2023

Assignment 2

DATA, INFERENCE & APPLIED MACHINE LEARNING (COURSE 18-785)

YASSER ALBERT CORZO

CARNEGIE MELLON UNIVERSITY

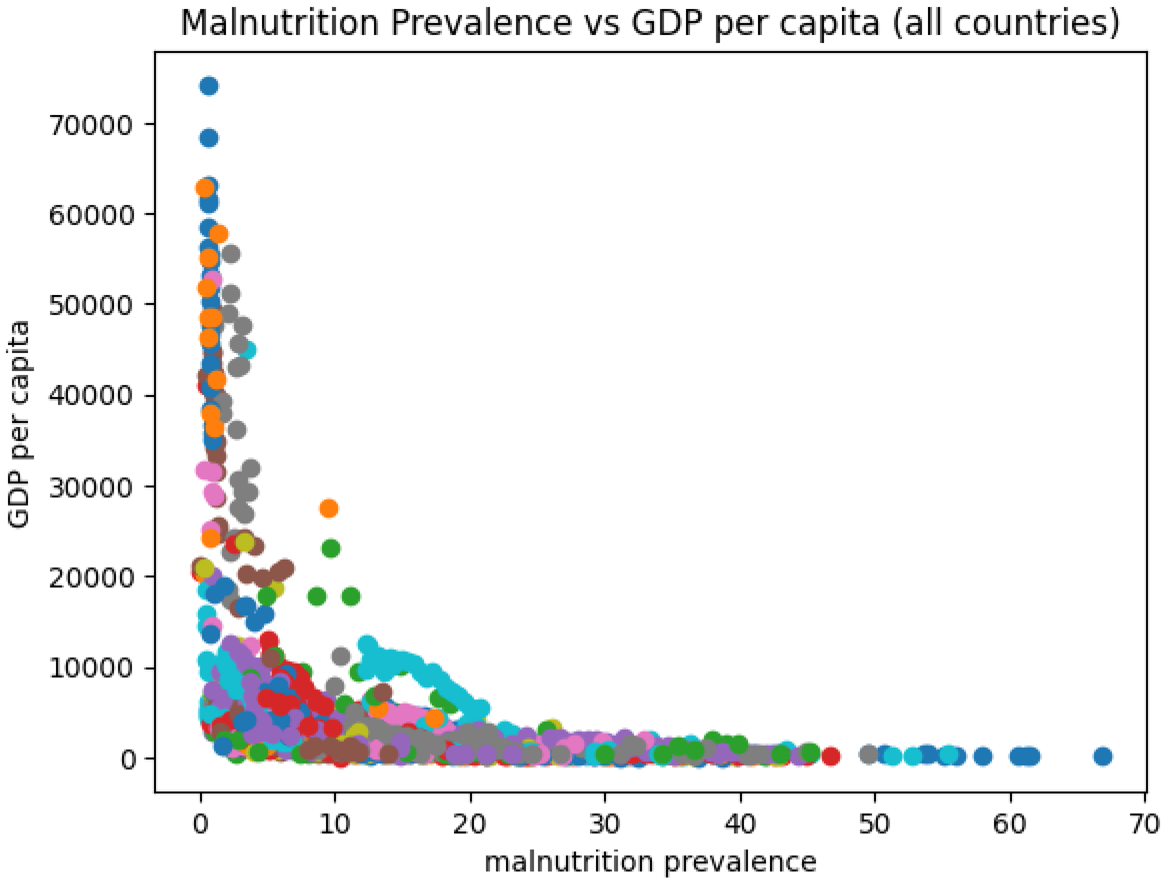
LIBRARIES USED

- Matplotlib
- Numpy
- Pandas
- Nasdaqlink
- Jupyter Notebook
- Tabulate

REPORT

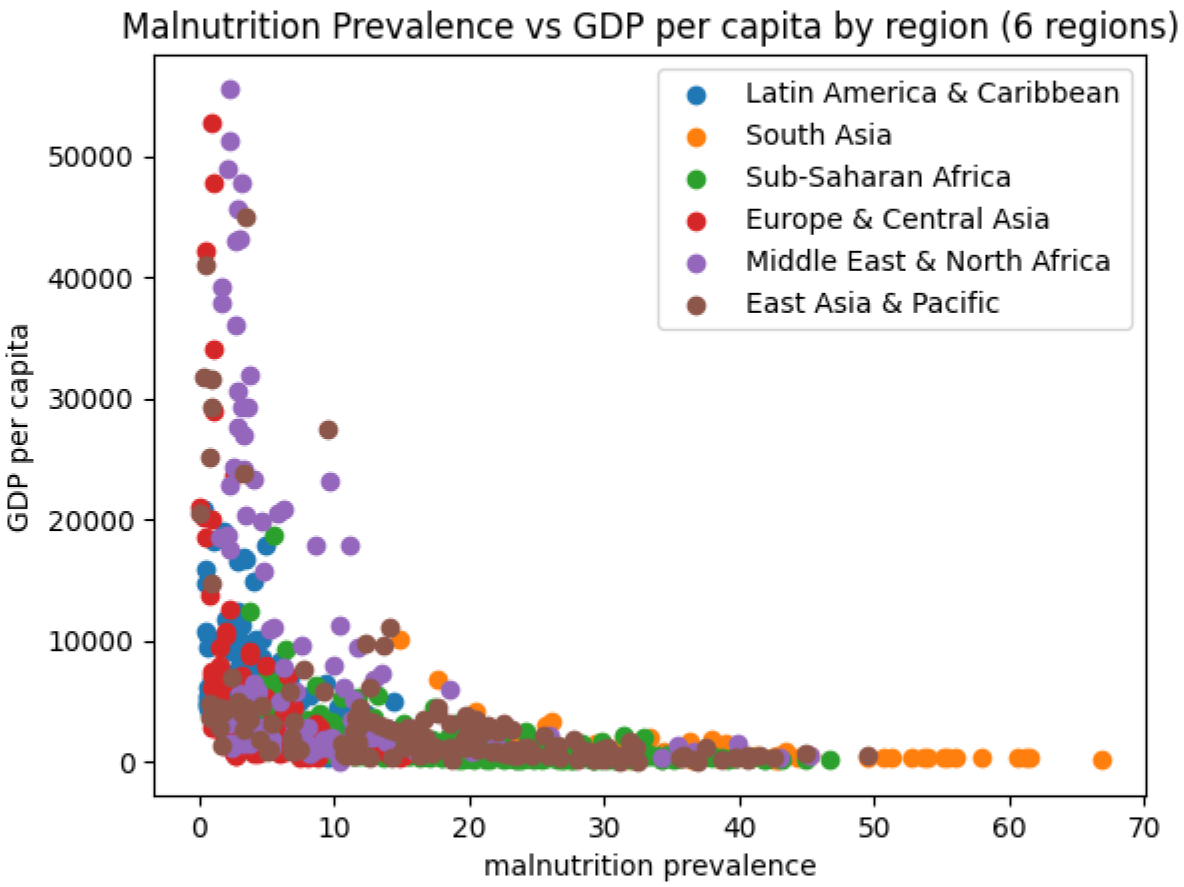
Q1

The kind of relationship that I expect between **GDP per capita** and **Malnutrition prevalence, weight for age (% ofchildren under 5)** would be the higher the GDP per capita of a country, the lower malnutrition prevalence that country would have and vice versa.



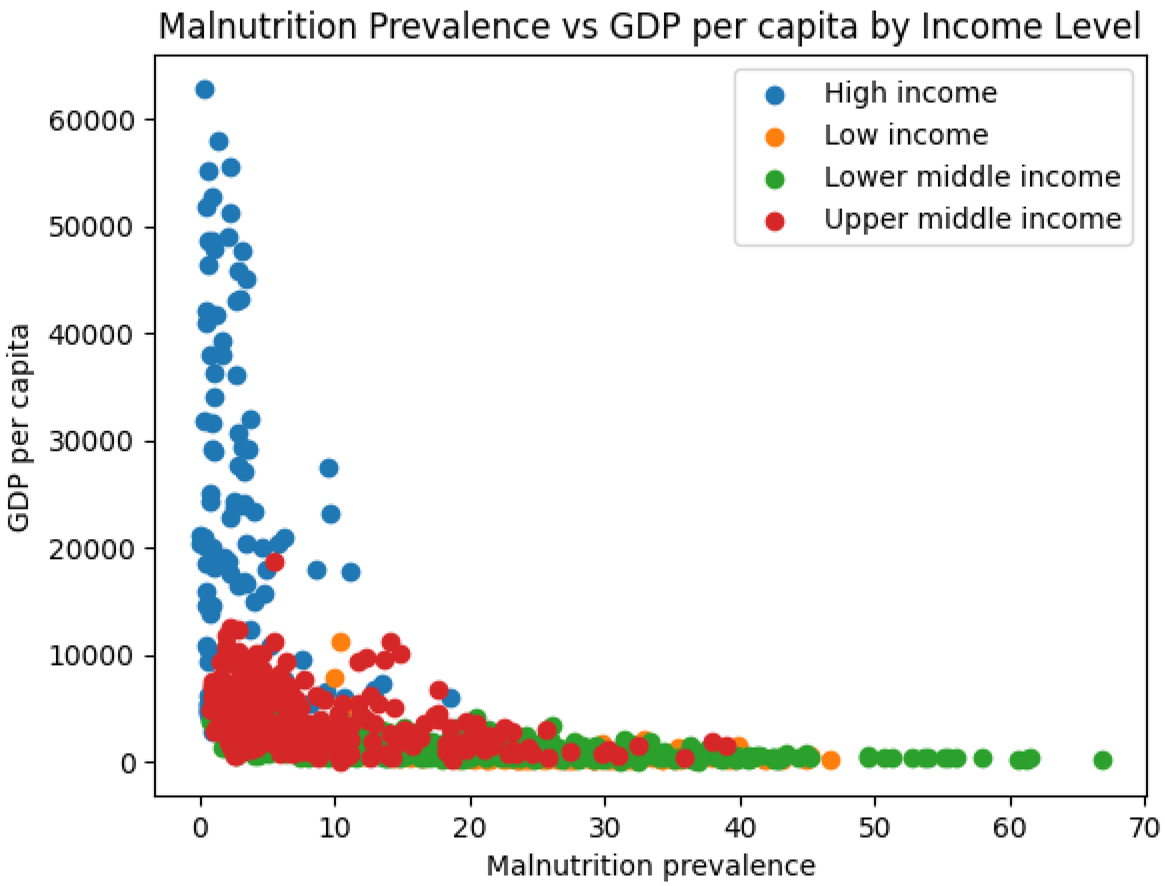
The relationship that I see in the graph of Malnutrition prevalence vs GDP per capita of all countries is that the lower the malnutrition prevalence, the higher the GDP per capita of that country is and vice versa. In other words, malnutrition prevalence is inversely proportional to the GDP per capita.

How I computed this graph was the first extract data only containing the GDP per capita of all countries and data pertaining to malnutrition, synchronizing these two datasets so they match up by country, then plotting every country's malnutrition and GDP across the years recorded.



Based on the above graph, the relationship I see here is between the regions and their GDP and malnutrition prevalence. Europe & Central Asia region have the lowest malnutrition prevalence with their GDP per capita being spread out, with Middle East & North Africa region having the second lowest malnutrition prevalence, meanwhile their GDP per capita is also spread out. Meanwhile, South Asia and East Asia & Pacific regions have the lowest GDP per capita, meanwhile their malnutrition prevalence varies. Sub-Saharan Africa and Latin America & Caribbean seem to have similar distributions in both malnutrition prevalence and GDP per capita.

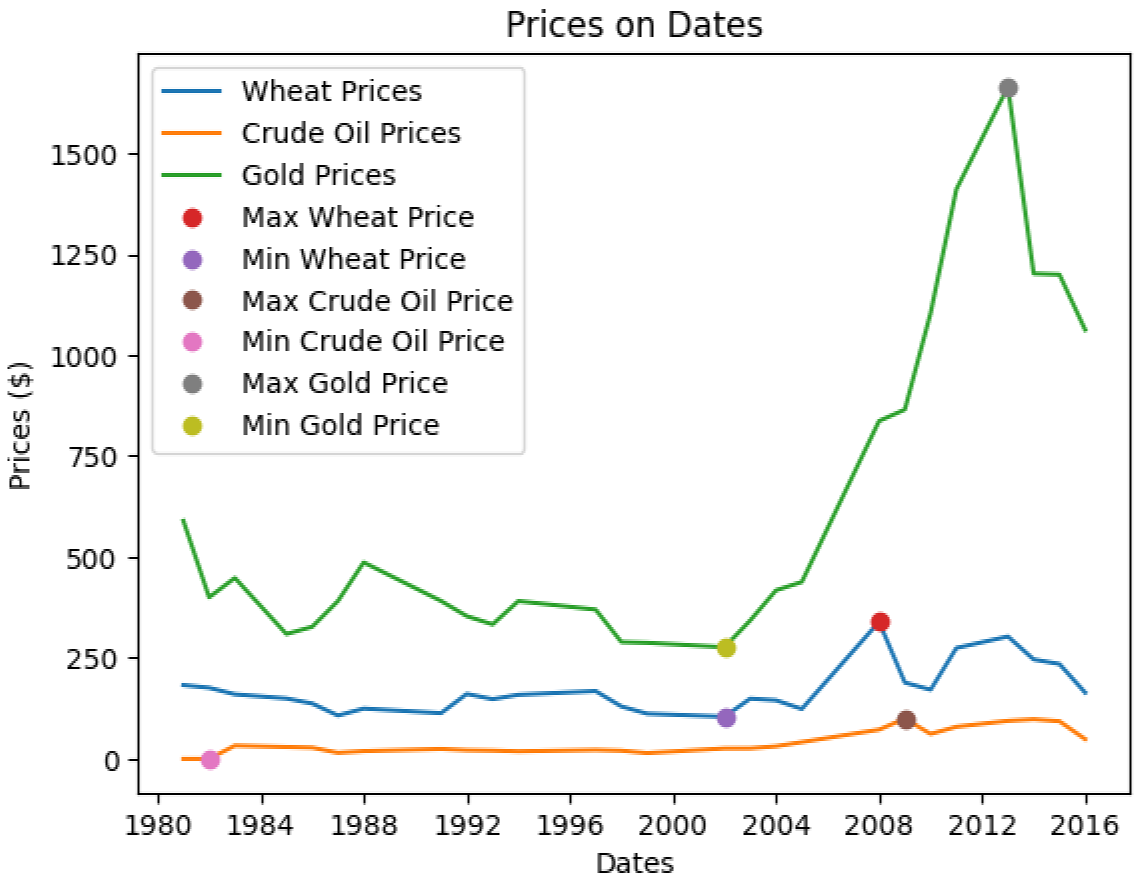
I created this graph of **Malnutrition Prevalence vs GDP per capita by region** was by extracting information relating regions and country's in those regions (by country code). There are country codes that do not pertain to a country. Rather, they pertain to a sub-region comprising of many countries but have no region placed. Therefore, data from these country codes could not be plotted as they don't pertain to one of the six regions that were requested to be plotted. From this I created a dictionary mapping regions to country codes. Since I already extracted information relating to country's GDP and malnutrition, I iterated through the regions, retrieving countries within that region (via the dictionary) as well as their malnutrition rate and GDP per capita and plotted these values.



Based on the graph above, higher income countries tend to have a lower malnutrition prevalence and higher GDP per capita. Lower income countries tend to have lower GPD per capita and higher malnutrition prevalence. Between lower and upper middle-income countries, GDP per capita does not vary much. However, upper middle-income countries have lower malnutrition prevalence compared to lower middle-income countries.

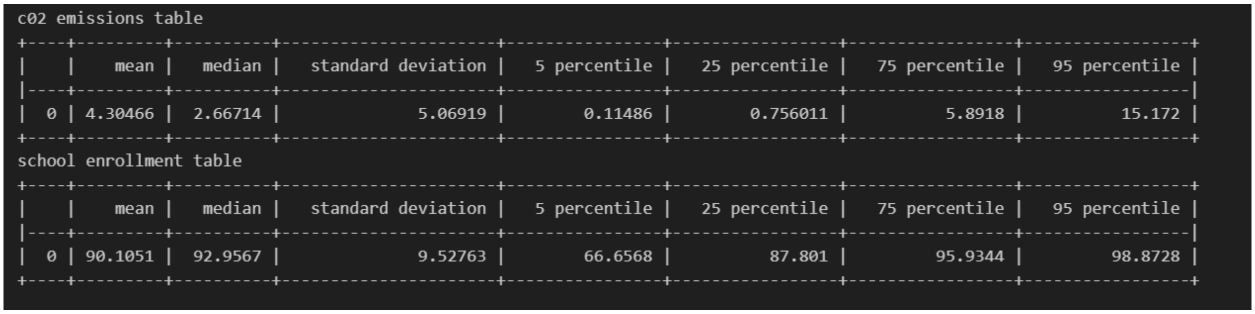
I created **Malnutrition Prevalence vs GDP per capita by income level** similar to the previous graph (i.e. **Malnutrition Prevalence vs GDP per capita by region).** Instead of a dictionary mapping region to countries, I created a dictionary mapping income level to countries. By iterating through the income levels and the respective countries within that income level, I could plot malnutrition prevalence vs GDP per capita by income level.

Q2



This problem involved extracting wheat, oil, and gold prices by date and merging all these datasets into a single data frame, matching all data points by date. Any dates that had missing wheat, oil, and/or gold prices were to be discarded and not considered to be plotted.

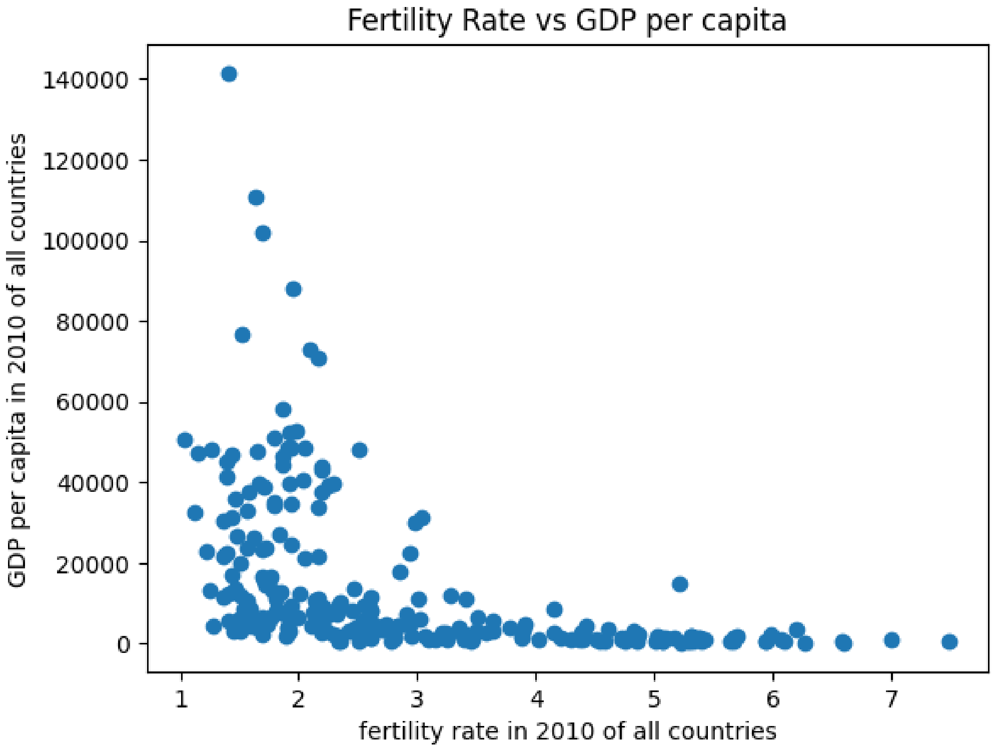
Q3



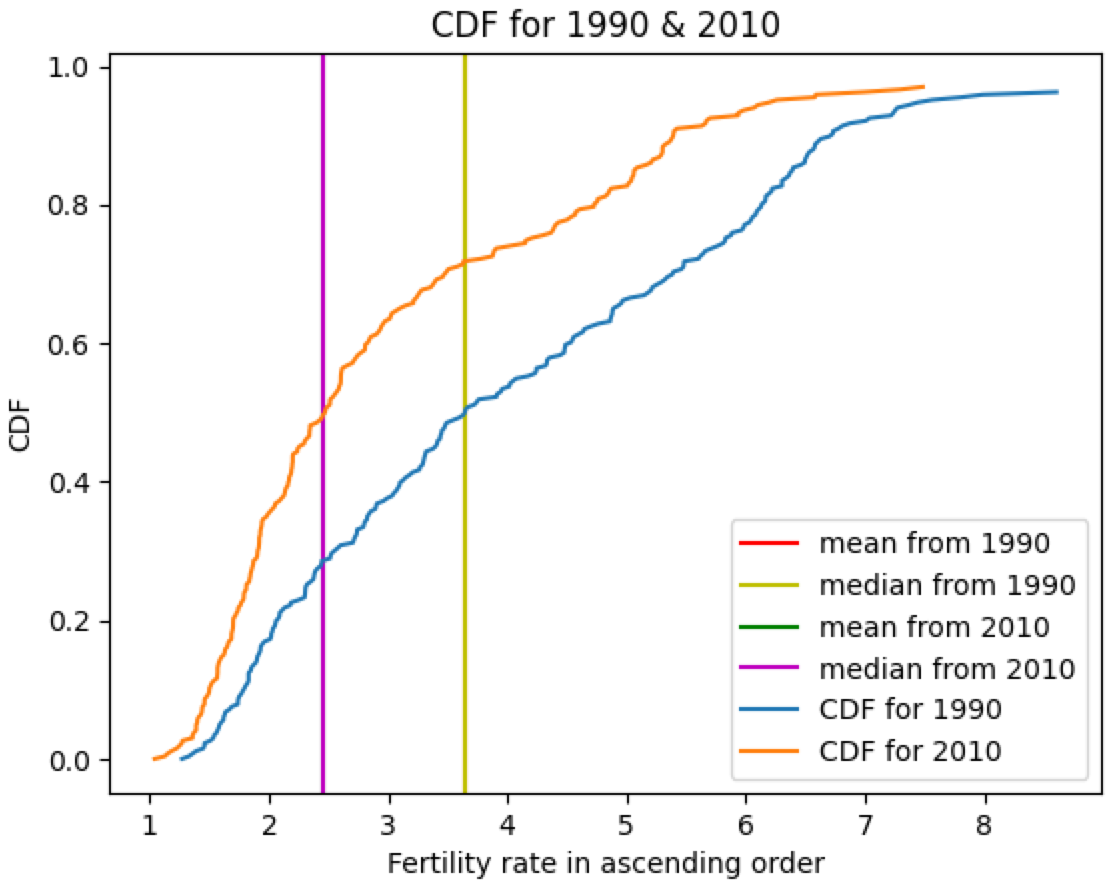
```
c02 emissions table
+----+---------+---------+---------------------+---------------+----------------+----------------+----------------+
|    |    mean |  median |  standard deviation |  5 percentile |  25 percentile |  75 percentile |  95 percentile |
|----+---------+---------+---------------------+---------------+----------------+----------------+----------------|
|  0 | 4.30466 | 2.66714 |             5.06919 |       0.11486 |       0.756011 |         5.8918 |         15.172 |
+----+---------+---------+---------------------+---------------+----------------+----------------+----------------+
school enrollment table
+----+---------+---------+---------------------+---------------+----------------+----------------+----------------+
|    |    mean |  median |  standard deviation |  5 percentile |  25 percentile |  75 percentile |  95 percentile |
|----+---------+---------+---------------------+---------------+----------------+----------------+----------------|
|  0 | 90.1051 | 92.9567 |             9.52763 |       66.6568 |         87.801 |        95.9344 |        98.8728 |
+----+---------+---------+---------------------+---------------+----------------+----------------+----------------+
```

The data presented in the tables were extracted by reading the C02 emissions excel file and extracting emissions from every country from 2010. There are ‘NaN’ values present as not every country has emission values. However, this didn't prove to be an issue as pandas [2] and numpy [3] methods for calculating the mean, media, and standard deviation ignore these values, thus the calculations are not corrupted.

Q4



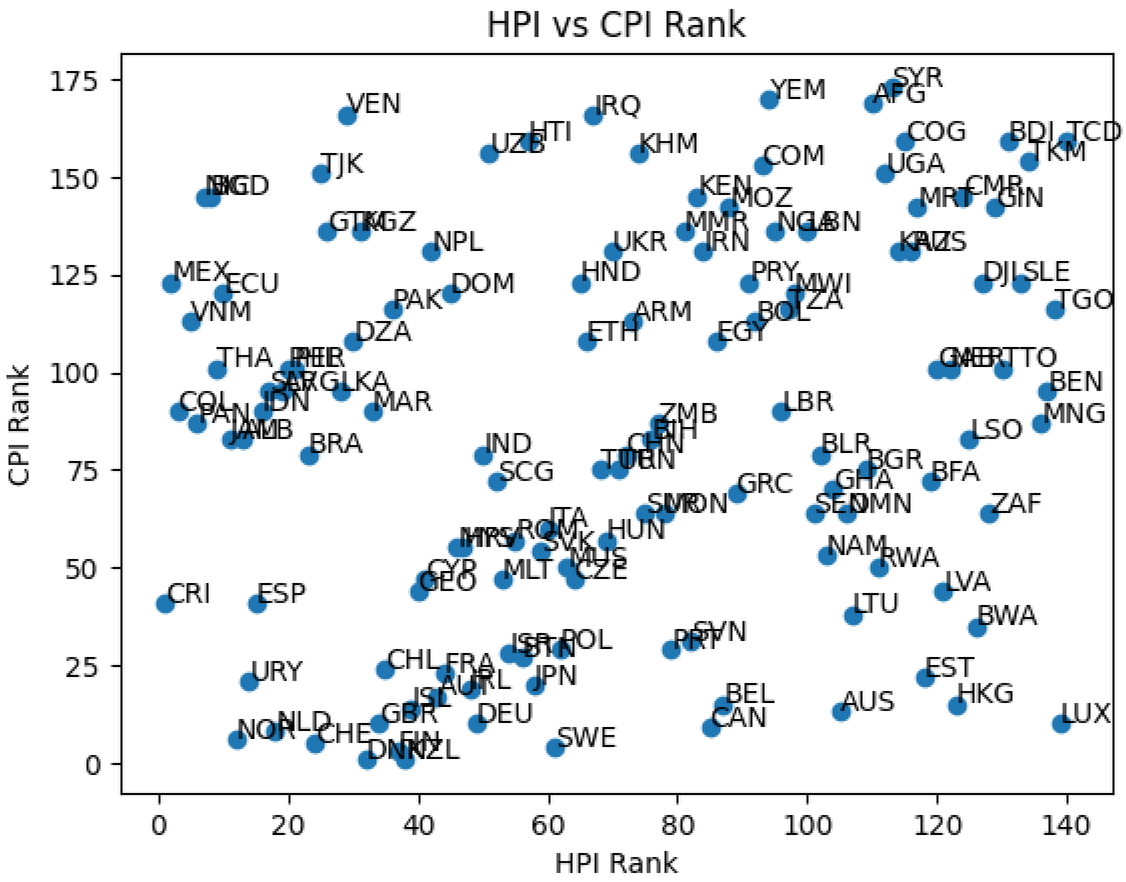
Plotting the above graph was very straightforward. From the 2010 column, all that was needed was to extract the GDP and fertility rate from 2010 and synchronize the data so that each country's GDP and fertility rate data were lined up and synchronized (by country).



Plotting the above graph required learning about the CDF function [1], involving sorting the data of fertility rates of all countries from 1990 and 2010 for the x-axis. Calculating the CDF for every fertility rate in the x-axis involved dividing the data number (order) of that point by the number of data points.

Over this 20-year period, fertility rates have not changed. The CDF functions of both 1990 and 2010 have similar slopes, meaning the distribution of sample points are similar in some respects. In addition, the mean and median of the distributions of both 1990 and 2010 are the same, indicating that birth rates haven't changed much.

Q5



Plotting the graph above involved extracting HPI and CPI rank of all countries listed, synchronizing the datasets so only countries that have HPI and CPI ranks recorded are plotted.

The country that struck me as unusual was Luxembourg (LUX). Having suck a low rank in CPI, one would predict that its HPI rank would be low as well as having a lower rank CPI implies less corruption, therefore a more functional government and more stability that could translate to a better living standard in the country and thus, increased happiness in the general population. However, that’s not the case for this country. Luxembourg’s HPI rank is higher than Syria (SYR), the latter of which is involved in a decades long civil war and whose CPI rank is much higher than Luxembourg, implying severe corruption in the government. It surprises me that the citizens in a war-torn country like Syria can be happier than citizens in a peaceful and less corrupt nation like Luxembourg.

REFERENCES

[1] “Take online courses. earn college credit. Research Schools, Degrees &amp; Careers,” Study.com | Take Online Courses. Earn College Credit. Research Schools, Degrees &amp; Careers, https://study.com/learn/lesson/cumulative-probability-distribution-formula-function-examples.html (accessed Sep. 10, 2023).

[2] Y. Follow, “Create the mean and standard deviation of the data of a Pandas series,” GeeksforGeeks, https://www.geeksforgeeks.org/create-the-mean-and-standard-deviation-of-the-data-of-a-pandas-series/ (accessed Sep. 10, 2023).

[3] “Numpy.percentile#,” numpy.percentile - NumPy v1.25 Manual, https://numpy.org/doc/stable/reference/generated/numpy.percentile.html (accessed Sep. 10, 2023).