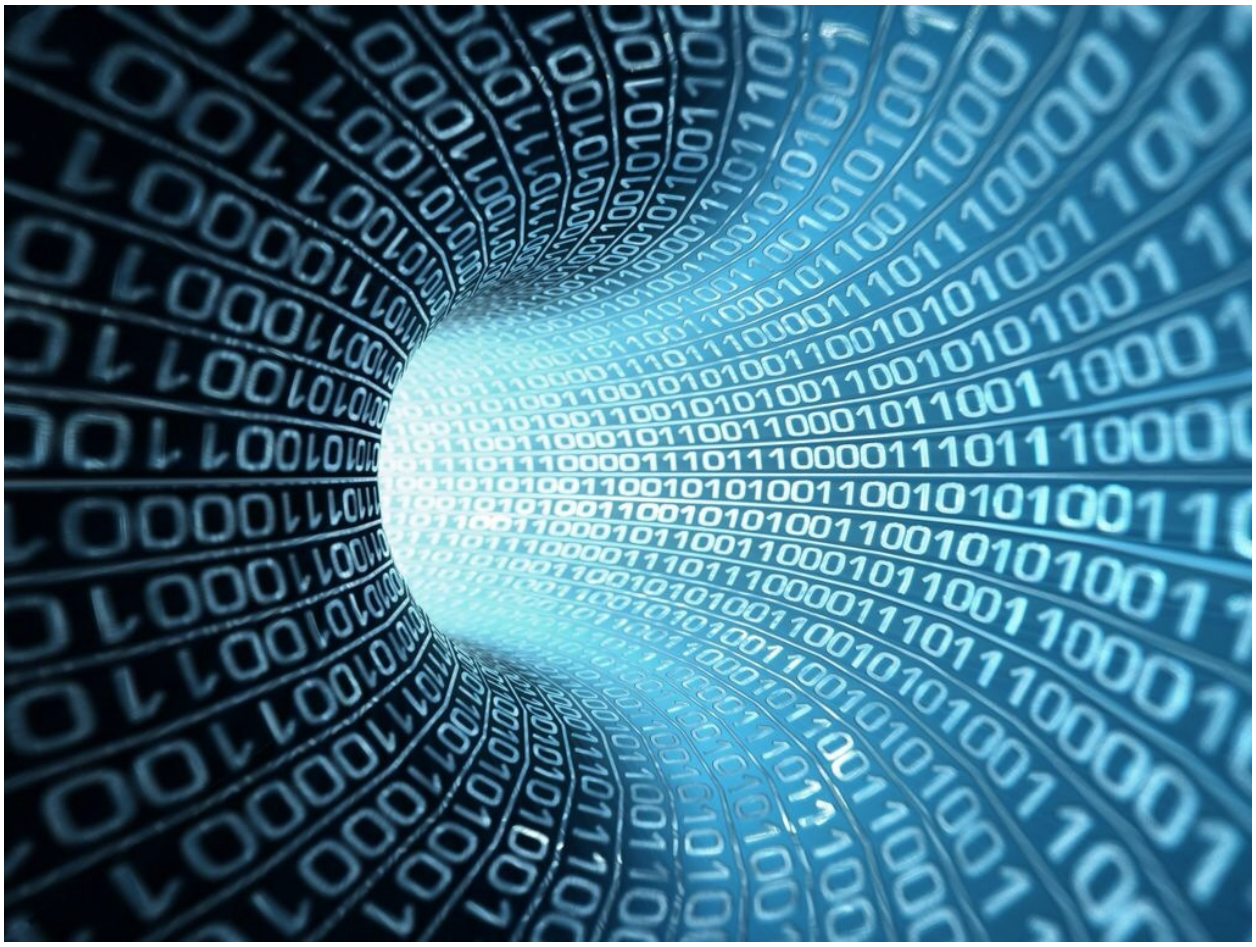
Report: DIAML Assignment 2



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: Tunga Tessema

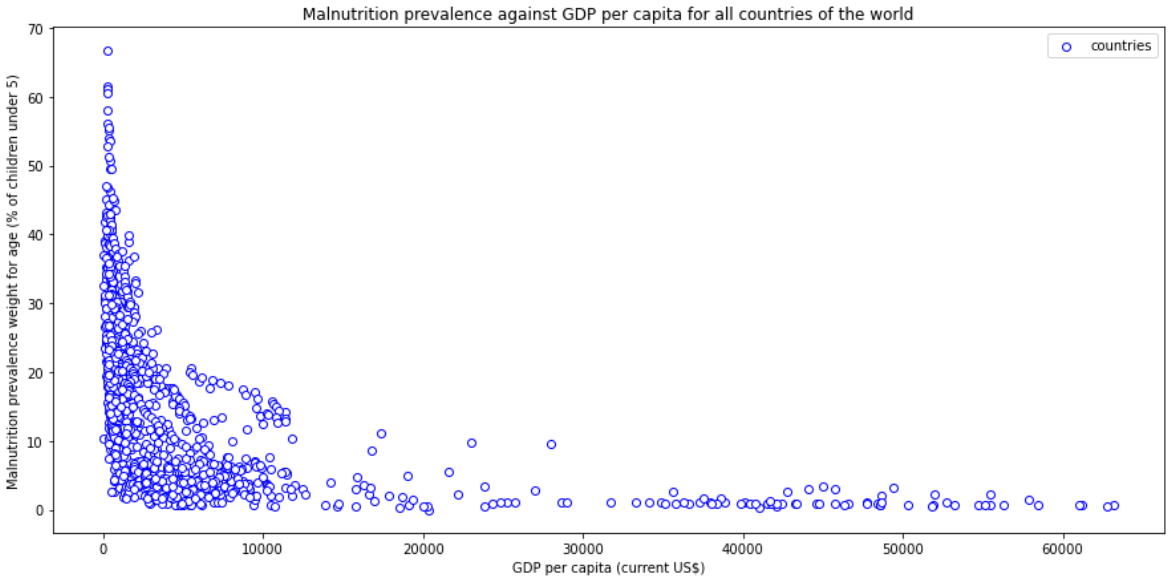
Andrew ID: tchamiss

Full Name: Tunga Tessema

Librairies

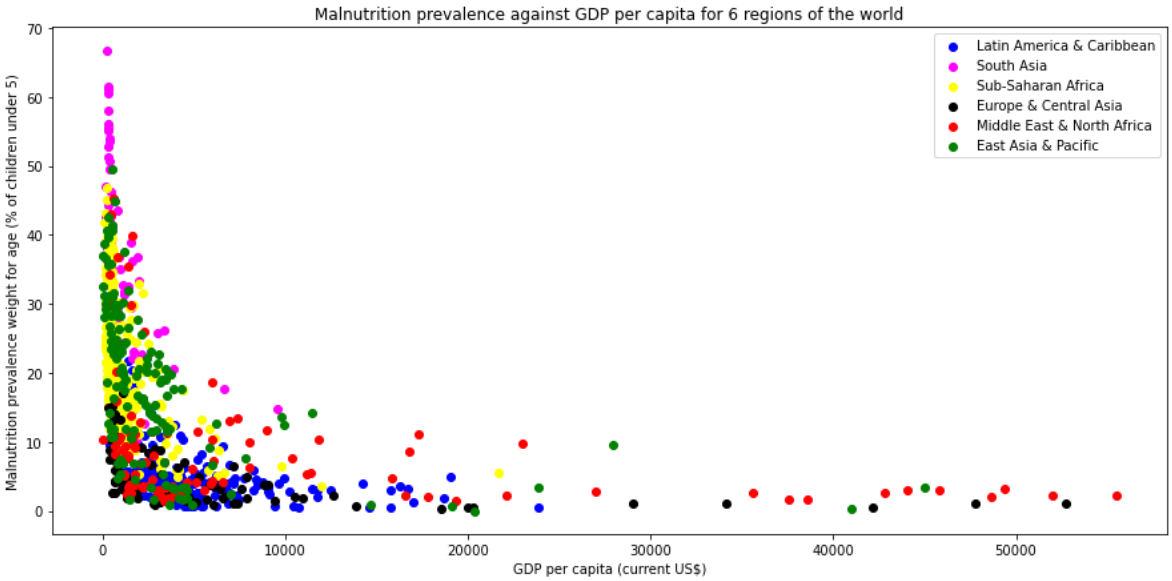
- **Matplotlib.pyplot:** to plot the graphs
- **Pandas:** to read excel files, convert them to dataframes and perform operations on them
- **Numpy:** to sort a dataframe column
- **Quandl:** to get data from the quandl api
- **Warnings**: to get the read of the warning for reading .xlsx files
- **Tabulate**: to create a table

Question 1



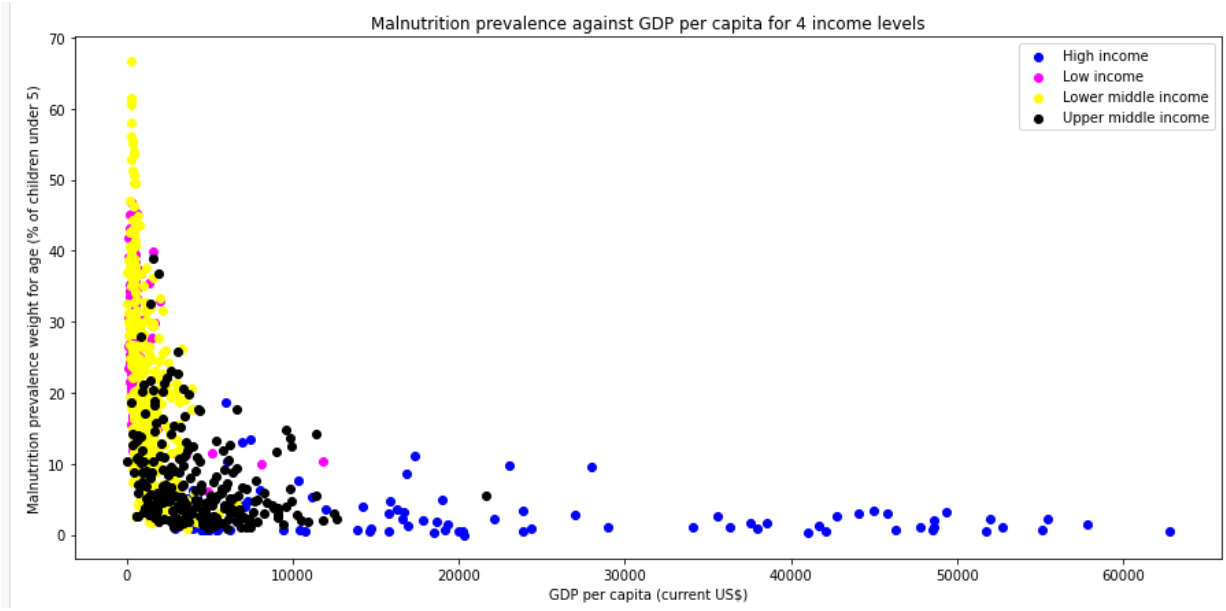
Steps: We were asked to plot malnutrition prevalence against GDP per capita (using all available years and countries). First, I loaded both excel files into dataframes using pandas. I removed rows which contained the data source and last updated date. I also set the first rows as header. I looped over every row to plot all the years of that country.

Insights: This is a J-shaped distribution. Before plotting the data, I expected malnutrition to be prevalent in countries that have low GDP.
After plotting the data, I indeed see that relationship. Countries who have GDP's greater than 20 000, have a malnutrition prevalence close to 0. When we see countries who have GDP's lower than 10 000, we can see that the malnutrition prevalence increases. The more we get closer to a 0 GDP, the more the malnutrition prevalence gets higher. It gets as high as 60 plus percent. However, we see that some countries have a low GDP and also a low malnutrition prevalence. This shows that GDP is not the only thing that affects malnutrition prevalence, there are other factors.
We can also see that most countries have a GDP less than 10 000, so they have higher percentages of malnutrition. This shows us that malnutrition is a serious problem in the world.



Steps: For the second graph, we were asked to plot the graph for malnutrition against GDP per capita using 6 regions of the world. I first got the metadata sheet of the GDP file and loaded it into a dataframe. I then merged the metadata dataframe with the GDP dataframe to get the region and income group columns. I then initialized empty dataframes of GDP and malnutrition for every region. I looped through the merged GDP dataframe and added every country to their respective regions' dataframe. I then proceeded to plot every region.

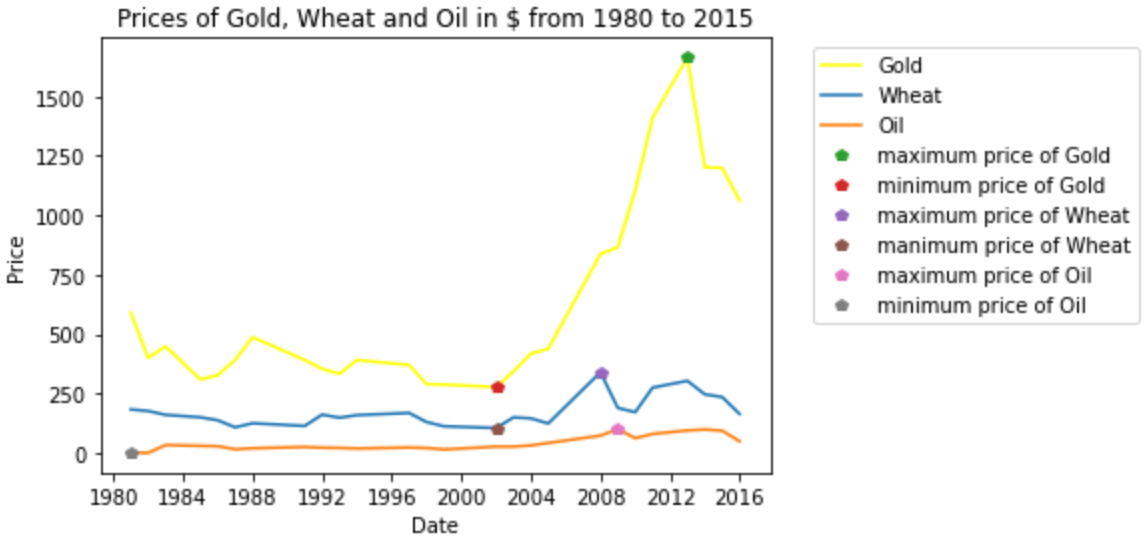
Insights: This is a J-shaped distribution. As we can see on the above graph, South Asia, which has a low GDP(mostly lower than 5000), is the region that has the highest malnutrition prevalence(in purple). Then comes sub-saharan Africa and east Asia & Pacific, whose malnutrition prevalence ranges from 5 to 45 percent.
Then we have Latin America where most countries have a GDP lower than 10 000 $ per capita and their malnutrition prevalence ranges from approximately 1 to 20 percent. We then have europe and central asia where most countries have a malnutrition prevalence of less than 10%. We also see that there are a few countries in the regions of Europe and central Asia, Middle East & North Africa and East Asia who have GDP's that are greater than 20 000 $ per capita and their malnutrition prevalence is close to 0.
Inside a region, even though the country's GDPs are close, we see that there is a variety in their malnutrition prevalence values. This shows that being part of a region doesn't determine a country's malnutrition prevalence, there are likely other factors that influence it..
We can also see that there is not a region among the 6 that doesn't have malnutrition issues.



Steps: For the third graph, we were asked to plot Malnutrition prevalence against GDP per capita for the 4 income levels. I first initialized empty dataframes of GDP and malnutrition for every income level. I looped through the merged GDP dataframe and added every country to their respective income level's dataframe. I then proceeded to plot every income level.

Insights: This is a J-shaped distribution. From the above graph we can see that low income countries have malnutrition prevalence that ranges from 5 to 45 percent. On the other hand lower middle income countries have malnutrition prevalences that range from 1 to approximately 68 percent. We then have upper middle income countries whose malnutrition prevalence mostly ranges from 1 to 20 percent. For high income countries we can see that most countries have a malnutrition prevalence lower than 10 percent.
We can therefore say that generally malnutrition prevalence is related to income level. The higher your income level, the less malnutrition prevalence you have. However, there are some lower middle income countries who have greater malnutrition rates than low income countries. This shows us that malnutrition prevalence doesn't just depend on the income level. There are likely other factors that influence it.

Question 2



Steps: For this question, I first created an account on Quandl and got an API key. I then downloaded the quandl python package. I downloaded each dataset by using their respective codes in the get request of the quandl python package. I then merged the datasets. I started with merging the gold and wheat datasets because the gold dataset had daily prices and wheat had monthly prices. I then merged the result of the previous merge with the oil dataset who had yearly prices.
I plotted all three datasets and marked their minimum and maximum values.

Insights: This graph shows us the price of gold, wheat and oil from 1980 to the end of 2015. From the beginning in 1980, we can see that gold is more expensive, followed by wheat then oil. In approximately 2002, gold and wheat had their minimum price. After that same time, prices of all three commodities started to increase. Gold reached its maximum in 2012 whereas wheat reached its peak in 2008 and oil in around 2009. After 2012, the price of gold started decreasing. At the end of 2015, we can see that once again gold is the most expensive commodity, followed by wheat then oil. However, when we compare with the price with which each commodity started in 1980, for oil we can see a slight increase, for wheat we can see a small decrease but for gold we see a large increase. Gold started at around 600$ in 1980 and at the end of 2015 it was around 1000$. If one was to invest, buying gold would yield the highest return.

Question 3

C02 emissions (metric tons per capita) summary statistics in 2010

Mean	4.33309
Median	2.68257
Standard deviation	5.01682
5%	0.112875
25%	0.721447
75%	6.08406
95%	15.5108

Steps: For this question, I first read the excel file of C02 emissions using pandas and loaded it into a dataframe. I then calculated the summary statistics using the describe function. I gave the describe function the 5th, 25th, 75th and 95th percentiles as arguments. I then created a nested array and stored the required statistics in it. I used the tabulate python package to display the nested array as a table.

Insights: From the above table we can see that on average every country has emitted 4.33 metric tons per capita. However, we can see that the median is 2.68 metric tons per capita. So 50% of countries emit below 2.68 metric tons per capita and the other 50% emit more than 2.68 metric tons per capita. The median is less than the average which tells us that there are some countries who emit more C02 than the other countries and are making the average greater. The distribution is right skewed, so if we want to know the “true average”, we use the median.
The standard deviation shows us how far the data is from the mean. The standard deviation is 5 and it is high. This shows us that the data has high variability.
Percentiles show us how countries did in comparison to the other countries. 5% of countries emitted less than 0.11 metric tons per capita, 25% of countries emitted less than 0.72 metric tons per capita, 75% of countries emitted less than 6.08 metric tons per capita and 95% emitted less than 15.51 metric tons per capita. In other words, this tells us that 5% of the countries which emit more than 15.51 metric tons per capita are the most polluting. The next most polluting countries are the ones that emit more than 6.08 metric tons per capita, they emit more than 75% of the other countries.

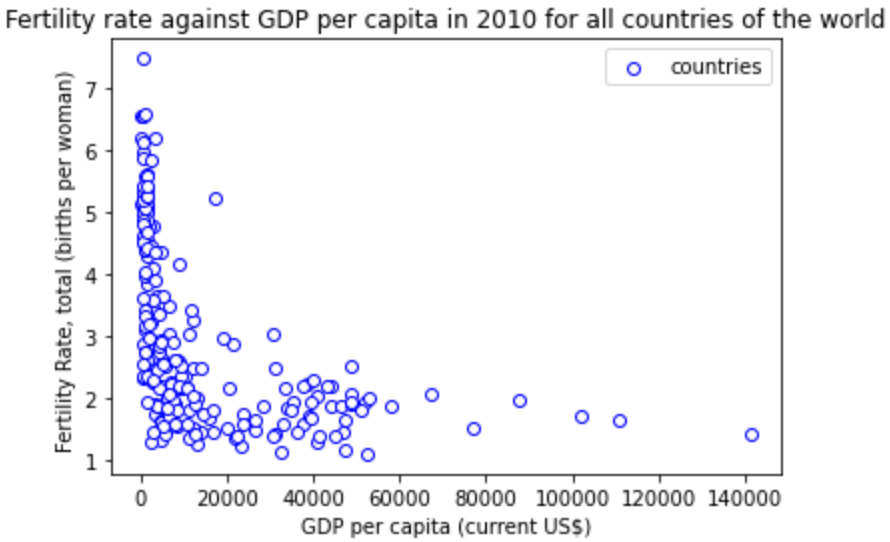
Primary Enrollment (% net) summary statistics in 2010

Mean	90.1051
Median	92.9567
Standard deviation	9.52763
5%	66.6568
25%	87.801
75%	95.9344
95%	98.8728

Steps: For this question, I first read the excel file of primary enrollment using pandas and loaded it into a dataframe. I then calculated the summary statistics using the describe function. I gave the describe function the 5th, 25th, 75th and 95th percentiles as arguments. I then created a nested array and stored the required statistics in it. I used the tabulate python package to display the nested array as a table.

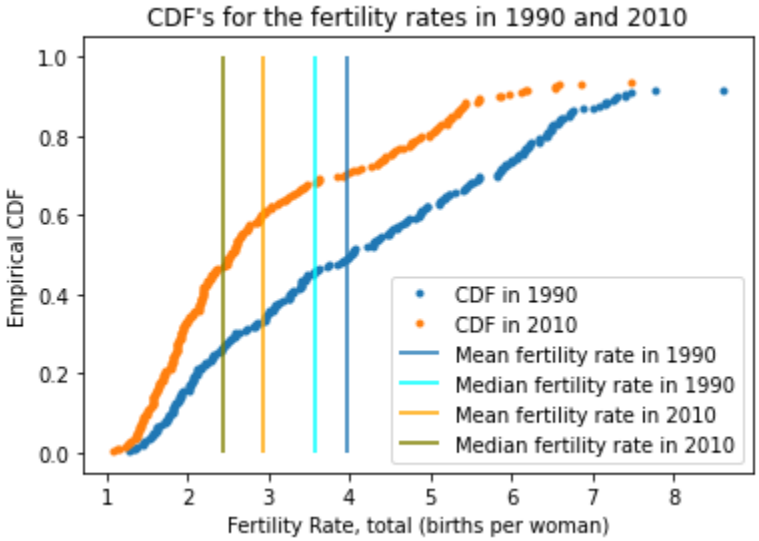
Insights: From the above table, we can see that on average countries have 90.10% of their population enrolled in primary schools. However, we can see that the median is 92.95%. This tells us that 50% of the countries have less than 92.95% primary school enrollment whereas 50% of the countries have more than 92.95% primary school enrollment. Since the median is higher than the mean, the distribution is left skewed.
The standard deviation shows us how far the data is from the mean. The standard deviation is 9.52 and it is high. This shows us that the data has high variability.
Percentiles show us how countries did in comparison to the other countries. 5% of countries have less than 66% primary school enrollment. 25% of countries have less than 87% primary school enrollment. 75% of countries have less than 95% primary school enrollment and 95% of countries have less than 98.8% primary school enrollment.
In general, the world is doing good in regards to primary education but 5% of the countries who have less than 66% enrollment should improve their primary enrollment programs.

Question 4



Steps: I first loaded the fertility excel file into a dataset using pandas. I then looped through every row and plotted the fertility rate against GDP per capita.

Insights: This is a J-shaped distribution. In general we can see that the more countries have higher GDP the less fertility rate they have. For example, countries with 40000 GDP per capita have a fertility rate of less than 2.5. There are also countries, who have a low GDP (less than 20 000) and have high fertility rates (more than 4). However, there are some countries who have low GDP (less than 20 000) and also low fertility rates (between 0.2 and 3). This shows us that there are other factors that affect fertility rates.

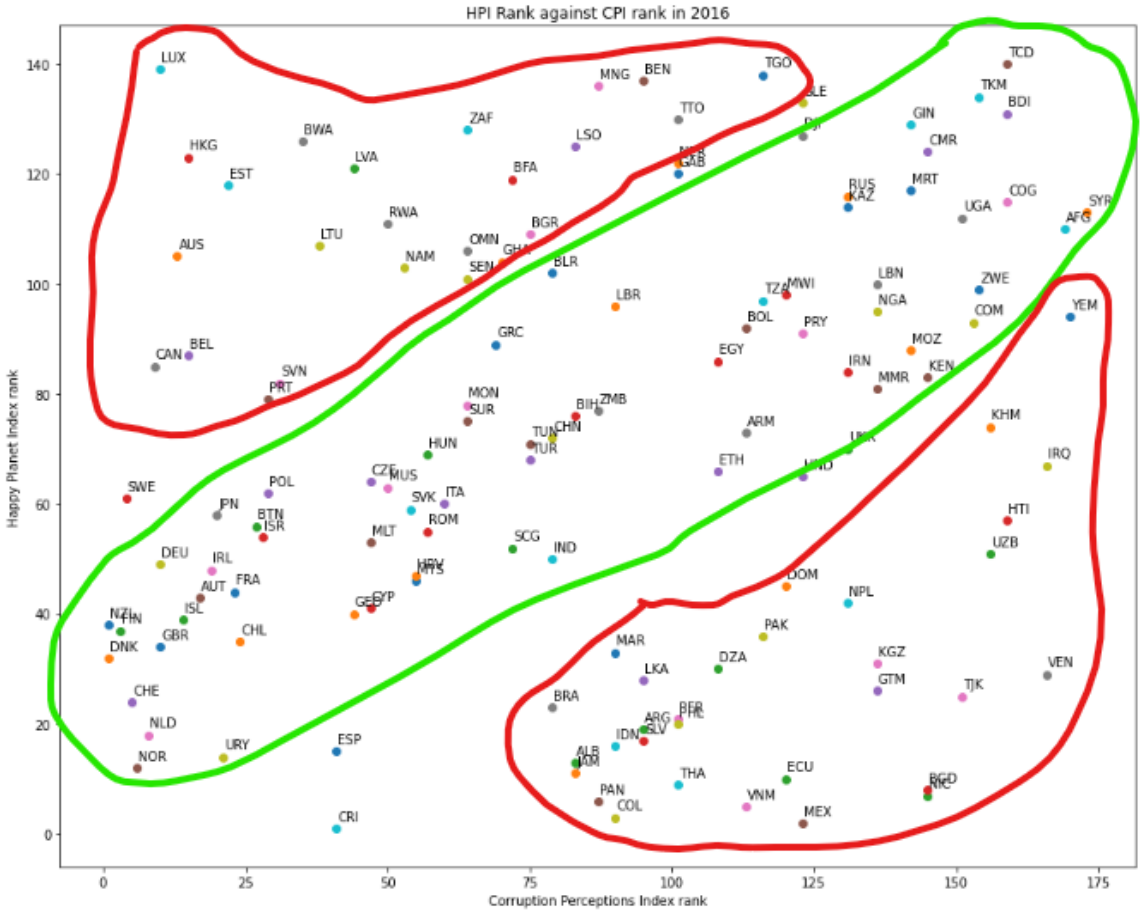


Steps: I first sorted the fertility rates for 1990. I then generated a list of numbers between 0 and 1. The list has the same number of elements as the list of fertility rates. I then plotted the values between 0 and 1 against the fertility rate. I took the same steps for the fertility data of 2010.

I also plotted the mean and median of the fertility rates in 1990 and 2010 using the mean and median function of pandas.

Insights: In 1990, the mean fertility rate was 3.9, whereas in 2010 the mean fertility rate was 2.9. So we can say that in 20 years the average fertility rate of the world has been reduced by 1. In 1990, the median fertility rate was 3.5 whereas in 2010 the median fertility rate was 2.4. This means that in 1990, 50% of the countries had a fertility rate of 3.5 or less whereas in 2010, 50% of the countries had a fertility rate of 2.4 or less.
A cumulative distribution function tells us the probability that a country has a fertility rate less than or equal to a specific value. If we see the fertility rate of 4 for example, we see that in 1990, approximately 50% of countries had a fertility rate less than or equal to 4. However in 2010, 70% of countries had a fertility rate less than or equal to 4. This shows us that in the 20 years that passed between 1990 and 2010, more countries have reduced their fertility rates.

Question 5



Steps: I first read the data from the excel files of CPI and HPI into pandas dataframes. I took the columns of interest which were Country, WB code and HPI rank and CPI rank. I then

merged the two dataframes based on country. I used the WB code to annotate each point on the graph.

Insights: In this graph, a low HPI rank is good and a low CPI rank is good. When a country has a low HPI index, it means it is ranked among the top in terms of happiness and when a country has a low CPI it means it is ranked among the top in terms of having low corruption.
I have circled in green, the countries for which happiness seems to have a relation with low corruption. However, there are many other countries for which happiness seems not to be related to low corruption (circled in red). Some of the countries that stand out are Luxembourg or Hong kong. These countries have low corruption but they are ranked among the lowest in terms of happiness. We can also take the example of Mexico or Bangladesh, who have high corruption but are among the top ten in terms of happiness.