**DATA, INFERENCE**

**&**

**APPLIED MACHINE LEARNING**

**(COURSE 18-785)**

**ASSIGNMENT 2**

**Mark Iraguha**

**(miraguha)**

**Libraries Used:**

Matplotlib – a python plotting library used to create animated, interactive and static visualizations.[1]

Pandas – another Python library used that provides data structures and functions used to carry out data analysis.[2]

Numpy – a simple yet powerful data structure provided in python.[3]

Tabulate – a python library that tabulates data to an output[4].

**Introduction:**

This report details the completion of Assignment 2. Assignment 2 requests answers to 5 critical thinking and data analytical questions.

**Question 1 Report:**

**Methodology**

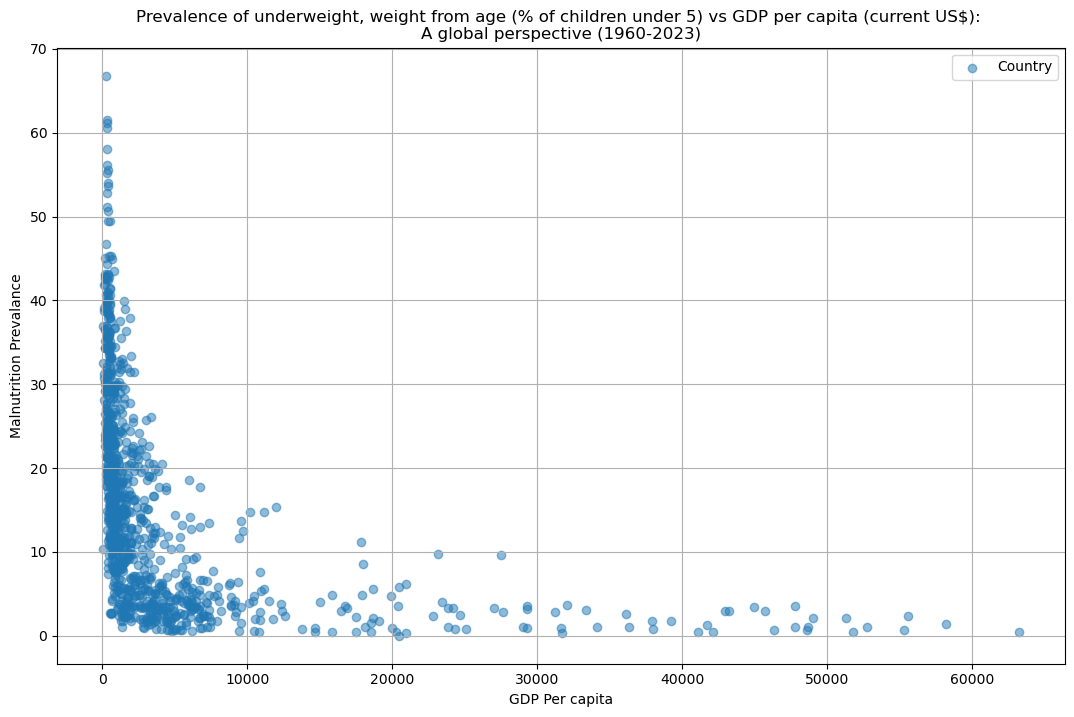
Making a scatter plot of malnutrition prevalence against GDP per capita.

Approach:

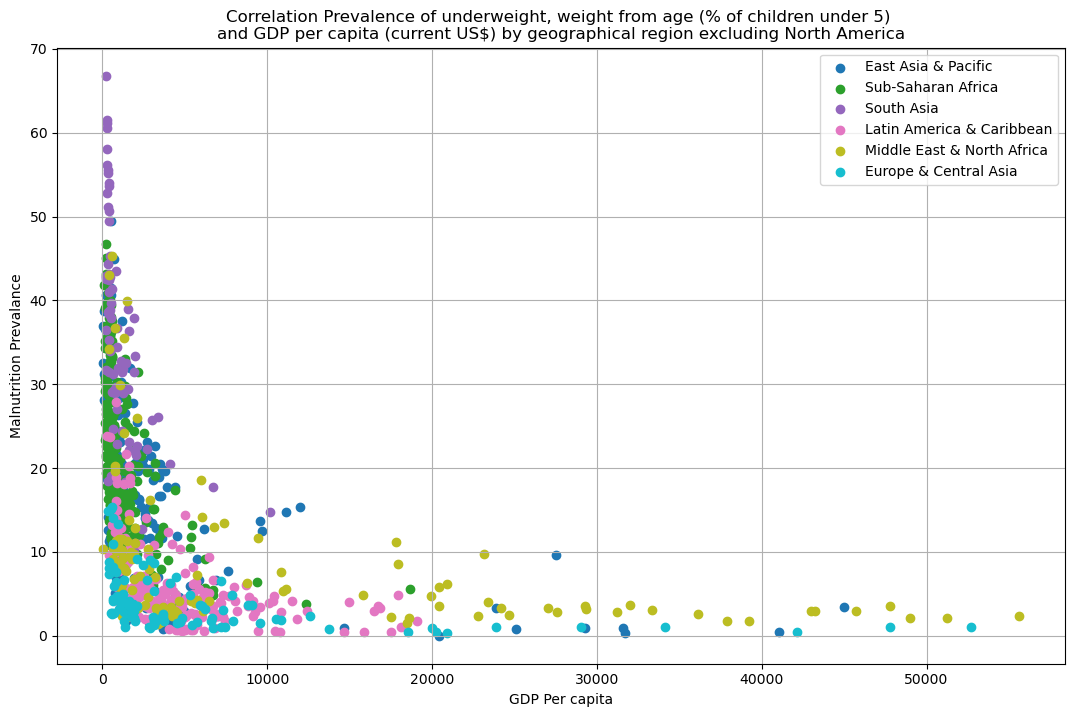
* Download data for “GDP per capita” and “Prevalence of underweight, weight for age”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as single column).
* Drop the NaN (Not a number) values.
* Configure and plot the graph.

**Results:**

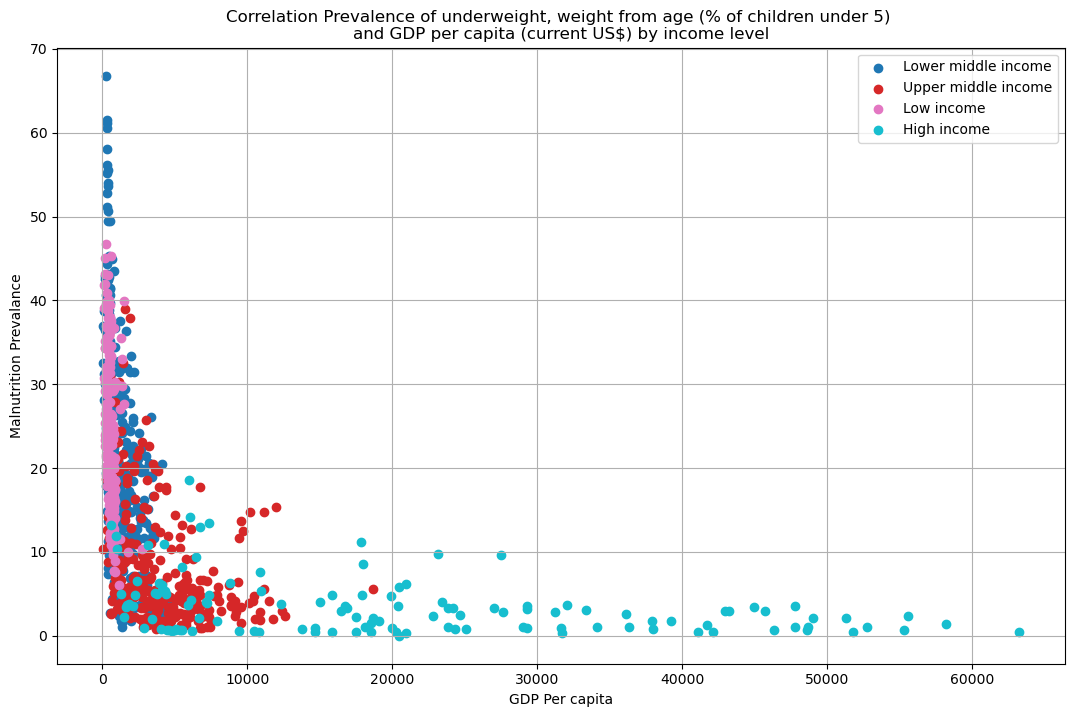
Three scatter plot graphs showing Malnutrition prevalence against GDP per capita.



*Scatter plot graph showing prevalence of underweight, weight from age (% of children under 5) vs GDP per capita (current US$) (1960-2023)*



*Scatter plot graph showing correlation of prevalence of underweight, weight from age (% of children under 5) and GDP per capita (current US$) by geographical region excluding North America*



*Scatter plot graph showing correlation of prevalence of underweight, weight from age (% of children under 5) and GDP per capita (current US$) by income level*

**Analysis and Insights**

Expected Relationship

* The kind of relationship we expect is a negative correlation. This is where we should observe that as GDP per capita increases, prevalence of malnutrition should decrease[5].

Malnutrition prevalence against GDP per capita analysis

* It’s observed that it’s indeed true that countries with a higher GDP per capita are associated with a lower Malnutrition prevalence. Hence forming a negative relationship. However, studies have shown that a higher GDP per capita might not be the sole reason for a low malnutrition prevalence as other contributing factors might play a larger part in lowering malnutrition in affected countries[5]. One of these factors include education[5].
* When a country falls into a lower middle-income group, it’s observed that the country potentially has a higher Malnutrition Prevalence. This is seen with countries in the South Asian region.
* Majority of the countries with a high-income level are seen with high levels of GDP per capita and low levels of Malnutrition Prevalence. This backs the idea that a high GDP per capita will positively affect the Malnutrition level of a country.

**Question 2 Report:**

**Methodology**

Making a line graph showing the maximum and minimum prices of Wheat, Crude Oil and Gold.

Approach:

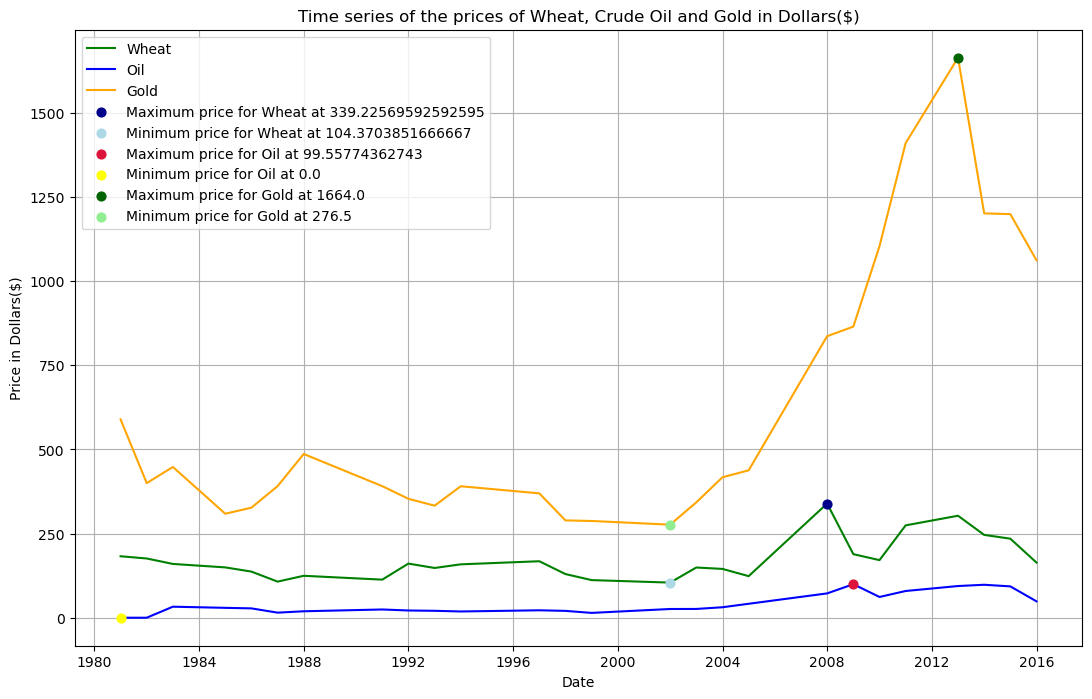
* Utilized Quandl API to download data for Wheat, Crude Oil and

Gold prices in dollars ($).

* Synchronized the time stamps.
* Formatted the data in the data frames.
* Set unique colors for each variable to be used in the graph.
* Configured and plotted graph while indicating the highest and lowest price for Wheat, Crude Oil and Gold.

**Results:**

Line plot graph showing the prices of Wheat, Crude Oil and Gold highlighting the lowest and highest prices.



*Line graph showing time series of the prices of Wheat, Crude Oil and Gold in dollars ($)*

**Analysis and Insights**

Dramatic Gold Price Increase

* The graph shows that once Gold reached a low price of $276.5 (lowest between a 1980 and 2016), it begun to steadily rise, and a sharp increase is observed around 2009. This rise would be attributed to the fact that investors at that time thought it would be the safest option to keep their money safe during the ongoing financial crisis[6].
* During the same time Gold prices were increasing, Oil and Wheat were facing the opposite behavior. The demand for Oil and Wheat sharply reduced and this was primarily driven by the ongoing financial crisis.

**Question 3 Report:**

**Methodology**

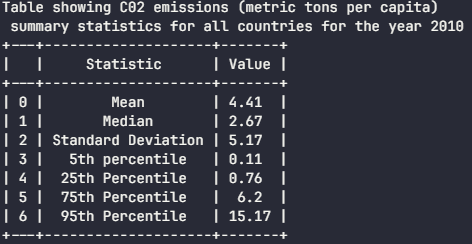
Using pandas library to calculate summary statistics[2]

Approach:

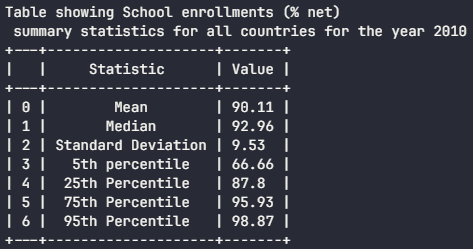
* Download data for “C02 emissions (metric tons per capita)” and “School enrolment, primary (% net)”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as a single column).
* Drop the NaN (Not a number) values.
* Calculate summary statistics.
* Display statistics in tabular format.

**Results:**

Finding Two tables showing summary statistics for C02 emissions and School enrollment for the year 2010.



*Figure showing C02 emissions (metric tons per capita) summary statistics for all countries for the year 2010*



*Figure showing School enrollments (% net) summary statistics for all countries for the year 2010*

**Analysis and Insights**

C02 emissions

* The low mean signifies countries have a low C02 emission.
* The median doesn’t stray so far from the mean signifying that outlier countries may not exist.
* The relatively high Standard Deviation is likely due to the differences among countries in areas such as energy sources, industrialization and population.
* The 5th, 25th and 75th percentile is relatively low while the 95th percentile signifies that many countries have high C02 emissions which could call for immediate intervention for the affected countries.

School Enrollments

* The high mean value signifies that many countries had a lot of school enrollments for the year 2010.
* The median value suggests that majority of the countries had a high number of enrollments, so a possibility of an outlier is low.
* The low relative standard deviation indicates that there isn’t a huge variation from one country to another in terms of number of school enrollments.
* Countries in the 5th percentile have relatively low school enrollments where as countries in higher percentiles (25th, 75th, 95th) signify that they are experiencing a high number of enrollments.

**Question 4 Report:**

**Methodology**

Utilizing numpy to produce cumulative distribution functions.

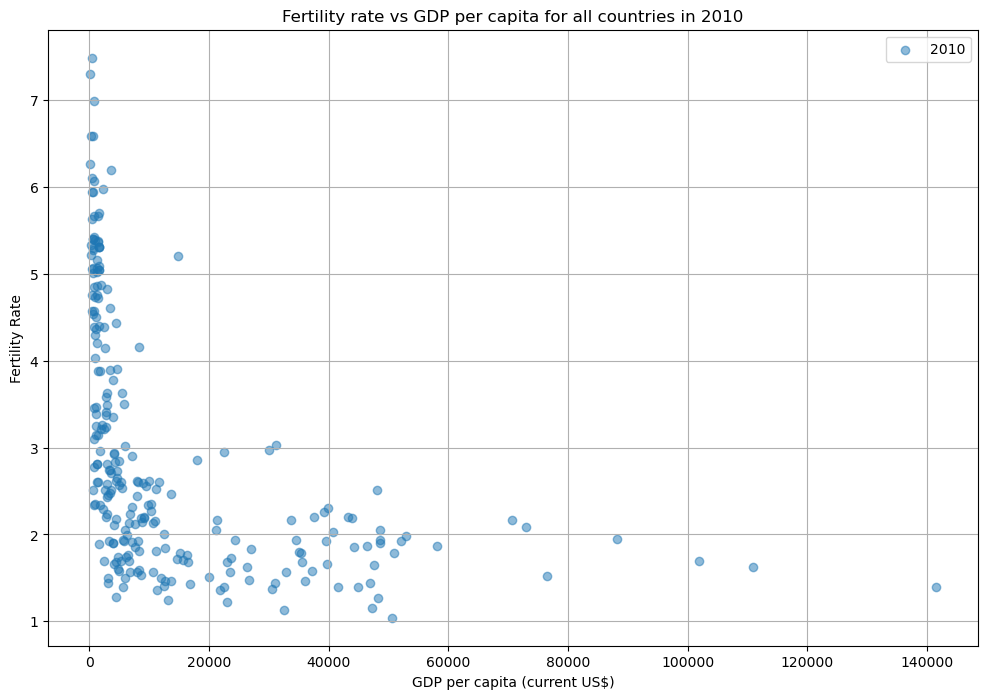
Approach:

* Download data for “Fertility rate, total (births per woman)” and “GDP per capita (current US$)”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as a single column).
* Configure and plot a scatter plot graph (2010 data only).
* Sort the data frames.
* Produce cumulative distribution functions.
* Calculate the median and mean for both years.
* Configure and plot line graph (1990 and 2010 data only).

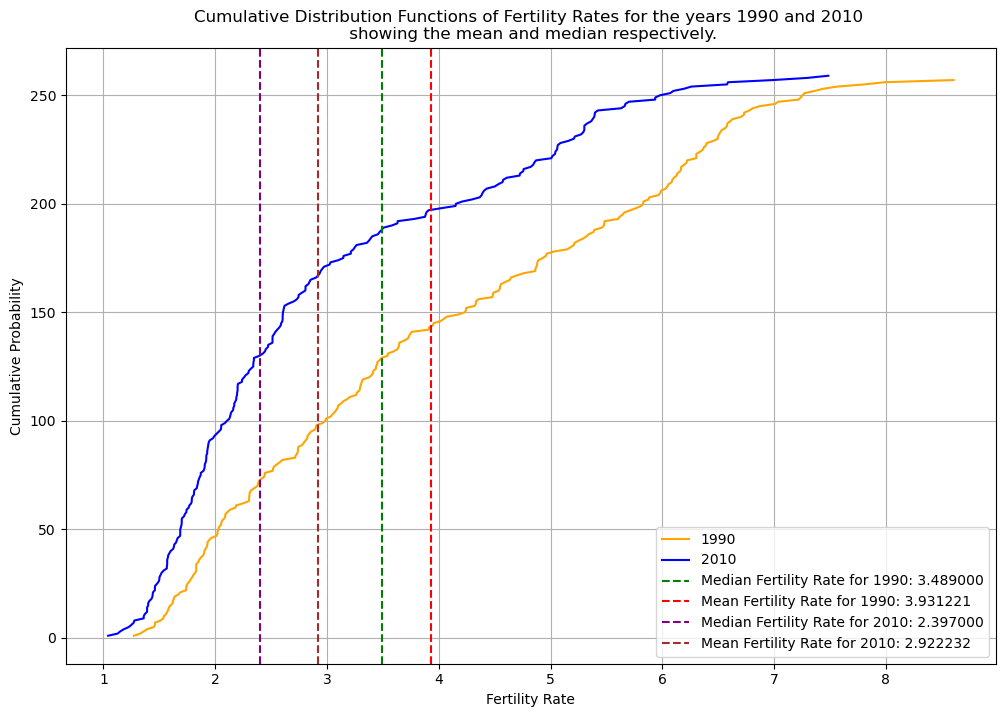
**Results:**

Scatter plot graph showing fertility rate against GDP per capita.

Line graph showing cumulative distribution function values.



*Fertility rate vs GDP per capita for all countries in 2010*



*Figure showing Cumulative Distribution Functions of Fertility Rates for the years 1990 and 2010 showing the mean and median respectively.*

**Analysis and Insights**

Fertility rate versus GDP per capita

* The j-shaped scatter plot signifies that, majority of the countries have a low GDP per capita and are likely to have high levels of fertility.

Cumulative distribution function graph

**Question 5 Report:**

**Methodology**

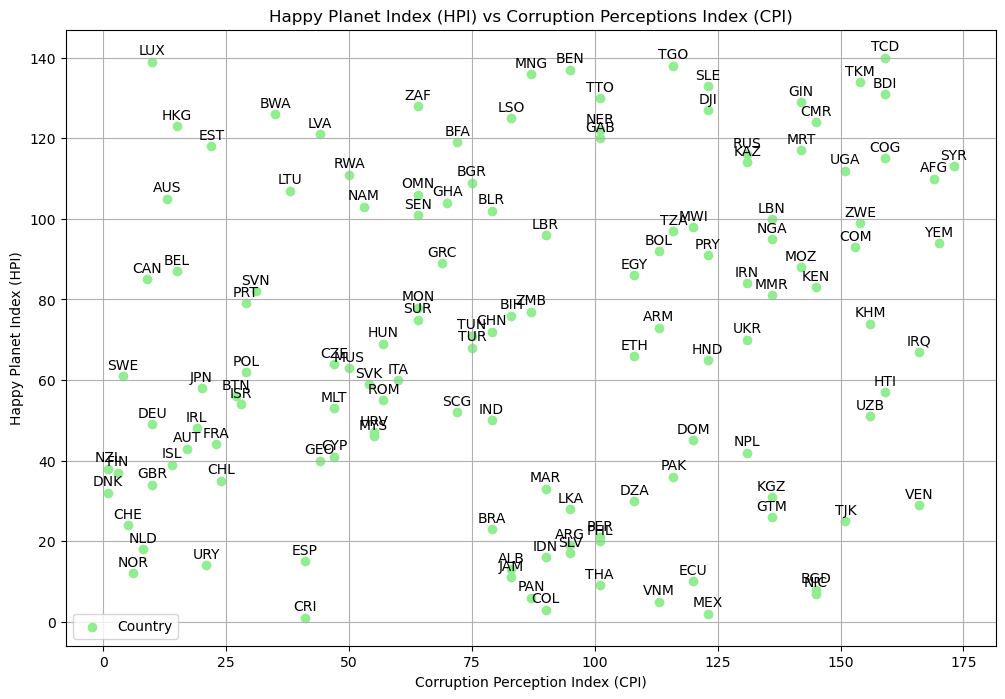
Scatter plot countries on Happy Planet Index vs Corruption Perception Index graph while annotating each one

* Approach:
* Download data for “Happy Planet Index” and “Corruption Perceptions Index”.
* Melt the data frames to long format (show the years as a single column).
* Configure and plot a scatter plot graph.

**Results:**

Plotting cumulated profits:

Utilizing the matplotlib[1] and pandas[2], we use the profits per day data to plot a line graph marking the initial investment and breakeven day.



*Figure showing scatter plot graph of Happy Planet Index (HPI) against Corruption Perceptions Index (CPI)*

**Analysis and Insights**

Country relationship

* Generally, one would expect to have all countries with a lower rate of corruption to have a high happy planet index. It can be observed that this is not the case as countries such as Norway who have a low corruption perception index, don’t have a high planet index. This indicates that a country like Norway is still unable to satisfy peoples lives.

References:

[1] “Matplotlib,” *Wikipedia*. Aug. 30, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Matplotlib&oldid=1243075914

[2] “pandas (software),” *Wikipedia*. Jul. 15, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pandas\_(software)&oldid=1234683004

[3] R. Python, “NumPy Tutorial: Your First Steps Into Data Science in Python – Real Python.” Accessed: Sep. 02, 2024. [Online]. Available: https://realpython.com/numpy-tutorial/

[4] *tabulate: Pretty-print tabular data*. Python. Accessed: Sep. 15, 2024. [OS Independent]. Available: https://github.com/astanin/python-tabulate

[5] N. Büttner, M. Heemann, J.-W. De Neve, S. Verguet, S. Vollmer, and K. Harttgen, “Economic Growth and Childhood Malnutrition in Low- and Middle-Income Countries,” *JAMA Netw. Open*, vol. 6, no. 11, p. e2342654, Nov. 2023, doi: 10.1001/jamanetworkopen.2023.42654.

[6] B. H. Program Producer Price Index, “Gold prices during and after the Great Recession,” Bureau of Labor Statistics. Accessed: Sep. 15, 2024. [Online]. Available: https://www.bls.gov/opub/btn/volume-2/gold-prices-during-and-after-the-great-recession.htm