Assignment for Data Analytics with Python

A data-driven analysis of CO_2 Emissions

This analysis examines global CO_2 emissions and their primary causal factors over time. There are three questions to be addressed:

Q1-What is the biggest predictor of a large CO_2 output per capita of a country?

Data Source

As an initial step, I identified the most relevant dataset containing CO_2 emissions data and associated contributing factors across countries over time. For this purpose, I utilized data from **GitHub**, originally sourced from **Our World in Data**.

Although this dataset was the most comprehensive available- covering over 100 years of CO_2 emissions alongside various contributing factors- it lacked data for some variables specified in the Winc Academy guidelines, such as dietary patterns, number of cars per capita, and mobility. Moreover, as other relevant data sets did not provide sufficient coverage across all countries over time, it was impractical to integrate them with the Our World in Data dataset.

Variable Selection

Addressing the question requires conducting a regression analysis to examine how much changes in CO_2 emissions (Y factor) can be explained by other factors (X factors). Accordingly, I selected the factors to be included in my analysis as follows:

Instead of utilizing the index of CO_2 emissions, I used CO_2 per-capita index to enable meaningful comparisons across countries, as this measure accounts for population size and reduces the bias introduced by differences in national populations.

Furthermore, regarding causal variables (X factors), I selected various factors from a multi-dimensional approach:

Dimension	Indicators	Definition
Economic	GDP and GDP per capita.	GDP is the total economic output of a country, whereas GDP per capita reflects the average economic output per person, indicating economic activity relative to population size.
Energy	Primary energy consumption and energy per capita.	It represents the total energy used from all sources, and energy per capita measures average energy use per person, capturing differences in energy demand across populations.

Other Gases	Methane and nitrous oxide per capita.	It quantifies average per-person emissions of these non-CO_2 greenhouse gases, highlighting their contribution to climate change beyond carbon dioxide.
Land Use	Land use change CO_2 per capita.	It measures average per-person emissions resulting from activities such as deforestation and land conversion

Data Cleaning Workflow

Subsequently, the dataset required a preparation phase through which the following steps are taken:

1. **Standardization:** Since the dataset included countries as well as regional and global totals, I used ISO country codes to filter out non-country entries and ensure that the analysis focused only on individual countries.
2. **Type Conversion:** I applied numeric coercion to convert any hidden text or corrupt data into NaN values, ensuring the code only attempts to calculate valid numbers. In other words, I carried out a code to ensure that all variables used in the regression analysis are numeric. It loops through the dependent variable and all independent variables, converting each column in the dataset to a numeric data type. Any values that cannot be converted (such as text or invalid entries) are replaced with missing values (NaN), preventing errors and ensuring the data is suitable for statistical analysis.
3. **Observation Filter:** I used the len() method to skip calculations if a country had fewer than 5 data points, as this provides insufficient data for reliable regression.
4. **Variance Check:** I implemented a check to ignore variables with zero variance. Since regression formulas require dividing by standard deviation (or variance), data that does not change causes mathematical errors (division by zero).

Analysis

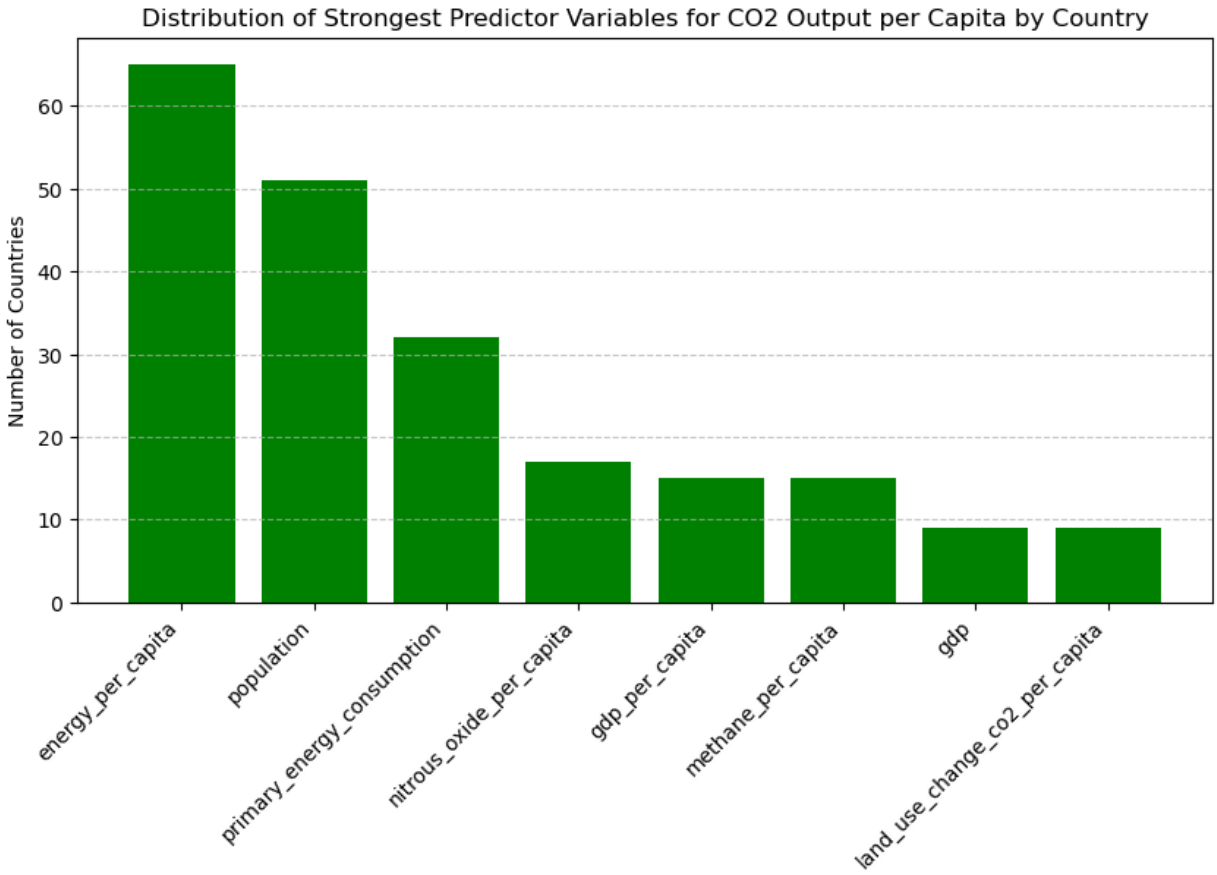
To examine the impact of multiple factors on CO_2 emissions per capita, I first conducted a correlation analysis to establish a foundation for the regression analysis. This step was necessary because separate regression models were required for each country in order to identify which factors most strongly contribute to national CO_2 emissions. The regression results were evaluated using key metrics such as regression coefficients, which indicate the magnitude and direction of each predictor's effect, and the R-squared value, which reflects how well the model explains variation in the dependent variable. In addition,

statistical significance and residual patterns were examined to assess the robustness and reliability of the results.

After performing regression analyses for all independent variables against the dependent variable, a Pandas script was used to identify, for each country, the factor most strongly associated with CO_2 emissions per capita.

Findings

In the graph below, countries are grouped according to their strongest contributing factor to CO_2 emissions per capita. As shown, “energy per capita” is the dominant factor for over 60 countries, while “population” is the most influential factor in more than 50 countries. It is important to note that “primary energy consumption” shows a weaker relationship with CO_2 emissions, as it does not account for population differences in energy use. Overall, the analysis suggests that investments aimed at reducing CO_2 emissions from energy consumption could have a significant impact on emissions in many countries.



Q2- Which countries are making the biggest strides in decreasing CO_2 output?

Addressing the question requires examining the changes of CO_2 emissions over years for each country to figure out which countries have had the biggest strides in decreasing CO_2 output.

Data Source

The most comprehensive dataset to address the question was accessed in **GitHub**, originally sourced from **Our World in Data**.

Variable Selection

Following the Winc Academy guidelines, CO_2 emissions per capita is used as the primary measure, since relying on total CO_2 emissions alone can be misleading due to changes in population size over time.

To compare changes in CO_2 emissions over time, the years 2015 and 2024 were selected as reference points for a clear comparison.

Since the dataset included countries as well as regional and global totals, I used ISO country codes to filter out non-country entries and ensure that the analysis focused only on individual countries.

Data Cleaning Workflow

The dataset was filtered to retain only CO_2 per capita data for the years 2015 and 2024. A separate copy of this filtered dataset was created to prevent changes to the original data. Additionally, all rows with missing values were removed to ensure a complete dataset for analysis.

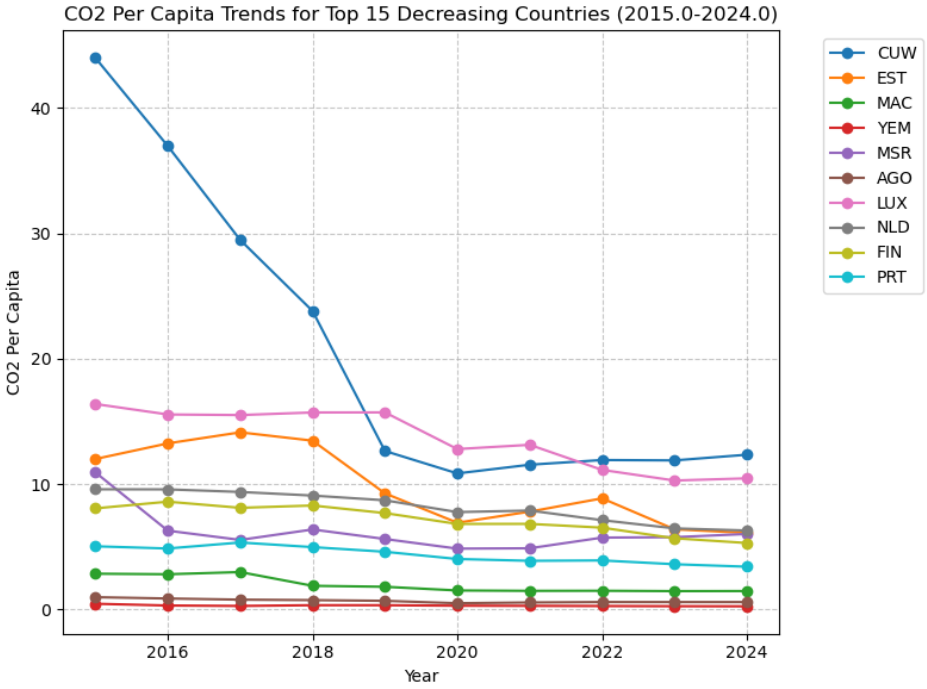
Analysis

The change in CO_2 emissions per capita was calculated by subtracting the 2015 value from the 2024 value and dividing the result by the 2015 value to obtain the percentage growth. Finally, the countries were ranked in descending order based on their percentage growth in CO_2 emissions per capita. Accordingly, countries with the largest positive change will appear at the top of the list, representing those that made the greatest progress in reducing CO_2 emissions.

Findings

The analysis demonstrates that the countries listed in the table below have made the biggest strides in decreasing CO_2 output:

iso_code	Country
CUW	Curaçao
EST	Estonia
MAC	Macau
YEM	Yemen
MSR	Montserrat
AGO	Angola
LUX	Luxembourg
NLD	Netherlands
FIN	Finland
PRT	Portugal



While most of the top ten countries in reducing CO_2 emissions experienced moderate changes, Curaçao achieved a dramatic positive reduction.

Q3- Which non-fossil fuel energy technology will have the best price in the future?

Answering the question requires a linear regressions analysis to predict the future price of non-fossil fuel energy technologies based on their values in the past.

Data Source

The most comprehensive dataset to address the question was accessed in **GitHub**, originally sourced from **Our World in Data**.

Variable Selection

The existing data included into analysis are listed in the table below.

Category	Indicator	Definition
Bioenergy	US$ per kilowatt-hour	Energy produced from organic materials (biomass), such as wood, crops, or waste, used to generate electricity or heat.
Geothermal	US$ per kilowatt-hour	Energy derived from the heat stored within the Earth, typically used to produce electricity or provide heating.
Offshore wind	US$ per kilowatt-hour	Electricity generated by wind turbines located in oceans or large bodies of water, taking advantage of stronger and more consistent winds.
Solar photovoltaic	US$ per kilowatt-hour	Energy generated by converting sunlight directly into electricity using solar panels composed of semiconductor materials.
Concentrated solar power	US$ per kilowatt-hour	Solar energy collected using mirrors or lenses to focus sunlight, generating heat that drives turbines to produce electricity.
Hydropower	US$ per kilowatt-hour	Energy produced from the movement of water, usually from rivers or dams, to drive turbines that generate electricity.
Onshore wind	US$ per kilowatt-hour	Electricity generated by wind turbines located on land, harnessing wind energy to produce power.

Data Cleaning Workflow

For this analysis, I filtered the dataset to include only the global-level rows, as examining individual countries is not required. To avoid the 'SettingWithCopyWarning' in Pandas and ensure that modifications are safely applied, I created an explicit copy of the filtered dataset.

I then renamed the columns to improve readability by shortening the column names. After sorting the data by year, I observed that only onshore wind technology contains data starting from 1984, while offshore wind technology has data beginning in 2000. For the remaining technologies, data is unavailable before 2010.

As a result, I decided to limit the analysis to the period 2010–2024, during which data is available for all technologies. Within this time range, only one value was missing, which I filled using linear interpolation. This method estimates the missing value while preserving continuity in the time series.

Finally, I displayed all rows of the modified Data Frame to verify that the interpolation was applied correctly. Handling missing values through interpolation ensures the dataset is complete and consistent, which is essential for reliable forecasting and analysis of global energy costs.

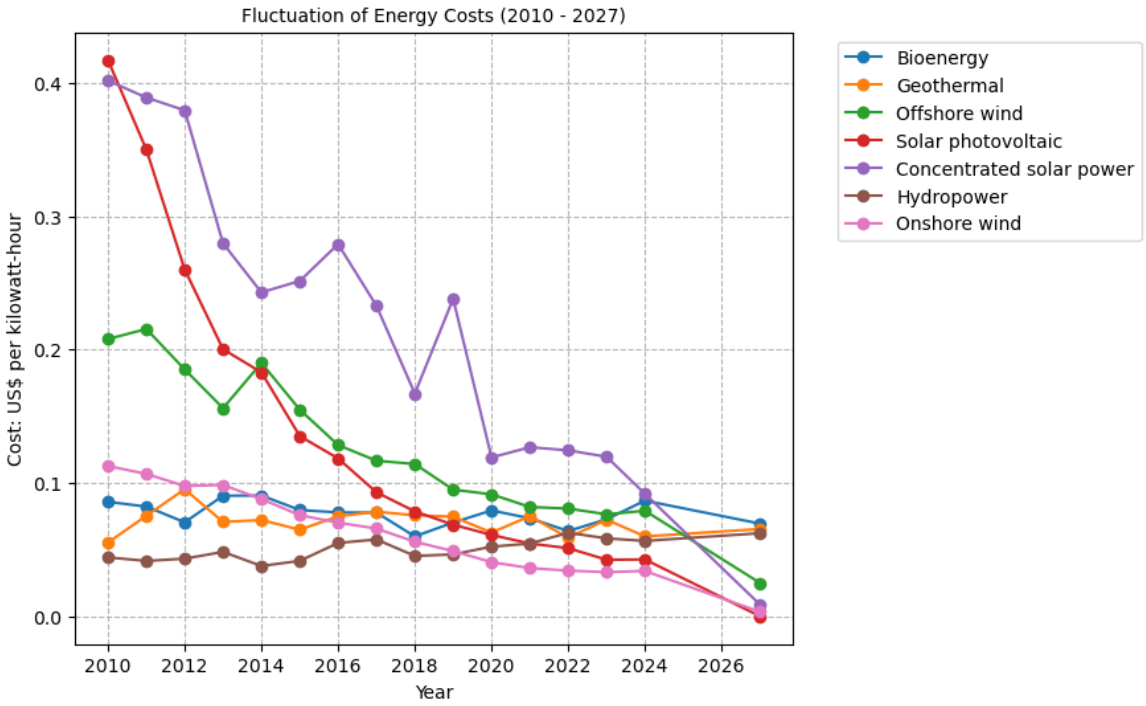
Analysis

A regression analysis was carried out to predict the future price of each non-fossil fuel energy technology (2027 as the target year) based on its earlier values (since 2010). However, I encountered a challenge as one of the predictions (Solar photovoltaic) resulted in a negative value, which is unrealistic for energy costs. Before proceeding to identify the lowest cost and visualize the results, I need to address this anomaly. I will consider zero value for Solar photovoltaic energy column.

Linear regression was applied to estimate and forecast the prices of energy technologies for the year 2027.

Findings

After the analysis, I added the 2027 year predicted value into the existing filtered table and visualized it- as can be seen in the table below.



The non-fossil fuel energy technologies with the most favorable predicted prices are in order:

	Non-fossil fuel energy technology	**Predicted levelized costs for 2027** (US$ per kilowatt-hour)
1	Solar photovoltaic	0
2	Onshore wind	0.0033
3	Concentrated solar power	0.0087
4	Offshore wind	0.0250
6	Hydropower	0.0622
7	Geothermal	0.0654
8	Bioenergy	0.0694