Carnegie Mellon University Africa

COURSE 18-785: DATA, INFERENCE & APPLIED MACHINE LEARNING

ASSIGNMENT 2

Nchofon Tagha Ghogomu

ntaghagh

September 16, 2024

LIBRARIES:

The following libraries were used:

- Math
- Pandas[1]
- Numpy[2]
- Pyplot from Matplotlib [3]
- Matplotlib.ticker from MultipleLocator - for defining axis spacing
- Quandl: The quandl library
- Tabulate: Library for table presentation

Programming Language:

- Python

INTRODUCTION

This assignment was made of 5 practical questions that portray real application of data analytics on world data, from a variety of sources. For every question, we had to come up with strategies to analyse and draw the insights required. In summary, throughout this assignment, we get to:

- Explore and analyse world data centred around development
- Download data from multiple sources including Quandl API
- Draw insight from these analysed data

Python was the programming language used and Jupyter notebook was the programming environment.

SOLUTIONS

Question 1:

1) Goal:

Analyse expected relationship before and after plotting malnutrition prevalence against GDP per capita data, and have 3 different plots that show general insights, regional and income level insights.

2) Steps:

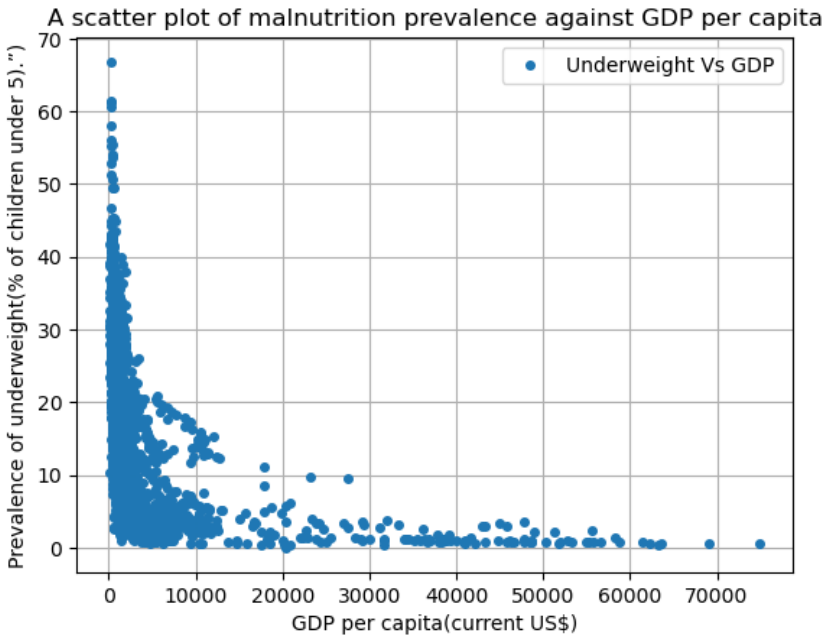
To arrive at the plots and insights, the following steps were used:

- The required data was downloaded from the World Bank Indicator data bank and imported (GDP and underweight prevalence).
- A list of years from 1960 to 2023 was made to specify columns that will be used.
- All data in the data from for both GDP and underweight prevalence for these years was looped through and stored in a list.
- Both lists were plotted against each other to produce a graph general graph
- To group by region and income levels, the meta data was used and a list of country code corresponding to income levels and regions was created.
- Using this list, the data corresponding to the cited classifications was extracted from the data frame and plotted, and annotated.

3) Results and observation:

Part A:

My expectation before plotting the graph was a clear invers relationship between GDP and underweight prevalence. This is because wealthier nations will have more income to feed their families properly and sometime feed to obesity.

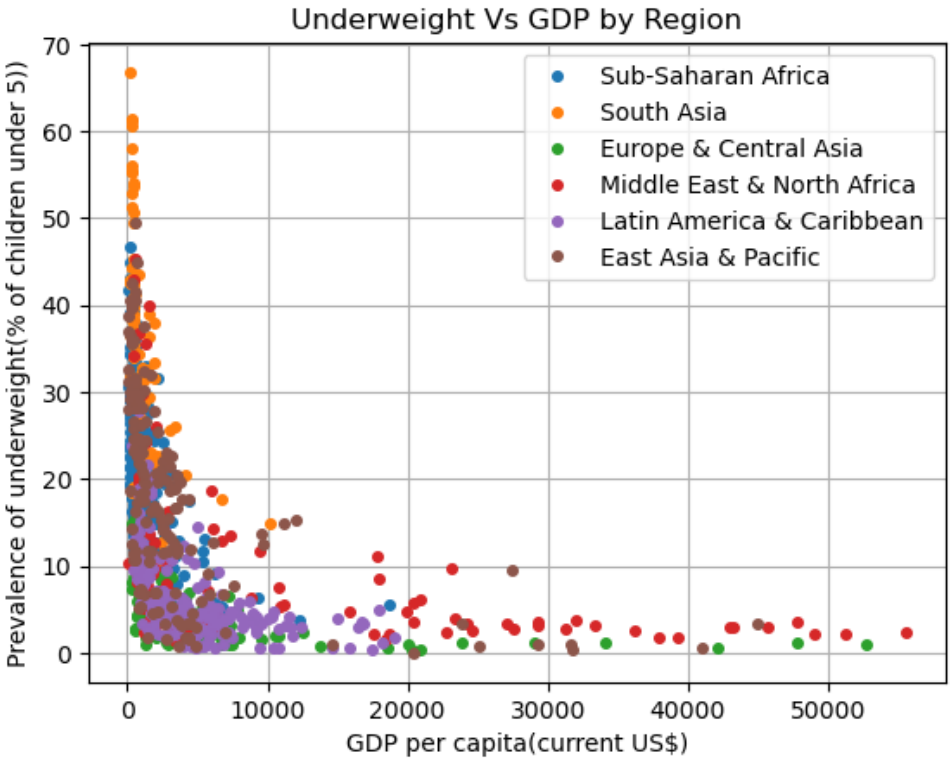


After plotting the graph, we observe a strong negative correlation between them. Except for a few countries, all countries with a GDP per capita of $15000USD tend to have a malnutrition rate of less than 10%. Similarly, countries with a GDP per capita of below $15000USD tend to have high malnutrition rates.

It is also worth noting that some a reasonable number of countries with low GDP per capita also have low malnutrition rate. Cuba for example, due to good health policies[4] is an exemplary country with surprisingly low GDP levels but also low malnutrition.

Part B:

As outlined in the steps, countries were grouped into regions notably: Sub-Saharan Africa, South Asia, Europe and Central Asia, Middle East and North Africa, Latin America and Caribbean, East Asia and Pacific and the results was as seen bellow.

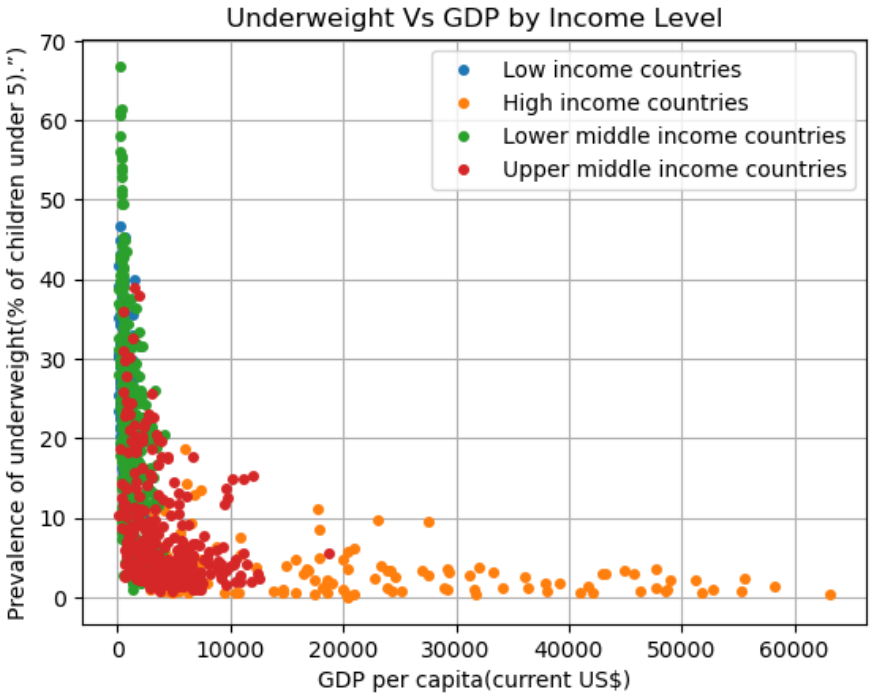


We notice that the highest GDP per capita comes from Middel East and North African, Europe and Central Asian, and East Asia and Pacific regions. For these countries with extreme GDP per capita, their malnutrition rates sit below 5%. These countries certainly have enough wealth live on a healthy diet. South Asian countries on the data available lead with the rate of malnutrition. They have a significantly low GDP per capita.

In summary, South Asian countries have the highest record of malnutrition rates with a corresponding low GDP per capita. Also, Middel East and North African, Europe and Central Asian, and East Asia and Pacific countries have the highest GDP per capita and correspondingly low prevalence of underweight.

Part C:

Grouping by income levels, it becomes clearer. Countries with higher income exhibit low malnutrition rates. How ever, Lower middle-income countries tend to have the highest malnutrition rates even higher than most low-income countries, that probably rely on their farms for food. Most upper middle-income countries sit around the low GDP low malnutrition rate area.



Question 2:

1) Goal:

Use data from Quandl API to plot the prices of 3 commodities over time: wheat, gold and crude oil.

2) Steps:

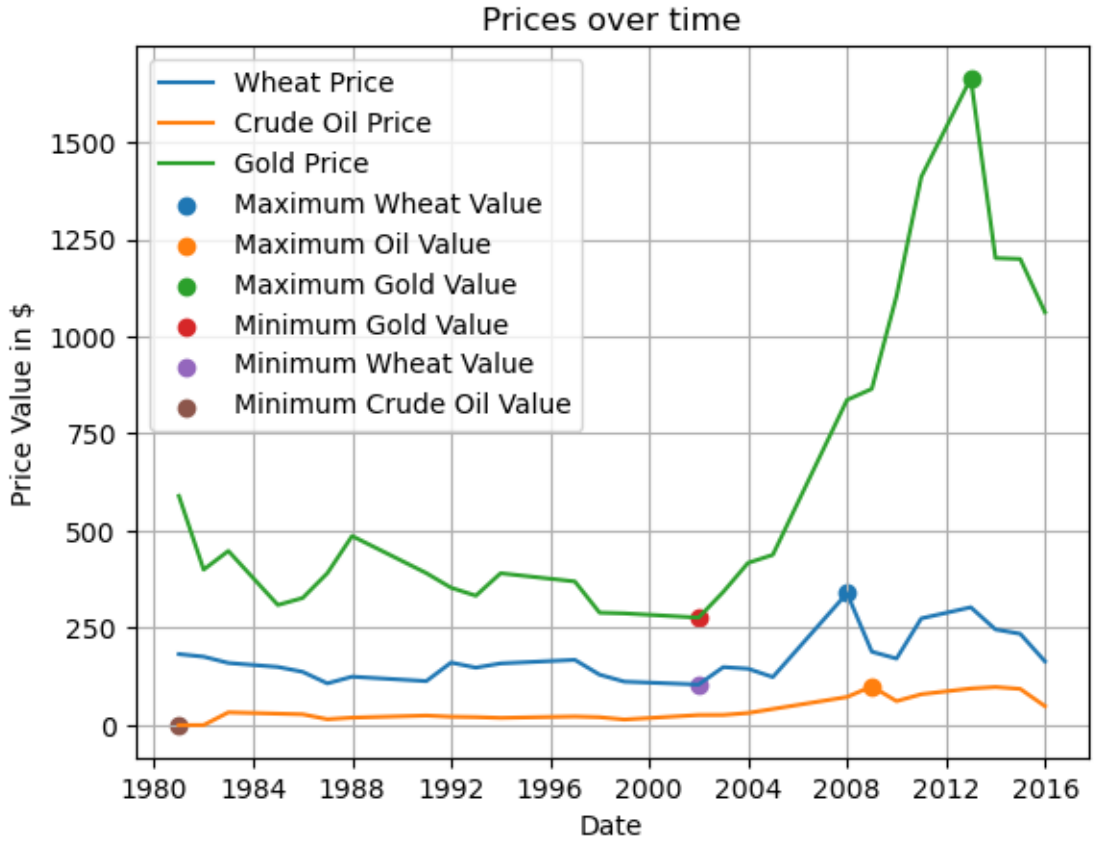
To arrive at the plots and insights, the following steps were used:

- The required data was downloaded from Quandl via their API.
- The data was merged into a data frame called price and the headers were renamed from default to Wheat Price, Crude Oil Price and Gold Price.
- The maximum and minimum prices for all commodities were extracted with their indexes to get the corresponding dates.
- These prices for each commodity were plotted, so were the maximum and minimum points.
- The graph was annotated.

3) Results and observation:

The results obtained can be seen bellow. It is observed that Gold has always been more valuable than crude oil and wheat. Also, all commodities had their peak between 2008 and 2013. This could be triggered by great recession that happened, which was the most significant economic crisis since the great depression of 1930. The years leading up to the crisis were characterized by an exorbitant rise in asset prices and associated boom in economic demand[5]. Towards 2016, all commodities tend to experience a fall in value.

The value of crude oil has been on a gentle rise since 1980. The minimum point for gold and wheat sits around 2002. This can be associated to the decline in economic activities in developed countries in the early 2000[6].



Question 3:

1) Goal:

Provide summery statistics for CO2 emissions (metric tons per capita), and School enrolment, primary (% net)"

2) Steps:

To arrive at the statistics and insights, the following steps were used:

- The required fields (Country code, years) are extracted into a data frame

- Using the inbuilt function: mean, median, std, and quantiles these values for the for CO2 emission and school enrolment for the given years is calculated.
- The data is presented using the tabulate function.

3) Results and observation:

After calculating the values for CO2 emission and enrolment, we obtain the following results

Part A:

For CO2 emission in metric tone per capita, it is observed that the mean value is 4.4076. However, there a are countries still producing a relatively high CO2 emission, up to 15.172 metric tone per capita. There is therefor much more effort that needs to be put into reducing CO2 emissions of some countries.

Statistics	Mean	Median	Standard Deviation	5th Percentile	25th Percentile	75th Percentile	95th Percentile
2010 Summary Statistics:	4.4076	2.66714	5.16505	0.11486	0.756011	6.20007	15.172

Figure 1 CO2 emissions (Metric tone per capita)

Part B:

The result of net school enrolment for primary is acceptably high. This could be cause by the efforts put in by the governments in many countries to both subsidise and encourage primary education. These values go up to, and beyond 98 %. In general, this are encouraging figures as for every child in a country, there is a significantly high chance of going through primary education.

Statistics	Mean	Median	Standard Deviation	5th Percentile	25th Percentile	75th Percentile	95th Percentile
2010 Stat Summary	90.1051	92.9567	9.52763	66.6568	87.801	95.9344	98.8728

Figure 2 School enrolment, primary (% net)

Question 4:

1) Goal:

Compare fertility rate, total (births per woman)" and "GDP per capita (current US$)", and compare fertility rate for the years 1990 and 2010.

2) Steps:

To arrive at the plots and insights, the following steps were used:

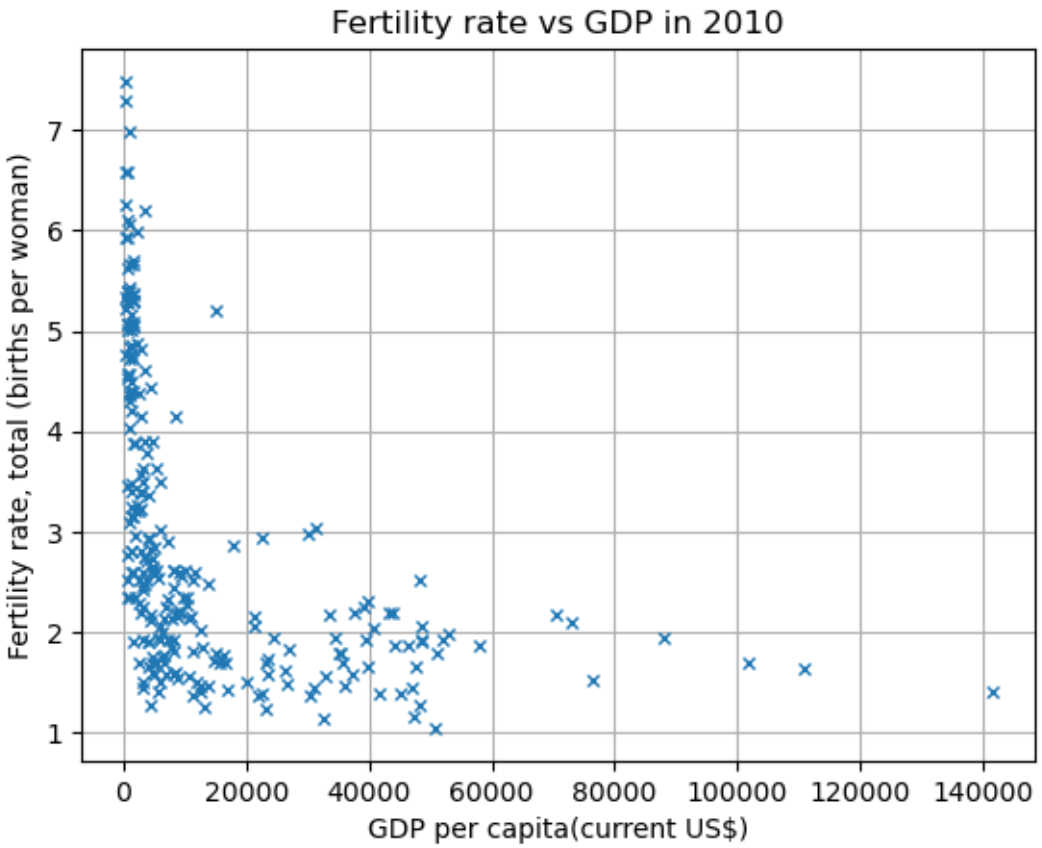
- The country code and years are imported form the files into respective data frames.
- The values of the various years were fed into a list.
- These values are plotted against each other.
- To get the CDF, we create an array of integers from 1 to the length of the fertility rate array and divide each value by the length of the fertility array.

- The CDF is plotted against the fertility rate to obtain the cumulative distribution function graph.
- The mean and median of fertility rates for both years are extracted and using the plt.axvline, we plot a vertical line with value the mean and median.
- The results as seen bellow is obtained.

3) Results and observation:

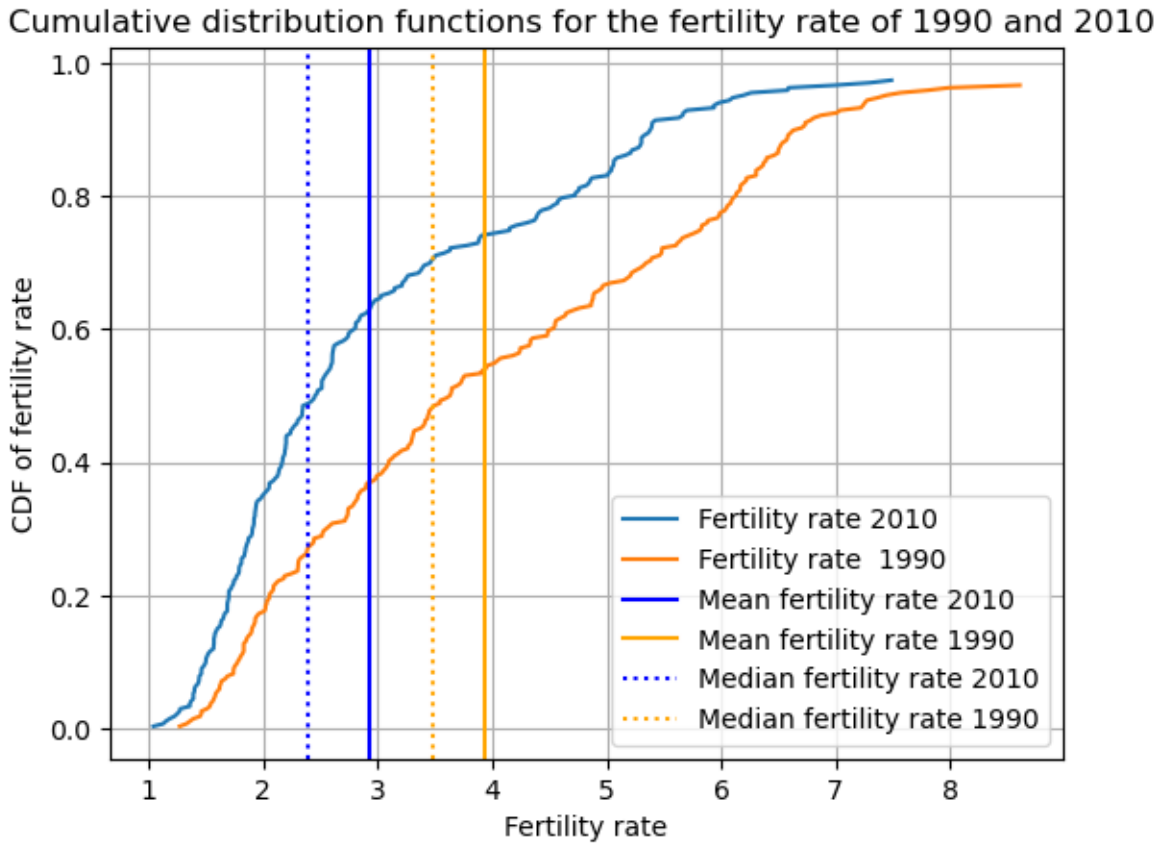
Part A:

The fertility rate and CDP present a strong negative correlation. Countries with the highest GDP levels present a fertility rate of less than 2 births per women. Also, countries with the highest fertility rates present a lower GDP level, less than 10000US$ per capita. Usually, low-income level countries are home to families who spend much time either in the home or farms. They have enough time for leisure and thus, have time to raise more children.



Part B:

The cumulative distribution function for fertility rates for 2010 and 1990 show significant differences. The rise in CDF is steeper in in 2010 than in 1990. 1990 has a higher mean and median than 2010. This suggests that 2010 converges to a lesser mean value thus a decline in average birth per women.



Question 5:

1) Goal:

The goal is to plot and Happy Planet Index and Corruption Perceptions Index rank per country and identify unusual countries.

2) Steps:

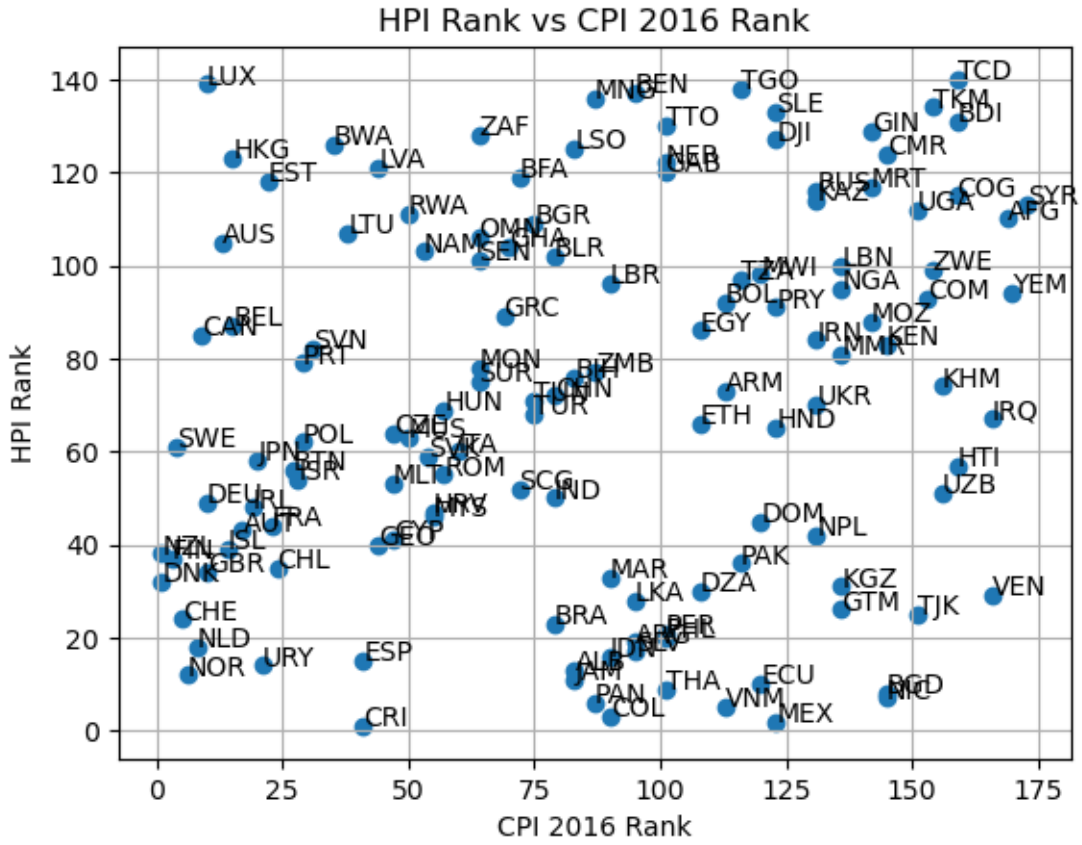
To arrive at the plots and insights, the following steps were used:

- Both data were extracted into a data frame from the downloaded sheet.
- The country, country code, and rank data were extracted from data fame.
- To find matching countries, countries from both data frames were compared against each other and countries that existed in both data frames were selected and the rest were dropped.
- The HPI was plotted against the CPI.
- The annotation was the country code column.

3) Results and observation:

The results of the plots are seen below. It is observed that there is very little or no correlation between the happy planet index and the corruption perception index of a country. Surprisingly, some countries with a high rank on the corruption perception index scale also

are high on the happy planet index rank. Some of them include Norway and the Nederland. Also, there are countries low on both the CPI and HPI index rank like Tchad and Burundi.



REFERENCES

[1] “pandas documentation — pandas 2.2.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/index.html

[2] “NumPy -.” Accessed: Sep. 02, 2024. [Online]. Available: https://numpy.org/

[3] “Matplotlib documentation — Matplotlib 3.9.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://matplotlib.org/stable/

[4] R. Pineo, “Cuban Public Healthcare: A Model of Success for Developing Nations,” *J. Dev. Soc.*, vol. 35, no. 1, pp. 16–61, Mar. 2019, doi: 10.1177/0169796X19826731.

[5] “Oil prices: George Soros warns that speculators could trigger stock market crash | Commodities | The Guardian.” Accessed: Sep. 16, 2024. [Online]. Available: https://www.theguardian.com/business/2008/jun/03/commodities

[6] “Early 2000s recession,” *Wikipedia*. Jul. 22, 2024. Accessed: Sep. 16, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Early_2000s_recession&oldid=1236063917#bodyContent