**DATA, INFERENCE**

**&**

**APPLIED MACHINE LEARNING**

**(COURSE 18-785)**

**ASSIGNMENT 2**

**Mark Iraguha**

**(miraguha)**

Table of Contents

[Libraries Used: 3](#_Toc177397439)

[Introduction: 3](#_Toc177397440)

[Question 1 Report: 4](#_Toc177397441)

[Methodology 4](#_Toc177397442)

[Results 4](#_Toc177397443)

[Analysis and Insights 7](#_Toc177397444)

[Question 2 Report: 9](#_Toc177397445)

[Methodology 9](#_Toc177397446)

[Results 9](#_Toc177397447)

[Analysis and Insights 10](#_Toc177397448)

[Question 3 Report: 11](#_Toc177397449)

[Methodology 11](#_Toc177397450)

[Results 11](#_Toc177397451)

[Analysis and Insights 12](#_Toc177397452)

[Question 4 Report: 14](#_Toc177397453)

[Methodology 14](#_Toc177397454)

[Results 14](#_Toc177397455)

[Analysis and Insights 16](#_Toc177397456)

[Question 5 Report: 17](#_Toc177397457)

[Methodology 17](#_Toc177397458)

[Results 17](#_Toc177397459)

[Analysis and Insights 18](#_Toc177397460)

[References: 19](#_Toc177397461)

# 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

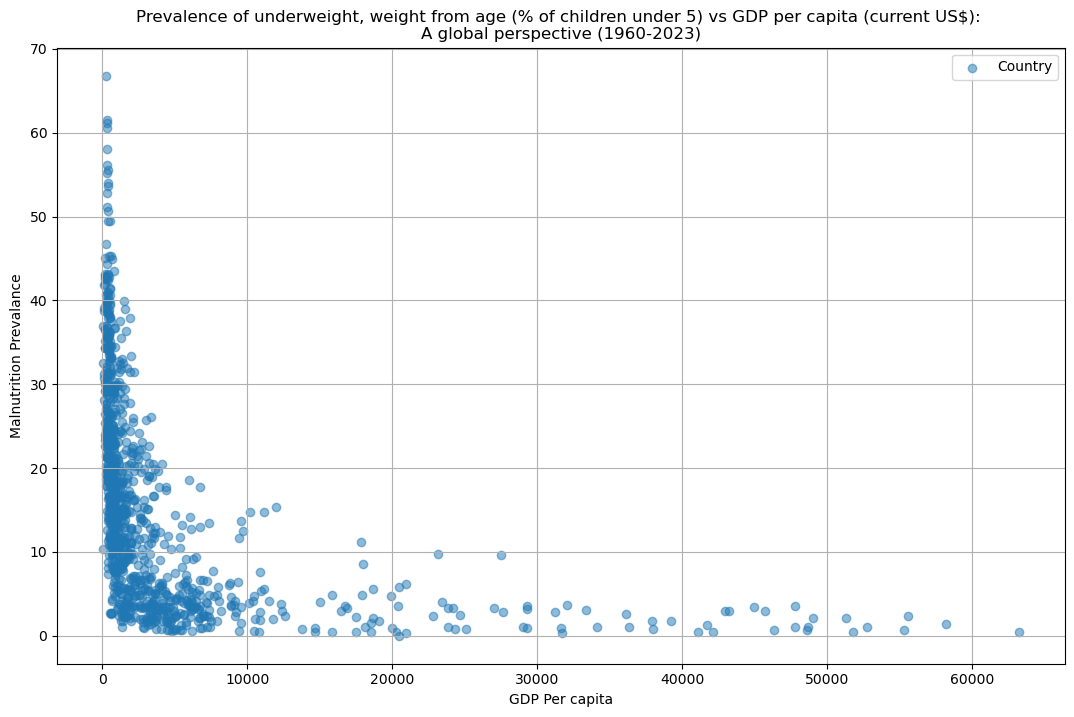
Making a scatter plot of malnutrition prevalence against GDP per capita.

Approach:

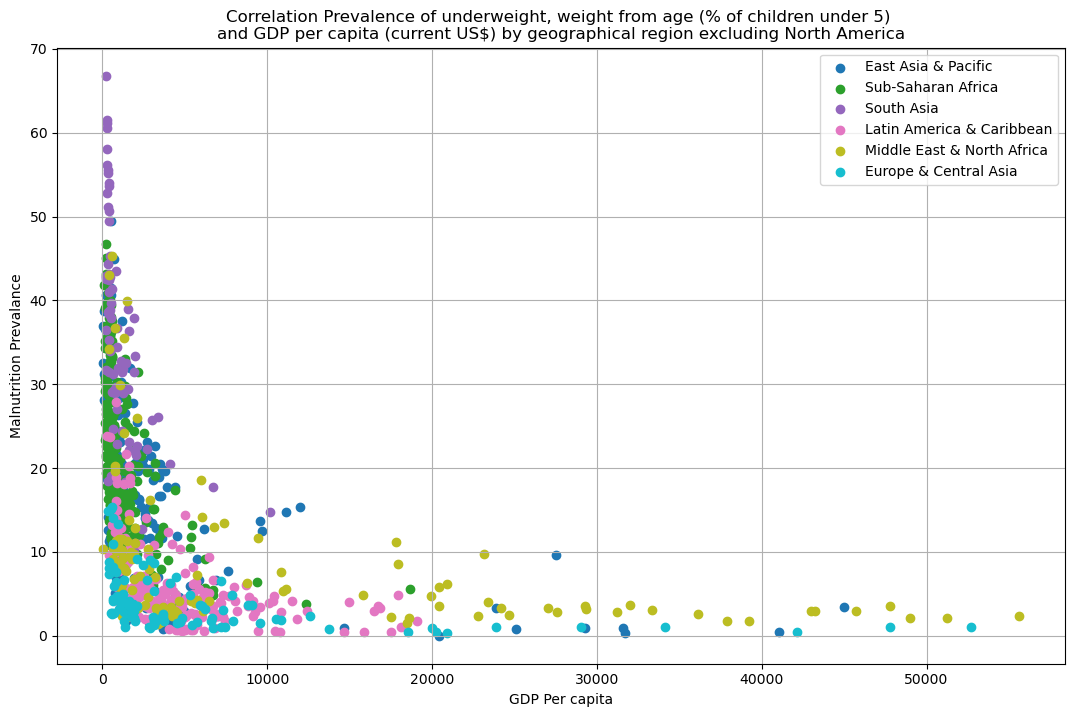
* Download data for “GDP per capita” and “Prevalence of underweight, weight for age”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as single column).
* Drop the NaN (Not a number) values.
* Configure and plot the graph.

## Results

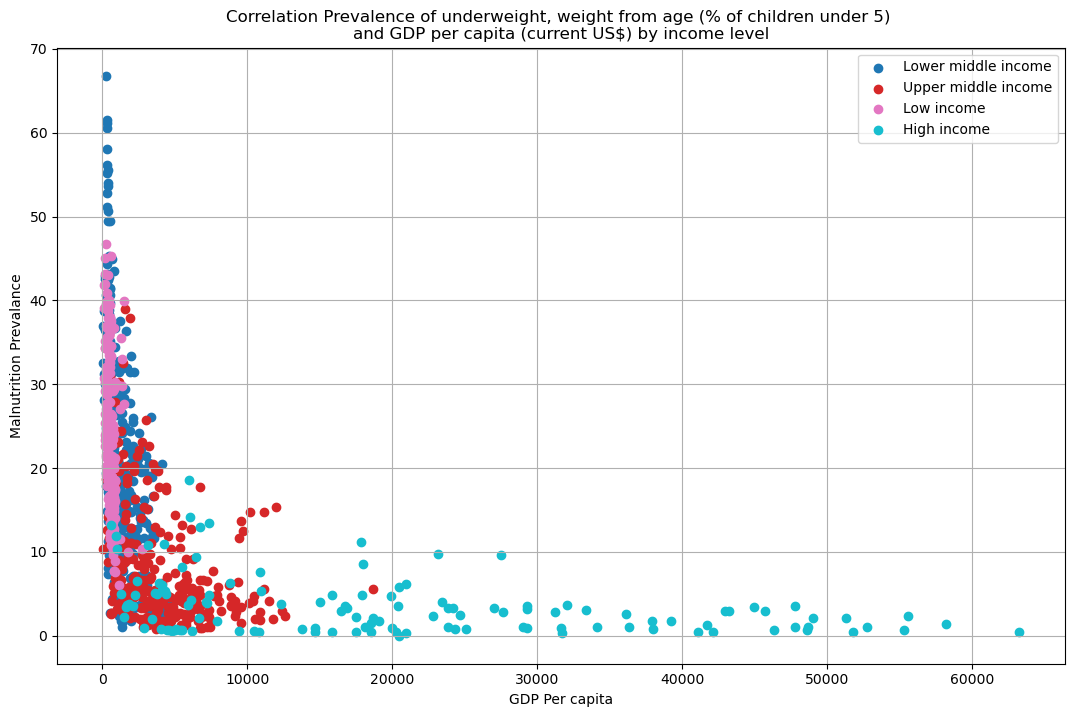
Three scatter plot graphs showing Malnutrition prevalence against GDP per capita.



*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:

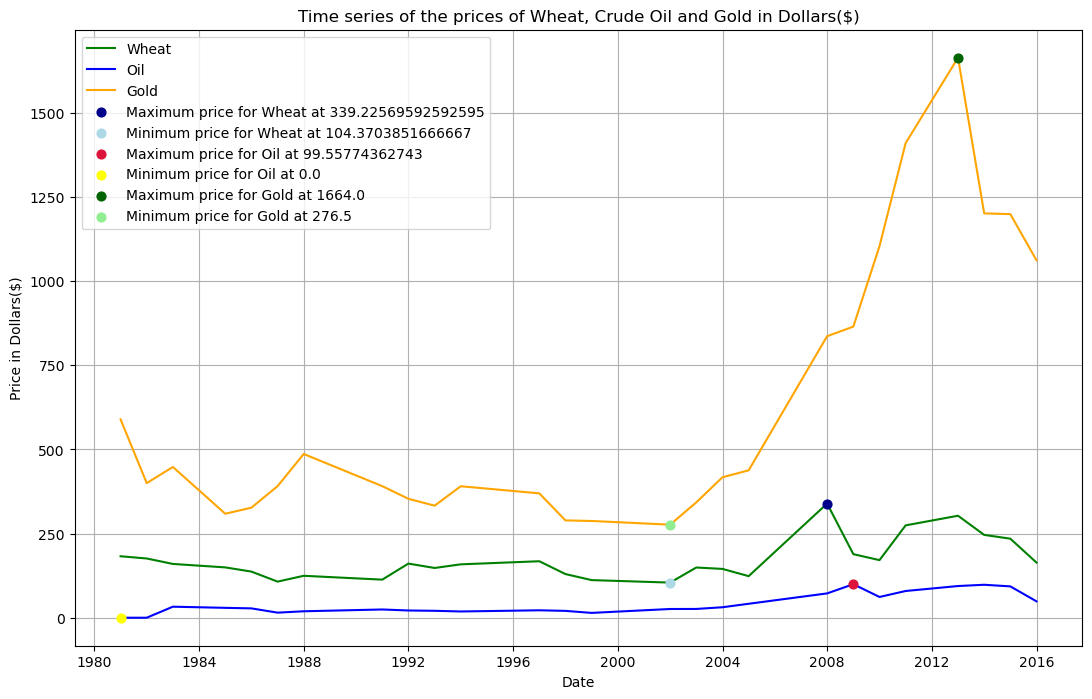
* Utilized Quandl API to download data for Wheat, Crude Oil and

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, which led to decreased consumer spending and industrial activity.

# Question 3 Report:

## Methodology

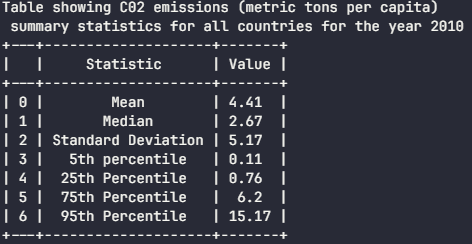
Using pandas library to calculate summary statistics[2]

Approach:

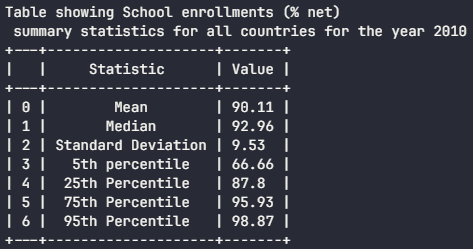
* Download data for “C02 emissions (metric tons per capita)” and “School enrolment, primary (% net)”.
* Clean the data and filter the data frames.
* Melt the data frames to long format (show the years as a single column).
* Drop the NaN (Not a number) values.
* Calculate summary statistics.
* Display 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 is significantly lower than the mean which strongly suggests the presence of outlier countries with very high emissions that are pulling the mean upward.
* The relatively high Standard Deviation is likely due to the differences among countries in areas such as energy sources, industrialization and population.
* The table indicates low values for the 5th, 25th and 75th percentiles, which signifies that a large portion of countries have relatively low emissions. However, the 95th percentile value of 15.17 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, on average, many countries had a lot of school enrollments for the year 2010. This signifies a positive trend in educational participation across the countries in the data.
* The median value suggests that majority of the countries had a high number of enrollments, so a possibility of an outlier is low. Although it doesn’t mean that there aren’t any outliers completely, there could still be some outliers but are not heavily influencing the mean.
* 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. In other words, this means that most countries have similar enrollment figures.
* Countries in the 5th percentile have relatively low school enrollments whereas countries in higher percentiles (25th, 75th, 95th) signify that they are experiencing a high number of enrollments. This indicates a trend towards higher enrollments as you move up through the percentiles.

# Question 4 Report:

## Methodology

Utilizing numpy to produce cumulative distribution functions.

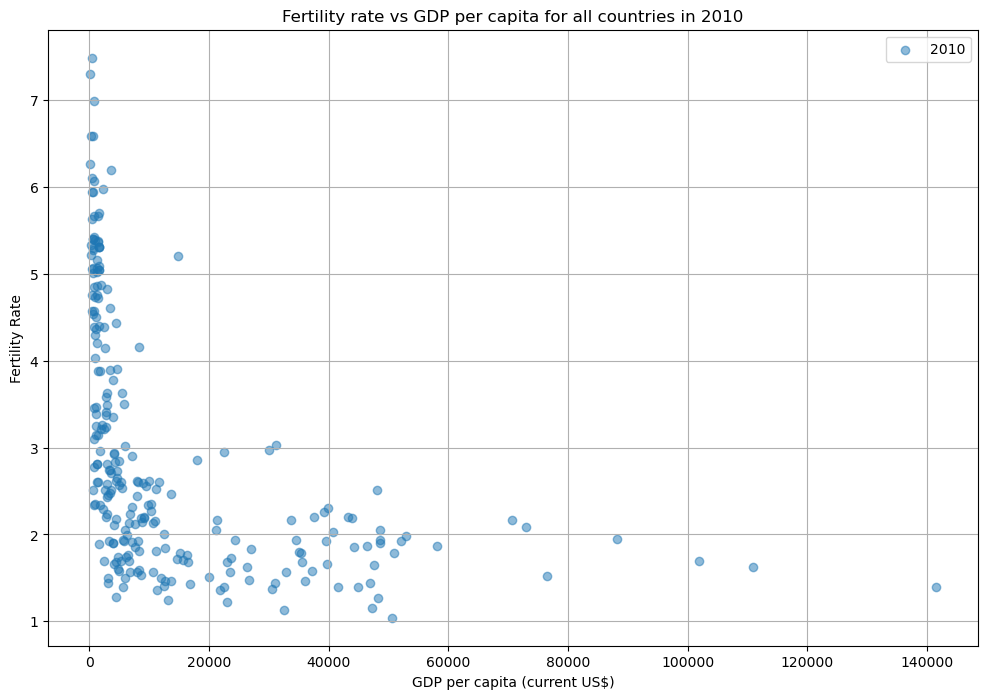
Approach:

* 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).

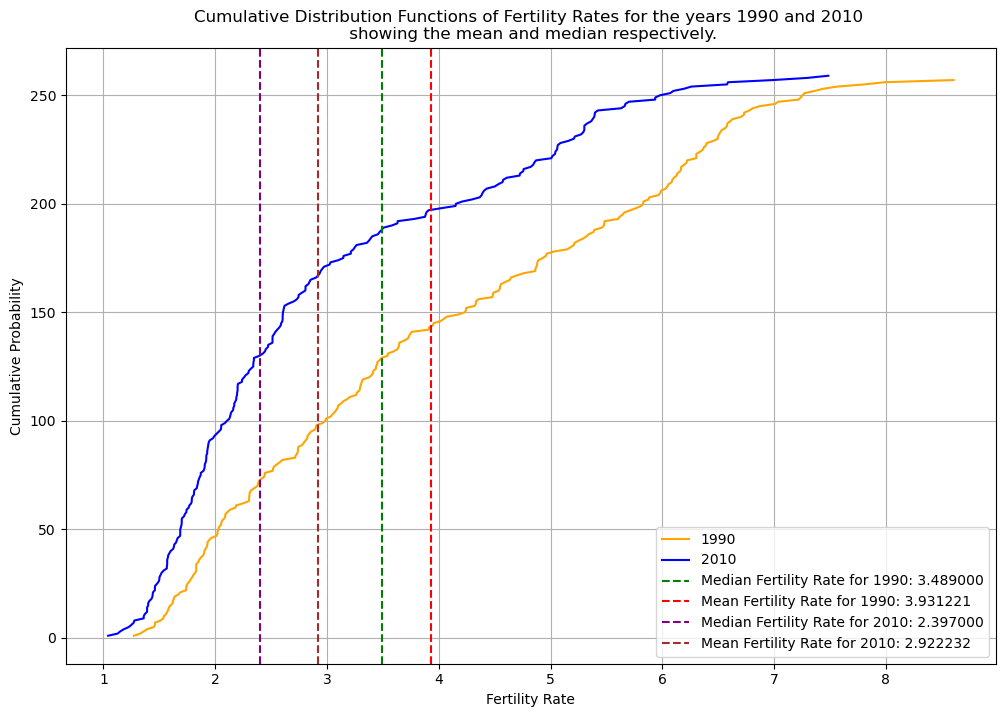
## Results

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 that have a low GDP per capita and are likely to have high levels of fertility. This pattern can be observed often in developing nations where socioeconomic factors, for example limited health care access and family planning resources contributes to higher birth rates. However, it is likely that at very high levels of GDP per capita, fertility begins to move upward again potentially due to developments in cultural shifts and financial stability. In general, the J-curve manages to highlight the non-linear relationship between Fertility rates and GDP per capita.

Cumulative distribution function graph

* We observe the initial steepness of the graph indicating that a significant number of countries have low to medium fertility rates. As you move from the low fertility rates, the probability that there will be the same high fertility rates increases. This means that the countries have successfully implemented policies to reduce the fertility rates with measures such as improved access to health care and family planning resources.
* The respective mean and median are observed to be in between the steepness which indicates that majority of the countries have high fertility rates. This applies to both the years (1990 and 2010).
* After the initial steep increase, the cumulative distribution function begins to flatten out. This signifies that fewer countries have higher fertility rates and those remaining countries with high fertility are becoming rare in the dataset.
* This pattern signifies the effectiveness of policies aimed at reducing fertility rates across countries.

# 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 people’s lives.
* Using the Quadrant Analysis method to identify unusual countries, we divide the graph into four quadrants[7]. These quadrants will include sections for countries with high HPI & high CPI, low HPI & low CPI, low HPI & high CPI and finally high HPI & low CPI. Countries with an unusual pattern fall into the high CPI & low HPI. Countries with low CPI & high HPI are seen with high well-being but low perceptions of corruption. The countries with high CPI score but low HPI scores, signify that corruption is not the only factor affecting people’s well-being. There are also cases of small island countries or countries with unique circumstances that lead to high HPI scores despite average CPI scores i.e. Burundi. In summary this graph plot has been observed to reveal a complex relationship where lower corruption correlates with high levels of happiness.

# 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

[7] J. Lynch, R. Carver, and J. M. Virgo, “Quadrant Analysis as a Strategic Planning Technique in Curriculum Development and Program Marketing,” *J. Mark. High. Educ.*, vol. 7, no. 2, pp. 17–32, 1996.