Carnegie Mellon University Africa

DATA, INFERENCE AND APPLIED MACHINE LEARNING

ASSIGNIMENT 2

Course code: 18-785

Name: Ange Izabayo

Email: aizabayo@andrew.cmu.edu

Submission date on Friday, 18th September 2023

INTRODUCTION

Assignment 3 Composite of 5 question, i have used jupyter notebook to Answer them. In the first cell I declared different libraries I used to compute the task. imported pandas to help in data analysis manipulation[1]. Second, I imported matplotlib as library which helped in data plotting. Imported NumPy library to help in dealing with numerical operation. Additionally, I used quandl library to help interaction with online data on the website without downloading them. This are the library I utilized, and I have updated along the way while performing the task.

QUESTION 1

Q_1. A

Asked to use GDP per capita (current US $) and Malnutrition Prevalence weight for age (% of children under 5) dataset from World Bank Indicators. Based on my understanding the countries with low GDP per capita tend to have high degree of malnutrition, because if a person cannot make even 5$ per day in year 1,800 it is not easy to find the proper meal of their child. And I expect most of the African country to have low GDP per capita and high percent of child below 5 year have malnutrition. I used pandas to read both data set and then I extracted the column from 1960 to 2022 data matrices to be used in scatter plotting the relation between two matrices array. The figure 1 shows the relation between Malnutrition prevalence (%of child under 5 years) Against GDP per capita (US $).

After accessing the data, I extracted the data from column spanning from 1960 to 2020 using NumPy to create array that would structure the data I obtained for both GDP per capita and Malnutrition prevalence (% of child under 5 years). Then, after I utilized Matplotlib library to plot scatter graph illustrating the relationship between Malnutrition prevalence and GDP per capita.

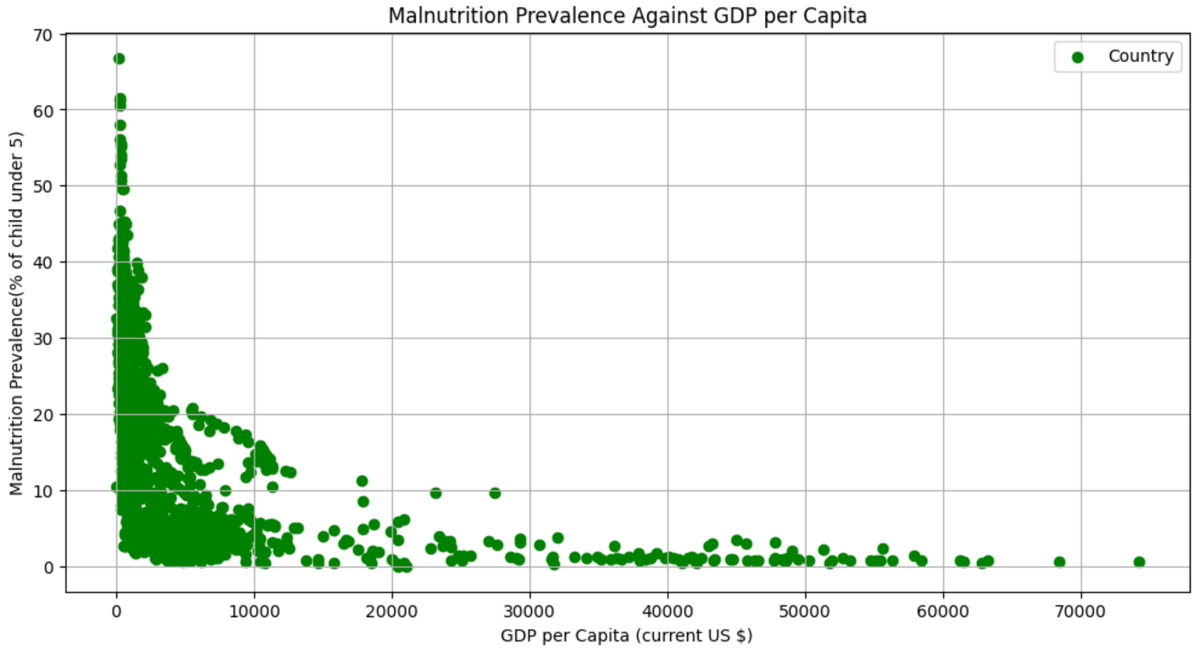


Figure 1.Malnutrition vs GDP per Capita

The figure.1 shows that the high number of countries, with populations earning between 0-10000 US $, have a malnutrition prevalence rate among child under 5 years ranging from 0-30%. This shows that most of the country's fall in this category population are in low-income level earn in range of 0-10000 US $ per year having low rate of malnutrition problem and some the malnutrition problem has been solved it is at zero. Countries with population earning above 60,000 US $ are categorized as high income, are less likely to face malnutrition issues. And on the graph, these countries malnutrition prevalence is at zero as shown in figure1. In summary, the graph illustrates that as the GDP per capita increases with malnutrition prevalence tend decreases, however it is important to note that there are still countries whose population faces the malnutrition issues for child below 5 year.

Q_1. B

The following question with the same dataset I was asked to include the Sheet called Metadata which contain other details data for both GDP per capita and Malnutrition prevalence. After loading both dataframe I merged both dataframe using pandas library to form single dataframe for GDP per capita and for malnutrition prevalence, then used code to form other excel datafile contain data and metadata on the same sheet for both malnutrition prevalence and GDP per capita. The task was to produce scatter plot of Malnutrition prevalence against GDP per capita containing all region excluding North America and show them within different color. After I used the merged dataframe and contain data and metadata dataframe to exclude north America from the column of Region. I assigned the color the region using dictionary after I created the loop that will iterate the dictionary of the color by mapping the Region with its designated colorm, where after removing north America I remained with six region (Latin America & Caribbean, Europe & Central Asia, South Asia, Sub-saharan Africa, Middle East & North Africa, East Asia & Pacific). Using matplotlib I created the graph shown in figure2.

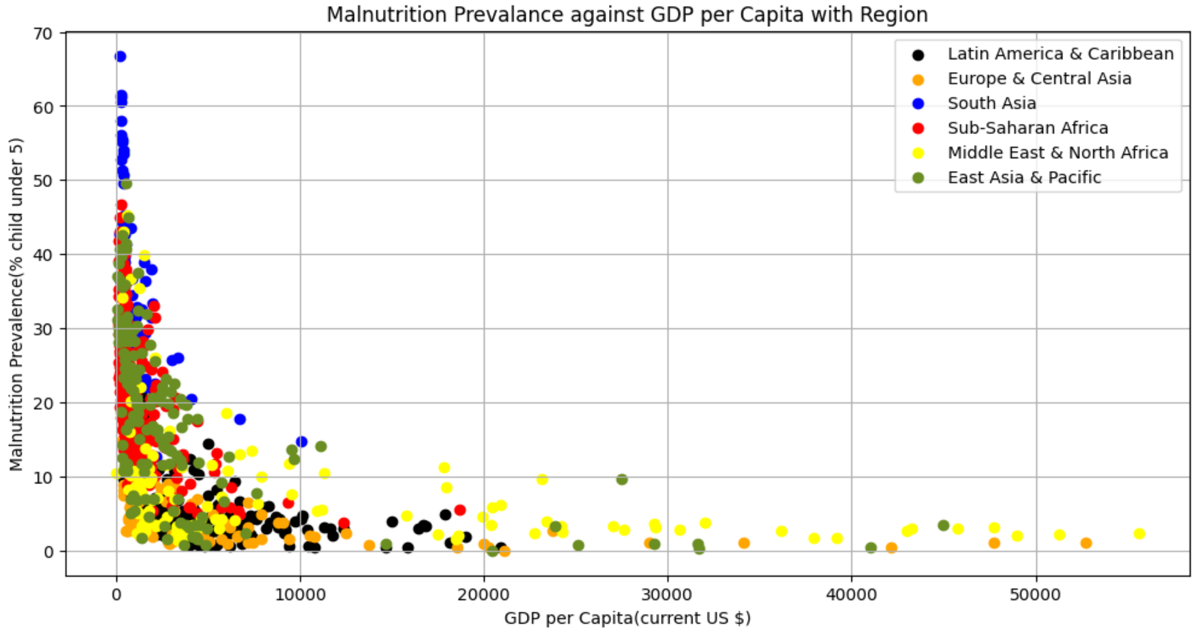


Figure 2. Malnutrition prevalence Vs GDP per capita with region

The figure2 shows that the region Sub-saharan Africa in red, South Asia in blue, Europe & Central Asia in orange, Middle East & North Africa in yellow, East Asia & Pacific in green. Based on the figure 2 the Sub-saharanAfrica, South Asia, East Asia & pacific, Latin America & Caribbean region are the most region where population earn between 0-10000 US $ and experience the malnutrition problem, where it is very hard in South Asia Region there have high degree experience of Malnutrition obove 60% which is the big issues, and Europe & and Central Asia most of the population earn between 0-10000 US $ but there are population have less or no Malnutrition problem compare to Sub-Saharan countries and South of Asia Region. For the above 50000 US $ earning in Europe, Few of East Asia & Pacific, and Middle East & North Africa have low or no number of Malnutrition prevalence for child under 5 years.

Q_1. C

Using the same dataset of Malnutrition Prevalence (% children under 5) and GDP per capita. I was tasked to produce a graph that categorizes countries based on income level, with four income level. To achieve this, I utilized a Merged dataframe from previous sub question, contain the Column of income level. I created a dictionary to map the income level to specific color. Then After I used for loop to iterate through the income level with marking the countries with the same income level colors. Finally, I used matplotlib library to create a scatter plot that illustrate Malnutrition prevalence against GDP per capita with legend showcasing the region and colors that differentiting them as shown in figure 3.

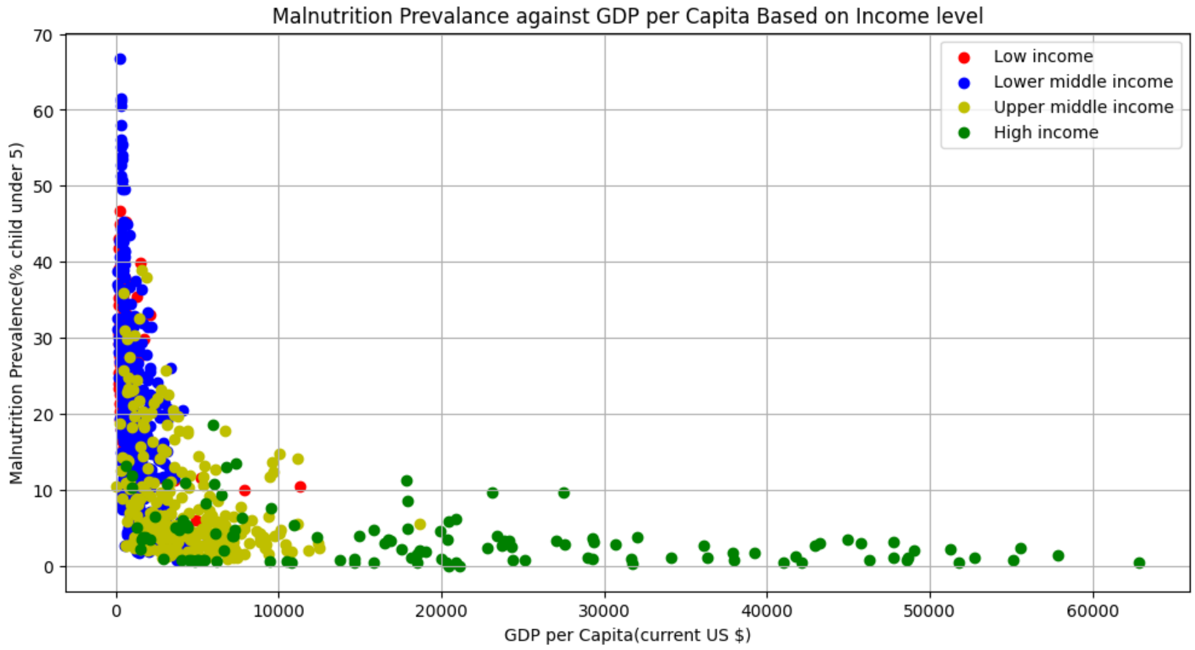


Figure 3. Malnutrition vs GDP per capita in relation to income level

The figure3 illustrates the distribution of countries population based on the income level in graph of malnutrition prevalence against GDP per capita. In the graph, countries with low incomes are represented by red color, indicating that they earn relatively small income and face a higher risk of Malnutrition prevalence among children under five years. Following closely are the countries colored in blue, representing the low middle income level who also fall under the same problem of facing malnutrition problem and has countries with malnutrition prevalence above 60% which is problem. Followed by countries colored by yellow color, representing upper middle-income level experience the malnutrition problem but at low rate. Figure 3 shows that these countries generally have malnutrition prevalence between 0-20. For high income countries experience a significant reduction in facing malnutrition problem where it is reduces below 10 percentage as represented by green color.

In conclusion, there is clear relation between income level and the degree of malnutrition. As the income rises, malnutrition prevalence reduces proportionally. Therefore, it is importance to put effort in mobilization and teaching how to combat malnutrition, particularly for the population in the low, lower, and middle income. And educating them benefit of how crucial is to avoid malnutrition.

QUESTION 2

To access data from Quandl, I first created an account on the platform to obtain an API key. This key served as the authorization to access the datasets I needed to perform the assigned task. I used individual code for accessing Wheat Prices, Crude Oil Prices, and Gold Prices, where each has its code to be used to retrieve these dataset. Once the datasets loaded in my Jupyter notebook, I read them and then merge them on date. I merged them to ensure full synchronisation to have common starting time which was 1982 for all Wheat prices, Crude oil prices, and gold prices using pandas. As the task was to visualize the time series for these prices and identify and then point out the minimum and maximum price across all three-column merged. To achieve this, I used built-in function idxmax and idxmin function to point out the indices or position in time when the prices reached their highest and lowest price, respectively. I utilized matplotlib library to create the time series variation plot for Wheat prices, Crude Oil prices, and Gold prices against Date, additionally I used scatter plot to mark the on the graph line the time when the price was minimum or maximum for all three lines as shown in figure4.

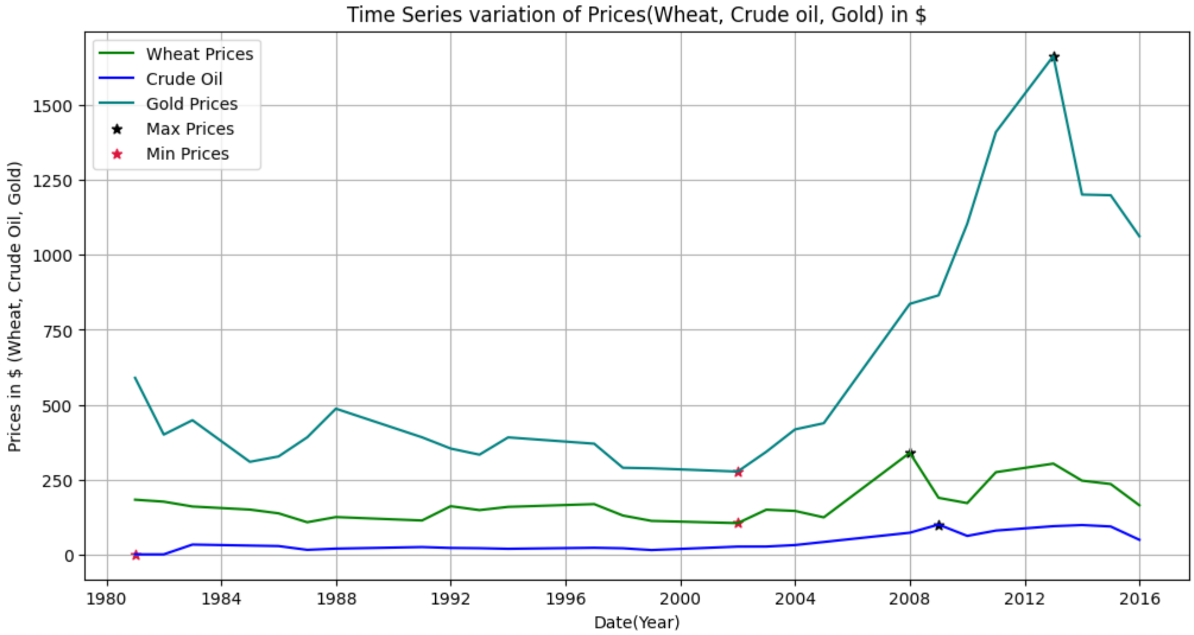
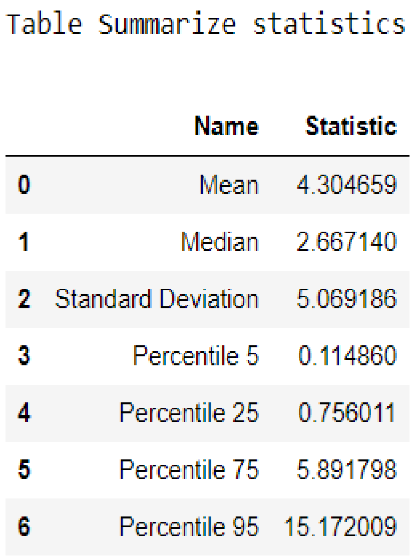


Figure 4 Prices vs variation on time

Synchronising timeline, as shown on the figure 4 starting from 1982, Blue color indicate the crude oil variation and as the time passes its price has been increasing slowly. Green color represent the Wheat prices variation with time which shows how it has been changing since 1982 where it has changed gradually sometimes prices increases other it fall little. The Gold prices compared to other prices has not been stables sometimes rises for example in between 2000-2004 is where we get minimum price of gold, and it is also the time its prices raised continued increasing until it reaches its maximum and then it fall again. According to the figure4 once the price of the either Wheat, Crude oil, or Gold reaches the minimum prices it's price automatically increases and once it has reached to the peak it falls directly. From this task I have learned how to mark on the graph and learned to use Quandl to access remote data without need of downloading them.

QUESTION 3

I begin by downloading CO_2 emissions (metric tons per capita) and School enrolment primary (%net) data from world bank databank, both in excel format. Using pandas, I imported these two dataset into my jupyter notebook, from excel sheet called Data. Since its two dataset I created two different dataframe to store those data. From dataframe of CO2 Emission and School enrolment I extracted the column of 2010 as instructed to be the one to use, then by return the data it contains some missing values (NaN). To calculate the average, standard deviation and percentile would not be possible due to adding nonvalues or dividing it. In that case I handled the missing value using built-in function dropna () to remove the row with that has NaN. Then to Compute the statistics of the data in 2010 column I used built-in function to calculate them. I calculated the mean, median, standard deviation, and percentile of different stage (5, 25,75, 95) on both CO2 Emission and school enrolment data in 2010 column. I used two cell to distinguish different dataframe but with the same methodology as shown in figure 5 it is the table containing Summary statistic of CO2 Emission for countries in 2010, while figure 6 is the table summarize the statistics of data collected of School enrolments primary(% net).

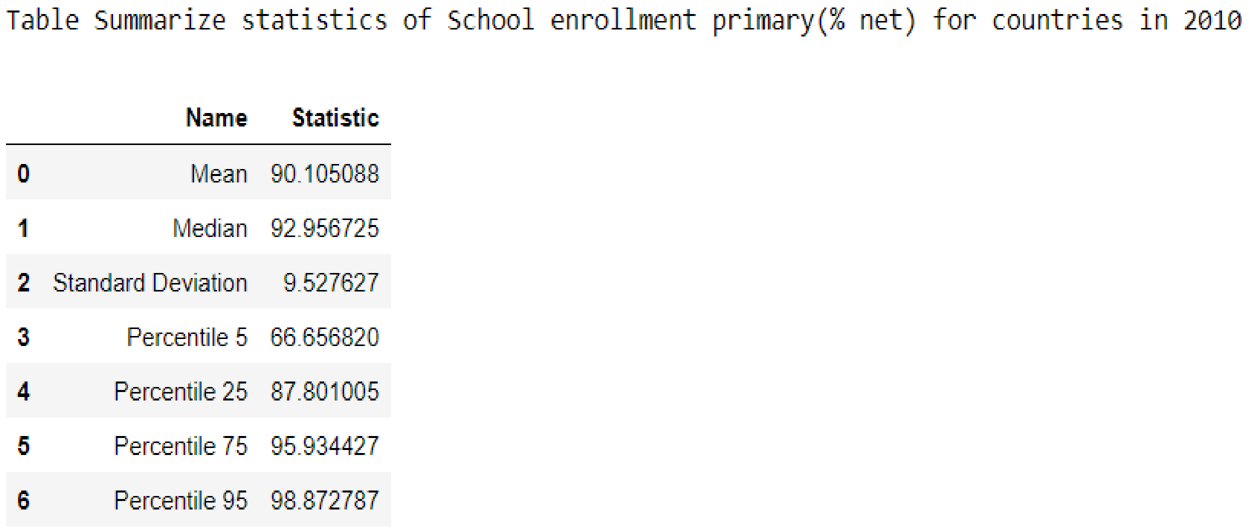


```
Table Summarize statistics of CO2 Emission for countries in 2010
```

	Name	Statistic
0	Mean	4.304659
1	Median	2.667140
2	Standard Deviation	5.069186
3	Percentile 5	0.114860
4	Percentile 25	0.756011
5	Percentile 75	5.891798
6	Percentile 95	15.172009

Figure 5.Statistics table of CO2 Emission 2010

The figure5 shows the summary statistics of Emission of CO2 in all country data has been collected in 2010. Where the average of CO2 emission for countries in the measurement is 4.3 mark the average of emitting in the atmosphere, while the median is 2.66 the middle value of emission. Standard deviation shows the spreading of the emitted CO2 percentile is showing by dividing into small part in the statistics in percentage for example 5 percent of CO2 emitted is 0.11. this statistics shows the level countries emit in 2010 where it continues to increases which lead to climate changes and respiratory diseases as standard deviation show it is very high countries where survey carried out tend to emit high level of CO2 and this was for 2010 for now it has raised.



```
Table Summarize statistics of School enrollment primary(% net) for countries in 2010
```

	Name	Statistic
0	Mean	90.105088
1	Median	92.956725
2	Standard Deviation	9.527627
3	Percentile 5	66.656820
4	Percentile 25	87.801005
5	Percentile 75	95.934427
6	Percentile 95	98.872787

Figure 6. Statistic of School Enrolment primary in 2010

The figure shows the summary statistics of School enrolments primary. How student enrolled in schools with specific in 2010 their statistics in different country survey has reached. Where the average of the student enrolled is 90.1 percent and the median was 92.95 percent. standard deviation or spread of enrolling in all country is 9.52. and by taking the part like 95 percentile of the data shows that 98.87 of enrolment. Since the Average of Student enrolment is very high to shows how children was encouraged to attend schools in different countries and its number was very high which is good thing and for having high number of student attending school.

QUESTION 4

Q_4. A

I Begin by downloading excel sheet of Fertility rate total (births per woman) and GDP per capita (current US $) from world bank indicators. I retrieve the excel file I downloaded to the Jupyter notebook using panda's library I imported in the first cell. The we had two task, the first is to make scatter plot of Fertility rate versus GDP per capita for all countries using data in 2010 column.

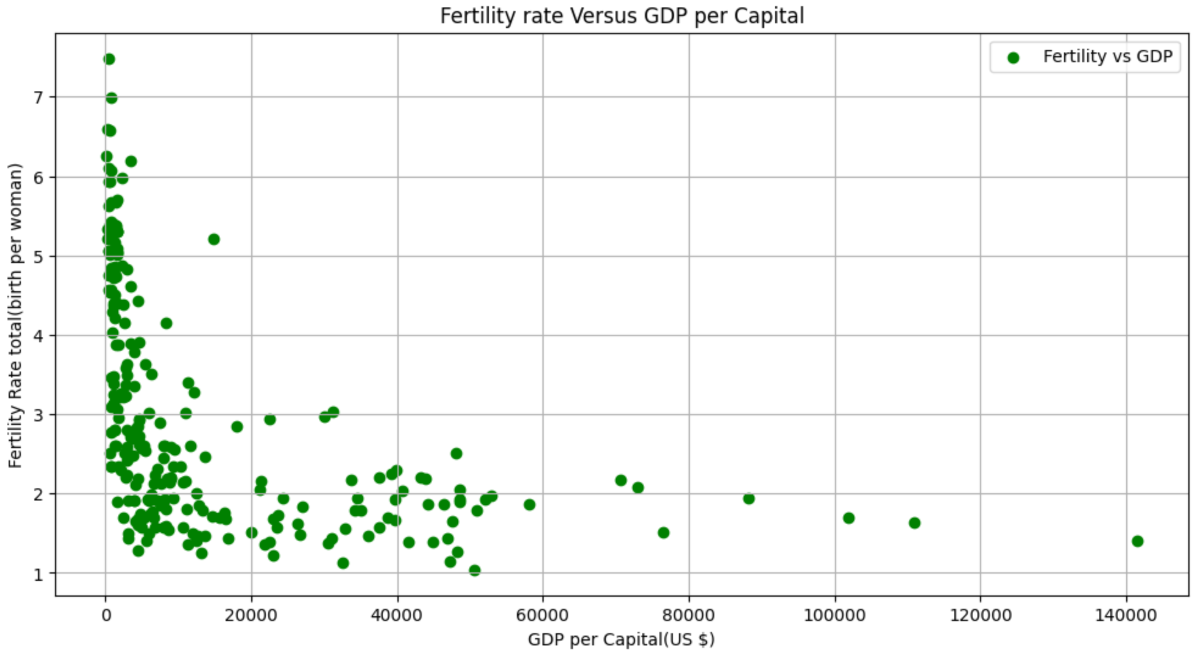


Figure 7. Fertility rate vs GDP per capital

This graph shows that for low-income people tend to give birth to many child as it shown in figure that the in the interval of 0-20000 GDP per capital countries tend to have many child birth compare to those with high income. For example for the developed countries the people tend to give birth to 1 or 2 children while here in developing countries mostly in sub-Saharan countries the family give birth to more than 5 children with no financial support there to help them. This was example to as shown in figure 7 for population whose GDP per capita is above 100000 US $ fertility rate is between 1 and 2.

Q_4. B

On the second task I was instructed to plot on the same axes Cumulative distribution function for fertility rate of both data of 1990, 2010. By Approaching this task, I used the same data extracted from fertility dataframe of 2010 and extract other of 1990 as instructed. By reading the data it contains the missing values which has to be removed in other to get the accurate Cumulative distribution function. Is used built in function dropna to exclude each row with NaN. Then I calculated the Cumulative distribution function for 1990 data and 2010 data using NumPy library to compute mathematical function.

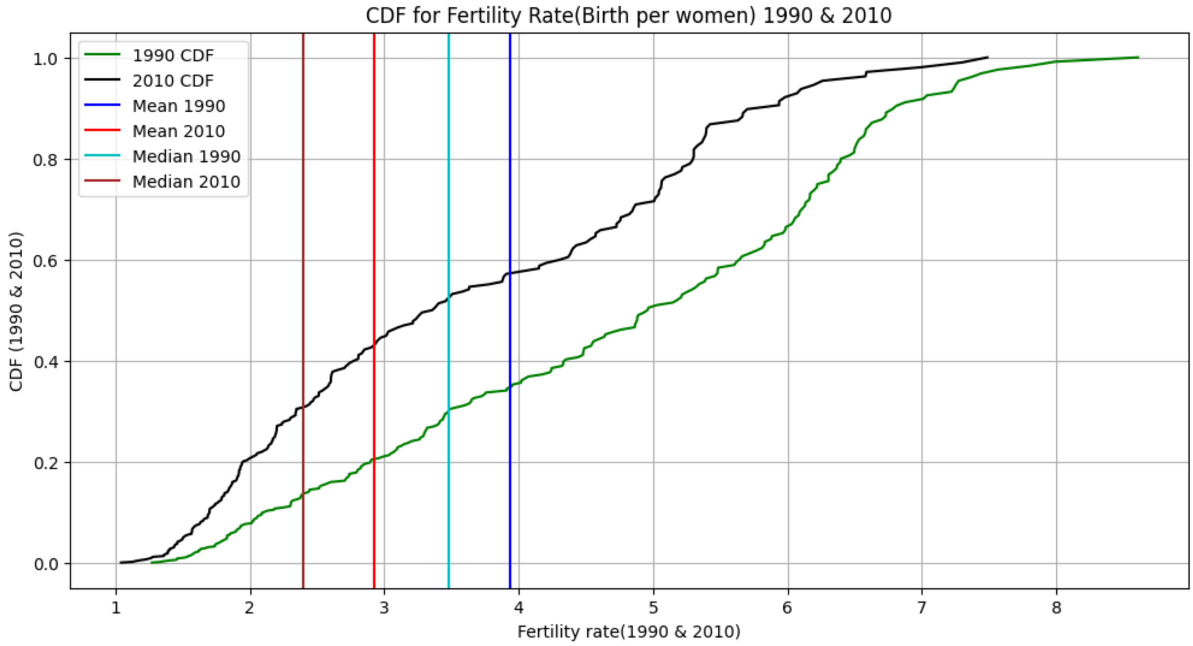


Figure 8. CDF (1990 &2010) against Fertility rate (1990 & 2010)

The figure 8 shows that after twenty years there is slight change where the fertility rate has reduced as shown in where the green line shows 2010 cdf against Fertility rate in 2010 which is low, and the average fertility rate is reduced from 3.9 in 1990 to 2.9 as shown in blue and read color of the mean.

QUESTION 5

I begin the question by downloading datafile of Happy Planet Index data and Corruption Perceptions index. Using pandas to read hpi and CPI excel file sheet specified in the notebook, due to each datafile has different sheet and for our task for HPI the sheet to use was Complete HPI Data and For CPI it was CPI 2015-2016. Through this dataframe loaded I extracted the column of Country for both dataset HPI Rank for HPI data and CPI 2016 Rank for CPI index. I used merging function to find the matching and combine 2 dataframe, merged on country and remain with single dataframe with 3 column. I proceeded by ranking the country based on the HP index data and CP index using the builtin function called rank(). After I utilized matplotlib library to make scatter plot shows the relation between HPI and CPI using rank. Using annotation to label the countries name in abbreviation and marking on the graph the point with country.

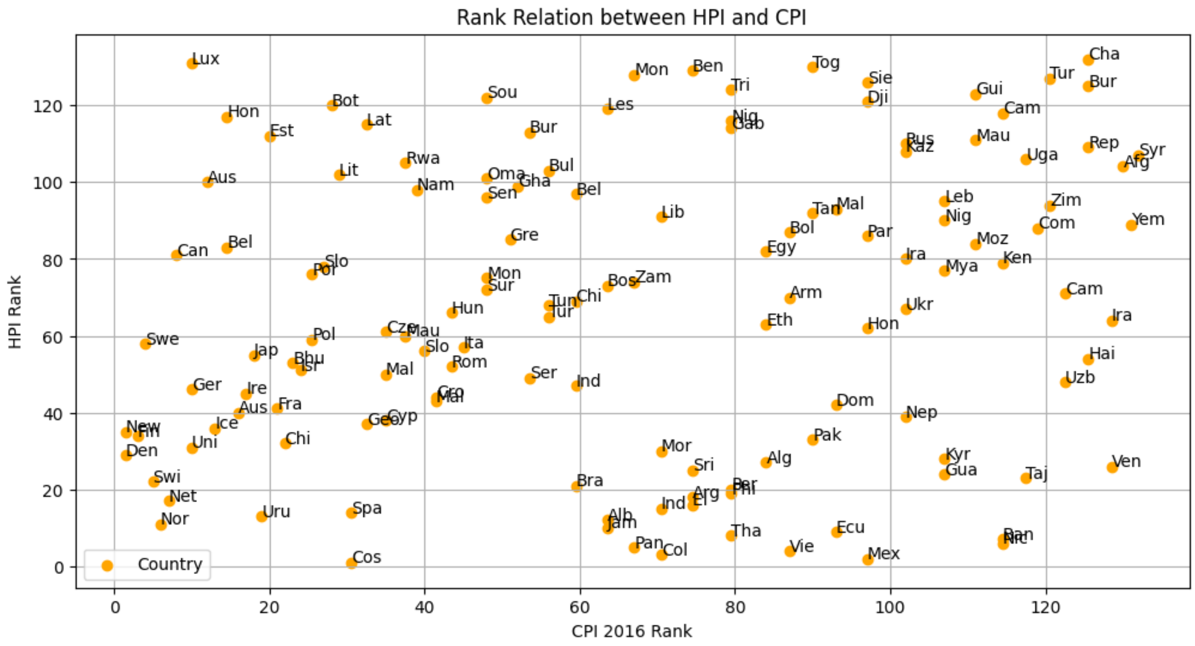


Figure 9. Relation Between HPI and CPI using Rank

The figure9 shows the ranking of the countries being happiness to statistic of corruption. Whereas indicated Luxembourg has happy population and low population, and Norway how low number of Corruption but in term of happiness it is still low. Sweden it stands to have low corruption rate and middle in happiness about 60. Chad, Syria Yemen   stand as unusual because on the plot shows their have happy population while there the country has wars in it which I think its unusual, country like Venezuela, iran has high degree of corruption.

Reference

[1] "Pandas Functions in Python: A Toolkit for Data Analysis." https://www.geeksforgeeks.org/pandas-functions-in-python/ (accessed Sep. 16, 2023).