Name: Niyomwungeri Parmenide ISHIMWE

Andrew-ID: parmenin

DATA, INFERENCE, AND APPLIED MACHINE LEARNING

18-785

ASSIGNMENT 2

19 SEPTEMBER 2022

I, the undersigned, have read the entire contents of the syllabus for course 18-785 (Data Inference and Applied Machine Learning) and agree with the terms and conditions of participating in this course, including adherence to CMU's AIV policy.

Signature: **Niyomwungeri Parmenide ISHIMWE**
Andrew ID: **parmenin**
Full Name: **Niyomwungeri Parmenide ISHIMWE**

The libraries used:

- numpy
- pandas
- matplotlib
- quandl

QUESTION 1:

The relationship I expected is that the malnutrition prevalence would increase when GDP per capita went down. This means that malnutrition would depend on the GDP per capita of a country.

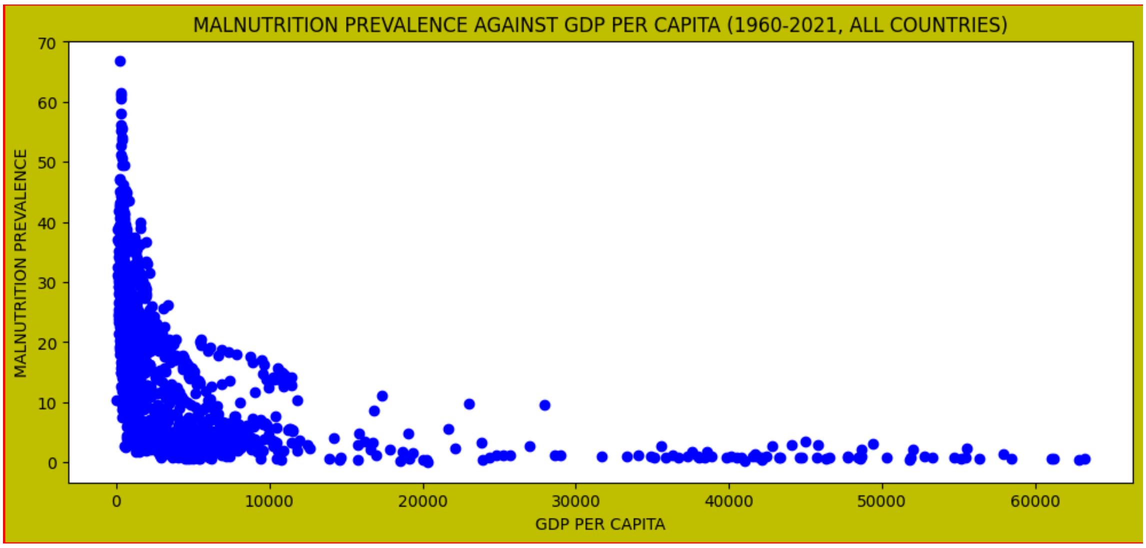


Figure 1: Scatter plot of malnutrition prevalence vs GDP per capita for all countries

As it was required to give a scatter plot showing the relationship between malnutrition prevalence and GDP per capita, by plotting all countries for available years from the datasets given, the first thing to be done after reading the data from the excel file sheet downloaded from World Bank Databank, then identify and extract the columns of years to use while plotting the data. After getting them, then they are scatter plotted year by year using a loop. After plotting, the relationship perceived is that the **increase in GDP per capita yields a decrease in malnutrition prevalence for children under 5**. This means that rich countries with good GDP per capita have children in good health with low levels of malnutrition.

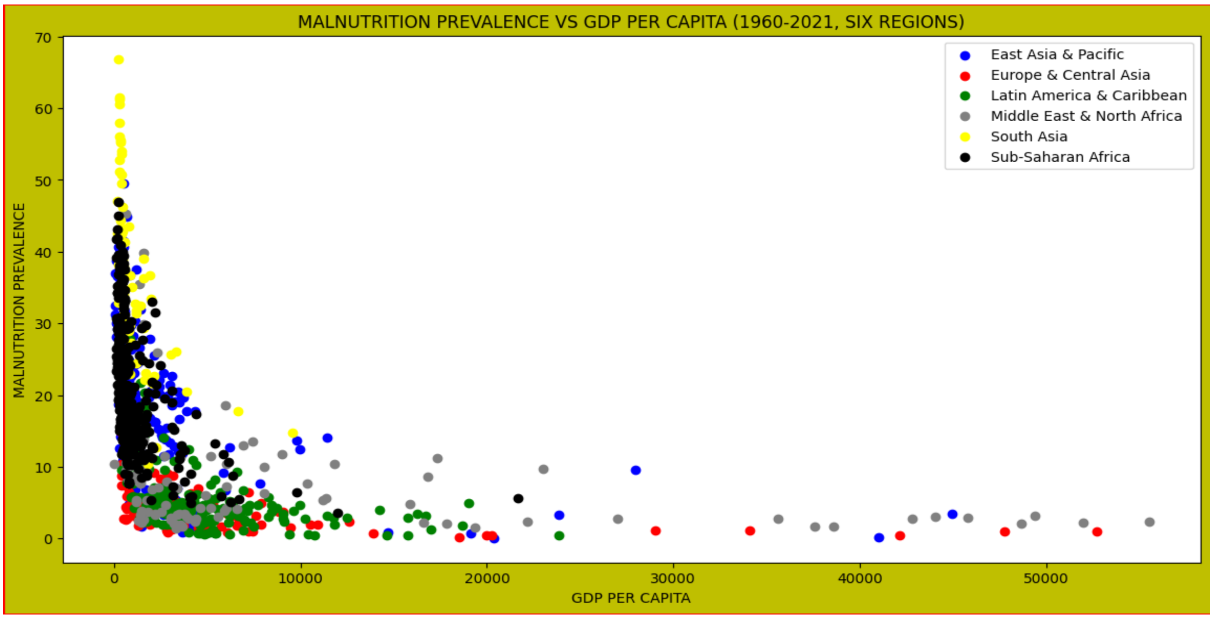


Figure 2: Scatter plot of malnutrition prevalence vs GDP per capita for regions

To give a scatter plot showing the relationship between malnutrition prevalence and GDP per capita by plotting for the six geographical regions excluding North America, using a different color for every region, the previous datasets were used and brought in the data from metadata which has the info about the regions to be used. The key here is to merge the previous datasets with corresponding metadata on the "**Country Code**" field to have relevant countries only. Next, columns of years are identified from merged datasets. The next step is to filter data related to each region and scatter plot them separately using a loop. Note that, the plot result is stored in a variable that helps to easily add the legend to each region. The graph above illustrates how several countries in the six developing regions including all **South Asian** countries have a low GDP per capita that is less than **$1,000. South Asia** and **Sub-Saharan** countries have greater levels of malnutrition prevalence with low levels of GDP. Contrarily, the **Middle East and North American** countries show high levels of GDP and a slight level of malnutrition prevalence.

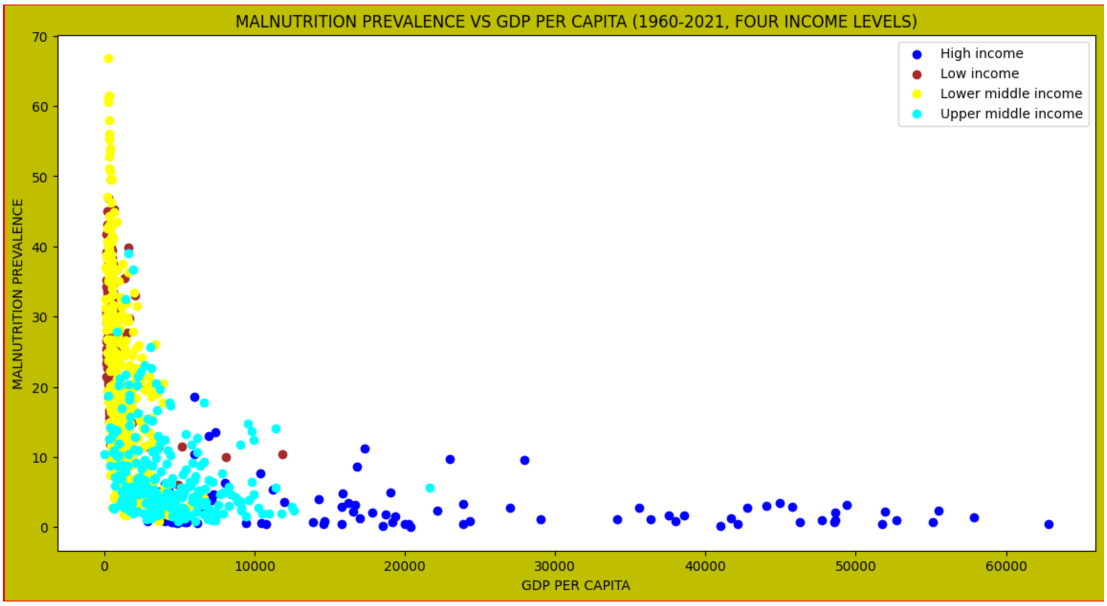


Figure 3: Scatter plot of malnutrition prevalence vs GDP per capita for income levels

It is also asked to give a scatter plot showing the relationship between malnutrition prevalence and GDP per capita by plotting for the four income levels using a different color for every income level. To achieve this, previous merged datasets and identified columns of years were used. The next step is to filter data related to each income level and scatter plot them separately using a loop. Note that, the plot result is stored in a variable that helps to easily add the legend to each income level. As shown on the graph, the **High-income** countries show a high GDP per capita and low malnutrition prevalence as expected. Thus, **lower-middle-income** and **low-income** countries show a high malnutrition prevalence. **Upper middle-income** countries have average malnutrition prevalence with also a medium GDP per capita.

QUESTION 2:

It was asked to plot time series for the prices of Wheat, Crude Oil, and Gold in $ on the same graph and indicate the maximum and minimum prices in all three-time series using coloured dots. This was addressed by first bringing in the data from the **quandl API** and then merging all three datasets on the date to have a dataset with synchronized time stamps. Next, the three datasets are plotted using the same graph with their prices against time series. To indicate their minimum and maximum values on the graph, the rows with minimum and maximum values were identified with corresponding dates and then were shown on the same graph using a scatter plot.

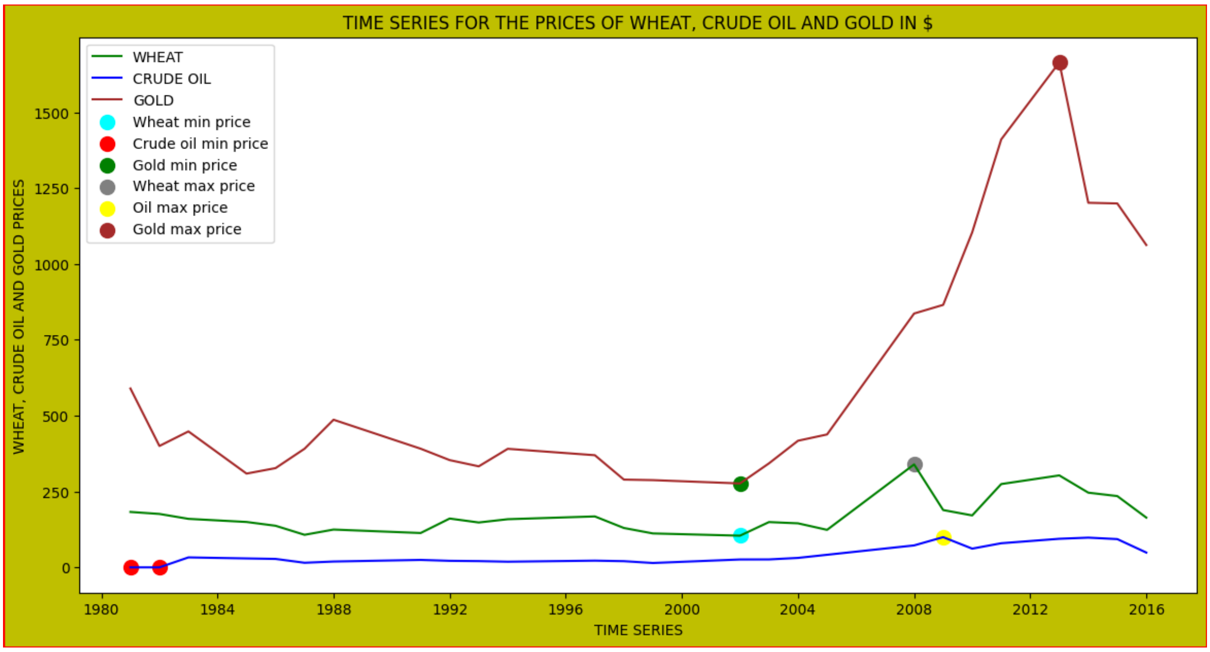


Figure 4: Plot of wheat, crude oil, and gold prices vs time series

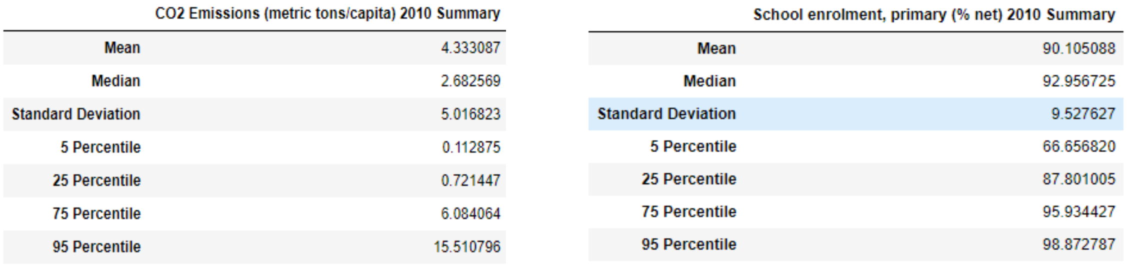
What can be inferred from this graph is that the price of **gold** had risen significantly between the years 1980 and 2016, starting at about **$620** in 1980 and reaching its bottom of **$270** at the end of 2001 and its peak of **$1600** at the end of 2012 before falling again to around **$1000** in 2016. In addition, **wheat's** price has been fluctuating below around **$250** in this period of 1980 and 2016, when its bottom value was in 2001 at **$104** and the maximum was **$339** in 2007. Finally, the price of **crude oil** has experienced a low increase compared to the price of gold that gold, reaching its peak of **$99** in 2008 and a minimum of **$0** in 1980 and 1981.

QUESTION 3:

It was required to calculate and provide summary statistics tables for CO2 emissions (metric tons per capita) and School enrolment, primary (% net) for all countries in 2010. This is done by first reading data from an excel file, extracting only the data needed to use (Country Name and 2010 columns), and then applying the **mean (), median (), std (),** and **quantile ()** to find the mean, median, standard deviation and percentiles for both CO2 emissions and School enrolment.

As the datasets have Nan values, they can be handled by skipping them, dropping them, filling them with zeros, or replacing them with the mean values. But here, it is found that filling them with zeros or replacing them with the mean would affect the summary statistics which can generate false results or errors. Therefore, skipping them or treating them as if they do not exist would yield correct summary statistics. Applying Python's summary statistics functions such as. **mean (), median (), std (), quantile**

() with or without **skipna=True** parameter would ignore the Nan values to provide accurate results. In addition, the results are summarized below in tables.



	CO2 Emissions (metric tons/capita) 2010 Summary
Mean	4.333087
Median	2.682569
Standard Deviation	5.016823
5 Percentile	0.112875
25 Percentile	0.721447
75 Percentile	6.084064
95 Percentile	15.510796

	School enrolment, primary (% net) 2010 Summary
Mean	90.105088
Median	92.956725
Standard Deviation	9.527627
5 Percentile	66.656820
25 Percentile	87.801005
75 Percentile	95.934427
95 Percentile	98.872787

Figure 5: Summary statistics tables for CO2 and school enrolment

QUESTION 4:

It was tasked to plot a scatter diagram of the Fertility rate versus GDP per capita for all countries in 2010. And then, to produce a graph for cumulative distribution functions for the fertility rate against GDP per capita using the 1990 and 2010 years and use vertical lines to indicate the mean and median. The approach here is to first import and extract the needed data set columns and then plot them with a scatter plot of Fertility rate versus GDP per capita for all countries in 2010 only.

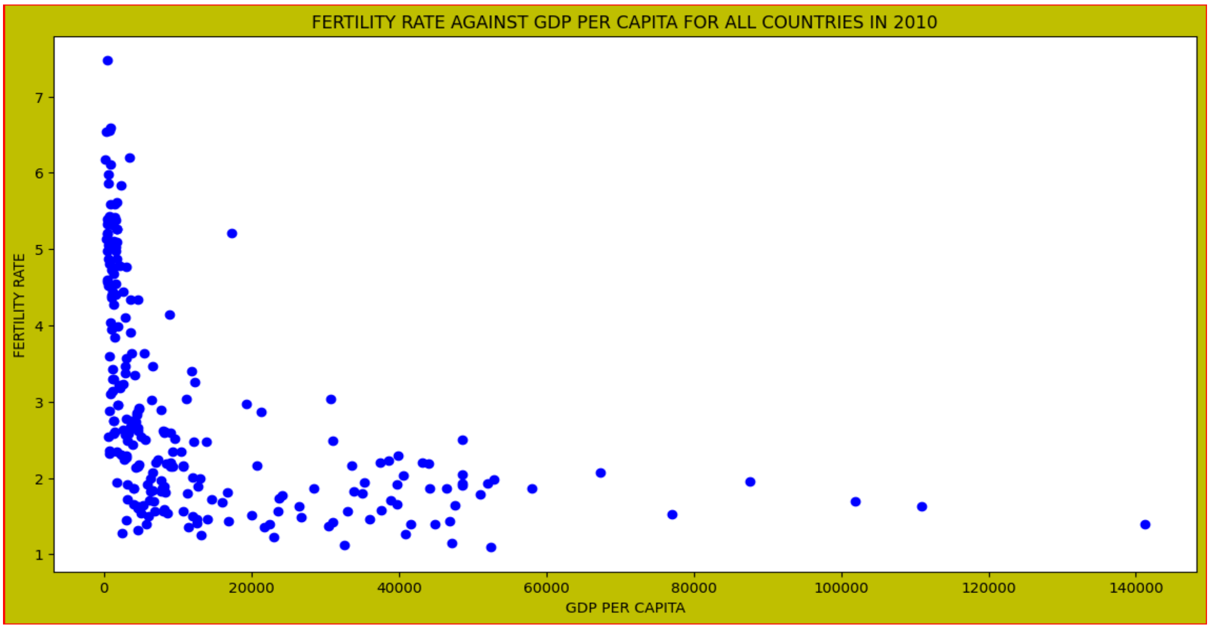


Figure 6: Scatter plot for fertility rate vs GDP per capita for all countries in 2010

What can be inferred from the plot, is that the fertility rate depends on the GDP per capita. Rich countries have a low fertility rate whereas poor countries have the highest ability to produce more children because their fertility rates are very high. Therefore, it may be assumed that as the wealth of the country increases, the fertility of the population falls.

Next, to plot the cumulative distribution functions, data for 1990 and 2010 also must be extracted for Country Name and corresponding year columns, sort the year's column, then calculate the **CDF**. To calculate it, the **arange ()** function from NumPy is used on the length of the column and divided its result by the number of items in the column minus 1. After that, it is plotted using the **sorted data** on the x-axis and the **CDF** on the y-axis. Then, calculated the mean and median for both year's columns to indicate them using the vertical lines on the plot by using the matplotlib **axvline ()** function. Thus, it is depicted in the next plot, that the fertility rate has dramatically changed over this period of 20 years.

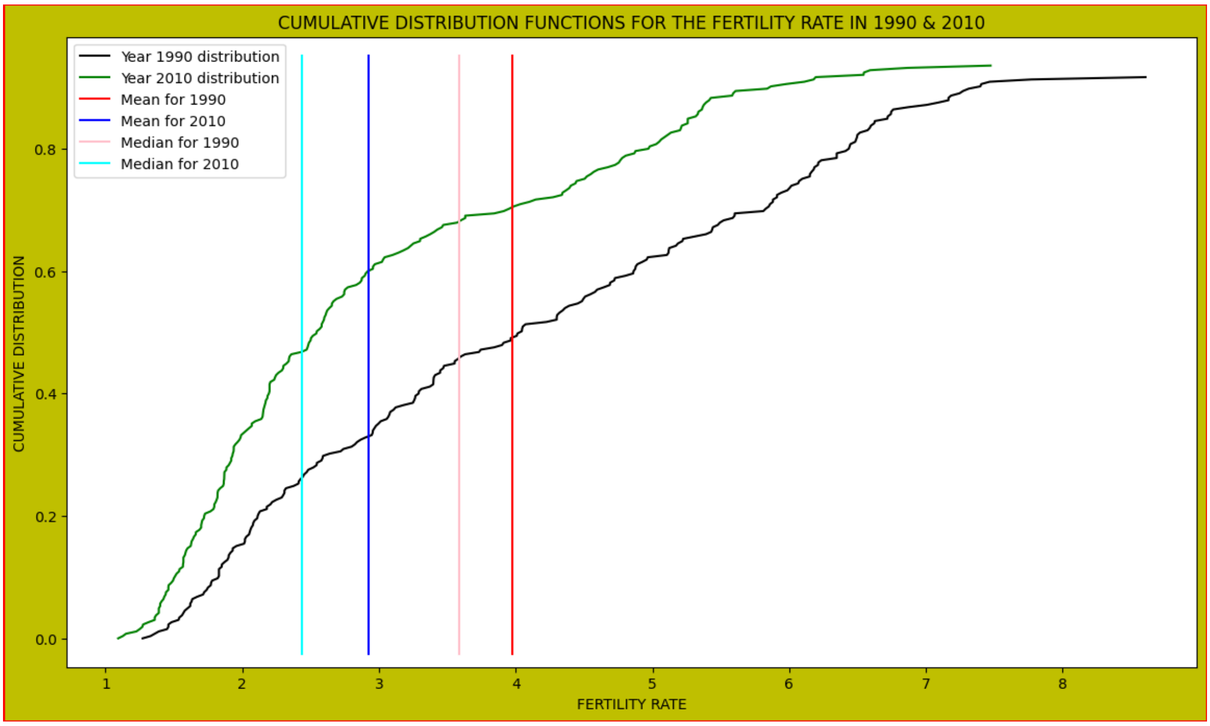


Figure 7: Plot for cumulative distribution functions for fertility rate in 1990 & 2010

What can be inferred here is that the fertility rate has been increasing as the cumulative distribution rose during the period of 1990 and 2010. However, the median and mean for both years show a decrease during this period where the mean fertility rate in the year 1990 was around **3.9** and had fallen to around **2.9,** and the median fertility rate in 1990 was around **3.6** before decreasing to around **2.4** in 2010.

QUESTION 5:

It has been asked to find matching countries for both indices and make a carefully labeled scatter plot of HPI against CPI to demonstrate the relationship using ranks in both cases. This had been approached by first bringing in the data from excel files downloaded from canvas and transparency websites. Then, merging them on the country field, and then extracting the ranks as they are our interest. After that, plotting them required a loop where enumerate function was used to return both the index and country code used to annotate each point on the scatter plot.

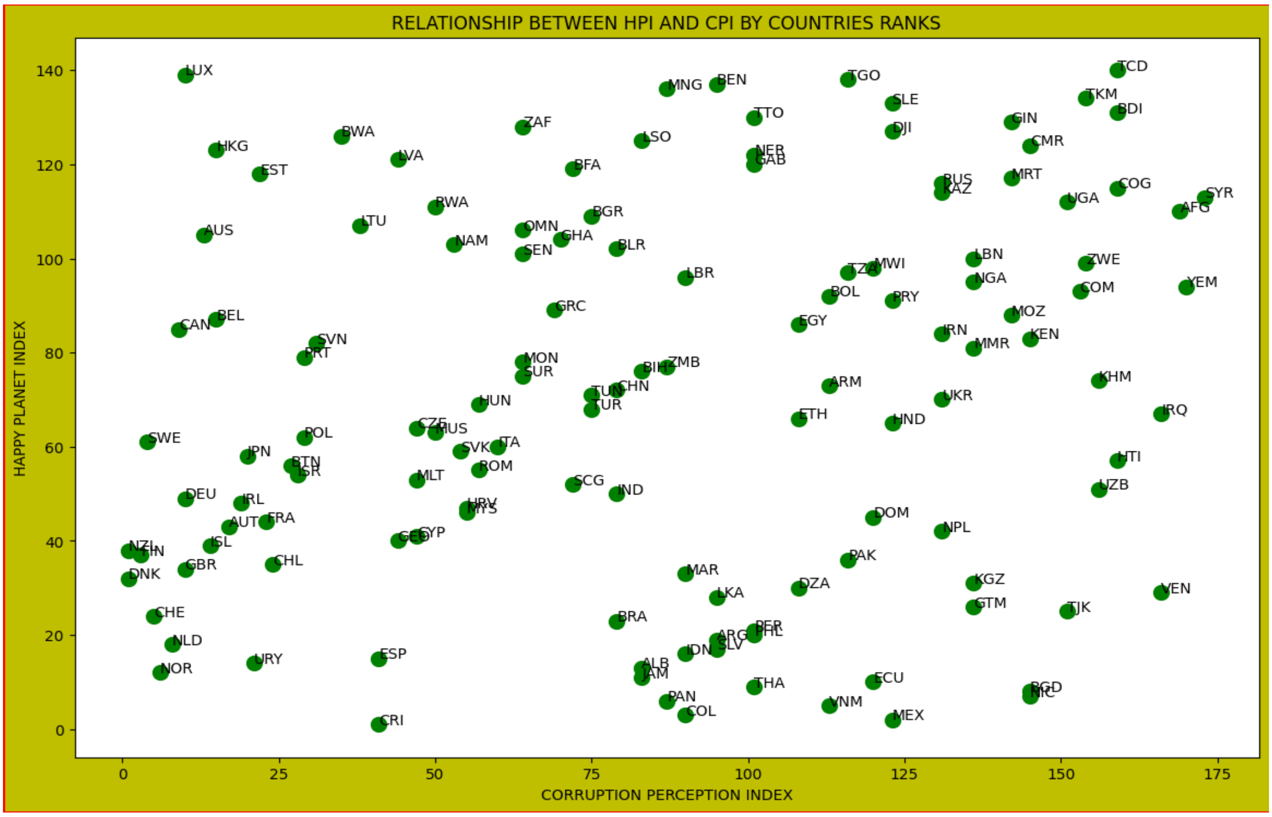


Figure 8: Scatter plot for the relationship between HPI and CPI by country ranks

The main results of this plot are that **Luxembourg** has an unhappy population even though it is not corrupt, **Venezuela** has more corruption but was well ranked on the happy planet index, **Costa Rica** is the happiest and **Tchad** is the least happy, **New Zealand** is the least corrupt and **Syria** has more corruption than any other country on the plot.