synchronizing Datasets using Date:** I used *1980-01-31* as the start date and *2016-04-18* as the end date, in a bid to synchronize the data. Both Crude Oil and Gold have start dates earlier than the 1980 of Wheat. To synchronize the data, I cut off the data before this time for both Crude oil and gold. Also, Wheat and Crude Oil have values beyond the end date I used but I cut them off to accommodate the earlier date of Gold which is *2016-04-18*.

**Merging Dataframes and Plotting the Graph:** To merge the dataframes, I renamed the “Value” column of each dataframe to real names of the items. I merged the three dataframes using the *merge()* function of the pandas library. I plotted the time series of Wheat, Crude Oil and Gold Prices from 1980 to 2016 thereafter, with each item having different colors, and using *dot(.)* to indicate crests or maximum data points and alphabet X (x) to indicate troughs or minimum data points, for the three items.

### Approach when using given Datasets

Aside from reading each of the given datasets and the common procedure to the two approaches, I merged the three dataframes on the ‘Date’ column, using inner join.

Here is the Time Series Graph with the legend at the top left.



**Interpretation of Time Series Graph:** The graph of prices (in \$) of Wheat, Crude Oil and Gold is plotted against the date in years, from 1980 to 2016. Gold has the highest prices overtime from 1980 to 2016, having its peak price as \$1800 shortly after 2012, and its lowest price at a price slightly above \$250. Wheat has its highest price of around \$350 in 2008 and its lowest price of about \$125 in 2002. Crude Oil has its highest price of around \$125 in about 2009, and its lowest price of \$0 around 1981.

**Insight:** The time series analysis, combined with highlighting the highs and lows in each commodity's price, helps us understand how global events, economic conditions, and market dynamics influence the prices of Wheat, Crude Oil, and Gold. These insights offer valuable information about the health of the economy, energy consumption, and food security across the world.



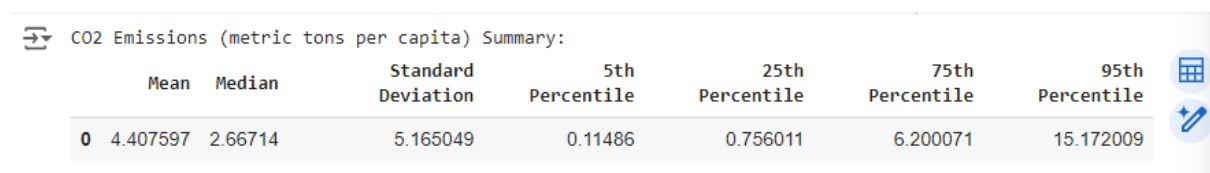
### Question 3.

#### Approach:

I imported *display* from *IPython.display* and read the CO2 data, skipping the first 4 rows. It's important to melt the dataframe to rightly display the data. I converted the 'Year' column in the dataframe to numeric to handle filtering, then I selected the data in *co2\_data\_melted* with the year 2010 and stored it in a dataframe, *co2\_data\_2010*.

Consequently, I defined a function called *summary\_statistics*. The function calculates the mean, median, standard deviation, 5th, 25th, 75th and 95th percentile, and returns a dataframe. The method takes in a dataframe and the column with the data to be used for computation. This function helps to manage the calculation of the measures of variation and measures of central tendency of CO2 and School enrolment. The table is then displayed using the display function from IPython library.

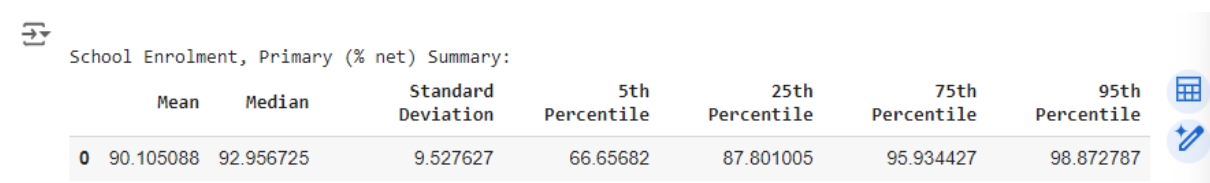
Here is the Table for CO2



	Mean	Median	Standard Deviation	5th Percentile	25th Percentile	75th Percentile	95th Percentile
0	4.407597	2.66714	5.165049	0.11486	0.756011	6.200071	15.172009

**Insight:** The difference between the 5th and 95th percentiles is large. This shows a major disparity in CO2 emissions. This likely indicates that wealthier or more industrialized countries are emitting far more CO2 per capita than developing countries. Regions with higher per capita CO2 emissions may need stricter environmental policies or greener technologies to reduce their impact, while low-emission countries may still be in early development stages and contributing less to global emissions.

Here is the table for School Enrolment



	Mean	Median	Standard Deviation	5th Percentile	25th Percentile	75th Percentile	95th Percentile
0	90.105088	92.956725	9.527627	66.65682	87.801005	95.934427	98.872787

**Insight:** A high mean and median close to 100% indicate that many countries have done so well with primary school enrollment, reflecting progress towards education-related development goals.

In addition, a large difference between the 5th and 95th percentiles could suggest huge inequality, with some countries still struggling to provide basic education to all children.

Consequently, countries with low enrolment rates (seen in the lower percentiles) likely face socio-economic challenges, including poverty or insufficient educational infrastructure.

**Question: How would you handle the NaN values?**

I dropped the NaN values because I don't want to introduce bias to the data. Also, dropping the NaN values ensures that calculations like mean, median, standard deviation, and percentiles are based only on available & complete data.

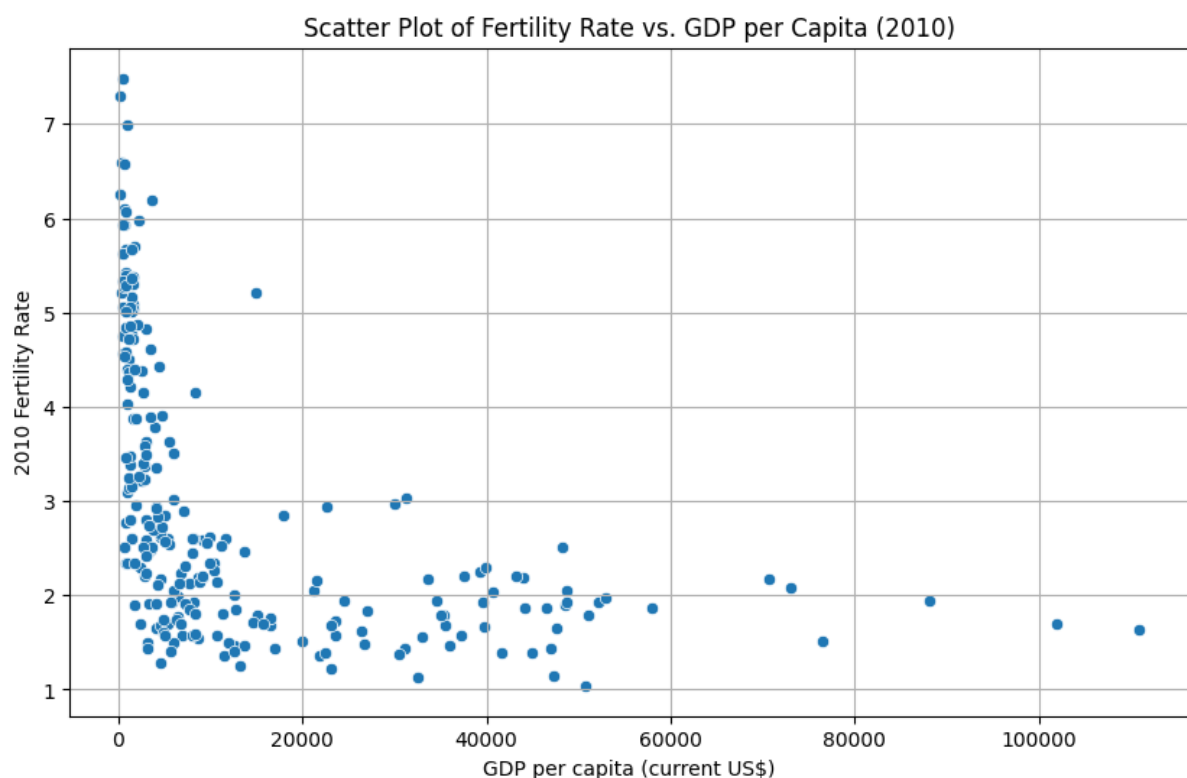
## Question 4

### Approach:

**Reading Data and Renaming Columns:** I read the data for fertility and skipped the first four rows since the values are not relevant for the analysis. I didn't have to read the GDP data again since I have read it in Question 1. Since the focus is on the year 1990 and 2010, I selected the following columns, Country Name, 1990 and 2010, and assigned the value to the dataframe, *fertility\_data\_2010*. I also renamed the columns of 1990 and 2010 to "1990 Fertility Rate, Total (births per woman)" and "2010 Fertility Rate, Total (births per woman)" respectively. I did the same thing for the GDP data, *gdp\_data\_2010*.

**Graph Plotting:** I merged the *fertility\_data\_2010* and *gdp\_data\_2010* dataframes and cleaned it by removing NaN values. After that, plotted the scatter plot graph of Fertility Rate against GDP per capita (2010), with the title "Scatter Plot of Fertility Rate vs. GDP per capita (2010)".

Here is the scatter plot:



**Insight:** Plotting the number of births per woman (the fertility rate) versus GDP per capita for every nation in 2010 reveals that nations with higher GDP per capita typically had lower fertility rates. i.e., families often have fewer children as nations get affluent. This is mostly due to factors like increased professional choices and improved access to education, particularly for women. On the other hand, because larger families are more

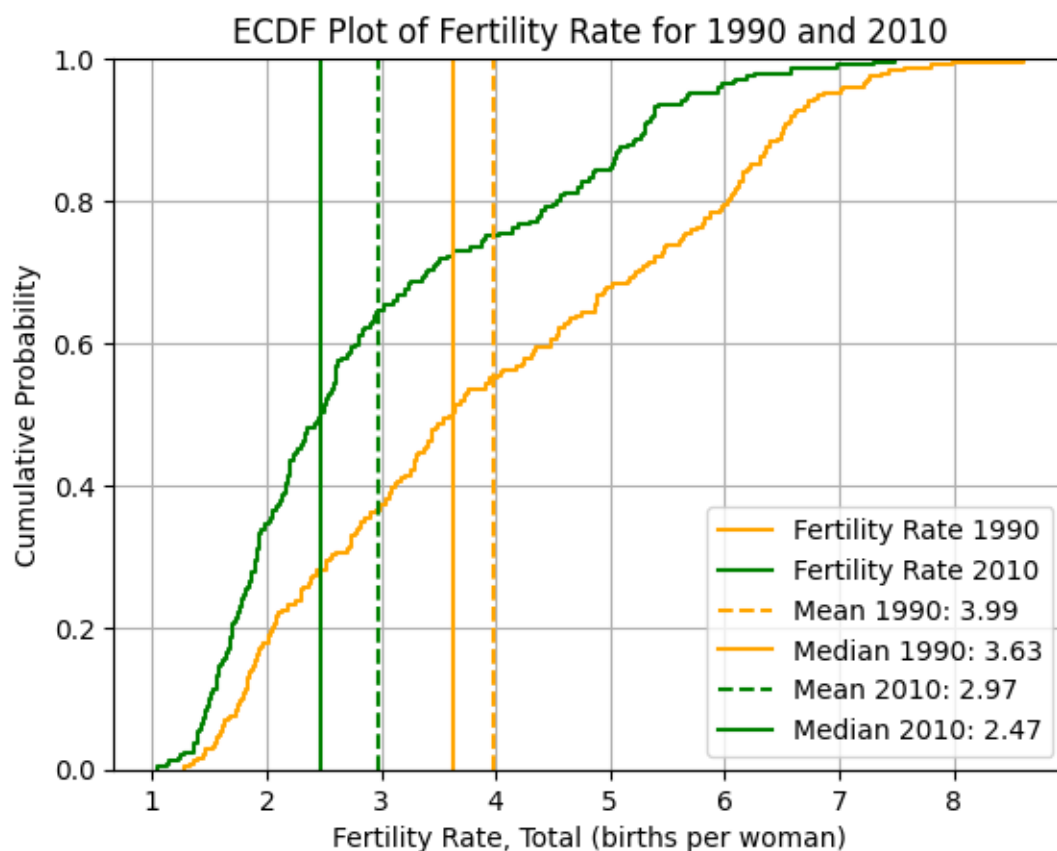
typical in low-income nations, fertility rates are often greater there. This could be because of certain cultural or economic considerations for the outliers.

### Plot the Cumulative Distribution Functions Plot

To Plot the ECDF plot of fertility rate for 1990 and 2010, I used the `ecdfplot()` method in the seaborn library to plot the graph using the column “1990 Fertility Rate, Total (births per woman)”. I gave this a label of “Fertility Rate 1990” and colored it **Orange**. In like manner, I used the `ecdfplot()` method to plot the column “2010 Fertility Rate, Total (births per woman)”, labelled it “Fertility Rate 2010” and gave it a color of **Green**.

To indicate the mean and median values with vertical lines on the plot, I had to calculate the mean for 2010, *mean\_2010*, median for 2010, *median\_2010*, and the mean for 1990, *mean\_1990*, median for 1990, *median\_1990*. I used the `axvline()` method to add vertical lines for the mean and median of 1990 and 2010, signifying the colors to differentiate each line.

Finally, I plotted the “ECDF Plot of Fertility Rate for 1990 and 2010”, with “Fertility Rate, Total (births per woman)” on the x-axis, and “Cumulative Probability” on the y-axis. The plot has a legend for the interpretation of the graph.



**Insight:** Comparing the fertility rates' cumulative distribution functions (CDFs) between 1990 and 2010 shows how much has changed. Given that many countries had greater fertility rates in 1990, the distribution might have been more dispersed. This distribution probably shifts to the left by 2010, suggesting that many nations have moved toward lower fertility rates.

Adding the vertical lines for the mean and median made it clearer. Here is it:

Most countries have fewer children than in 1990 if the mean and median fertility rates in 2010 have changed to lower levels. Also, a narrower difference between the mean and median may also indicate a trend toward more national consistency in fertility rates.

**Statistics:** I displayed the mean and median values for the Fertility Rate of 1990 and 2010 in tabular form.

	Fertility Rate 1990	Fertility Rate 2010
Statistic		
Mean	3.985405	2.968362
Median	3.630000	2.474000


### Have Fertility Rates changed over time?

Yes, throughout the 20 years between 1990 and 2010, fertility rates have decreased. This is due to a number of causes, including improved family planning access, rising urbanization, and shifting social standards regarding the size of families. Family sizes tend to decline as nations' economies grow, the graph illustrates this.

## Question 5

**Reading & Data Cleaning:** I read the Happy Planet Index, and the Corruption Perceptions Index data as excel files. I used the 'Complete HPI data' sheet from the happiness index file and skipped 5 rows, and removed the first and last column, as they did not contain any useful information for the analysis. For the Corruption Perceptions Index data, I used the 'CPI2016\_FINAL\_16Jan' sheet for the analysis.

**Filtering and Scatter Plot:** The columns needed to demonstrate the relationship between the happiness index and the corruption perception index are the Country, HPI Rank columns from the Happy Planet Index dataframe, *hpi\_data\_selected*, and Country, WB Code, and Rank columns from the Corruption Perceptions Index dataframe, *corruption\_data\_selected*. The two dataframes were merged into *hpi\_corruption\_merged* dataframe using the 'Country' column.



	Country	HPI Rank	WB Code	Rank
0	Afghanistan	110.0	AFG	169
1	Albania	13.0	ALB	83
2	Algeria	30.0	DZA	108
3	Argentina	19.0	ARG	95
4	Armenia	73.0	ARM	113

Fig: First 5 rows of the merged dataframes.

I plotted a graph of HPI Rank (Happy Planet Index) against CPI 2016 Rank (Corruption Perception) using the **HPI Rank** and **Rank** columns from *hpi\_corruption\_merged*, to demonstrate the relationship between happiness and corruption across countries. The scatter plot is well annotated using the country codes.

Fully annotated Scatter Plot showing relationship between Happy Index Rank and Corruption Rank (2016)

