Final assignment CO2 emissions

Introduction

For the final assignment of the course “Data Analytics with Python”, the following questions need to be answered.

1. What is the biggest predictor of a large CO2 output per capita of a country?
2. Which countries are making the biggest strides in decreasing CO2 output?
3. Which non-fossil fuel energy technology will have the best price in the future?

For this analysis, data from the website www.ourworldindata.com is used. This website contains data about the large problems in the world. The world faces many problems like poverty, disease, hunger, climate change, war, existential risks, inequality and many more. Our World in Data collects data about these problems free for analysts and other interested parties to use. To answer the questions asked, I have browsed this website to obtain relevant CSV files about different variables and used them to make an analysis.

For every analysis, the collected data is reviewed, recalculated, cleaned and visualised. The data and visualisation is explained on the next pages. Tables and figures are added to make this explanation visual.

Below are the concise conclusions per question, more details about the data, methods, analysis and full conclusion are found further in this document.

Question 1. What is the biggest predictor of a large CO2 output per capita of a country?: This question is answered with a statistical approach. I used different variables to compare to the CO2 output. By calculating the correlation coefficients, I could see if there is a relation between the variables and the CO2 output per capita. By visualising this in a scatterplot with a linear regression line, it is obvious that the variables have a relation to the CO2 output. But only the variable for fossil fuel use in kWh per year per capita shows a very high correlation and almost linear values to the regression line. The conclusion is: a large use of fossil fuels per capita in combination with a great prosperity is the biggest predictor of a large CO2 output per capita of a country.

Question 2. Which countries are making the biggest strides in decreasing CO2 output?: For this question the relative change in CO2 output per capita per 5 years was calculated. It is calculated per capita, because this accounts for growing or shrinking populations. The relative change was calculated every five years with reference to 1961 and the endpoint is 2019. These values are made into a dataframe, sorted and the top 10 countries are set in a line plot. The line plot shows that Zambia made the biggest strides in decreasing CO2 output.

Question 3. Which non-fossil fuel energy technology will have the best price in the future?: For this last question I needed to predict future prices. I decided to calculate the linear regression and visualise it in a plot to see how the prices will develop. The plot shows that solar energy will have the best price in the nearest future (before 2030). Wind power will have a better price too, but a few years later.

Analysis of the questions

Question 1

What is the biggest predictor of a large CO2 output per capita of a country?

Data

For this analysis, fourteen files were collected from www.ourworldindata.com, uploaded to GitHub and imported in Google Colab using pandas library. After renaming columns and recalculating some values I selected the relevant columns and merged this in a new dataframe. This dataframe contains values for 64 countries and 27 years to compare. The variables below are chosen, because these would be best to answer the question asked.

The used variables are (the type of recalculation is explained per variable):

- CO2 output (file contains total CO2 output in tonnes per year, recalculated to kg/per year/per capita)
- Total population
- Population density
- Share of urban population (file contains absolute numbers, recalculated to share of total population)
- Share of rural population (file contains absolute numbers, recalculated to share of total population)
- GDP and GNI per capita
- Gini coefficient
- Human development index (not recalculated, calculation based on life expectancy, GNI and education)
- Share of population under the poverty line
- Share of employees working in industry
- Use of different energies in kWh per capita (file contains use of energy in tWh, recalculated to kWh)
- Agricultural output in US$ per capita
- Food supply in Kcal per capita per day
- Share of land area used for agricultural purpose

The next variable is dropped from the comparison, because it only contains values for 12 years and 34 countries, this reduced the data too much. The calculations and plot were analysed and the outcome was not significant to answer the question)

- Amount of cars owned per capita

Method

Step one is to calculate the correlation. To answer the question, there needs to be a relation between the CO2 output per capita and the variables. In statistics correlation coefficients are used to show a relation between values. It does not say what relation (causal or not), but in this case if there is no or only a small connection, the variable cannot be used. The correlation is calculated using pandas library and the results are collected in a new dataframe. The dataframe is cleaned (dropping rows and columns) and sorted to get the values that will be used in a plot. Figure 1 shows this plot.

I used the following information about the degree of correlation from www.statisticssolutions.com:

- *High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.*
- *Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.*
- *Low degree: When the value lies below ± .29, then it is said to be a small correlation.*

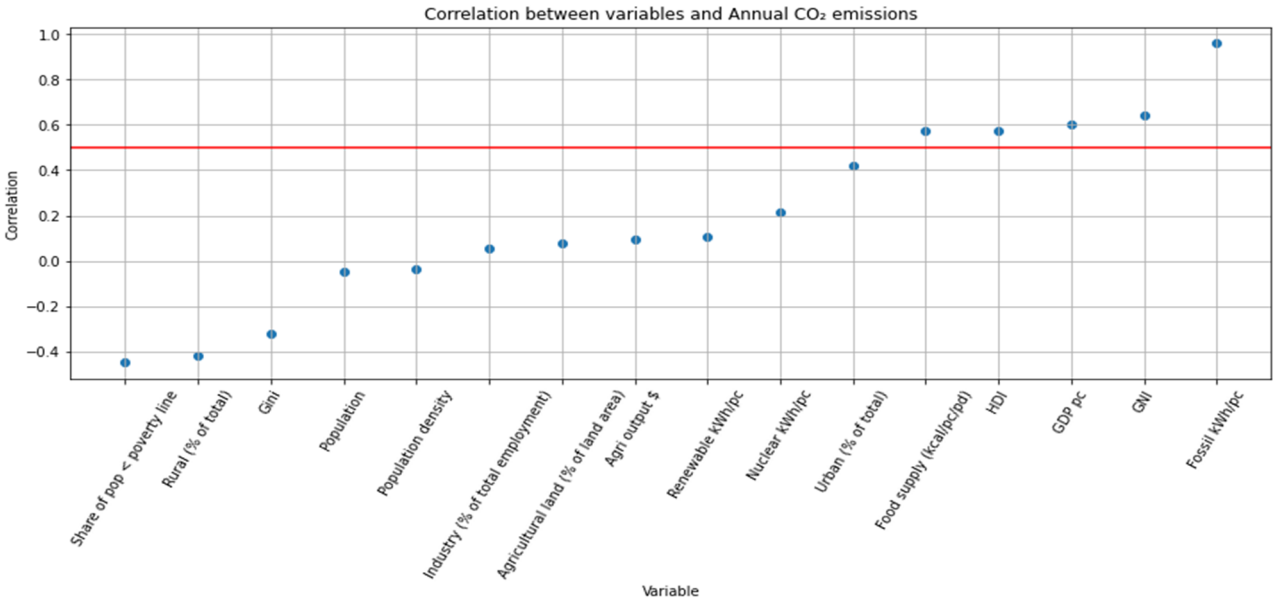

Figure 1: Correlation between variables and annual CO_2 emissions

Step two is to visualize the CO2 output compared to the variables in scatterplots with an added linear regression line (see figure 2 for the plots). These are the variables with a high correlation degree (above 0.5, see the red line in figure 1). Visualising the values makes it easier to see how the values are scattered and how the regression line relates to these values. The libraries used for the calculations, plots and regression lines are pandas, NumPy and matplotlib.

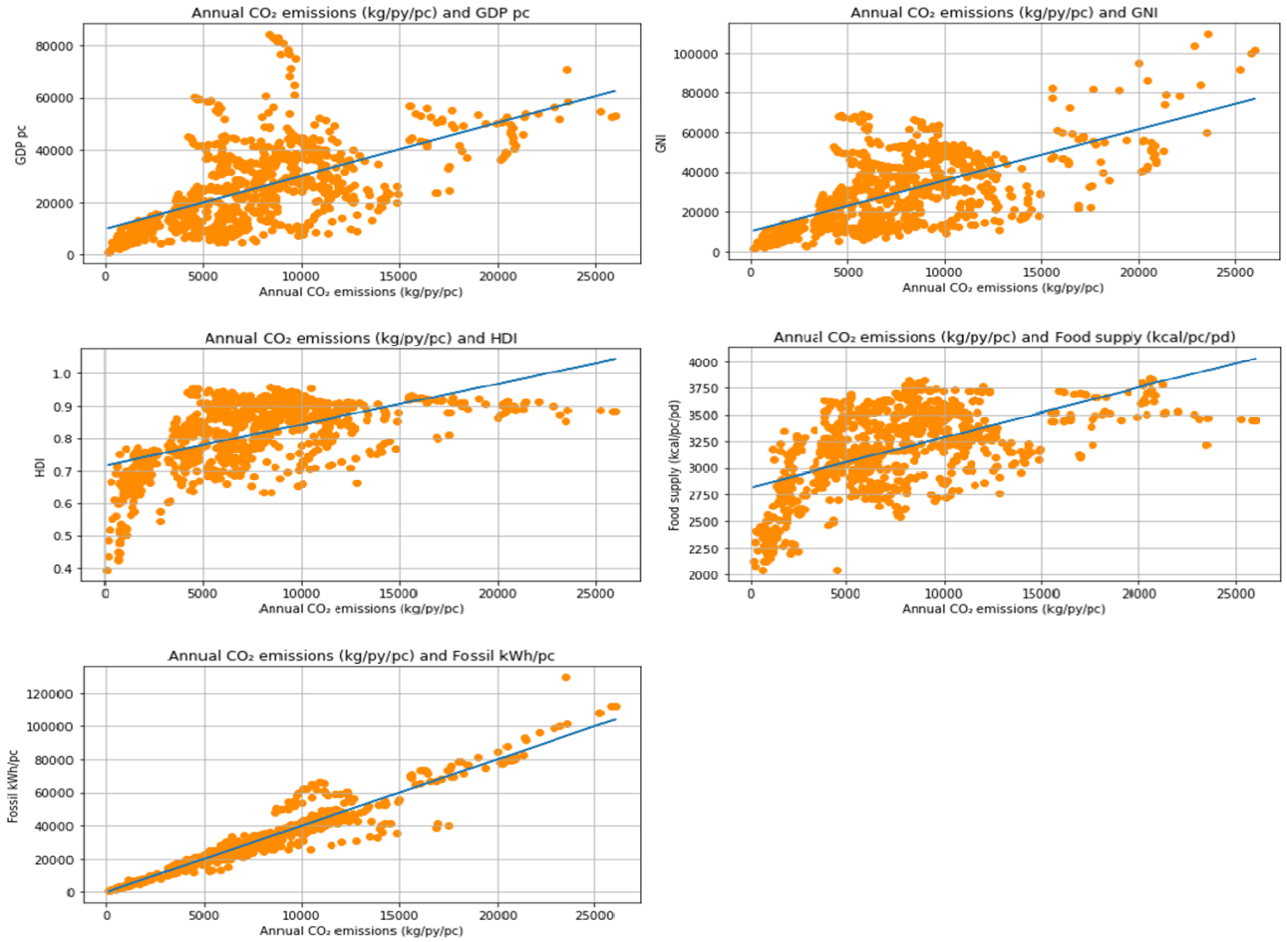

Figure 2: Plots with linear regression line for Annual CO_2 emissions against the variables

Analysis

For the analysis, we need to look at the plots in figure 2. There is a moderate to steep inclining slope on each regression line. But because the values are very much scattered or clotted together in one place, the regression line does not really show a relation to the values. Only the variable for fossil fuel use in kWh per capita shows almost linear values without major outliers and a regression line that follows the values very close.

Conclusion

Eleven of the variables compared to the CO2 output per capita have just a small relation to it or none at all. Five variables showed moderate to high correlation (above 0.5, see the red line in figure 1). The plots of GDP, HDI, GNI and food supply show a moderate to steep slope in the regression line, but with scattered and clotted values. These are not the best predictors of a large CO2 output per capita on their own. They are related to each other, because they represent the prosperity of a country.

Only the variable for fossil fuel use in kWh per capita shows almost linear values without major outliers and a regression line that follows the values very close. This is the best variable to use to predict. In combination with the other 4 variables, I can conclude the following:

To answer the question: What is the biggest predictor of a large CO2 output per capita of a country?
A large use of fossil fuels per capita in combination with a great prosperity is the biggest predictor of a large CO2 output per capita of a country.

Question 2

Which countries are making the biggest strides in decreasing CO2 output?

Data

The data consists of two csv files and was downloaded from www.ourworldindata.com, uploaded to GitHub and imported in Google Colab using pandas library. The variables below are chosen, because these would be best to answer the question asked. To take shrinking or growing population into account, the CO2 output in tonnes per year per capita is calculated and compared, instead of the total amount of CO2 output in tonnes per year. The dataframe contains rows for “non-countries” (e.g. Continents, Income related, Territories). They are not relevant and therefore removed. The data is transformed to get a dataframe with the years set as columns and countries as rows. After dropping rows with missing values, the resulting dataframe contains values for 134 countries and 58 years.

The used variables are:

- CO2 output (file contains total CO2 output in tonnes per year, recalculated to kg/per year/per capita)
- Total population

Method

To see which country made the biggest strides in decreasing CO2 output, a visualisation of the relative CO2 output per period is needed. The first measurement is from 1961. This is point zero with a value of 100%. The next relative datapoints are calculated every 5 years with reference to 1961 and added in new columns. To make the dataframe smaller, the columns for the original years are dropped. The last measurement is from 2019 and is the endpoint. The calculated data is merged into a dataframe (see table 1), sorted from low to high (based on the values in column for relative CO2 output in 2019) and the best 10 are plotted in a line plot (see Figure 3). The libraries used for the calculations and plot are pandas and matplotlib.

Table 1: Merged dataframe of relative CO_2 output in % per 5 years

	rel1961	rel1966	rel1971	rel1976	...	rel2006	rel2011	rel2016	rel2019
Zambia	100	80.38	74.70	68.68	...	16.18	20.52	30.32	36.59
United Kingdom	100	101.73	106.14	95.54	...	83.87	66.65	54.64	49.05
Venezuela	100	92.73	87.00	68.79	...	96.29	98.53	80.77	50.12
Zimbabwe	100	102.12	126.30	133.90	...	65.21	61.71	56.53	56.15
Sweden	100	143.21	161.65	165.24	...	91.40	80.59	67.36	61.64
...	...	...	...	...	...	...	...	...	...
South Korea	100	184.66	323.14	469.32	...	1925.60	2330.07	2284.72	2291.98
Thailand	100	196.45	345.21	451.15	...	2306.27	2494.02	2703.90	2693.36
Laos	100	184.09	293.87	139.07	...	564.88	944.41	4565.06	5193.78
Nepal	100	213.59	199.15	254.67	...	1191.11	2461.93	4504.86	5988.76
United Arab Emirates	100	180.25	78809.46	82740.35	...	31875.50	28868.84	30283.85	29033.24

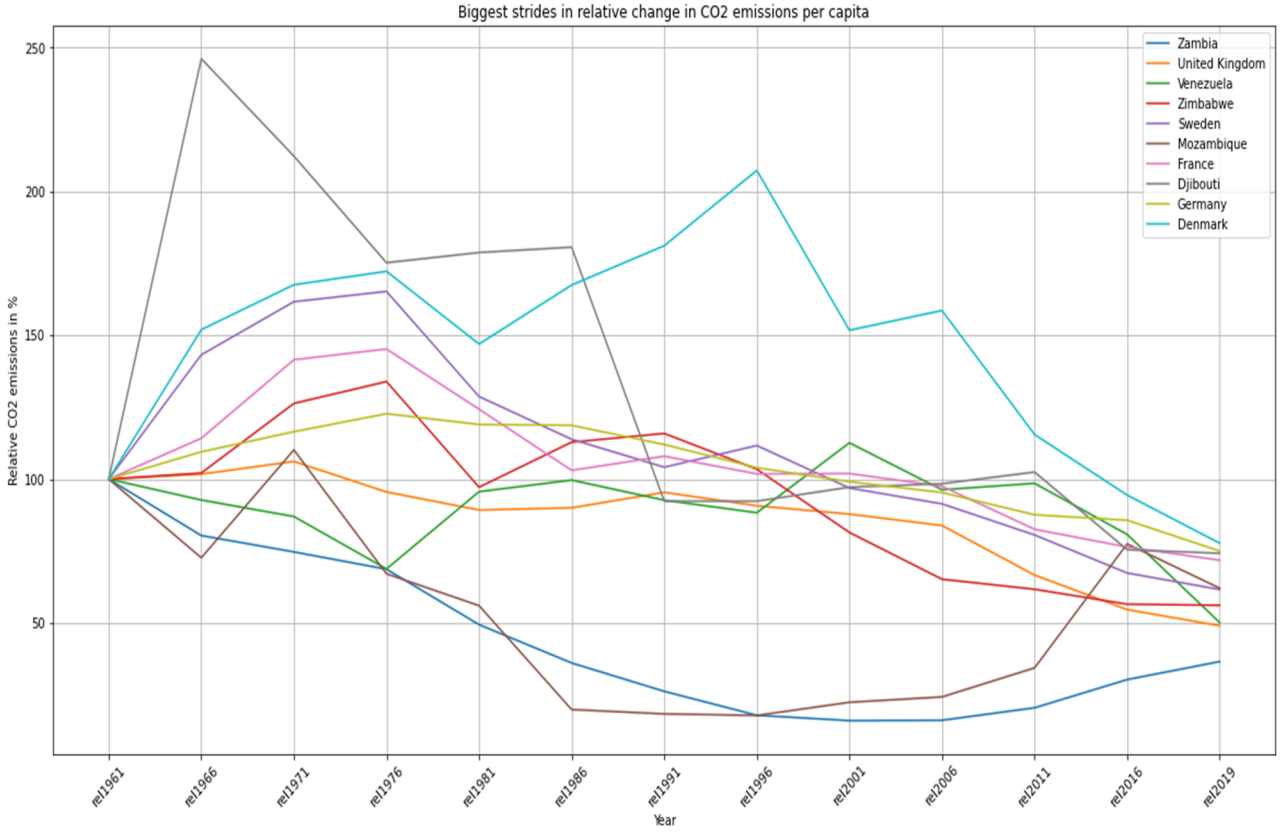

Figure 3: Line plot of the top 10 countries that made the biggest strides in decreasing CO_2 output

Analysis

The line plot in figure 3 shows the top 10 countries that have made the biggest strides in decreasing CO2 output. This makes it easier to see which countries are the best.

Conclusion

To answer the question: Which countries are making the biggest strides in decreasing CO2 output?
The top 10 of countries that made the biggest strides in decreasing CO2 output are (in order from greatest to least):

1. Zambia
2. United Kingdom
3. Venezuela
4. Zimbabwe
5. Sweden
6. Mozambique
7. France
8. Djibouti
9. Germany
10. Denmark

Question 3

Which non-fossil fuel energy technology will have the best price in the future?

Data

The used data is collected from www.ourworldindata.com. There was only one csv file containing energy prices. It was uploaded to GitHub and imported in Google Colab using pandas library. There are some values per country, but they are not complete. Only the rows for “World” are mostly filled with values. That is why I used the values for “World” instead of per country. Some variables don’t have values for all years. This is not a problem, because the plot is based on the years available. I did not have to recalculate values.

The used variables are:

- Bioenergy
- Geothermal energy
- Solar photovoltaic power
- Concentrated solar power
- Offshore wind power
- Onshore wind power
- Hydropower

Method

To predict the future prices for the various energy sources, the current data is used to create a linear regression line that goes further than the last measuring point. The missing values (NaN in the dataframe) are not replaced, because the course of the price development is not linear. The plot is made based on the available values. Table 2 shows a part of the dataframe, figure 4 shows the plot. . The libraries used for the calculations, plots and regression lines are pandas, NumPy and matplotlib.

Table 2: Part of merged dataframe with prices per kWh per energy source

index	Year	Bioenergy	Geothermal energy	Offshore wind power	Solar photovoltaic power	Concentrated solar power	Hydropower	Onshore wind power
0	1983	NaN	NaN	NaN	NaN	NaN	NaN	0.327851
1	1984	NaN	NaN	NaN	NaN	NaN	NaN	0.320074
2	1985	NaN	NaN	NaN	NaN	NaN	NaN	0.297221
3	1986	NaN	NaN	NaN	NaN	NaN	NaN	0.264194
4	1987	NaN	NaN	NaN	NaN	NaN	NaN	0.256420
…	…	…	…	…	…	…	…	…
34	2017	0.071070	0.070917	0.106152	0.083660	0.206213	0.050722	0.059959
35	2018	0.055360	0.067583	0.100049	0.071139	0.149103	0.039837	0.050880
36	2019	0.063933	0.067305	0.086388	0.062119	0.211831	0.041409	0.044592
37	2020	0.072473	0.054264	0.086266	0.055444	0.106653	0.045966	0.037137
38	2021	0.067343	0.067616	0.075167	0.048346	0.114242	0.048300	0.033123

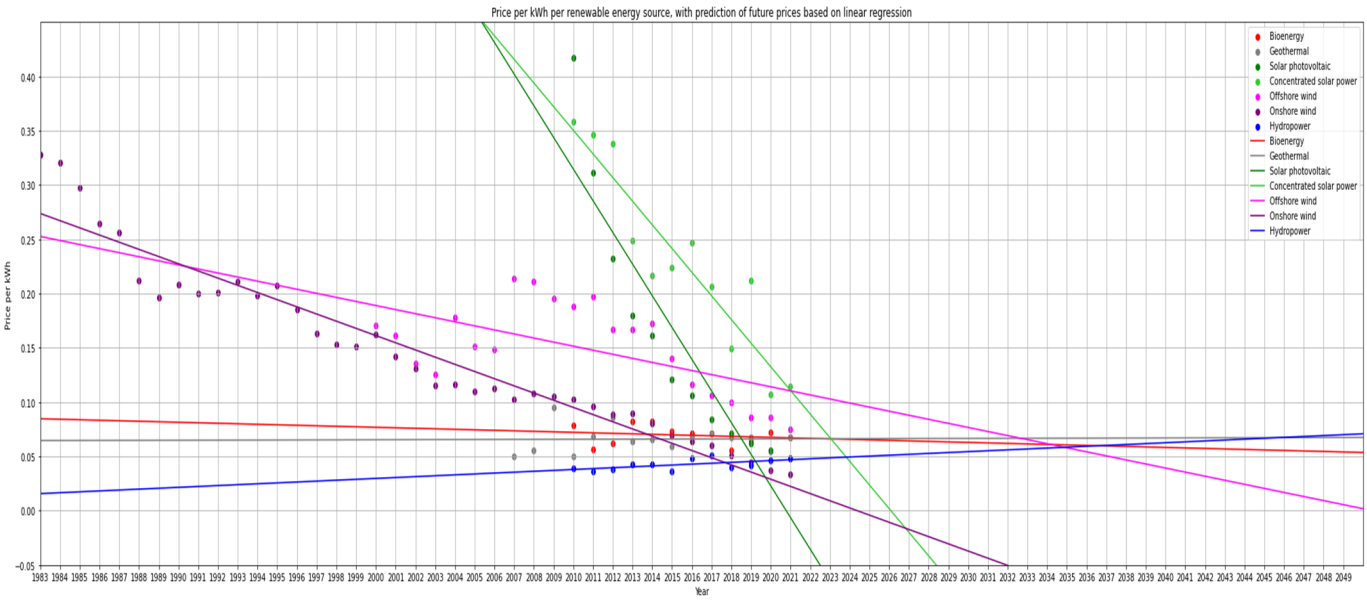

Figure 4: Plot of linear regression for the prices per kWh of various renewable energy sources

Analysis

Although the prices of energy will not go below zero, the plot in figure 4 shows which energy source will have a lower price per kWh in the future. Bioenergy, geothermal energy and hydropower will roughly stay the same or increase in price. Solar photovoltaic power, concentrated solar power, offshore and onshore wind power will decrease in price. Where both solar powers will reach the lowest point earlier than both wind powers.

Conclusion

To answer the question: Which non-fossil fuel energy technology will have the best price in the future?
Solar power will have the best price in the nearest future (before 2030). Wind power will have a better price too, but a few years later.

Appendix

All raw data that is used in this analysis is found at https://github.com/Twoltinge/Data.

The full python code that was used is found at https://github.com/Twoltinge/Final_assignment_CO2_emissions.

When executing the code in python, the dataframes that are used for this analysis are fully visible. It also includes all plots that are shown in this document.

There is no other relevant information to include.