REPORT

Exercise 1.

In brief, this question focused to study the relationship between malnutrition prevalence and GDP per capital.

Steps followed:

1. First thing is to have data, for this question data were obtained from world bank DataBank[1] were by visiting the site (https://data.worldbank.org/indicator) I downloaded the following Indicators:
    - GDP per capita (current US $)
    - Malnutrition prevalence, weight of age(% of children under 5)
2. After having the data downloaded the **first question** was asking **which relationship that could be expected between GDP per capita and malnutrition prevalence**.

Answer: the relationship I predict to be between those datasets is to follow a decreasing exponential function. Because I predict that as the increase in gross product per capita will stimulates and leads to high population's standard of living which in turn we contribute to downfall of malnutrition as people will afford to access quality of life.

3. After downloading the data, I defined the indicator code for both GDP per capita and malnutrition prevalence and I defined data frame from the indicators code.

   I could use to read excel files such as using pandas, the concern with them is that they provide unordered data which require some more cleaning and preparation.

4. I defined the pandas dataframe by using getcountry() so as to get countries information that later helped to remove aggregates from data. In the data given for countries were mixed with aggregates so I had to first remove aggregates so that the information about countries will not be biased.
5. Then, important targeted data were extracted from the data frames. In this case, countries and their malnutrition values or their GDP value on another hand.
6. Then next step was to sketch the graph of GDP values against malnutrition values as the period passed by and then observe and analyse the relationship.

   I used matplotlib module and its scatter function to plot the diagram

   Interpretation of the graph

   The graph can be viewed on the next page. The graph shows the relationship between malnutrition prevalence and GDP per capita. It just answers the question about how GDP affect malnutrition prevalence.

According to the graph, I interpret the relationship between malnutrition prevalence and GDP per capita as L- shaped relationship. The graph shows that malnutrition prevalence decreases with increase in GDP per capita. This is reasonable because as GDP raises, people obtain enough money and resources to improve their standard of living which in turns combats malnutrition.
The graph is L-shaped because at the same GDP levels countries may have different malnutrition level. Thus, around zero GDP many countries have different malnutrition level but note that the maximum malnutrition prevalence is observed at zero GDP
As the GDP rose from around zero countries malnutrition prevalence fell dramatically and malnutrition kept decreasing gradually as the GDP increased.

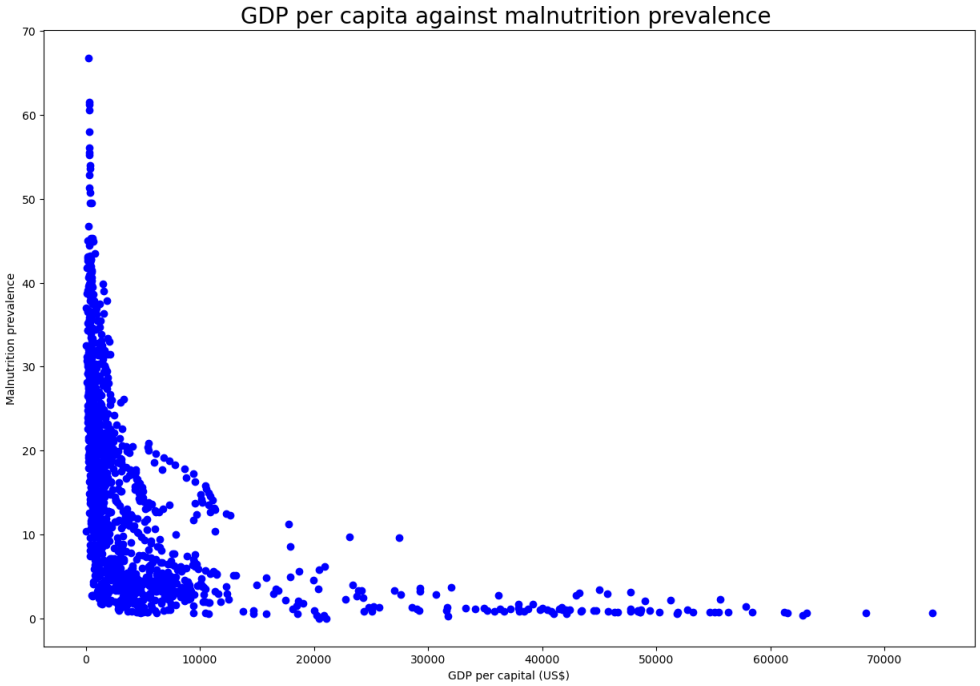


Figure 1: malnutrition prevalence against GDP

7. After this, the next task was to differentiate countries in their regions and show the malnutrition prevalence against GDP of those countries but classifying them in their respective regions but North America that is to mean that data from north America was not included in the analysis and in the graph as well.

8. The approach is used is to first read GDP 's and malnutrition metadata sheet and extract the columns of regions but excluding north America and them merge GDP values with the extracted column of regions and do the same for malnutrition values.

9. Then, all regions excluding North America were put in one list and the list of the colours to be used were initialised in the list.

10. Next, I used for each loop to iterate through each region in the region list while filtering out the data which does not belong to that region and then plot the data which belong to the

region and denote that scatter plot with a certain colour and repeat the process for all regions in the region list.

11. Finally, the graph was decorated by adding titles and axes label and then displayed on the screen. This is the obtained graph:

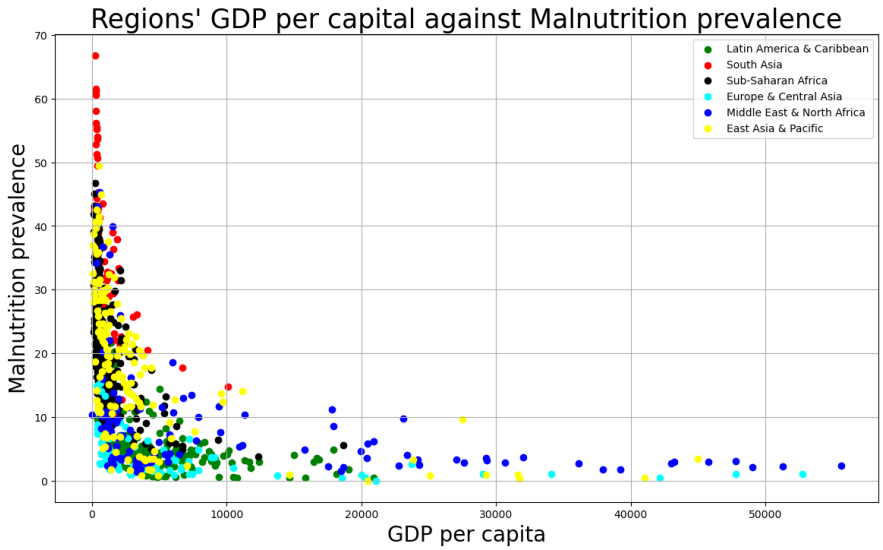


Figure 2: Malnutrition prevalence against GDP per capita for different regions

Interpretation

The graph shows that the countries in south Asia had the lowest GDP per capital as compared to other regions hereby the south Asian country with highest GDP per capital had approximately 10000 US $. Other than the GDP south Asian when the its GDP was almost zero, south Asian countries had the maximum malnutrition prevalence as compared to other countries.

Generally, as GDP per capital increases the malnutrition prevalence falls. Countries from middle east and north Africa distinguished them to be the countries with the highest GDP per capital and smallest number of malnutrition prevalence about 1 or 0.

Sub-Saharan countries also had the low GDP per capital and thus it had high malnutrition prevalence. The GDP per capital of Latin America and Caribbean countries was low but their malnutrition prevalence is not very.

The results that were obtained from this data is reasonable just as the GDP per capital increase the malnutrition prevalence reduce because people will be highly occupied by the economic activities to earn enough money which strongly will contribute to promotion of their living standard, hence the downfall in malnutrition prevalence.

12. The next step is now to classify the countries in their respective income level group. By accessing metadata the income levels were accessed and merged with GDP dataframe and malnutrition prevalence. The country in their income level were shown on the graph. The obtained graph was given as follows:

Figure 3: malnutrition prevalence against GDP per capita of different income level countries

Interpretation of the graph

High income countries had the highest GDP per capita and the lowest malnutrition prevalence. In contrast, lower income countries had the lowest GDP but the highest malnutrition prevalence. Lower middle and upper middle are in between high income countries and low income countries in terms of GDP per capita and malnutrition prevalence. The results from data is reasonable because in high income countries people acquire enough money and resources (this raise GDP per capita) to satisfy their needs hence raising their living standards thus mitigating malnutrition. In contrast in lower income countries, their population do not have enough money and resources to satisfy their needs thus contributing to high malnutrition prevalence.

Exercise 2

This was demanding to get data from quandl and then synchronize the data then, plot the time series of data in same graph showing maximum and minimum points on the graph.

Steps followed:

1. Download data: I accessed the data from quandl by using python code where by quandl module was imported and used get function from quandl module which required the indicator code as the parameter. I then cast the data from the quandl to csv file to be accessed anytime without limit exceeded errors.
2. Clean and extract the required columns from data frames. I dropped all NAN value by using dropna() method, changed the column called value in oil price data frame, gold price dataframe and in wheat price data frame to different names so as to be able to differentiate them in the combined dataframe including all the dataframes.
3. Synchronization: for the sake of synchronization, I merged all dataframes into one combined dataframe.
4. Find the maximum and minimum oil price, gold price and wheat price and their dates so as to be able to denote show them on the graph.
5. Plot the synchronized gold prices, wheat prices and oil prices against time period. This is the obtained graph

6. Using scatter plot show the maximum and minimum prices on the graph with different colour to easily recognize them.

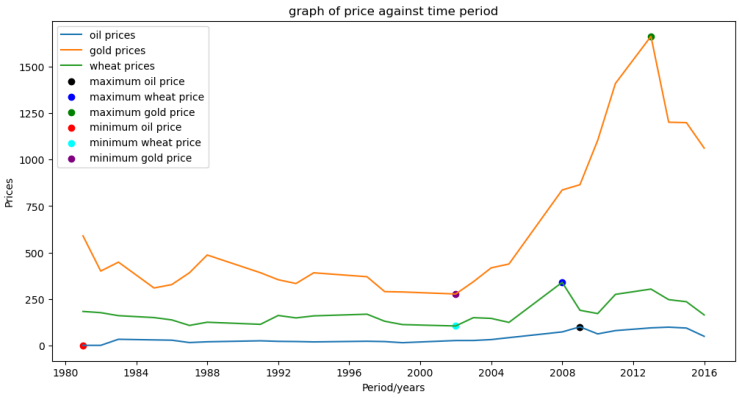


Figure 4: oil, gold and wheat prices over time period

Interpretation of the graph

This graph shows how gold price, oil price and wheat prices has raised over the past years from 1990 up to 2016.

It shows that gold has been expensive than oil and wheat since 1990. from 1990 to about 2002 gold price was fluctuating about 300, going a little higher and a little lower sometimes. From then, the price started to rise with approximately a recognizable increase in the rate per each year. And there was a sharp increase in gold price in the period ranging between 2008 and 2013. This was mainly due to economic and financial crises of the Great Recession which was dominant in some developed countries including USA[2]. When the economy is struggling, many investors turn to gold to protect their principal. Later from 2013, there was a sharp decrease in gold price but there after the decrease rate reduced. The overall trend for gold price is upward price meaning that on average, the price was rising.

In contrast, over the whole period wheat price did not rise or fall too much there was only a fluctuation of price of the period from 1990 up to 2016 though there has been a short-term slight increase of the price in 2006 up to 2008. The wheat price was lower than that of gold but also higher than that of oil.

The graph shows that the oil price was the lowest as compared to that of gold and wheat. Over the period oil price did neither rise nor fall significantly. There was only a very slight increase in the time period up to 2010 and then after there was a fluctuation of price.

The graph also shows the minimum and maximum value of oil price, wheat price and gold price. Oil price was zero in December 1980 and in December 1981, amazing to have oil price at zero!, but the data might be questionable. Note that there was two dates on which there was minimum oil price which is zero

The maximum and minimum point prices of oil, gold and wheat was shown in the table below

maximum and minimum prices and their dates

feature	date	price
maximum oil price	2008-12-31 00:00:00	99.5577
minimum oil price	1980-12-31 00:00:00	0
maximum gold price	2012-12-31 00:00:00	1664
minimum gold price	2001-12-31 00:00:00	276.5
maximum wheat price	2007-12-31 00:00:00	339.226
minimum wheat price	2001-12-31 00:00:00	104.37

Exercise 3:

The focus of this question is determine the summary statistic in the table of the given data

The summary steps that I used are discussed as below:

- ✓ From world data bank, download indicator of CO2 emissions and school enrolment in excel format.
- ✓ read the excel file in notebook by using pandas read_excel function and select the data for 2010.
- ✓ Prepare the data by skipping the first 3 rows for the sake of having column titles as the first row of the data frame.
- ✓ To get the summary statistics I used describe method of the data frames.
- ✓ Then I eliminated the value displayed from describe() method which were not added and I specified percentiles as they were given
- ✓ **Note that 50% percentile is same as mean.**

The obtained results for CO2 emissions are:

summary statistics for CO2 emissions in 2010

parameter	value
mean	4.30466
std	5.06919
5%	0.11486
25%	0.756011
50%	2.66714
75%	5.8918
95%	15.172

The average number of CO2 emissions in 2010 according to data is estimated to 4.30466 metric tons per capita and with all the characteristics as shown in the table. This predicted average is expected to rise for the following years up to now the average number of Co2 emission is forecasted to be above that one because as the time passes by many vehicles are manufactured and different industries and company are created, in turn cars and industries contribute to high co2 emissions in the atmosphere that is why co2 is expected to rise for the following years from then.

The summary statistics for school enrolment in primary was shown as in the table below:

summary statistics for school enrollment in 2010

parameter	value
mean	90.1051
std	9.52763
5%	66.6568
25%	87.801
50%	92.9567
75%	95.9344
95%	98.8728

The average value of student enrolment in 2010 is found to be 90.105 as shown in the summary statistics in the table on the right side.

Exercise 4.

Steps followed

1. For data , I was able to access fertility rate(birth per woman) and GDP per capita
2. I used indicators and their names as the parameter for wbdata's get module which permits the access to the data from world bank indicator and then I defined fertility rate data frame.
3. Then I defined countries information from wbdata.get_country() function and changed the country name from name to country for the sake of having same country name that I will use to combine with others.

4. Select fertility rate for 1990 and 2010 and sort them

5. Define a function for calculating cumulative density function, cdf: this function take the dataframe as a parameter the function will look under the column called "Fertility rate" from the inserted data frame the cdf will be the taking obtained from the following formula

Cdf= $\frac{df[\text{'Fertility rate'}].rank()}{len(df)}$ where len(df) is length or size of dataframe,

df =a dataframe

cdf= cumulative density function

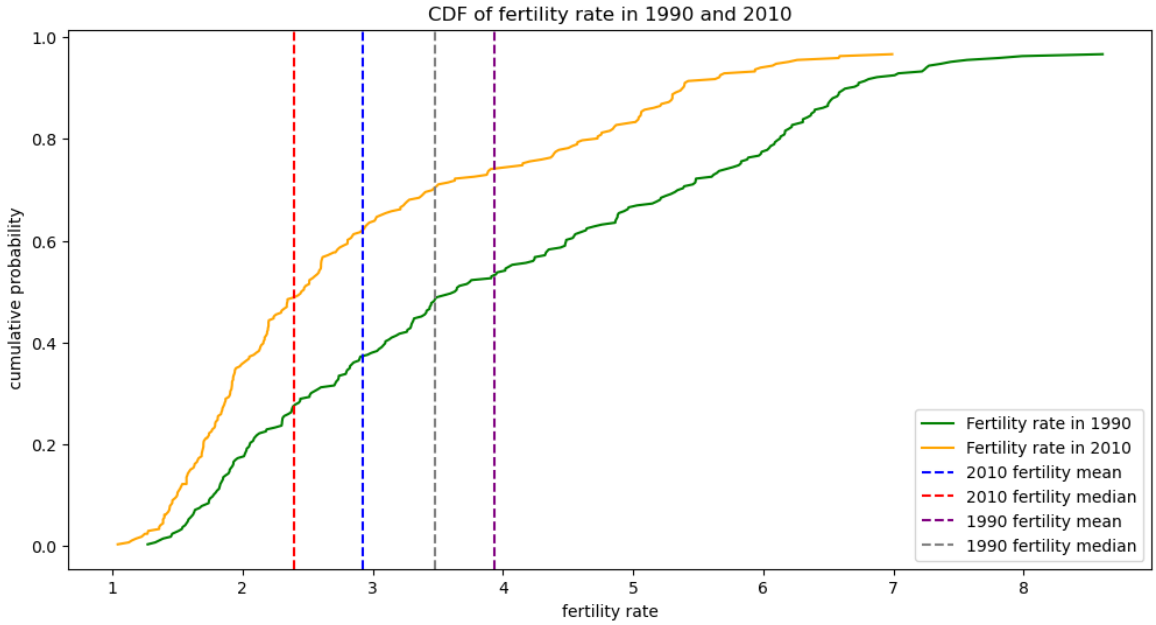
$df[\text{'Fertility rate'}].rank()$ calculates the rank of every element of column of fertility rate in data frame.

6. After defining the cdf function I used it to plot the cdf of fertility rate in 1990 and cdf of fertility rate in 2010. By using the following line of codes
7. I also used dashed vertical line to show mean and median of the distribution in 2010 and in 1990.

```
def cdf(df):
    return df["Fertility rate"].rank()/len(df)

plt.figure(figsize=(12,6))
plt.title("CDF of fertility rate in 1990 and 2010")
plt.plot(sorted1990_df['Fertility rate'], cdf(sorted_df),c="green", label="Fertility rate in 1990")
plt.plot(sorted2010_df['Fertility rate'], cdf(sorted1990_df),c="orange", label="Fertility rate in 2010")
plt.xlabel("fertility rate")
plt.ylabel(" cumulative probability")
plt.axvline(x= sorted2010_df['Fertility rate'].mean(), c="blue",linestyle="--", label="2010 fertility mean")
plt.axvline(x= sorted2010_df['Fertility rate'].median(),c="red",linestyle="--", label="2010 fertility median")
plt.axvline(x= sorted1990_df['Fertility rate'].mean(), c="purple",linestyle="--", label="1990 fertility mean")
plt.axvline(x= sorted1990_df['Fertility rate'].median(),c="gray",linestyle="--", label="1990 fertility median")
plt.legend()
plt.show()
```

Results



Interpretation

First of all recall that the cumulative distribution provides the probability that random variable is less than or equal to a certain value x and it is obtained by summing up all the probabilities below that points (for discrete variables) or applying integration on the probability distribution function on the range less than that x [3]

The graph shows the rate at which probability of having children from one child is changing

The orange curve shows that the rate of having children in 2010 was increased dramatically from 1 up to 3.5 and then the rate decreased it only rose slightly. This implies that probability for a woman to produce few children like 2, 3 children in 2010 is higher than probably of that she was tending to produce 5 or6 children. It shows that probability of having more than 6 children is approximately zero. Therefore, As shown by the vertical blue dashed line, the average number of births per woman was about 3 in 2010.

In contrast, going back in 1990 the data shows different things as they can be visualized on the green curve. The increase rate in cumulative probabilities of births for few number of children (2,3) is close that of many number of children like 6 or 7 children. This is to mean that probability of having 5 is quite the same with probability of producing 2 children per woman. The average number of children per woman in 1990 was predicted to be about 4 as shown by the purple dashed line. Here the probability of that a woman bear more than 6 children per average is not zero like it was in 2010 but for here the probability of having more than 7 children is approximately 0

Briefly, the curve in 2010 is more steep and the cummulative density function get to 1 for fewer children than in 1990. This implies that as years went by fertility rate is diminishing whereby the average number of children per woman tends to be smaller. For 1990 The cumulative density gets to 1 at approximately 7 fertility rate but for 2010 the cumulative density function get at 1 at around 5.8 fertility rate.

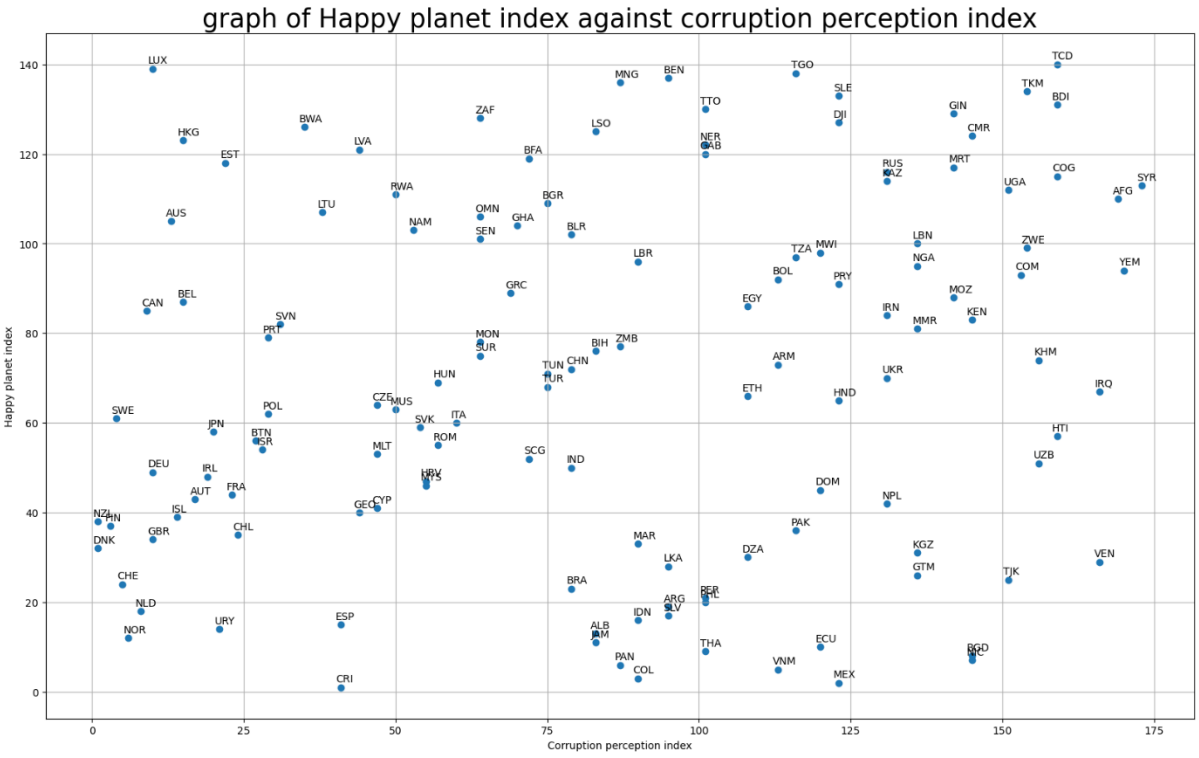
Yes, the fertility rate has decreased as the years passed by.

Exercise 5.

For this question we used data for happy planet Index [4]and that for corruption perception index and we plotted HPI against CPI[5] scatter plot to demonstrate the relationship between ranks in both CPI and HPI.

Results:

Let ‘s analyse the following obtained graph



Insights

The graph shows costarica is the first on happy planet index that is to mean it is the first country with happiest country in the country samples used in the study and there is no much corruption in the country.

Chad is proven to be one of the most unhappy countries and with the great percentage of corruption

Bangladesh is one of the happiest countries even though it has very large corruption percentage

Somalia and South sudan are the first two country with too much corruption and its people are quite unhappy may be due to wars in that are frequent in those countries. but Somalia was not shown on the graph because it is not include in country which has data on HPI index.

On the graph Syria is the first country with highest corruption and its population are quite unhappy.

Although Luxembourg is among countries with low corruption it is among the three top unhappy countries

DNK IS the first country with less corruption and it is among happiest countries in the world. According to the data provided.

References

[1] 'World Bank Open Data', *World Bank Open Data*. https://data.worldbank.org (accessed Sep. 18, 2023).
[2] B. Hergt, 'Gold prices during and after the Great Recession', 2013.
[3] 'Numeracy, Maths and Statistics - Academic Skills Kit'. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/distribution-functions/cumulative-distribution-function.html (accessed Sep. 18, 2023).
[4] 'Files'. https://canvas.cmu.edu/courses/36029/files/folder/Assignments/Assignment%202?preview=9906024 (accessed Sep. 18, 2023).
[5] '2016 Corruption Perceptions Index - Explore New Zealand's results', *Transparency.org*, Jan. 31, 2023. https://www.transparency.org/en/cpi/2016 (accessed Sep. 18, 2023).