Assignment 1: Introduction to Data Science and Python

Laboratory Report
in DAT405 / DIT405 Introduction to data science and AI

Matilda Sjöblom
Simon Duchén

Group 7

November 10, 2020

1 Download some data related to GDP per capita and life expectancy.

A) GDP per capita and Life expectancy visualisations

We selected the life expectancy and GDP per capita for every country from the website our-worldindata, the life expectancy from the following link:

- Life expectancy.

The GDP per capita data was collected from the following link:

- GDP

As there were a lot of entries for many years in both the files, we decided to use the most recent year that both the files contained, which were 2016. We then extracted every row for every 2016 entry the files. After that, they still had a different amount of countries recorded, where life expectancy had more entries. Therefore, we filtered the files on matching country names. As we knew that the files ordered the entries in alphabetic ordered, we could be sure that the entries would match. The resulting scatter plot can be seen in figure 1
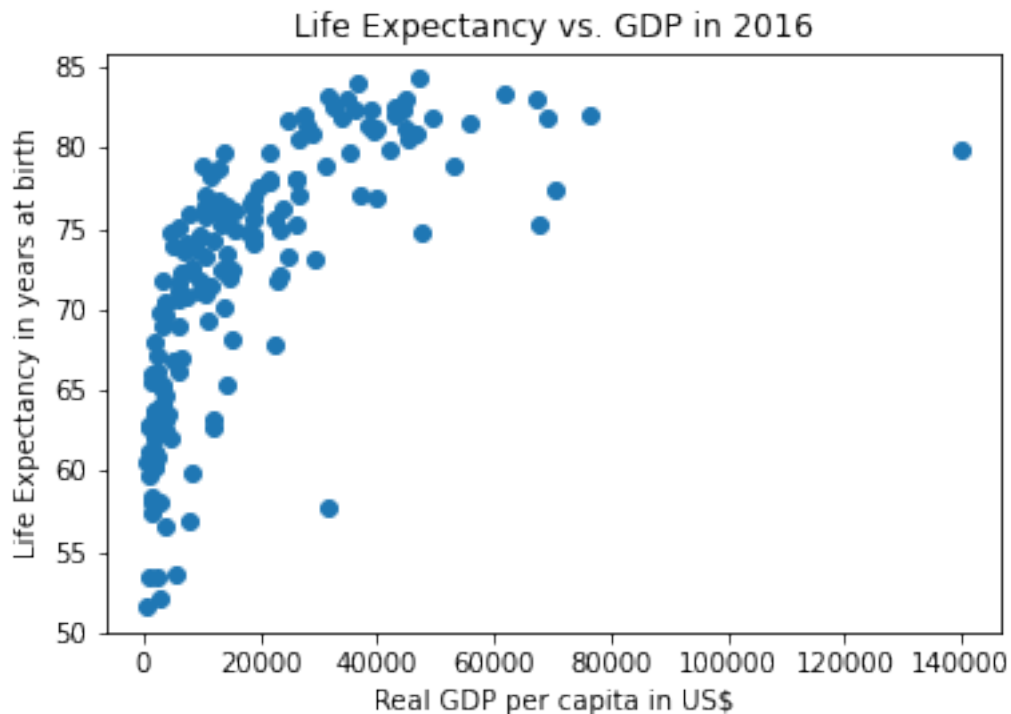

Figure 1: A scatter plot generated in matlib pyplot showing the correlation between life expectancy and GDP per capita over 165 different countries in the world at the year 2016.

B) Discussion

As can be seen in figure 1, the life expectancy increases in correlation with increasing GDP per capita. The result seems reasonable. A country that has a higher GDP is probably more likely to be more developed, i.e. have better healthcare, infrastructure, clean water, and such. If people are able to get healthcare, food, and clean water, they are more likely to live longer.

C) Data cleaning

We did, as mentioned in a, some data cleaning in order to fit the different data sets together. We did not remove anything that we did not think was useful, as every country that appeared in both the sets in the same year seemed useful and interesting to include in the plot, as with more data points you are more likely to see a correlation and trends.

D) Countries with high life expectancy

The table below shows the countries with life expectancy higher than one standard deviation added to the mean.
The standard deviation of the life expectancy data set is: 8
The mean of the life expectancy data set is: 72

Countries	Life expectancy (years)
Australia	82.959
Austria	81.258
Belgium	81.139
Canada	82.117
Cyprus	80.513
Denmark	80.583
Finland	81.382
France	82.328
Germany	80.902
Greece	81.727
Hong Kong	84.277
Iceland	82.601
Ireland	81.615
Israel	82.503
Italy	83.008
Japan	84.09
Luxembourg	81.803
Malta	82.063
Netherlands	81.862
New Zealand	81.859
Norway	82.029
Portugal	81.444
Singapore	83.083
Slovenia	80.852
South Korea	82.379
Spain	83.145
Sweden	82.382
Switzerland	83.31
United Kingdom	81.111

Figure 2: A table containing countries with life expectancy one standard deviation above mean.

E) Countries with high life expectancy but low GDP

The mean of the GDP is \$18,429.65 and the standard deviation is \$19,462.79. The country with the highest life expectancy and GDP below mean is Costa Rica, with a life expectancy of 79.738 years and \$13,986. Other countries with one half standard deviation (i.e. 4) above mean (i.e. 72) and GDP below mean can be seen in table 3

Countries	GDP (US Dollar)	Life expectancy (years)
Costa Rica	13986	79.738
Barbados	10160	78.888
Lebanon	12683	78.8
Albania	11285	78.194
Bosnia and Herzegovina	10576	76.998
Colombia	12963	76.732
Montenegro	18244	76.568
Sri Lanka	11118	76.482
Thailand	14341	76.403
Ecuador	10536	76.365
Algeria	13328	76.298
China	12320	76.21
Tunisia	10621	76.115

Figure 3: A table containing countries with a GDP below mean and life expectancy above one half standard deviation above mean. The table is sorted by highest life expectancy.

F) Strong economy correlation to high life expectancy

We chose all countries that had higher GDP per capita than one standard deviation added to the mean and assumed that these countries can be seen as strong economies. Below there is a table showing all these countries, their GDP per capita and life expectancy.
The mean of the GDP per capita data set is: $18429
The standard deviation of the GDP per capita data set is: $19464

Countries	GDP (US Dollar)	Life expectancy (years)
Qatar	139783	79.868
Norway	76397	82.029
United Arab Emirates	70284	77.47
Luxembourg	69057	81.803
Kuwait	67798	75.22399999999999
Singapore	67180	83.083
Switzerland	61844	83.31
Ireland	55653	81.615
United States	53015	78.885
Netherlands	49254	81.862
Saudi Arabia	47474	74.76100000000001
Hong Kong	47043	84.277
Germany	46841	80.902
Denmark	45141	80.583
Austria	45010	81.258
Australia	44783	82.959
Sweden	44371	82.382
Iceland	42980	82.601
Canada	42969	82.117
Taiwan	42304	79.92699999999999
Bahrain	39773	76.899
Belgium	39733	81.139
United Kingdom	39162	81.111
France	38758	82.32799999999999
Finland	38335	81.382

Figure 4: A table containing countries with a GDP one standard deviation above mean, and their corresponding life expectancy. The table is sorted by highest GDP.

As can be seen in the table above the countries with highest GDP per capita doesn't have the highest life expectancy.

G) Discussion

The result of F) Strong economy correlation to high life expectancy and E) Countries with high life expectancy but low GDP shows that life expectancy is not necessarily always higher in strong economies, take for example United states and Albania that have a similar life expectancy around 78 years but have a great differ in GDP per capita as Albania has \$11285 and United States have \$53015.

United States have almost five times more GDP per capita but still have the same life expectancy as Albania. A reason for this correlation could be that there are a small amount of the population that are very wealthy and therefore drives figures up higher. In the United States the average wage is around \$50000 (2018) but the median wage is \$33000 (2018) [1]. This suggests that their are outliers with great wealth that raises the economical state of the country which points to that the population over all doesn't have a strong economy, only a few. So a high GDP per capita doesn't raise the standard of living and life expectancy in a country necessarily, it depends on how the economy is distributed over the population.

---

[1] https://howmuch.net/articles/richest-countries-in-the-world, 2020-11-09

2 Download some other data sets

We chose to collect the following data in addition to GDP per capita and life expectancy: happiness, CO2 per capita. These are collected from the following sources:

- Happiness data set source

- CO2 per capita data set source

A) New data and visualisations

As our new data, we chose to compare and combine CO2 emissions per capita against self-reported life satisfaction and against GDP per capita. All the datasets are taken from 2014. The means and standard deviations of the datasets are:

- **CO2:** Mean: 5.12 Standard deviation: 6.2

- **GDP:** Mean: 18367.99 Standard deviation: 19026.29

- **Self-reported life satisfaction:** Mean: 5.40 Standard deviation: 1.12

Happiness vs. CO2 per capita
The data we chose to compare first was self-reported life satisfaction, in other words happiness, and CO2 emissions per capita. We wanted to see if there is any correlation between the two data sets. We thought that more emissions would result in happier people and, as can be seen in graph 5 below, there is a positive trend with happiness for higher emissions.

- Are people happier in countries with high CO2 emissions?

- Are there any countries with low CO2 emissions and a happy people?

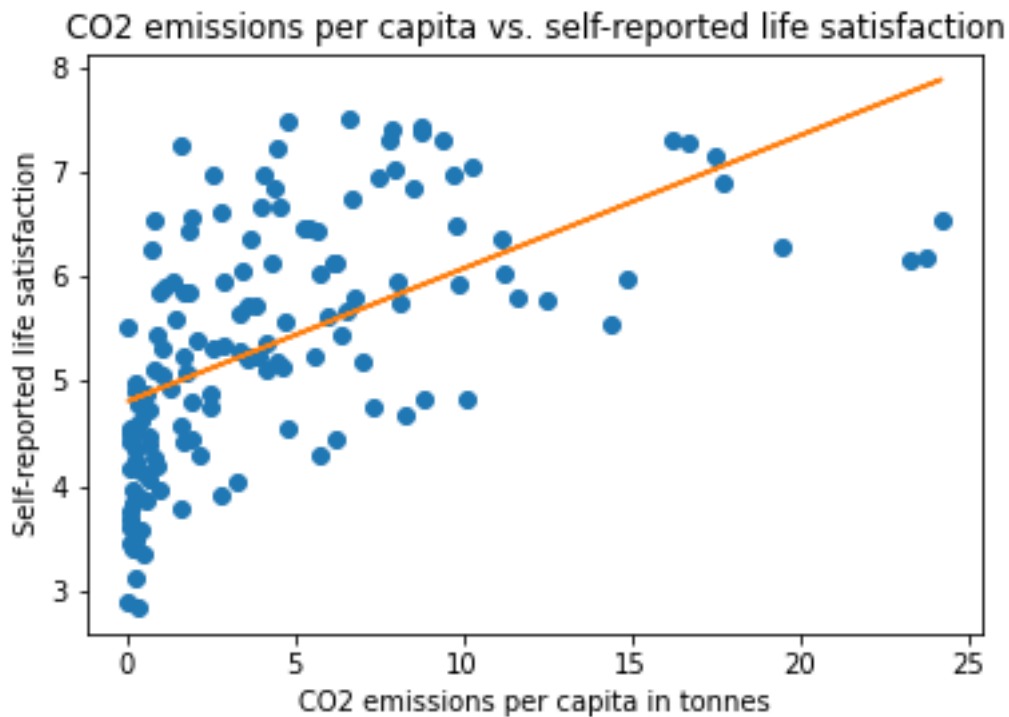- Is it more common to have low CO2 emissions and a happy people, or is it the other way around?

Figure 5: A scatter plot generated in matlib pyplot showing the correlation between self-reported life satisfaction and CO_2 emissions per capita in 2014.

We removed the values for Qatar because it was an outlier with a CO_2 emission per capita value at 43 tons which made the resolution of the graph very low and other data points cluttered.

In the graphs on the next page, we want to investigate our questions regarding the datasets. Therefore, we made one scatterplot with all the countries with a low CO_2 emission but a happy people (fig 6), and one scatterplot with high CO_2 emissions and low happiness (fig 7). In the figure with low CO_2 emissions and high happiness, there are 12 countries, and in the figure with low happiness and high CO_2 emissions, there are 6 countries.
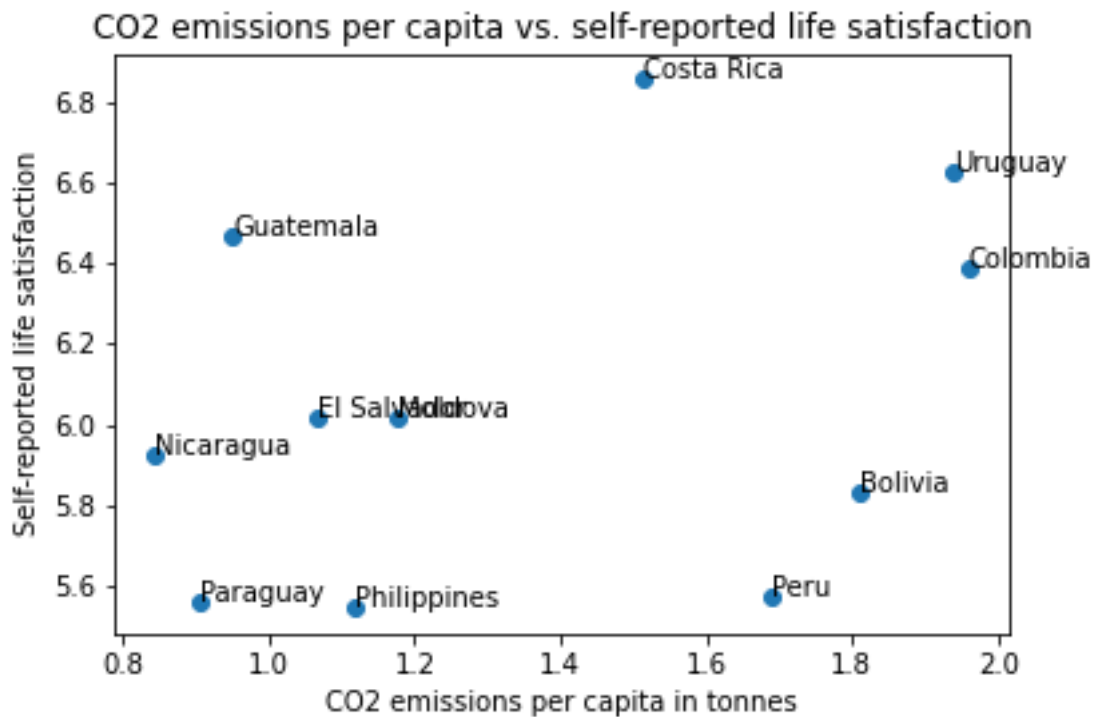
Figure 6: A scatter plot generated in matlib pyplot showing countries with self-reported life satisfaction above mean and CO_2 emissions below one half standard deviation below mean.
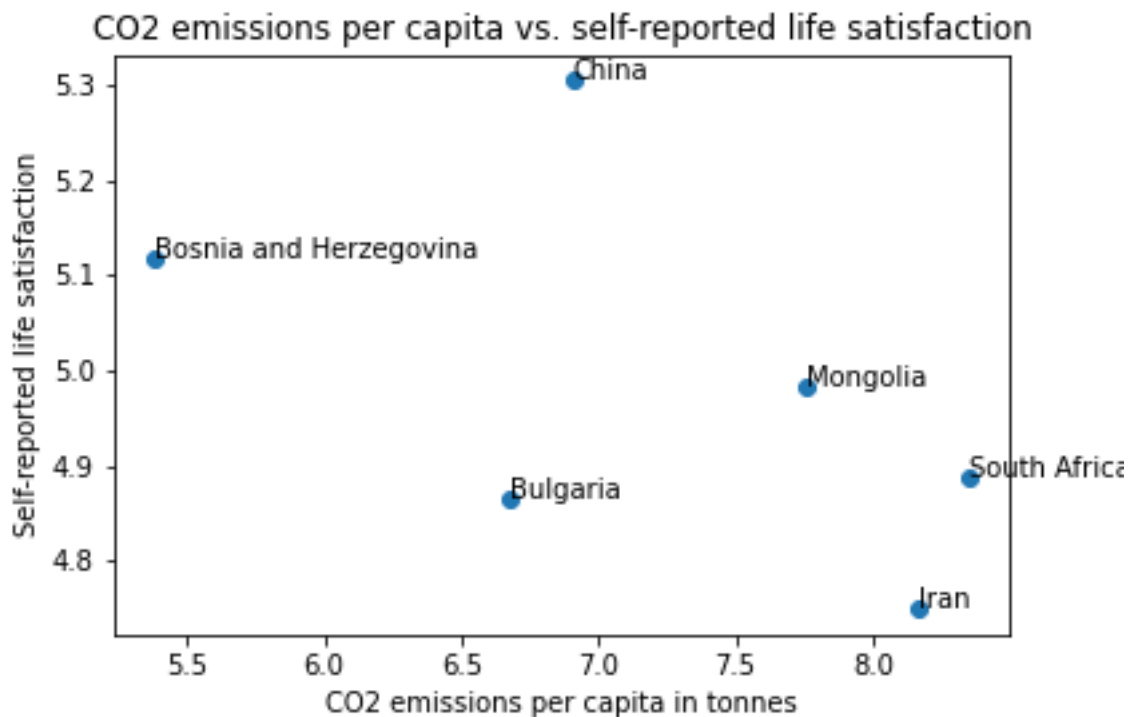

Figure 7: A scatter plot generated in matlib pyplot showing countries with self-reported life satisfaction below mean and CO_2 emissions above mean.

CO2 per capita vs. GDP per capita

Wealthier countries should contribute to higher CO2 emissions, at least that is what we think. As higher GDP countries should have a higher amount of production and consumption.

- Do citizens with more money automatically contribute more to climate change?

- Are there any countries that breaks the trend that can be seen in figure 8, in other words, countries with high GDP and low CO2 emission per capita.
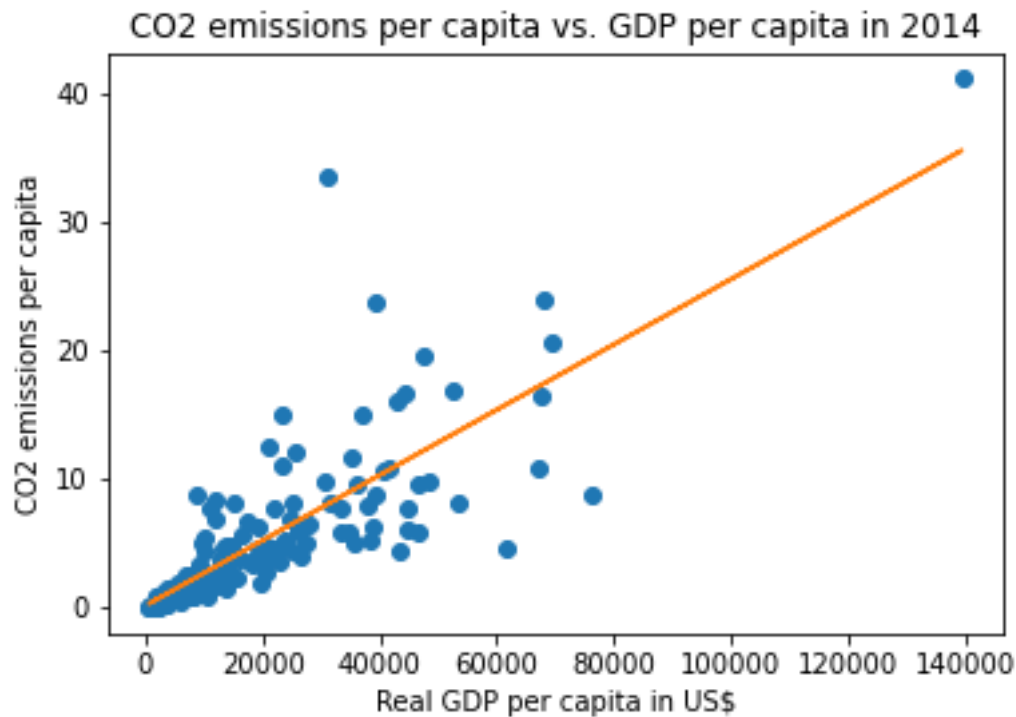

Figure 8: A scatter plot generated in matlib pyplot showing the correlation between CO2 emissions per capita and GDP per capita.

In the graphs on the next page, we investigate which countries have low GDP and high CO2 emissions (fig 9) and which countries have high GDP but low CO2 emissions (fig 10). There are eight countries in the figure with high CO2 emissions and low GDP, and two countries in the figure with low CO2 emissions but high GDP.
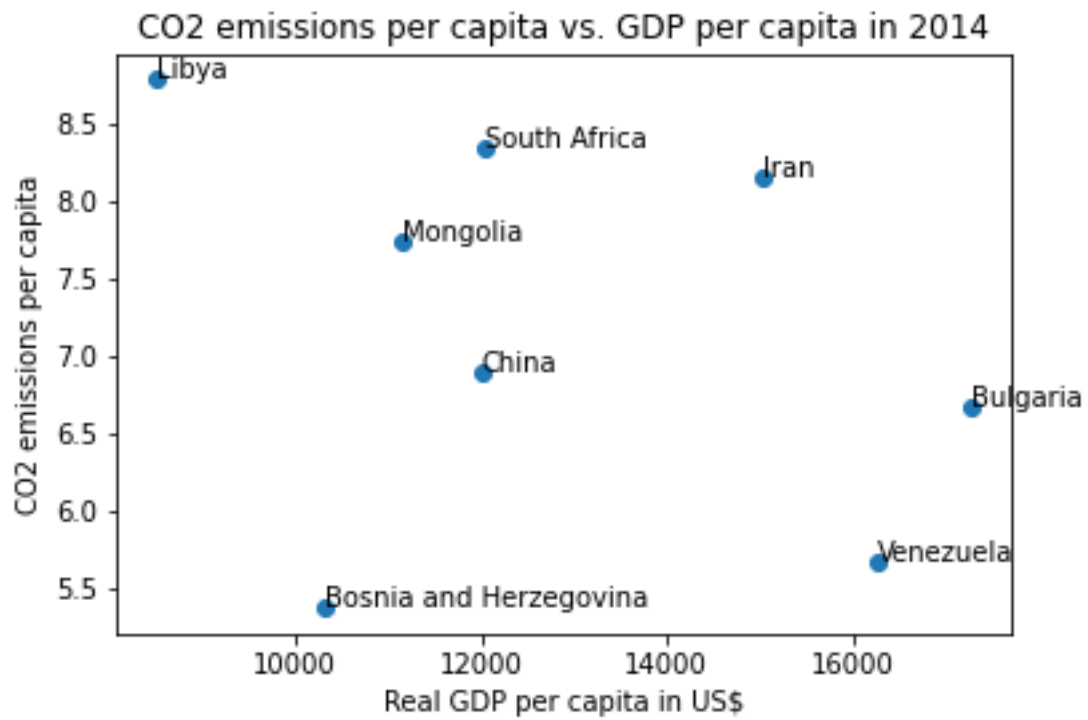
Figure 9: A scatter plot generated in matlib pyplot showing countries with CO_2 emissions higher than mean and GDP lower than mean.
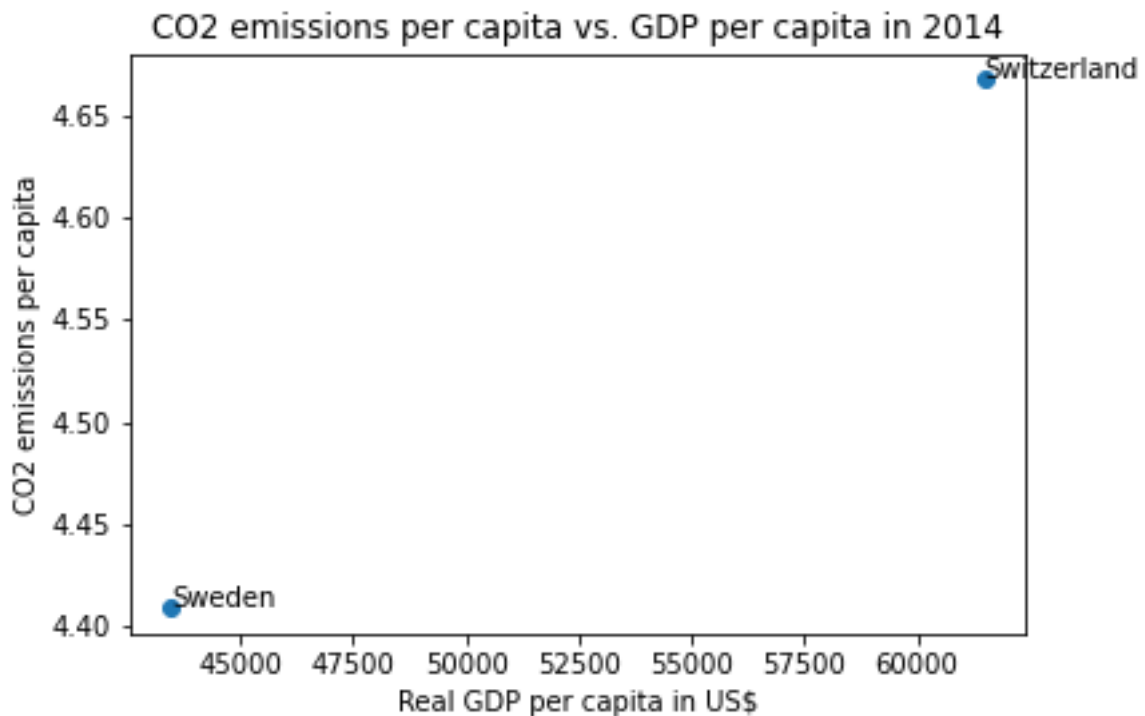

Figure 10: A scatter plot generated in matlib pyplot showing countries with CO_2 emissions lower than mean and GDP higher than one standard deviation above mean.

B) Discussion

CO2 per capita vs Happiness
There is a positive trend that agrees with our assumption that countries with higher happiness index have produce more CO2 emissions. A theory that can be applied here is that people that consumes a lot of goods emits more CO2 and are more happy, since materialism is something that grows in the modern world [2].

But consumption does not necessarily correlate to CO2 emissions since a lot of countries mostly consume imported goods and services. Then there are a lot of other factors that applies to the happiness index, for example there are a few countries that have very high CO2 per capita but low happiness index such as China and Iran. Iran for example has the majority of it's emission rate from oil which maybe doesn't affect the happiness index a that much, and the factor that Iran heavily suppressed the rights to freedom of expression, association and assembly [3].

CO2 per capita vs GDP per capita
The figure 1 shows that countries with higher GDP also have higher CO2 emissions in general. Since GDP per capita defines as the economical production value per citizen a higher value probably means higher production which in a lot of cases adds up to higher emissions. We also think that the consumption is higher in countries with higher GDP per capita since people over all have more money to spend, this raises the production which can contribute to more emissions.

But there are exceptions such as Sweden that have a high GDP value (over the mean + one standard deviation) but lower CO2 emissions than the mean, this can be seen in figure 1. A reason for this could be that Sweden imports a lot of goods and the production and transport of these goods is not taken into account in the CO2 emissions result[4]. Sweden also self produces most of the electricity from sources like nuclear power and hydropower plants which don't emit that much CO2.

---

[2] https://springerplus.springeropen.com/articles/10.1186/s40064-015-1494-5, 2020-11-10

[3] https://www.amnesty.org/en/countries/middle-east-and-north-africa/iran/, 2020-11-10

[4] https://oec.world/en/profile/country/swe, 2020-11-10