DIAML Assignment 1 Report

Georgina Nyokabi Njoroge
gnjoroge

September 2025

Libraries Used

`pandas, numpy, matplotlib, tabulate, stats from scipy, math, ploty, seaborn`

Compiled on September 25, 2025

Question 1

Methods

For this sample, I created an array to hold the values. I used built-in Python methods to calculate the mean and standard deviation. These functions are by default for samples; hence, I did not adjust the parameters.

Seeing as this is a two-tailed test, I divided the significance level by two to account for critical values at the lower and upper ends. I used the built-in statistics library to get the T-table value when calculating the critical value and the p-value.

Results

The null hypothesis for this experiment is

$$H_0 : \mu = 7725kJ \tag{1}$$

The alternative hypothesis for this experiment is the following.

$$H_1 : \mu \neq 7725kJ \tag{2}$$

Since we are testing for both low and high values, greater or less than, we will use a two-tailed test. This will inform us of any deviation in either direction.

The following are the values calculated from the provided sample.

```
Sample Mean: 6753.636363636364
Sample Standard Deviation: 1088.9717646166182
Standard Error Mean: 328.3373409567765
t-statistic: -2.9584318175114594
Degrees of Freedom: 10
p-value: 0.014326498687519784
Reject Null Hypothesis
```

Figure 1: This image shows the descriptive statistics and final decision on the null hypothesis for the sample provided

Discussion

The summary statistics show that the data is fairly spread as the standard deviation is high. The difference between the sample mean and the population mean is large relative to the sample error mean, 328.

From the results, we can see the following.

$$p{-}value < \alpha \tag{3}$$

$$0.0143 < 0.05 \tag{4}$$

Therefore, the null hypothesis, H_0, will be rejected, as there is enough evidence for the alternative hypothesis, i.e. the mean energy intake is different from the initial recommended value of 7725kJ.

Question 2

CO2 emissions excluding LULUCF(tons per capita), 2023

Methods

For this data set, I filtered the 2023 data from the World Data Indicators website. I converted the 2023 values from string objects to numeric values.

For the missing and malformed values, I handled them by dropping the affected rows. 5 of the rows had only comments, but 15 of them had malformed numeric values that were cast into NaN values.

Summary statistics, mean, median, and standard deviation were calculated using built-in Python functions. The values were recorded in a dataframe and printed using the tabulate library for visual enhancement.

Results

The table below shows the summary statistics for this data set.

	Statistic	Value
0	Mean	4.458792891604376
1	Median	2.60244663572093
2	Standard Deviation	7.1667785095483545
3	5th Percentile	0.0897049692485516
4	25th Percentile	0.724544016838403
5	75th Percentile	5.442895727960955
6	95th Percentile	14.1816381032734

Table 1: This table shows the summary statistics for the CO2 emissions excluding LULUCF(tons per capita), 2023

Discussion

From the results, we can tell that the median is lower than the mean. Half of the values are below 2.60 and 75% of the emissions are below 5.44. Although most values are small, there are large values that are raising the mean, as evidenced by the value of the 95th percentile. From this we can deduce that the data is skewed to the right.

The standard deviation is also high, meaning that the emissions vary a lot

between countries.

From this we can conclude that in 2023, most of the countries had little CO2 emissions; however, there were a few that had very high values causing the mean to be high. From the data, these high-emission countries include Palau with the highest emission at 81.2 followed by Qatar at 48.15.

Primary completion rate, total(% of relevant age group)

Methods

For this dataset, I downloaded the entire data set, including previous years. This is because the number of missing values in 2023 was much higher than in the CO2 emissions data set.

I imputed the missing values by forward filling as it was time series data. Most of the missing values were filled out from the values after 2014. The values in the dataset do not seem to have varied much in the last 10 years and therefore I believe that the data will still serve to inform us on the estimate for primary completion rate in 2023. Then 15 rows that still had missing values were dropped after imputation.

For the summary statistics, I calculated them similarly to what I did with the previous dataset.

Results

The table below shows the summary statistics for this data set.

	Statistic	Value
0	Mean	88.68434194739598
1	Median	93.4889602661133
2	Standard Deviation	16.552657861654403
3	5th Percentile	57.8701572418213
4	25th Percentile	80.45396041870114
5	75th Percentile	98.35221862792969
6	95th Percentile	106.8583488464355

Table 2: This table shows the summary statistics for the primary completion rate as a percentage of the relevant age group

Discussion

From the results, we can tell that the median is higher than the mean. 75% of countries have completion rates above 80%. However, there are some that have very low values in comparison, as evidenced by the fifth percentile, which is at 57.87%.This implies that the data is skewed to the left, as there are small values

that pull the mean down.

The standard deviation is high, showing that the data are widely spread and not clustered around the mean.

From this we can conclude that by 2023 the completion rate for primary education in most countries was above 80%. However, there are some countries such as South Sudan and Equatorial Guinea that are still behind at 20.55% and 39.06%, respectively.

Question 3

Methods

For this problem, I downloaded the complete data sets for both GDP per capita and the prevalence of underweight children.

Then I converted the years for both datasets to two columns, one for year and the other for their respective values. This allowed me to have a unique combination for each country and year that would make it easy to plot all these data points on the scatter plot. Then I did an inner merge of the new data frames on the country name, country code, and year. In the combined data frame, I did a merge on the left with the country metadata to add the region and income group data. Some countries did not have this data; however, I did not wish to lose data on the variables we were interested in over metadata.

Finally, I plotted the graphs using the combined data frame.

Results

For these two variables, I expect that they have a negative linear relationship. This means that if a country has a high GDP per capita, I expect that the prevalence of underweight children under the age of 5 will be low. This is based on the assumption that a rich country is able to feed its citizens.

The following are the graphs plotted for this dataset.

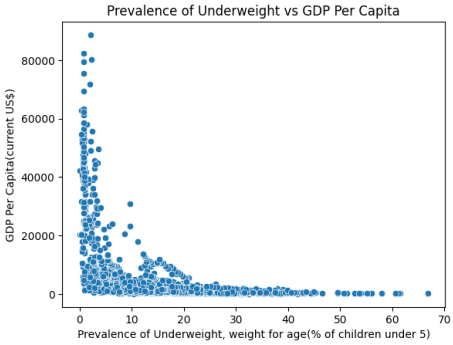

Figure 2: This scatter plot shows the relationship between the prevalence of underweight children below the age of 5 and the GDP per capita for different countries in 25 years

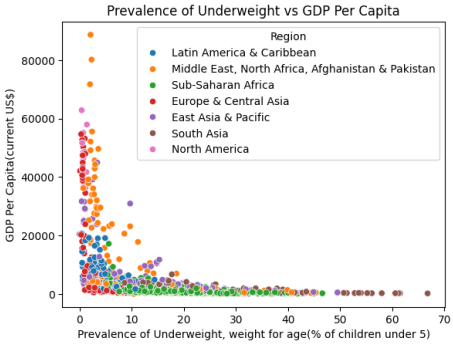

Figure 3: This shows the same data but with color coding for the different regions represented in the dataset

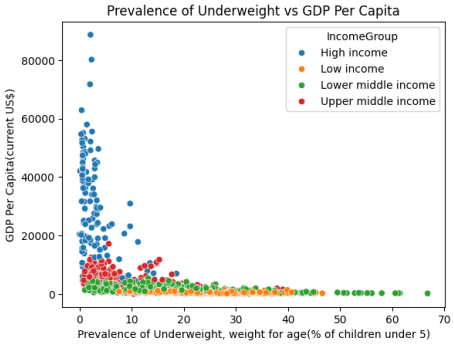

Figure 4: This shows the same data but with color coding for the different income levels for the countries represented in the dataset

Discussion

The first graph shows that many countries cluster close to the origin where both the prevalence of underweight children and the GDP per capital are low. The data points spread out in either direction, with few countries on the extremes where GDP per capita is high and the percentage of underweight children is low and vice versa. From the graph, we can deduce that these two variables have a moderate negative linear relationship, where for some of the data points, an increase in one variable means a decrease in the other.

The other two graphs paint a clearer picture based on the region and income level labels. It is evident that lower income countries have a high prevalence of children under the age of 5 years who are underweight. Most of these countries are in Sub-Saharan Africa and South East Asia. Most high-income countries with a low prevalence of malnourished children are in Europe, Central Asia, and North America.

Question 4

Methods

For this problem, I loaded the data from the respective csv files. For both datasets, I converted the date column from a string object to a date time object for better visualization of the x-axis as the range would be automatically determined by the scatter plot, unlike with a string object where it would try displaying all the data points.

Then I normalized the Adj Close value as requested by dividing the entire column by the first value and multiplying it by 100. This was done for both SPY and TLT. I plotted the time series with the dates tilted for better visibility.

I used a built-in Python method to calculate the daily returns from the original adjusted closing price. Using the new column for the daily returns, the minimum, maximum, and average were obtained for the daily returns. I multiplied these values by 100 to express them as percentages.

Results

Below is a graph showing the time series for two ETFs, SPY and TLT.

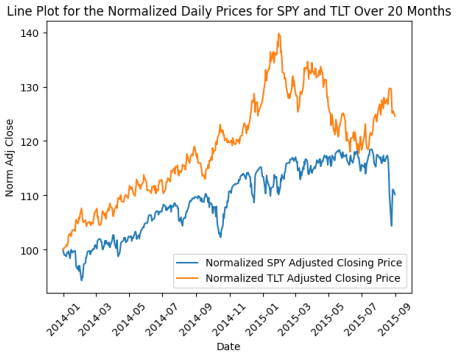

Figure 5: This line plot shows the change in the daily prices for two EFTs, SPY and TLT, over 20 months

The following values were calculated for the daily SPY returns.

```
Min SPY Daily Return: -4.211%
Max SPY Daily Return: 3.839%
Average SPY Daily Return: 0.026%
```

Figure 6: This image shows some descriptive statistics on the SPY daily returns

The following values were calculated for the daily TLT returns.

```
Min TLT Daily Return: -2.432%
Max TLT Daily Return: 2.647%
Average TLT Daily Return: 0.056%
```

Figure 7: This image shows some descriptive statistics on the TLT daily returns

Discussion

From the graph, it is evident that TLT did much better than SPY over those 20 months. SPY increased at a steady rate with a few dips, while TLT had dramatic changes over the 20 months. TLT increased with a large peak from

October 2014 that lasted until June 2015 when it declined before peaking again but less dramatically.

The calculated averages further support what is clear from the graphs that the TLT had better returns than SPY as evidenced by the higher average daily return. Furthermore, the min and max values for both also indicate how stable or volatile each of them was. TLT has a larger difference between the minimum and maximum daily returns compared to SPY, which further supports that TLT changed a lot while SPY remained stable.

Question 5

Methods

For these datasets, I loaded both of them leaving out the first two rows. I carried out an inner merge for the two data sets on country name and code. In the new combined data frame, I only included the year 2023 for both the fertility rate and the GDP other than the country name and code.

From this combined data frame, I plotted a scatter plot for the two columns, fertility rate versus GDP per capita PPP.

Finally, I calculated the correlation coefficient for these two columns using a built-in Python method.

Results

The graph below represents the relationship between fertility rate versus GDP per capita PPP for different countries in 2023.

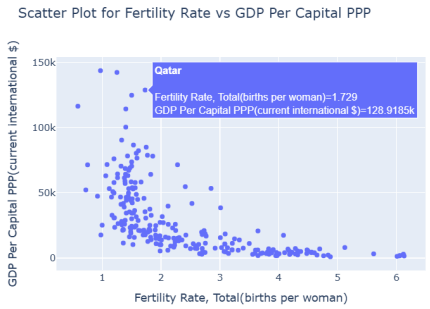

Figure 8: This scatter plot shows the relationship between the number of births per woman versus the GDP per capita PPP.

The calculated correlation coefficient below better tells us the direction and strength of the relationship between these variables.

```
Correlation coefficient: -0.6188248795954212
```

Figure 9: This image shows the results for the correlation coefficient for the fertility rate and GDP per capita PPP

Discussion

The graph shows that the fertility rate increases with a decrease in GDP per capita PPP. This implies that in richer countries, women have fewer children than in poorer countries. Furthermore, the corelation coefficient of -0.618 suggests that these two variables have a moderate negative linear relationship. This further supports what we see in the graph.

Question 6

Methods

These two data sets were in excel files and thus, in order to get the required data, I specified the names of the sheets to be used, the header row, and the columns I was interested in. For both of them, I was only interested in the country, rank, and index. In the CPI dataset, I also included the country code since it would be useful for labeling the plots.

Then I carried out an inner merge on the two datasets on country, renamed some columns, and reordered the columns for better readability. Finally, I plotted a scatter plot using the combined data frame; HPI rank against CPI Rank. I used Ploty to create an interactive plot where you hover over a point to see the country code label. This was to reduce visual clutter on the plot.

Results

Below is the scatter plot for the Happy Planet Index Rank against the Corruption Perceptions Index Rank.

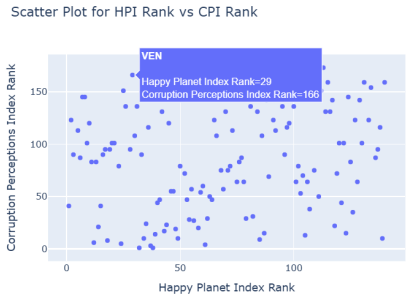

Figure 10: This is a scatter plot showing the relationship between countries' HPI ranks and CPI ranks

Discussion

Based on how spread the plots are, we can infer that these two variables have no linear relationship. We cannot infer one's behavior by the other. They are **uncorrelated.**

Although we were unable to relate these two variables, some countries have interesting results. An example is Tajikistan, which ranks 25th in the Happy Planet Index but is at position 151 when it comes to corruption. Another example is Venezuela, which ranks 29th on the Happy Planet Index, but 166th in the corruption index. These countries, despite being among the happiest, have high corruption rates.