DATA, INFERENCE, AND APPLIED MACHINE LEARNING

ASSIGNMENT ONE REPORT

NAME: EZE, NNAMDI SHEDRACK

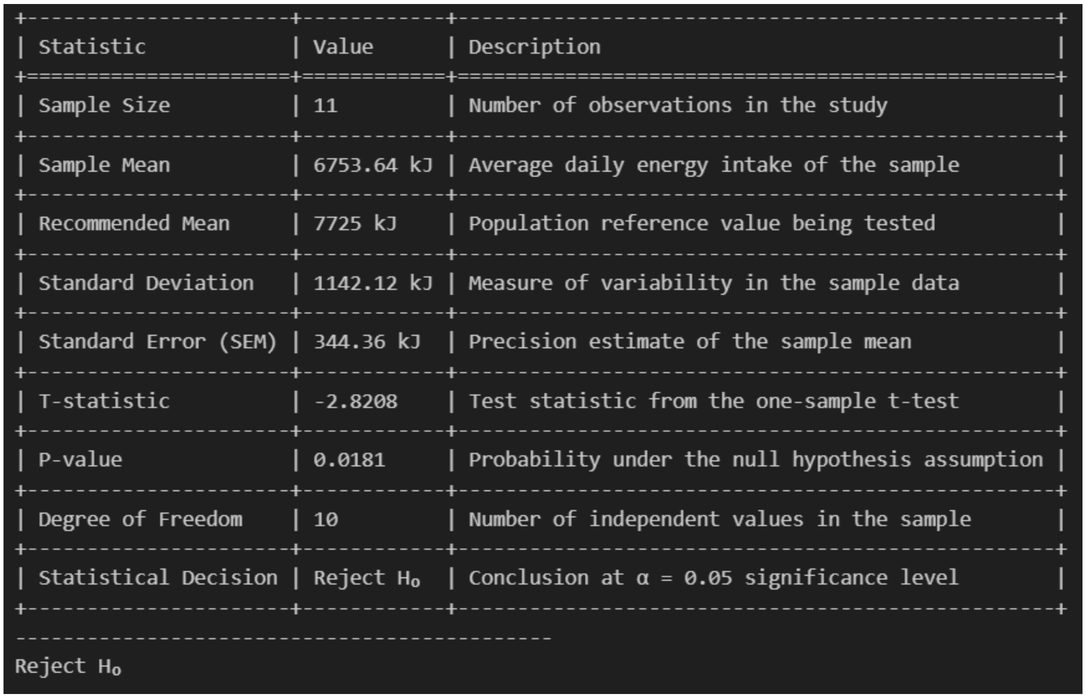
ANDREW ID: SEZE

DATE: 23rd SEPT 2025

Libraries used: Numpy, Pandas, Matplotlib, Openpyxl, Seaborn, Scipy, requests, tabulate

QUESTION (1): INVESTIGATING ENERGY INTAKE IN WOMEN

The purpose of this analysis is to see whether the average daily energy intake of a group of 11 women is different from the recommended value of 7725 kilojoules. The starting assumption, which in statistics we call the **null hypothesis**, was that there is no real difference between the average energy intake of our group of 11 women and the recommended value of 7725 kJ. The alternative hypothesis however is that there is significant difference between the average energy intake and the recommended energy intake. The goal of the test is to see if we have enough evidence to challenge the null hypothesis. The method used in this analysis is called a one-sample t-test. This test compares the average energy intake from my sample to the recommended value. It also tells if any differences observed are likely real or they happen just by chance. Because the assumption wasn't whether intake would be specifically higher or lower; therefore, a two-tailed test was used. This kind of test checks for all differences not just whether the average is above or below the target. The result of the analysis is shown in figure one below.



```
+----------------------+------------+--------------------------------------------------+
| Statistic            | Value      | Description                                      |
+======================+============+==================================================+
| Sample Size          | 11         | Number of observations in the study              |
+----------------------+------------+--------------------------------------------------+
| Sample Mean          | 6753.64 kJ | Average daily energy intake of the sample        |
+----------------------+------------+--------------------------------------------------+
| Recommended Mean     | 7725 kJ    | Population reference value being tested          |
+----------------------+------------+--------------------------------------------------+
| Standard Deviation   | 1142.12 kJ | Measure of variability in the sample data        |
+----------------------+------------+--------------------------------------------------+
| Standard Error (SEM) | 344.36 kJ  | Precision estimate of the sample mean            |
+----------------------+------------+--------------------------------------------------+
| T-statistic          | -2.8208    | Test statistic from the one-sample t-test        |
+----------------------+------------+--------------------------------------------------+
| P-value              | 0.0181     | Probability under the null hypothesis assumption |
+----------------------+------------+--------------------------------------------------+
| Degree of Freedom    | 10         | Number of independent values in the sample       |
+----------------------+------------+--------------------------------------------------+
| Statistical Decision | Reject H₀  | Conclusion at α = 0.05 significance level        |
+----------------------+------------+--------------------------------------------------+
--------------------------------------------
Reject H₀
```

Fig 1.0 Energy intake of Women

To do this, I imported NumPy and SciPy, two python libraries that are used for calculations and statistical computations. From the libraries, I calculated the sample mean using Numpy as show in figure 1 above, which came out to be 6753.64 kilojoules. This is of course lower than the recommended value. Next step was to also calculate the standard deviation using Numpy as well to see how much the individual values varied and the result came out to be 1142.12 kilojoules. This indicates that the energy intake levels were quite spread out. The standard error of the mean came out with 344.36 kilojoules. This gives us an idea of how accurate the sample's mean value is as an estimate for the population's true average.

Using these values, next we computed the t-statistic using SciPy, and the result came out to be -2.82. This number tells how far the sample mean is from the recommended value in terms of standard errors. The degrees of freedom for this test was 10, just 1 less than the number of people from which our sample was drawn. Finally, to find the p-value, the calculation was done with t-statistic, a module within the SciPy library. The calculated p-value returned a result of 0.0181. The p-value tells us how likely we are to see a difference of this size just by chance assuming the recommended value is correct. Since our p-value of 0.018 is smaller than 0.05, we have strong enough evidence to reject the null hypothesis because the difference is statistically significant.

Conclusion: Reject the null hypothesis because the difference is statistically significant. Women in this sample need to consume more energy than they are and this is unlikely simply due to random variation. The statistical result provides strong evidence that on average their intake is indeed lower than the recommended 7725 kJ. This could be due to a variety of factors such as dietary habits, the way of life, or even cultural influences. Although the size of the sample was limited, the findings are clear enough for us to reject the null hypothesis.

QUESTION 2: ANALYSIS OF GLOBAL CO2 EMISSIONS AND PRIMARY COMPLETION RATES

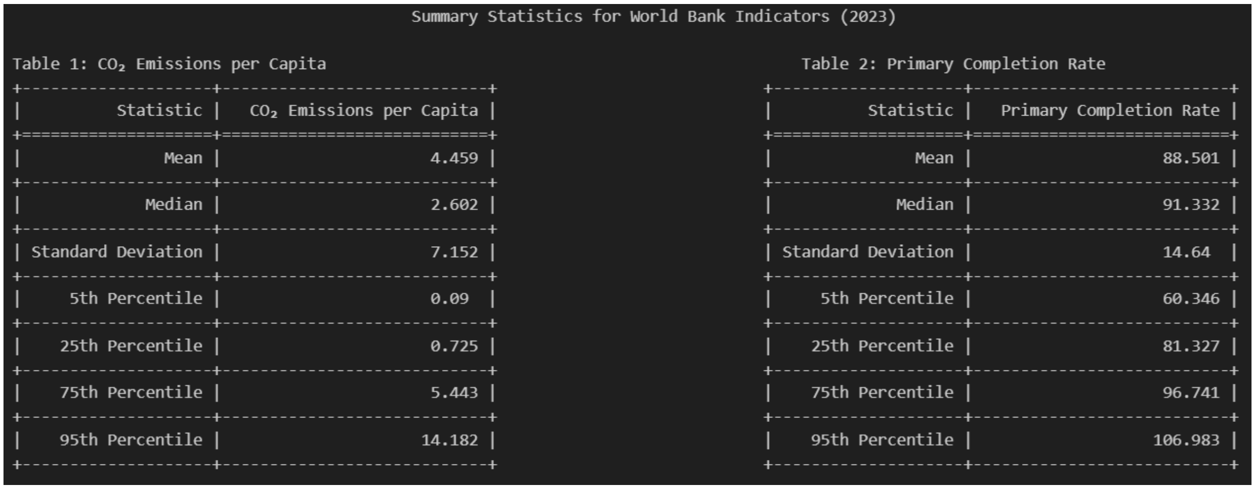
For the second question of the assignment, the goal was to perform a statistical analysis of two key global indicators from the World Bank for the year 2023. The first indicator was "CO2 emissions per capita," which reflects how much carbon dioxide is emitted per person in each country, excluding emissions from land use and forestry. and the second was the "Primary completion rate," which shows the percentage of children who complete primary school in each country. This process began with using Python and the World Bank API to fetch the data for all countries [1] [2]. To ensure the integrity of the analysis, the dataset was cleaned by removing any entries that had missing values for 2023. This step was crucial for calculating accurate summary statistics. Next, the summary statistics for each indicator was computed. These statistics include the mean, median, standard deviation, and percentiles at the 5th, 25th, 75th, and 95th levels. This helps give a clearer picture of how these values are distributed globally.

Upon analyzing the dataset, it was found that the mean, or average, emission level was **4.46 tons** per capita. However, to get a more representative picture of a typical country, I also calculated the median, which was significantly lower at **2.60 tons**. The large difference between the mean and the median strongly suggests that the data is skewed. Specifically, this indicates that a number of countries with very high emission rates are pulling the average up, while the majority of countries have emissions below this average. To quantify the spread of the data, the standard deviation of the dataset was calculated, which was **7.15**. Such a high standard deviation confirms that there is a vast range of emission levels globally. The percentiles further clarified this distribution; the 75th percentile was **5.44 tons**, while the 95th percentile jumped to **14.18 tons**, underscoring the disproportionate impact of the highest-emitting nations.

The process was repeated for the primary completion rate. The statistical profile of this dataset was markedly different. Upon analyzing this dataset, the mean completion rate was found to be **88.5%**, and the median to be **91.3%**.

The closeness of these two values indicates a much more symmetrical distribution compared to the CO_2 data. This suggests that most countries have similarly high completion rates, without extreme outliers pulling the average down. An interesting finding was the 95th percentile, which stood at **107%.** At first, I thought this was strange or an error in my calculation. However, upon further evaluation, I found that a value over 100% is possible because the World Bank's calculation method can include students who are older than the typical primary school age but have completed their primary education in that year [3]. This detail reflects a positive trend, showing that educational systems are successfully accommodating students who may have started late. The result for both dataset is shown in table 1 below.

Table 1: Summary Statistics for CO_2 Emissions per Capita and primary completion rates in 2023



```
                               Summary Statistics for World Bank Indicators (2023)

Table 1: CO₂ Emissions per Capita                                       Table 2: Primary Completion Rate
+--------------------+---------------------------+                     +--------------------+--------------------------+
|          Statistic |  CO₂ Emissions per Capita |                     |          Statistic |  Primary Completion Rate |
+====================+===========================+                     +====================+==========================+
|               Mean |                     4.459 |                     |               Mean |                   88.501 |
+--------------------+---------------------------+                     +--------------------+--------------------------+
|             Median |                     2.602 |                     |             Median |                   91.332 |
+--------------------+---------------------------+                     +--------------------+--------------------------+
| Standard Deviation |                     7.152 |                     | Standard Deviation |                   14.64  |
+--------------------+---------------------------+                     +--------------------+--------------------------+
|     5th Percentile |                     0.09  |                     |     5th Percentile |                   60.346 |
+--------------------+---------------------------+                     +--------------------+--------------------------+
|    25th Percentile |                     0.725 |                     |    25th Percentile |                   81.327 |
+--------------------+---------------------------+                     +--------------------+--------------------------+
|    75th Percentile |                     5.443 |                     |    75th Percentile |                   96.741 |
+--------------------+---------------------------+                     +--------------------+--------------------------+
|    95th Percentile |                    14.182 |                     |    95th Percentile |                  106.983 |
+--------------------+---------------------------+                     +--------------------+--------------------------+
```

In conclusion, my analysis of these two datasets revealed two contrasting global narratives. The CO_2 emissions data shows significant inequality and a widely dispersed distribution, whereas the primary completion rate data indicates a more consistent and positive global performance.

QUESTION 3: RELATIONSHIP BETWEEN GDP PER CAPITA AND PREVALENCE OF UNDERWEIGHT CHILDREN UNDER FIVE

This task required an exploration of how a country's economic status, measured by GDP per capita in current US dollars, relates to the prevalence of underweight children under five years old. GDP per capita is a common indicator of economic well-being because it reflects the average income per person in a country. The prevalence of underweight children is a key health and nutrition indicator, often linked to poverty, food insecurity, and limited access to healthcare. Understanding this relationship helps us see how economic growth can influence child health outcomes.

Before analyzing the data, my expectation is that as countries become wealthier, the number of underweight children should decline. This is because I believe higher income levels should mean better access to nutritious food, healthcare services and improved living conditions. However, the real world is complex, and we need to check if the data supports this hypothesis. To do that, I will be analyzing the world bank data showing how GDP per capita relates to underweight prevalence, then we can confirm if my hypothesis is correct or not.

3.1 Data and Methodology

The data comes from the World Bank's World Development Indicators [4] [5]. Two main indicators were used are GDP per capita (current US$), which measures the average income per person and Prevalence of underweight, weight-for-age (% of children under 5) which measures the percentage of children under five who are underweight for their age. I combined these indicators for all available countries and years. Additional metadata from the World Bank was used to classify countries by geographical region and income level. Next, Scatter plots were created to visualize the relationship between GDP per capita and underweight prevalence. GDP per capita was plotted on both a linear and a logarithmic scale because income differences between countries are very large, and when plotted on a linear scale, I found that it wasn't easy to see patterns, but when I switched to a log scale of GDP, it was clear that this scale makes patterns easier to see, as shown in the image below where I compare linear and log scale visualization. To do this, I used Python alongside some of its most commonly used libraries for data manipulations, such as Pandas for data wrangling, Matplotlib for plotting, and Seaborn for data visualizations. Previously I had used the API to pull the data but this time, I downloaded both the required datasets from the World Bank databank manually.

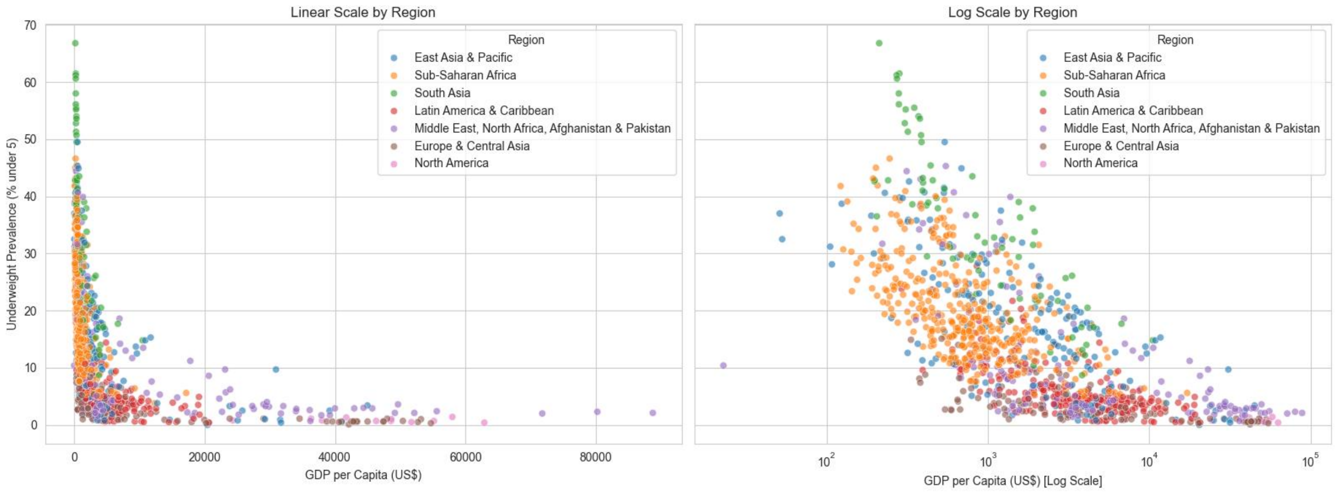


Fig 2.0: Linear versus Log scale of GDP against Underweight in children

I carried out exploratory data analysis using pandas to see some summary statistic and get a feel of the data. I used pd.describe(), pd.shape, pd.info(), pd.head(), and other functions to quickly see what the dataset looked like. Pandas allowed me to merge both datasets and filter out missing values efficiently. Then I merged the metadata with the already merged dataset to create the final required dataset for analysis “merged_df”. Upon reviewing the GDP dataset before merging, I could see the date in GDP per capita started from 1960 but Underweight prevalence started from 1983 so I had to remove the empty rows and columns when I was merging the dataset so I had a combined dataset that starts from 1983 which I then used for analysis.

Seaborn was particularly useful for adding color coding by region and income level, while matplotlib gave me control over plot details like axis labels and legends. Using these libraries together made the process efficient and I came up with different plots for GDP per Capita versus Prevalence of underweight children, GDP per Capita versus Geographical region, and GDP per Capita versus Income level, all of which I have shown in the charts below.

3.2 Results and Discussion

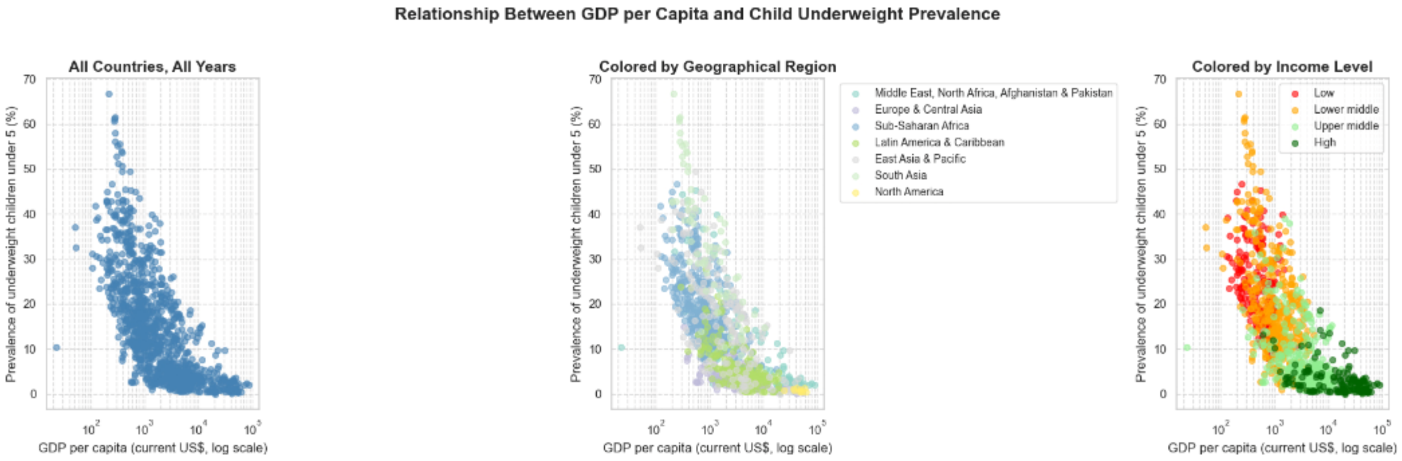


Fig 3.0 Combined graph of GDP per Capita versus Underweight prevalence

The first scatter plot in figure 4 below shows all countries across all available years. The pattern is clear that as GDP per capita increases for each country, the prevalence of underweight children decreases. This aligns with my initial hypothesis that the richer a country is, the better the average weight of the children from that country. However, despite this negative relationship being very strong, it is not perfectly linear. At very low income levels, the prevalence of underweight children seems to varies significantly which could mean that beyond just the GDP of a country, other factors other than income, such as governance, conflict, or cultural practices, also play a role. At high income levels, underweight prevalence is close to zero which is expected because wealthier countries have better nutrition, diet, food availability, lower hunger, and healthcare systems.

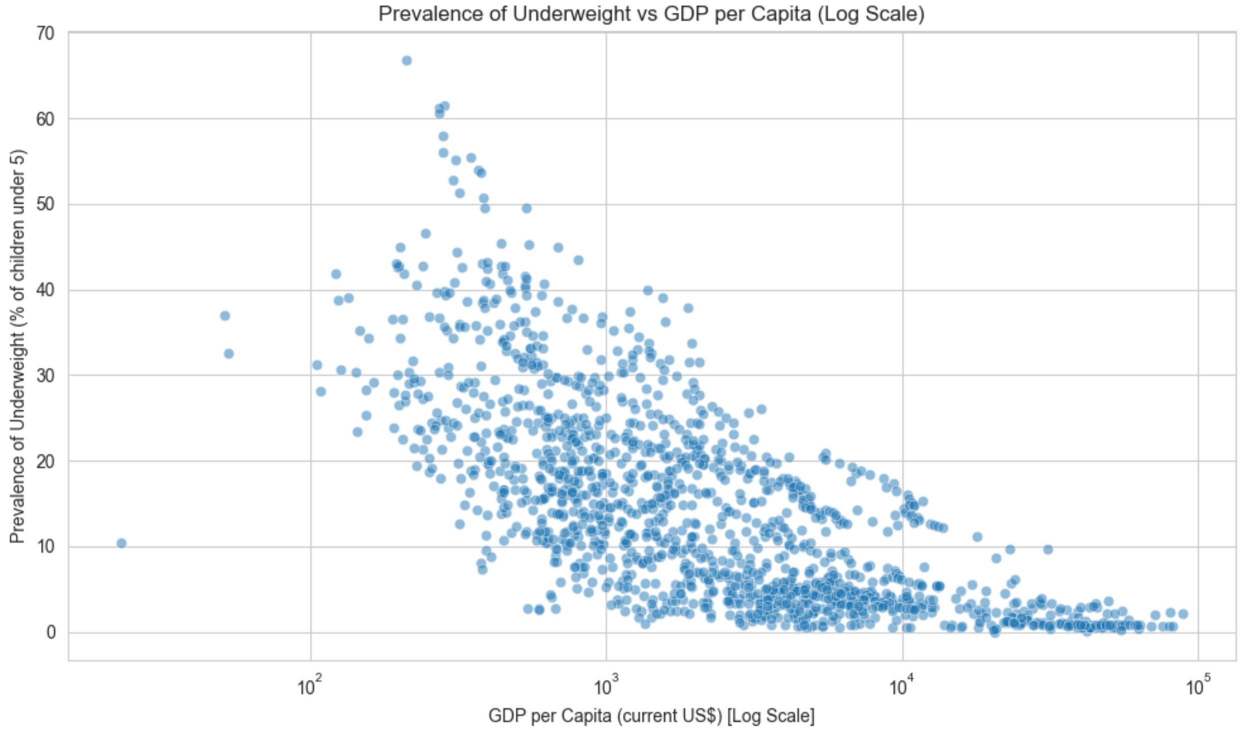


Fig 4.0: GDP Per capital vs Underweight children over the years (1983 – 2024)

The second plot in figure 5, colors the data points by geographical region. Here, we see major distinct clusters. South Asia and Sub-Saharan Africa dominate the low-income, high-underweight zone. The reason for this might be because these regions face persistent challenges such as poverty, food insecurity, illiteracy, environmental, and limited healthcare infrastructure. In contrast, regions like Europe and Central Asia, North America, and Latin America are clustered near the bottom right, where GDP per capita is really high and underweight prevalence is very low. Again, this is because these regions are usually the richest in terms of infrastructure, education and healthcare. This visualization highlights regional disparities and shows that economic development alone does not fully explain health outcomes as regional policies and social and economic factors matter too.

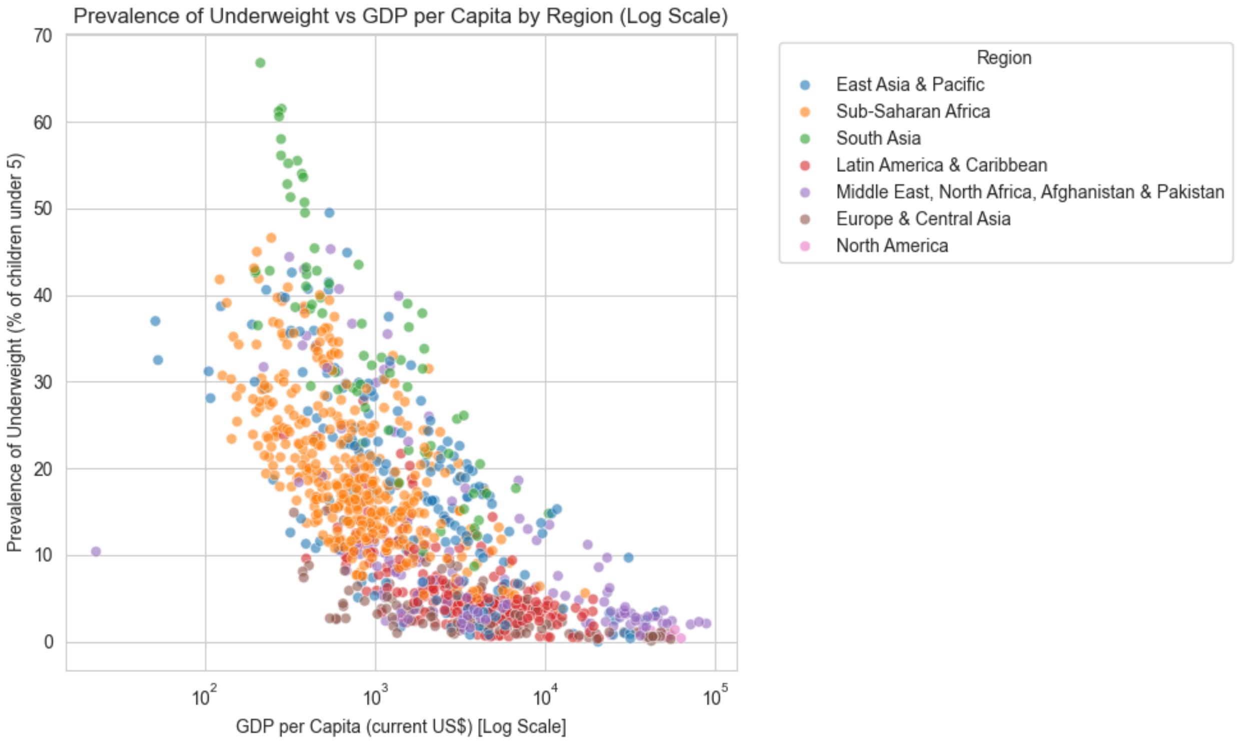


Fig 5.0: GDP per capita colored by regions.

The third plot in figure 6 groups countries by World Bank income classifications, we can see the low income, lower-middle income, upper-middle income, and high income regions color coded for easy separation. This plot again validates my earlier hypothesis and shows that there is a strong disparity in income from region to region. Low-income countries clearly have the highest prevalence of underweight children globally, while high-income countries clustered in the bottom right of the image, have almost none.

Lower-middle and upper-middle income countries fall somewhere in between, and we can also see that as income rises, the number of underweight children decreases. This confirms the expected pattern and emphasizes the role of income in reducing child malnutrition.

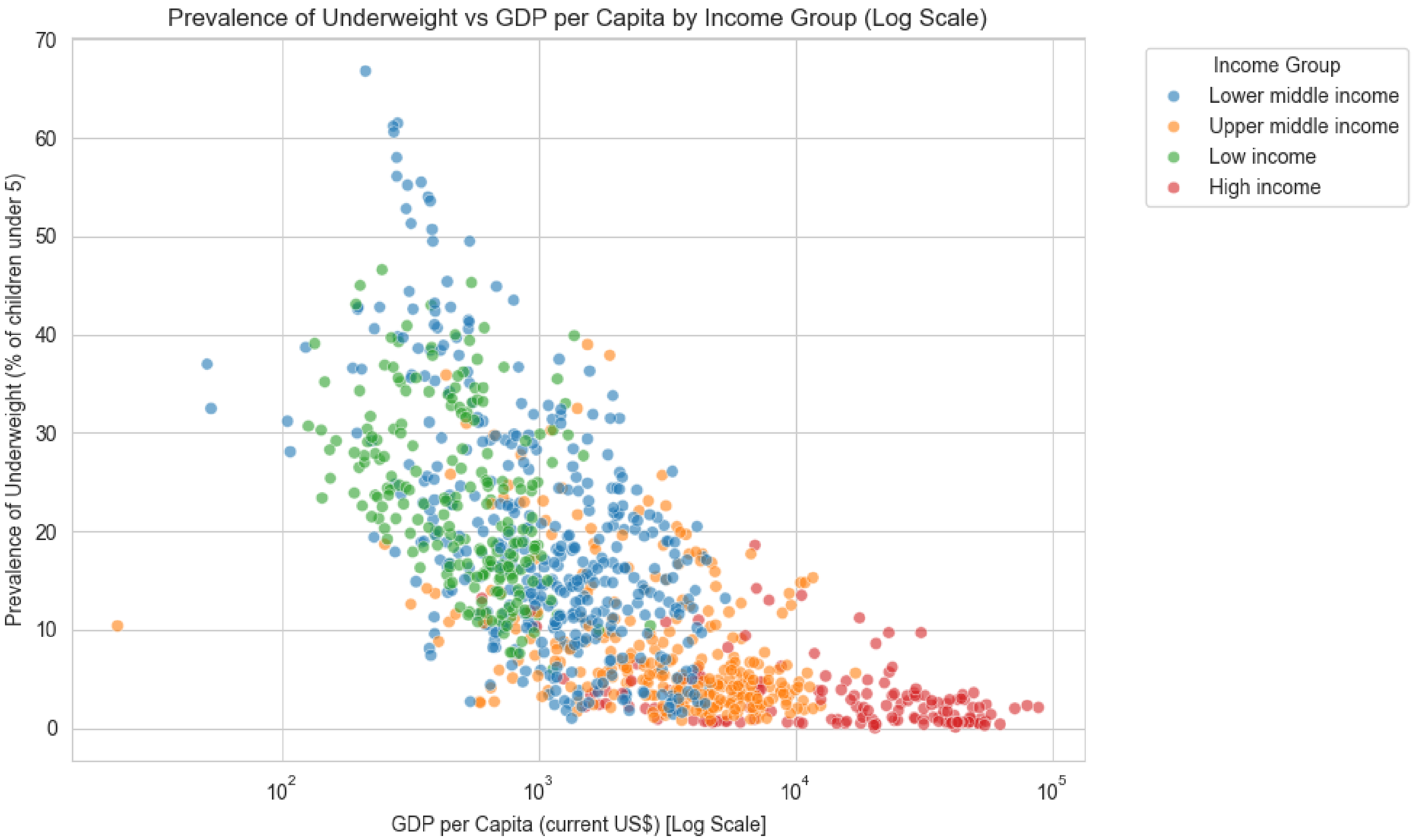
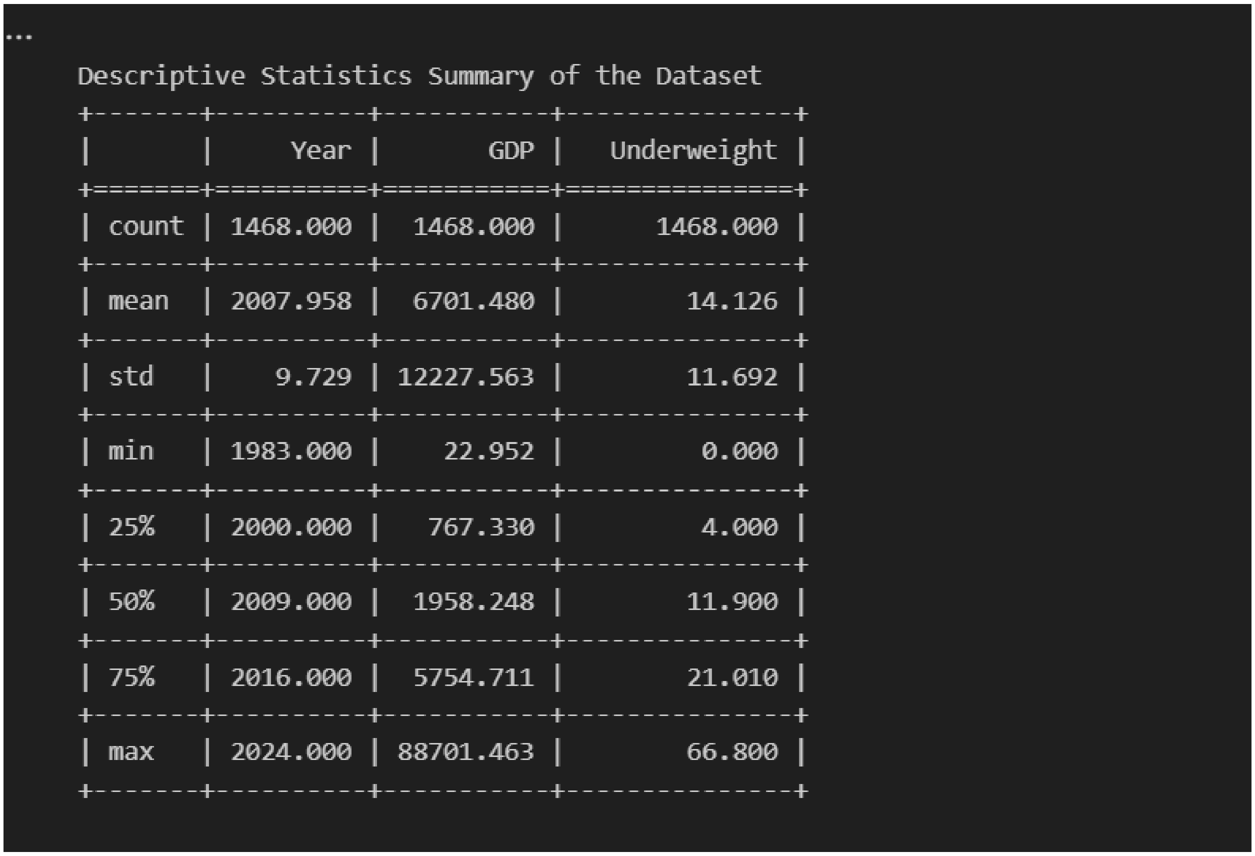


Fig 6.0: GDP per capita vs Underweight children colored by income level

The results strongly support the hypothesis that higher income levels are associated with better child nutrition outcomes. However, the relationship is not perfect. Some countries with similar income levels have very different underweight rates, which suggests that other factors, such as inequality, education, health systems, and cultural practices also matter. Additionally, the underweight data is based on household surveys, which are not always conducted every year, so some variation may reflect data gaps rather than real changes. Finally, GDP per capita in current

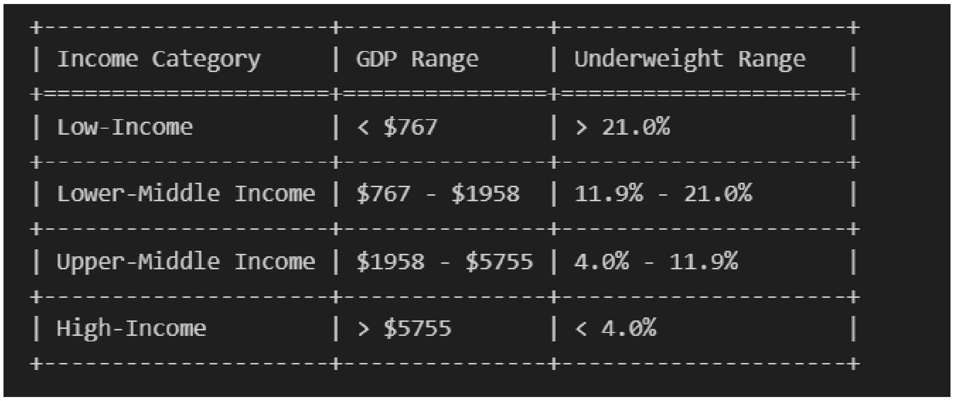
US dollars can be influenced by exchange rates and inflation, so it is not a perfect measure of living standards.

Table 2: Summary stats for GDP Per Capital and Underweight Prevalence (% of Children <5 years)



```
...
    Descriptive Statistics Summary of the Dataset
    +-------+----------+-----------+---------------+
    |       |     Year |       GDP |   Underweight |
    +=======+==========+===========+===============+
    | count | 1468.000 |  1468.000 |      1468.000 |
    +-------+----------+-----------+---------------+
    | mean  | 2007.958 |  6701.480 |        14.126 |
    +-------+----------+-----------+---------------+
    | std   |    9.729 | 12227.563 |        11.692 |
    +-------+----------+-----------+---------------+
    | min   | 1983.000 |    22.952 |         0.000 |
    +-------+----------+-----------+---------------+
    | 25%   | 2000.000 |   767.330 |         4.000 |
    +-------+----------+-----------+---------------+
    | 50%   | 2009.000 |  1958.248 |        11.900 |
    +-------+----------+-----------+---------------+
    | 75%   | 2016.000 |  5754.711 |        21.010 |
    +-------+----------+-----------+---------------+
    | max   | 2024.000 | 88701.463 |        66.800 |
    +-------+----------+-----------+---------------+
```

Table 3: Income Category Analysis Based on Quartiles



```
+---------------------+---------------+---------------------+
| Income Category     | GDP Range     | Underweight Range   |
+=====================+===============+=====================+
| Low-Income          | < $767        | > 21.0%             |
+---------------------+---------------+---------------------+
| Lower-Middle Income | $767 - $1958  | 11.9% - 21.0%       |
+---------------------+---------------+---------------------+
| Upper-Middle Income | $1958 - $5755 | 4.0% - 11.9%        |
+---------------------+---------------+---------------------+
| High-Income         | > $5755       | < 4.0%              |
+---------------------+---------------+---------------------+
```

3.3 Conclusion

The analysis shows a strong negative relationship between GDP per capita and the prevalence of underweight children under five. Looking at the results of our analysis in details, we see a very clear relationship between GDP and Underweight prevalence in children under 5years old). Evidently, how much money a country has is deeply connected to how healthy its youngest children are. We can group countries into four simple levels based on how much money they have per person. The poorest countries which we consider the low-income countries, where people live on less than a few hundred dollars a year, with a GDP per capita under $767, face severe malnutrition, with over 21% of children under five being underweight. More than one in every five children is underweight and doesn't have enough nutritious food to grow properly. As economies transition into the lower-middle income bracket ($767 - $1,958), we start to see a glimmer of hope. Nutritional outcomes begin to improve, here the number of underweight children begins to drop, but it's still high, affecting more than one in ten kids, ranging from 11.9% to 21.0%.

The most significant improvements are observed in upper-middle income countries ($1,958 - $5,755), where rapid development is coupled with a substantial decline in underweight rates to between 4.2% and 11.9%. With more money, countries can build better healthcare systems and families can afford healthier food, leading to a major drop in child underweight problems. Finally, the wealthiest, high-income nations (GDP > $5,755) have mostly overcome this challenge, consistently maintaining underweight prevalence below 4.0%. This shows us a massive gap between the richest and poorest parts of our world. The difference in wealth is large, and so is the difference in children's health.

QUESTION 4: COMPARATIVE ANALYSIS OF SPY AND TLT ETFS (2013–2015)

This task required an analysis of the behavior of two major exchange-traded funds (ETFs): SPY, which tracks the S&P 500 index, and TLT, which tracks long-term U.S. Treasury bonds. The period under review is from December 31, 2013, to August 31, 2015. The goal is to understand how these two financial instruments performed during this period, using Python to analyze and visualize their price movements and daily returns. I started by importing the necessary Python libraries. Pandas was used to handle and clean the data, numpy to help with calculations, and matplotlib to create graphs. First step involved downloading and loading the SPY and TLT datasets from the links provided to us. These files contain daily trading data, including prices and volumes. Specifically, the focus was on the "adjusted closing price" because it gives the most accurate picture of an asset's value. This price accounts for dividends and other adjustments, making it ideal for comparing performance over time.

Once the data was loaded, exploratory data analysis was used to quickly see what kind of data it is and get a quick overview of what I'm working with. I saw that there is a date column, so this date column had to be converted into a format that Python understands as actual calendar dates. This step is important because it allows us to filter the data by time and plot it correctly. Then the data was filtered to include only the dates between December 31, 2013, and August 31, 2015 as instructed. By narrowing the time frame, we could better observe trends and patterns without being distracted by data from other periods.

After filtering, I extracted the adjusted closing prices and created a new column called "Normalized." This column was then used to compare SPY and TLT. This normalization makes it easier to see which one performed better relative to the other. They were then plotted as normalized prices on the graph shown below using matplotlib.

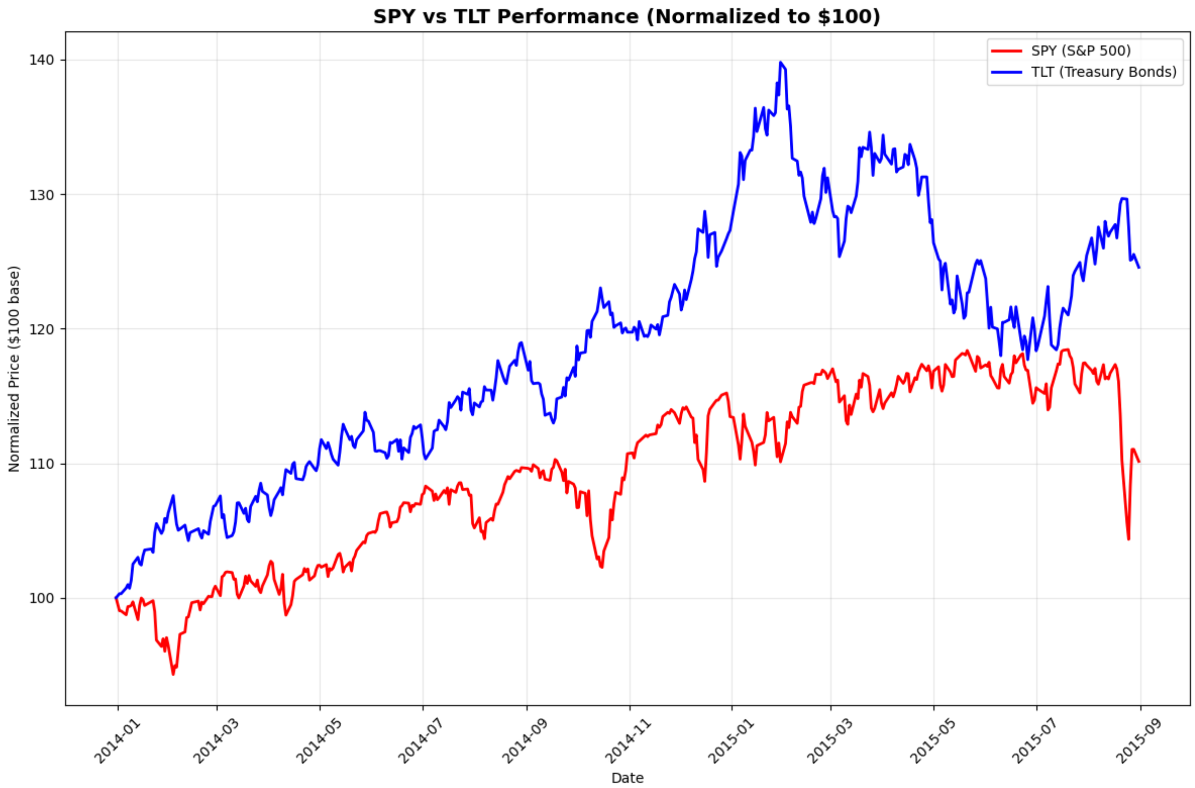


Fig 7: SPY vs TLT Starting at $100 between 31-12-2013 to 31-12-2015

This visual comparison in figure 7 above shows that SPY generally trended upward, reflecting the growth of the U.S. stock market. However, it also experienced some sharp dips, especially in mid-2015. TLT, on the other hand, showed a steadier climb. As a bond ETF, TLT is influenced by interest rates and investor demand for safer assets. Its smoother trajectory suggests that it was less affected by market volatility compared to SPY. After visualizing the price movements, the daily returns was calculated. A daily return tells us how much the price changed from one day to the next, expressed as a percentage. The formula we used is **r(t) = [(p(t) / p(t−1)) − 1] × 100**

Where:

r(t) is the daily return expressed as a percentage.

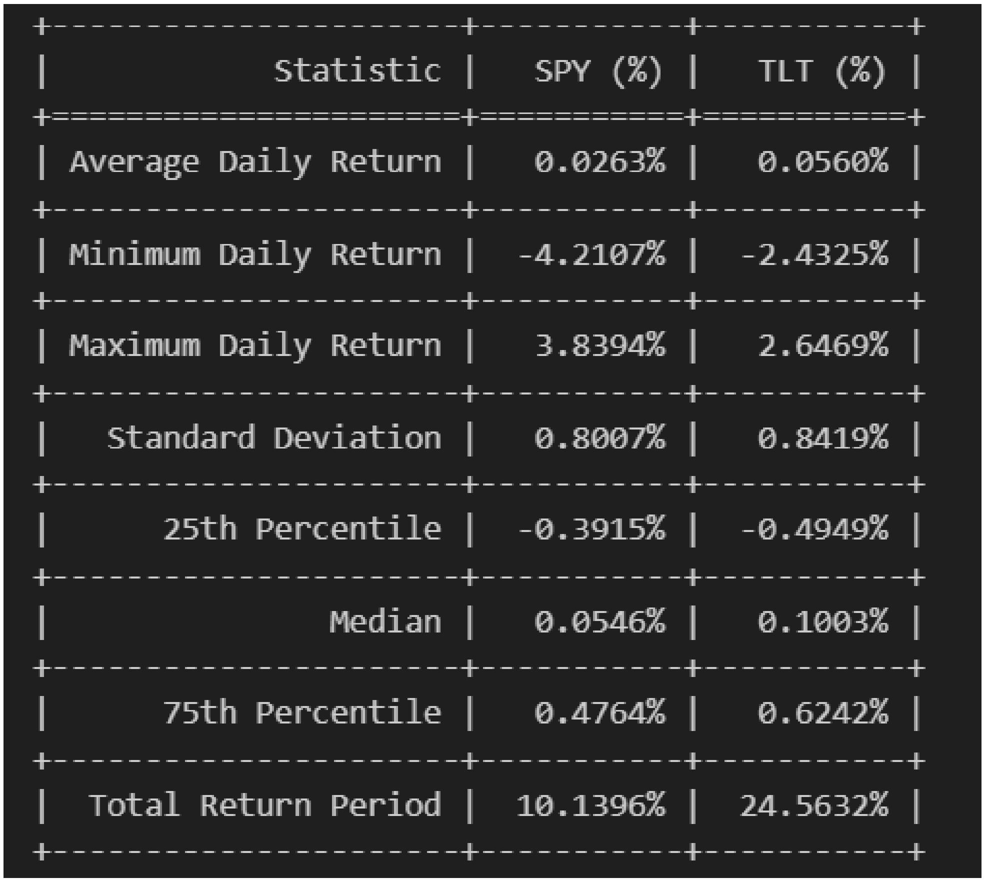
p(t) is the adjusted closing price on day *(t)*.

p(t−1) is the adjusted closing price on the previous trading day.

This gives us a clear view of how volatile each ETF was on a day-to-day basis.

The summary statistics for each ETF's daily returns was then computed, these include the average return, the minimum (worst) return, and the maximum (best) returns.

Table 4: ETF (SPY vs TLT Summary Statistics



```
+----------------------+----------+----------+
|            Statistic |  SPY (%) |  TLT (%) |
+======================+==========+==========+
| Average Daily Return |  0.0263% |  0.0560% |
+----------------------+----------+----------+
| Minimum Daily Return | -4.2107% | -2.4325% |
+----------------------+----------+----------+
| Maximum Daily Return |  3.8394% |  2.6469% |
+----------------------+----------+----------+
|   Standard Deviation |  0.8007% |  0.8419% |
+----------------------+----------+----------+
|      25th Percentile | -0.3915% | -0.4949% |
+----------------------+----------+----------+
|               Median |  0.0546% |  0.1003% |
+----------------------+----------+----------+
|      75th Percentile |  0.4764% |  0.6242% |
+----------------------+----------+----------+
|  Total Return Period | 10.1396% | 24.5632% |
+----------------------+----------+----------+
```

Results:

For SPY, the average daily return was 0.0263%, the worst day saw a drop of 4.2107%, and the best day saw a gain of 3.8394% while for TLT, the average daily return was 0.0560%, the worst day saw a drop of 2.4325%, and the best day saw a gain of 2.6469%. These results show that between December 31, 2013, and December 31, 2015, TLT had a slightly higher average daily return than SPY, but SPY experienced more extreme ups and downs. This probably means that stocks (SPY) are riskier that bonds, I then did some search and found that generally, stocks offer higher potential returns but come with greater risk, while bonds (TLT) provide more stability, this aligns with what the result of the analysis is showing.

QUESTION 5: FERTILITY RATE VS GDP PER CAPITA PPP (2023)

For this part of the assignment, our objective was to explore the relationship between a country's economic prosperity and its demographic trends. Specifically, we were to investigated the connection between the Fertility Rate (total births per woman) and the GDP per capita (adjusted for PPP) for the year 2023. To carry out this analysis, Python programming language was used along with several of its powerful data science libraries.

The process began with setting up the environment by importing the necessary libraries. Pandas was used for the core data loading and manipulation while Numpy was used for numerical operations, Matplotlib and Seaborn were used in combination for creating the visualization, and the pearsonr function from scipy.stats for a more detailed statistical analysis.

To prepare the data for exploration, the dataset was downloaded from the world bank data repository[6] [7], then I loaded the two datasets from the provided CSV files using the pandas.read_csv function. A crucial first step was using the skiprows=4 argument to bypass the metadata at the top of the files, ensuring the data was read into a DataFrame correctly. From there, only the required columns had to be extracted for us to be able to analyze for 2023, it was then renamed to to Fert_Rate and GDP_PPP for easier access. A key challenge I experienced was how to handle the inconsistent data types. To address this, I used the pd.to_numeric function with the errors='coerce' parameter, which robustly converted the data columns to numbers and automatically handled any non-numeric entries by turning them into missing values. Then the two datasets was merged into a single DataFrame using the 'Country Code' as the common key, which is a more reliable identifier than 'Country Name'. Finally, to remove the missing rows and columns, the .dropna() method was used, resulting in a clean and complete dataset for my analysis.

For the visualization, matplotlib was used to create a scatter plot as shown in figure 8 below. I deliberately set the x-axis (GDP per capita) to a logarithmic scale. I did this because the first plot created, somehow had the countries clustered weirdly, later it was found that this happened because the dataset was so large and the scale had to be logarithmic, like in question 3, so I implemented it [8]. Having implemented the logarithmic scale, it made the chart more easily comprehensible and one can now really see patterns start to emerge. This made it easier to effectively visualize the relationship across the vast range of national incomes without having the lower-income countries clustered together. To better illustrate the trend, a trend line was calculated and plotted using Matplotlib and Numpy. This was a multi-step process where I first took the natural logarithm of the GDP data using numpy.log, then used np.log and numpy.polyfit to calculate the slope and intercept of the best-fit line for this log-transformed data. Finally, the trend linen was plotted over the scatter plot, which clearly highlighted the curved nature of the relationship as shown in figure 8 below. From the chart in the figure below, we can see there is generally a negative correlation between GDP and Fertility rate where an increase in GDP typically leads to lower fertility rates.

Further checks shows some outliers such as Chad, Somalia, Mali, Congo, all of whom have low GDP and low fertility rates. Likewise we can see that Israel also is an outlier because they defy the trend with a high GDP and higher than the usual fertility rate for their GDP.

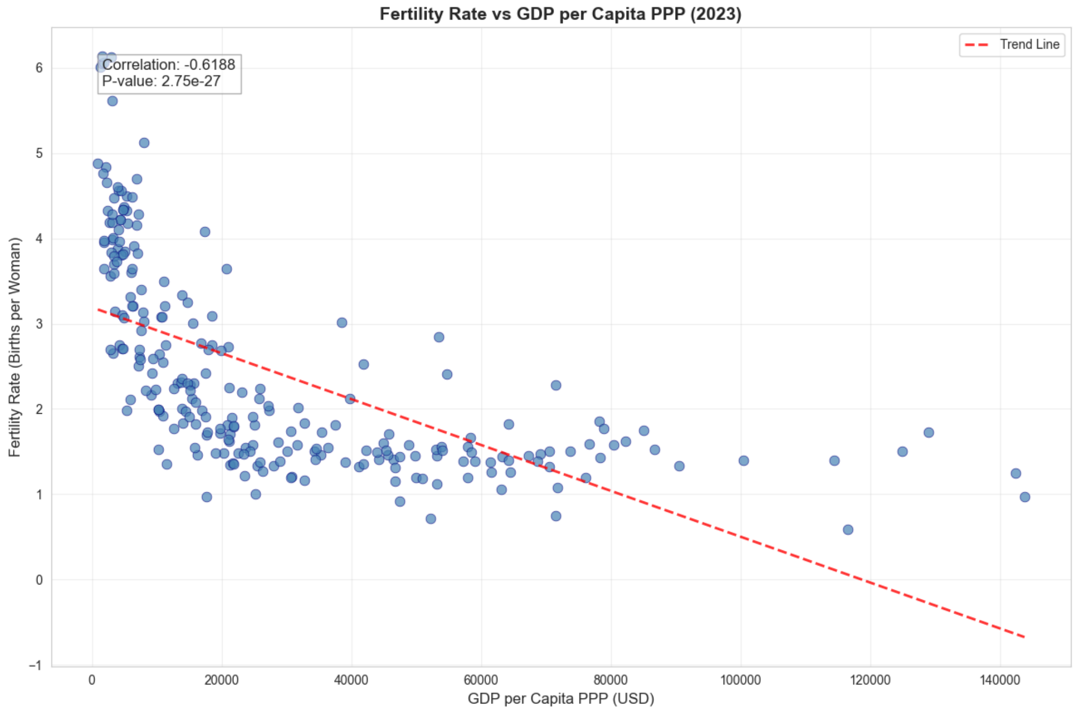


Fig 8.0: GDP Per Capita PPP vs Fertility rate per woman

To numerically quantify this relationship, two correlation coefficients were calculated. The standard linear Pearson correlation, calculated using the pandas .corr() method, was **-0.62**. This value indicates a moderately strong negative relationship. However, to better match the visual evidence from the plot, it made sense to also use the scipy.stats.pearsonr function to calculate the correlation between the fertility rate and the logarithm of GDP, and the result is shown in the figure 8.1 below. This log-linear correlation yielded a much stronger value of **-0.84**. This result confirms that the relationship is very strong, but non-linear. The p-value for this correlation was an extremely small **1.40e-65**, which provides overwhelming statistical evidence that this strong negative relationship is not due to random chance. In conclusion, this analysis, conducted using Python's data handling and visualization tools, demonstrates a very strong and statistically significant negative correlation between a nation's economic output and its fertility rate which implies that as GDP increases, fertility rates decreases.

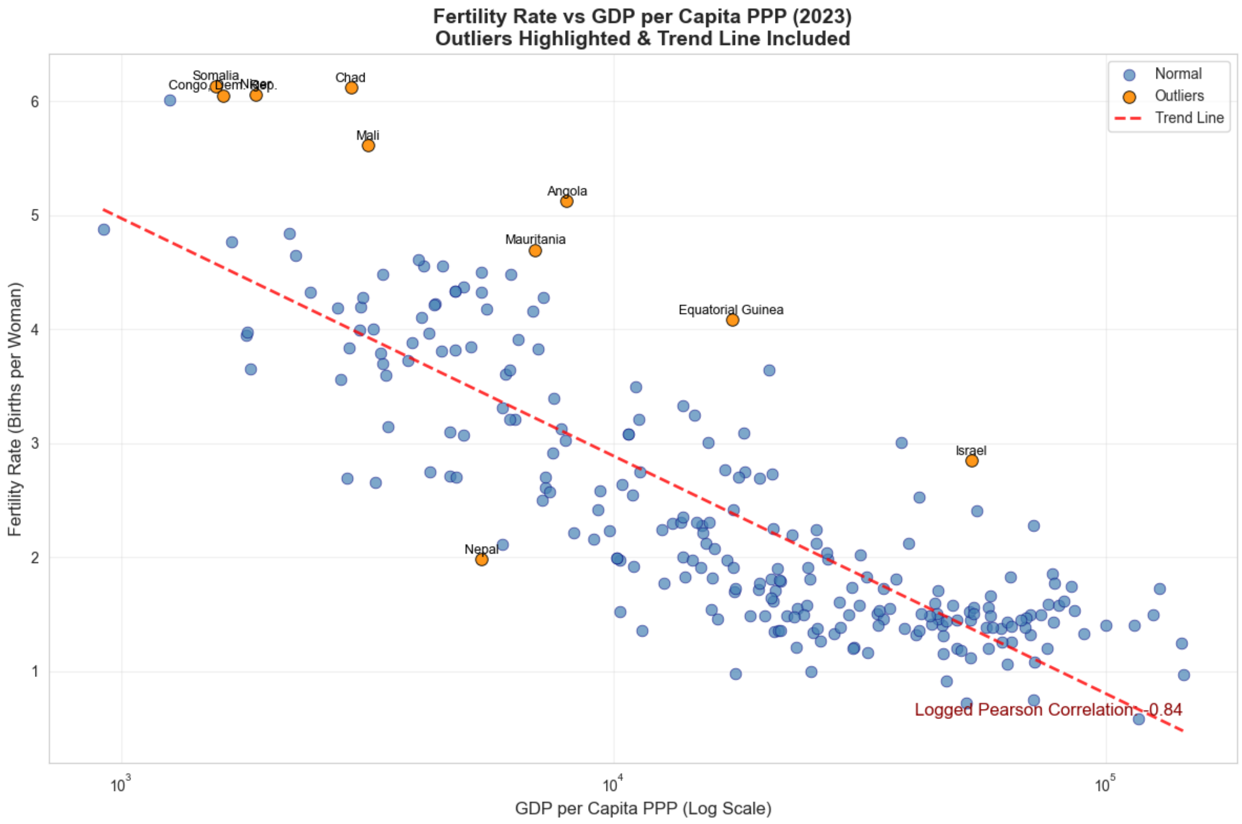


Fig 8.1: Logged Fertility vs GDP

QUESTION 6: RELATIONSHIP BETWEEN THE HAPPY PLANET INDEX AND THE CORRUPTION PERCEPTIONS INDEX (2016)

This part of the assignment was for us to explore the relationship between two global indicators: the Happy Planet Index (HPI) and the Corruption Perceptions Index (CPI), both from the year 2016. The HPI measures how well countries deliver long, happy lives to their citizens while considering environmental sustainability [9]. CPI ranks countries based on how corrupt their public sectors are perceived to be. These two indices reflect very different aspects of national performance, and the goal of this analysis was to understand whether there is any meaningful relationship between them.

To begin, the official website for these two indicators had to be accessed and from there, he datasets were downloaded, specifically for the year 2016. The Happy Planet Index data was downloaded from happyplanetindex.org, and the Corruption Perceptions Index was obtained from transparency.org. Each dataset came in the form of an Excel spreadsheet. Upon opening the files, I explored their structure to identify the relevant sheets. The HPI file contained several sheets, but the one titled "Rank order" was the most useful because it listed countries along with their HPI ranks and other indicators. The CPI file also had multiple sheets, and the one named "CPI2016_FINAL_16Jan" was selected because it included country names and their CPI ranks.

To analyze the data, Python was again used, along with some key libraries. The first was pandas, which was imported as pnd. This library is essential for reading and manipulating tabular data, especially Excel files. Then matplotlib.pyplot was also imported, and this was used for creating the scatter plot that visualizes the relationship between the two indices. Since the Excel files were in .xlsx format and contained multiple sheets, there was need to use the openpyxl engine to read them properly. For a beginner like me, working with Excel files in Python can be a bit confusing, especially when dealing with multiple sheets and metadata rows. I found that using the skiprows parameter helped remove unnecessary header information in the HPI file. Tutorials from Real Python [10] and GeeksforGeeks [11] were particularly helpful in understanding how to use pandas.read_excel() effectively.

After loading the data, I selected only the columns that was needed: country names and ranks. These two were then merged based on country names to find the countries that appeared in both. This gave a clean dataset with both HPI and CPI ranks for each country. With this merged data, A scatter plot shown as figure 9 below was created to visualize the relationship, where the x-axis represented CPI rank and the y-axis represented HPI rank. Each dot on the graph represented a country, and each dot was labeled using the first three letters of the country's name. This made the graph readable and allowed for easy identification of individual countries. The scatter plot revealed that there is no simple or direct relationship between the two indices. Some countries that ranked high in happiness and sustainability had very low CPI ranks, meaning they were perceived as corrupt. Others, like Luxembourg, had excellent CPI ranks but poor HPI scores. To explore this further, it made sense to calculate the absolute difference between each country's HPI and CPI rank. This helped identify the countries with the largest mismatch between the two rankings.

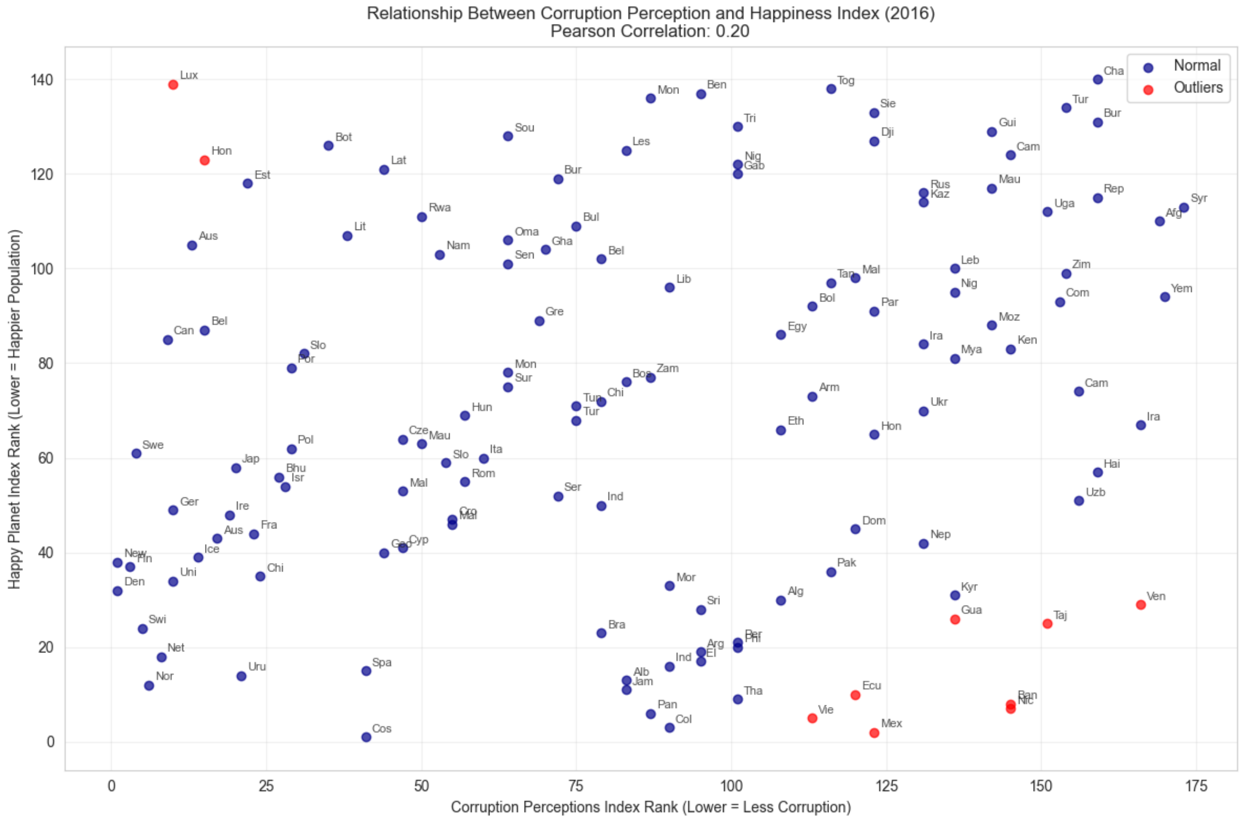
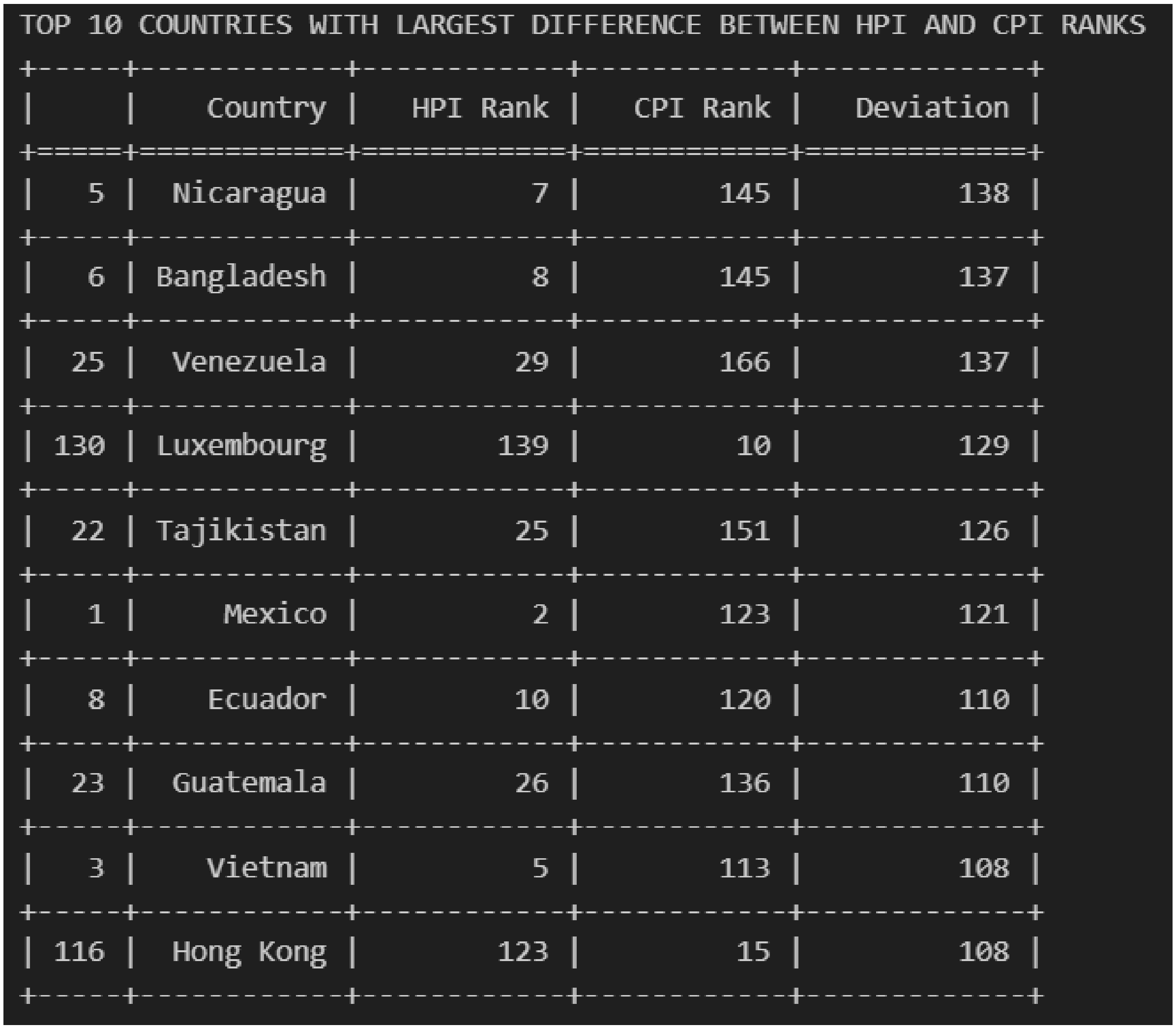


Fig 9.0: CPI index vs HPI index (2016)

As shown in Table 5 below, the top 5 most unusual countries were Nicaragua, Venezuela, Bangladesh, Luxembourg, and Tajikistan. Nicaragua ranked 7th in HPI but 145th in CPI, suggesting that despite being perceived as corrupt, it scores well in terms of delivering happy and sustainable lives. Bangladesh showed a similar pattern. Luxembourg, on the other hand, ranked 10th in CPI, indicating low corruption, but 139th in HPI, which was surprisingly low. This could be due to factors like high ecological footprint or inequality, which are considered in the HPI but not in the CPI. Tajikistan also showed a significant mismatch, ranking 25th in HPI and 151st in CPI.

Table 5: Top 10 countries with Largest Deviation Between HPI and CPI Ranks



```
TOP 10 COUNTRIES WITH LARGEST DIFFERENCE BETWEEN HPI AND CPI RANKS
+-----+------------+------------+------------+-------------+
|     |    Country |   HPI Rank |   CPI Rank |   Deviation |
+=====+============+============+============+=============+
|   5 |  Nicaragua |          7 |        145 |         138 |
+-----+------------+------------+------------+-------------+
|   6 | Bangladesh |          8 |        145 |         137 |
+-----+------------+------------+------------+-------------+
|  25 |  Venezuela |         29 |        166 |         137 |
+-----+------------+------------+------------+-------------+
| 130 | Luxembourg |        139 |         10 |         129 |
+-----+------------+------------+------------+-------------+
|  22 | Tajikistan |         25 |        151 |         126 |
+-----+------------+------------+------------+-------------+
|   1 |     Mexico |          2 |        123 |         121 |
+-----+------------+------------+------------+-------------+
|   8 |    Ecuador |         10 |        120 |         110 |
+-----+------------+------------+------------+-------------+
|  23 |  Guatemala |         26 |        136 |         110 |
+-----+------------+------------+------------+-------------+
|   3 |    Vietnam |          5 |        113 |         108 |
+-----+------------+------------+------------+-------------+
| 116 |  Hong Kong |        123 |         15 |         108 |
+-----+------------+------------+------------+-------------+
```

These results suggest that happiness and corruption are not always linked in the way we might expect. A country can be perceived as corrupt but still provide a good quality of life and environmental sustainability.

On the other hand, a country with clean governance might struggle with wellbeing or sustainability. These mismatches highlight the importance of looking at multiple indicators when evaluating a country's performance. It also shows that global indices are useful tools, but they must be interpreted together to understand the full picture.

In conclusion, the relationship between HPI and CPI is complex and not easily defined by a single trend. The scatter plot and rank deviations highlight that some countries defy expectations, and understanding why requires looking beyond the data to consider cultural, economic, and environmental factors. This kind of analysis is a reminder that progress is multidimensional, and we need to use a variety of lenses to truly understand what makes a country thrive.

REFERENCES

[1] "Carbon dioxide (CO2) emissions (total) excluding LULUCF (Mt CO2e)," World Bank Open Data. Accessed: Sept. 20, 2025. [Online]. Available: https://data.worldbank.org/indicator/EN.GHG.CO2.MT.CE.AR5

[2] "Primary completion rate, total (% of relevant age group)," World Bank Open Data. Accessed: Sept. 20, 2025. [Online]. Available: https://data.worldbank.org/indicator/SE.PRM.CMPT.ZS

[3] "How can school enrollment and completion indicators be over 100 percent? – World Bank Data Help Desk." Accessed: Sept. 21, 2025. [Online]. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/1986157-how-can-school-enrollment-and-completion-indicator

[4] "GDP per capita (current US$)," World Bank Open Data. Accessed: Sept. 20, 2025. [Online]. Available: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

[5] "Prevalence of underweight, weight for age (% of children under 5)," World Bank Open Data. Accessed: Sept. 20, 2025. [Online]. Available: https://data.worldbank.org/indicator/SH.STA.MALN.ZS

[6] "Fertility rate, total (births per woman)," World Bank Open Data. Accessed: Sept. 20, 2025. [Online]. Available: https://data.worldbank.org/indicator/SP.DYN.TFRT.IN

[7] "GDP per capita, PPP (current international $)," World Bank Open Data. Accessed: Sept. 20, 2025. [Online]. Available: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD

[8] "Log scale — Matplotlib 3.10.6 documentation." Accessed: Sept. 20, 2025. [Online]. Available: https://matplotlib.org/stable/gallery/scales/log_demo.html

[9] "Homepage - Happy Planet Index." Accessed: Sept. 21, 2025. [Online]. Available: https://happyplanetindex.org/

[10] "pandas: How to Read and Write Files – Real Python." Accessed: Sept. 21, 2025. [Online]. Available: https://realpython.com/pandas-read-write-files/

[11] "Reading an excel file using Python - GeeksforGeeks." Accessed: Sept. 21, 2025. [Online]. Available: https://www.geeksforgeeks.org/python/reading-excel-file-using-python/