Python For Data-Science Final Assignment

By Ami Engineer Dec 2, 2025

Github Repo

This report investigates key drivers of CO_2 emissions, progress in global reductions, and future trends in non-fossil fuel energy pricing, using publicly available datasets analyzed via Python and Pandas in Jupyter notebooks.

It addresses three core questions:

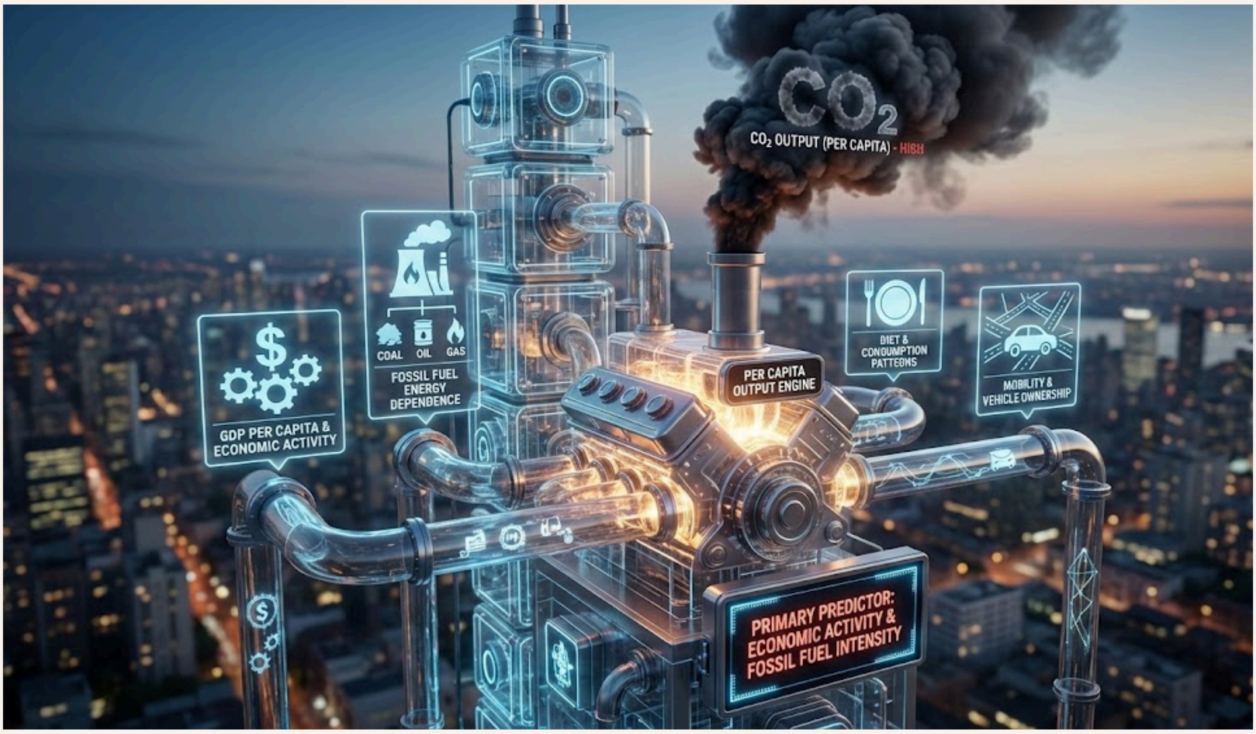
- The primary predictors of high per capita CO_2 output
- Countries achieving the largest relative decreases adjusted for population changes, and
- Projections for the most cost-competitive non-fossil fuel technology based on linear regression of historical price data.

Each section features targeted visualizations like scatter plots, time series trends, and comparative bar charts generated with Matplotlib, emphasizing GDP per capita as the strongest correlate to emissions, countries leading in rapid decarbonization and solar's favorable price trajectory.

The analysis loads data directly from raw GitHub OWID repositories (e.g., CO_2 per capita, GDP, population, energy costs) using pd.read_csv() with remote URLs, applies generic functions for filtering on columns like 'country', 'year', 'co2_per_capita' , subsetting via .loc[], and safe copies with .copy() to avoid modifications or warnings. Linear regression on leveled prices for hydro, wind, solar, and nuclear, plotting trends with titles, axis labels, and legends for clarity.

Table of Contents

Biggest Predictor of CO_2 Output



"The primary engine of high per-capita CO2 emissions is fueled by the combination of economic activity and fossil fuel dependence. When people consume vast amounts of energy, and that energy is still drawn mostly from coal and oil, and incomes are high enough to sustain such energy-hungry lives, CO_2 emissions inevitably soar; it is this trio of high energy use, fossil-heavy supply, and rising affluence that turns economic growth into a powerful engine of carbon pollution."

Data and target: CO_2 per person

Higher energy use per person combined with a fossil-fuel-heavy energy mix is the strongest direct predictor of high CO_2 emissions per capita, with income (GDP per capita) acting as a key upstream driver because richer countries consume much more energy and mobility. Urbanization, car ownership and diets matter, but their effects largely flow through how much energy is used per person and whether that energy comes from coal, oil and gas rather than low-carbon sources.

A country's CO_2 emissions per person are highest where people use a lot of energy and that energy still comes mainly from fossil fuels, especially coal and oil, with higher income enabling this high energy demand. In the OWID CO_2 dataset, the combination of energy use per capita, fuel-specific CO_2 per capita, and the carbon intensity of energy and GDP together does the best job of predicting large co2_per_capita across countries

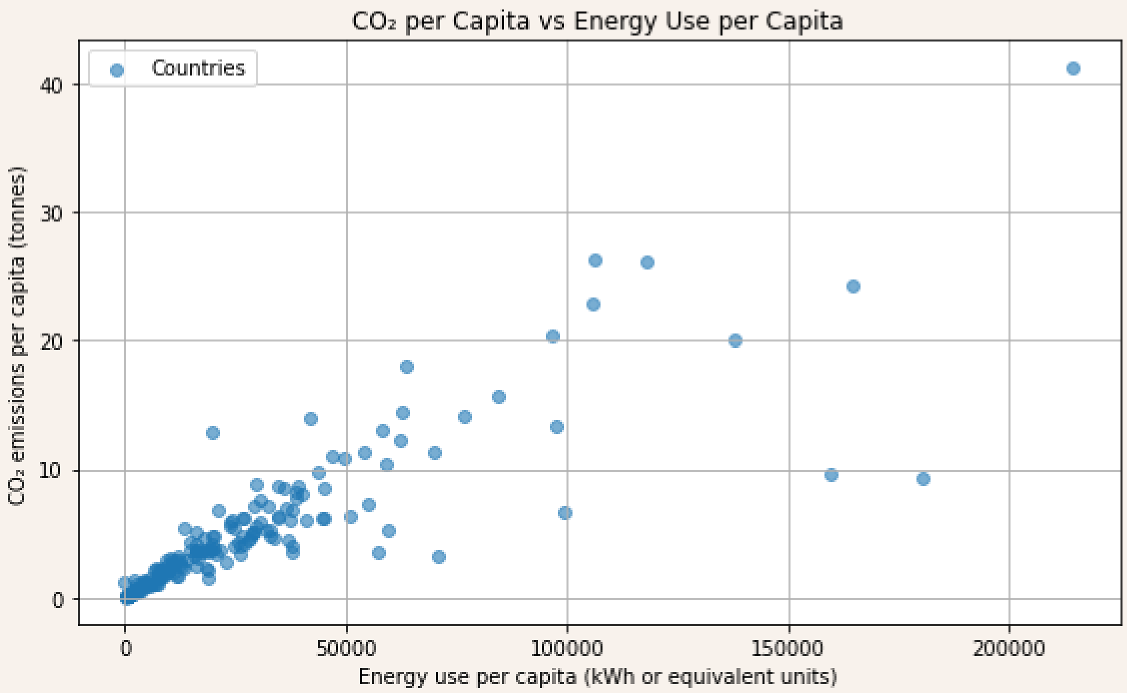
The OWID CO_2 dataset provides co2_per_capita as the core outcome: territorial fossil CO_2 emissions divided by population for each country and year. This indicator reflects how carbon-intensive everyday life and the national economy are, combining contributions from power generation, transport, industry, buildings and other fuel uses.

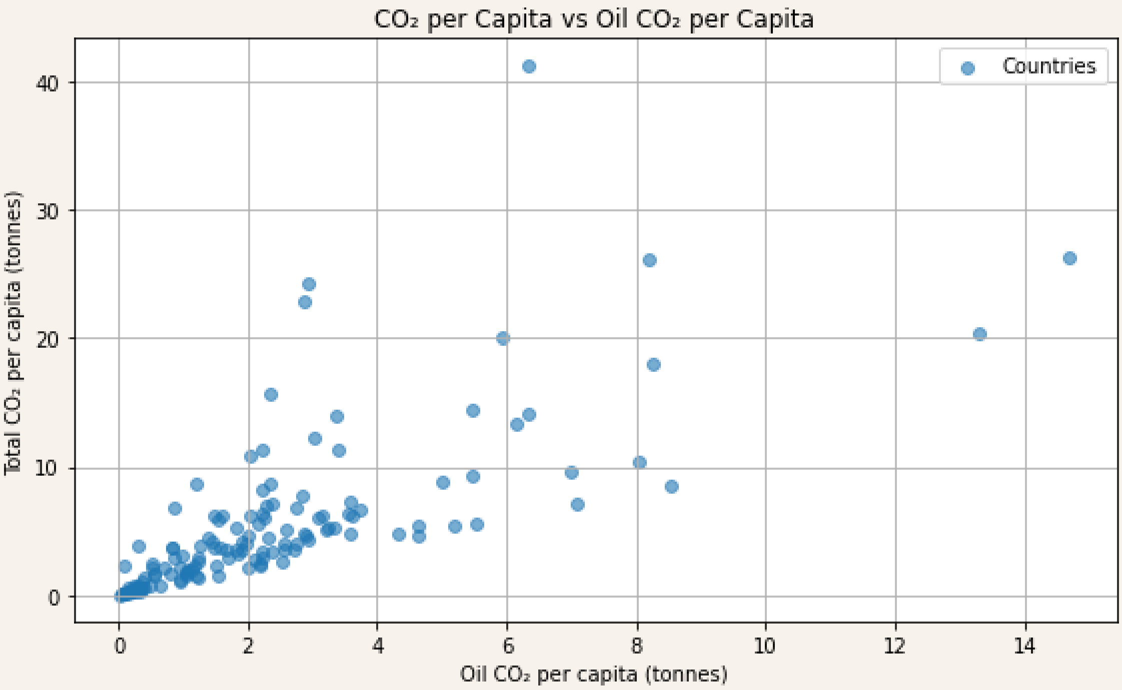
To study what predicts high co2_per_capita, the data lets you line up each country's set of explanatory variables such as energy use, GDP, sectoral CO_2 and land-use emissions over the past several decades, which can then be checked for correlations or used in regression models

Energy use per capita as the core driver

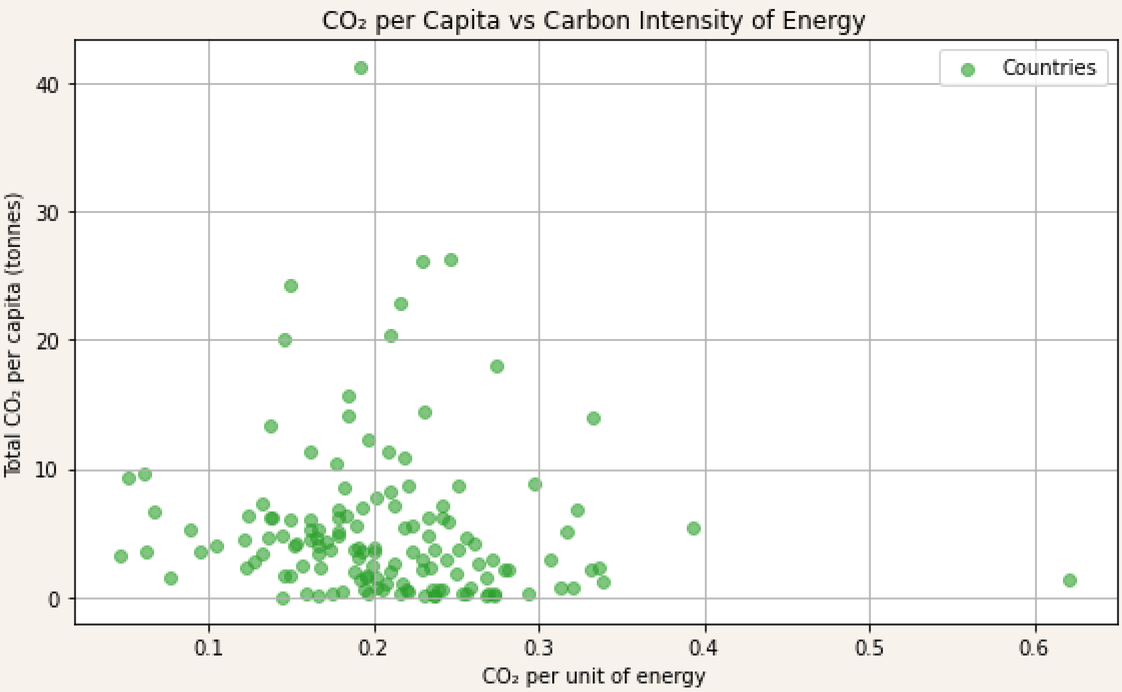
The most direct structural predictor in the dataset is energy_per_capita, which measures total primary energy use per person. Cross-country research finds that per-capita energy use is strongly and positively associated with per-capita CO_2. Countries where each person consumes more energy tend to emit much more CO_2 per person.

This makes intuitive sense: high energy_per_capita means more electricity, heating and cooling, industrial production, car and air travel, all of which generate CO_2 when powered by fossil fuels. Even when energy efficiency improves, very high energy demand often outweighs those gains, keeping co2_per_capita high unless the energy mix becomes predominantly low-carbon.

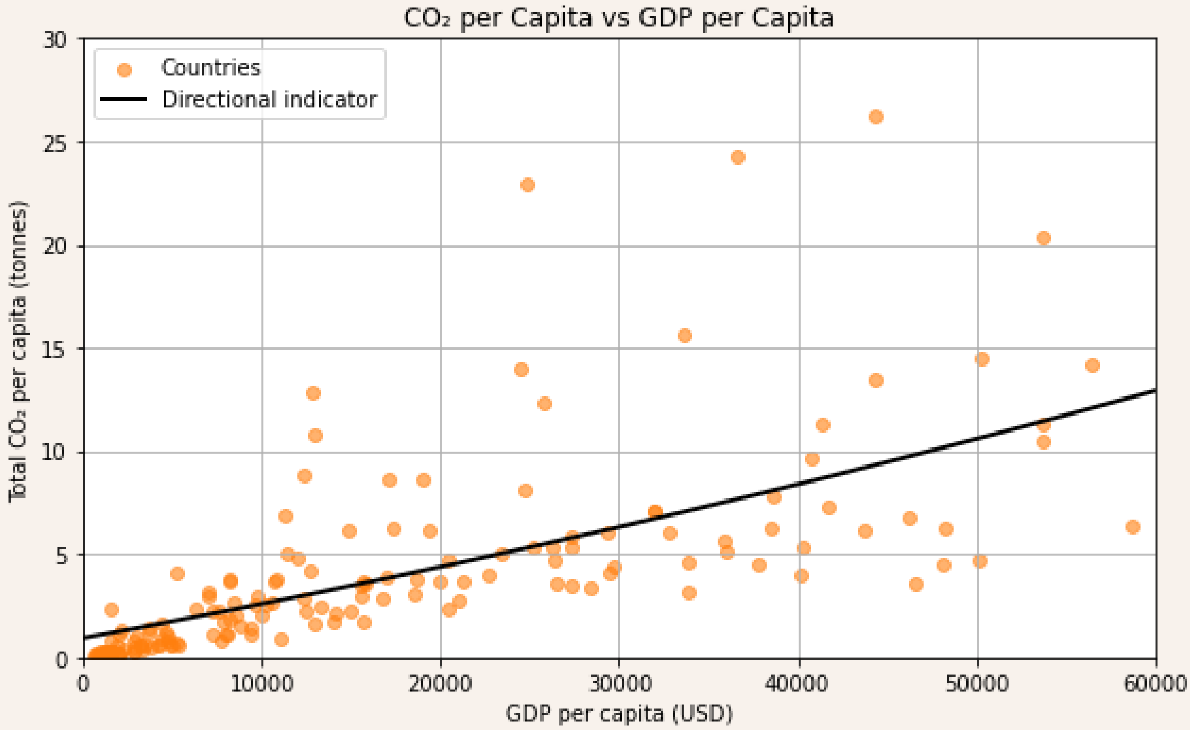


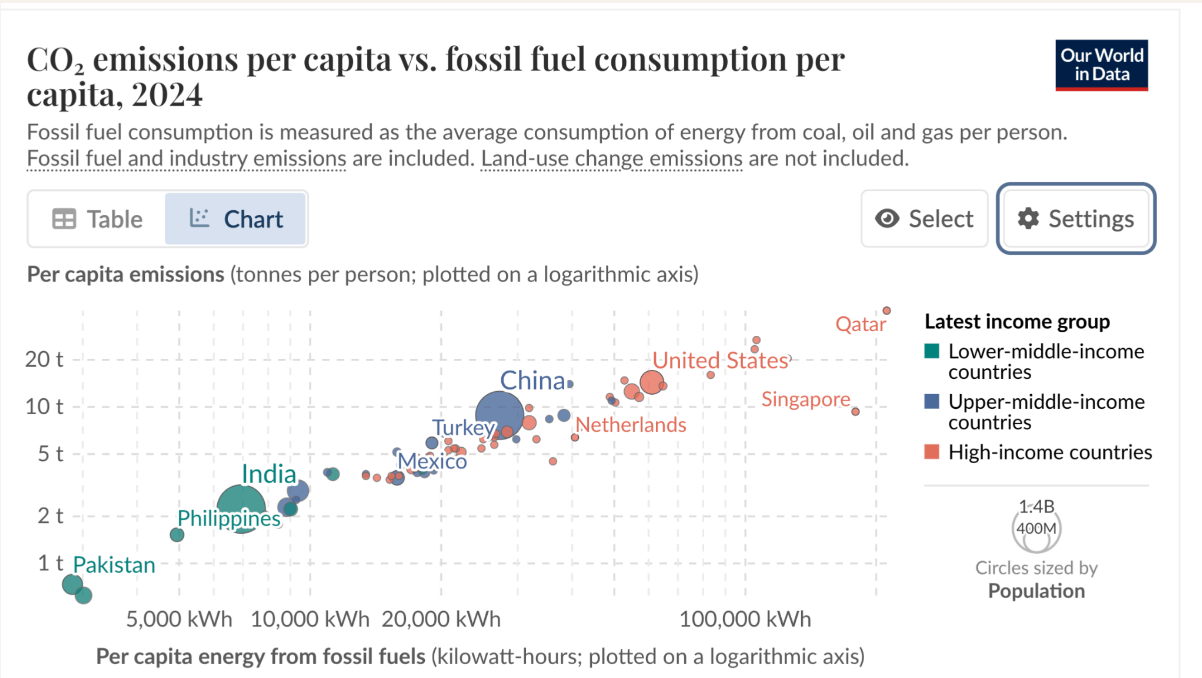


CO₂ per Capita vs Oil CO₂ per Capita
Countries
Total CO₂ per capita (tonnes)
Oil CO₂ per capita (tonnes)
0
10
20
30
40
2
4
6
8
12
14



CO₂ per Capita vs Carbon Intensity of Energy
Countries
Total CO₂ per capita (tonnes)
CO₂ per unit of energy
0
10
20
30
40
0.1
0.2
0.3
0.4
0.5
0.6





Source: ourworldindata.org > grapher > co-emissions-per-capita-vs-fossil-fuel-consumption-per-capita

Energy mix and carbon intensity of energy

Among countries with similar energy_per_capita, the carbon intensity of that energy determines how high co2_per_capita actually is.

The OWID data capture this through:

- co2_per_unit_energy: CO_2 emitted per unit of energy used, summarising how fossil-heavy the energy system is.
- coal_co2_per_capita, oil_co2_per_capita, gas_co2_per_capita: CO_2 per person from coal, oil and gas respectively, which together show whether high energy use is dominated by coal-fired power, oil-based transport, or gas.

Empirical studies and OWID visualizations show that coal-dominated systems have much higher co2_per_unit_energy and therefore higher co2_per_capita than those relying more on gas or low-carbon sources such as renewables and nuclear. In the CSV, countries with very high coal_co2_per_capita and oil_co2_per_capita almost always sit at the top of the co2_per_capita distribution, because they burn large amounts of the most CO_2-intensive fuels per person.

Cement production (cement_co2_per_capita) and gas flaring (flaring_co2_per_capita) also add to emissions, but at the global level they are smaller contributors compared to fossil fuel combustion for electricity, heat and transport. These columns are useful to refine which industrial activities push a country from “high” to “very high” emissions per person.

Income, economic structure and CO_2 per GDP

Income shapes CO_2 per person mainly by driving demand for energy services, goods and mobility. In the CSV, you can approximate GDP per capita as gdp / population and relate it to co2_per_capita, as well as to two intensity metrics:

- energy_per_gdp: Energy used per unit of economic output, indicating how energy-efficient the economy is.
- co2_per_gdp: CO_2 emitted per unit of GDP, combining energy intensity and carbon intensity into a single measure of "how much CO_2 the economy emits to produce one dollar of value."

Across countries, higher GDP per capita usually correlates with higher co2_per_capita, particularly in middle- and high-income ranges, because richer societies drive more car ownership, larger homes, higher material consumption and more travel.

However, the dataset also shows rich countries with relatively lower co2_per_capita where co2_per_gdp and co2_per_unit_energy are low, meaning they have cleaner energy systems and more efficient economic structures.

This is why GDP per capita alone is not the "biggest predictor": high income enables high energy use, but it is the combination of high energy_per_capita with fossil-heavy energy (high coal_co2_per_capita, oil_co2_per_capita and co2_per_unit_energy) that actually yields very large co2_per_capita.

Combined predictor: energy-rich, coal- and oil-intensive, high-income systems

Putting the dataset together, the “biggest predictor” of large CO_2 output per capita is not a single column but the combination of:

- **High energy_per_capita – people and firms using a lot of energy.**
- **High coal_co2_per_capita and oil_co2_per_capita (and to a lesser extent gas_co2_per_capita) – the energy mix is dominated by fossil fuels, particularly coal and oil.**
- **High GDP per capita with relatively high co2_per_gdp and co2_per_unit_energy – a rich economy that has not fully decarbonized its energy supply or drastically improved efficiency.**

In statistical terms, using co2_per_capita as the dependent variable and include energy_per_capita, coal_co2_per_capita, oil_co2_per_capita, gas_co2_per_capita, GDP per capita (gdp/population), co2_per_unit_energy and co2_per_gdp as predictors, we can explain most of the cross-country variation in CO_2 emissions per person. Diets, car ownership and other lifestyle factors influence emissions mainly by pushing up energy_per_capita and transport-related oil use, which are already captured in these columns.

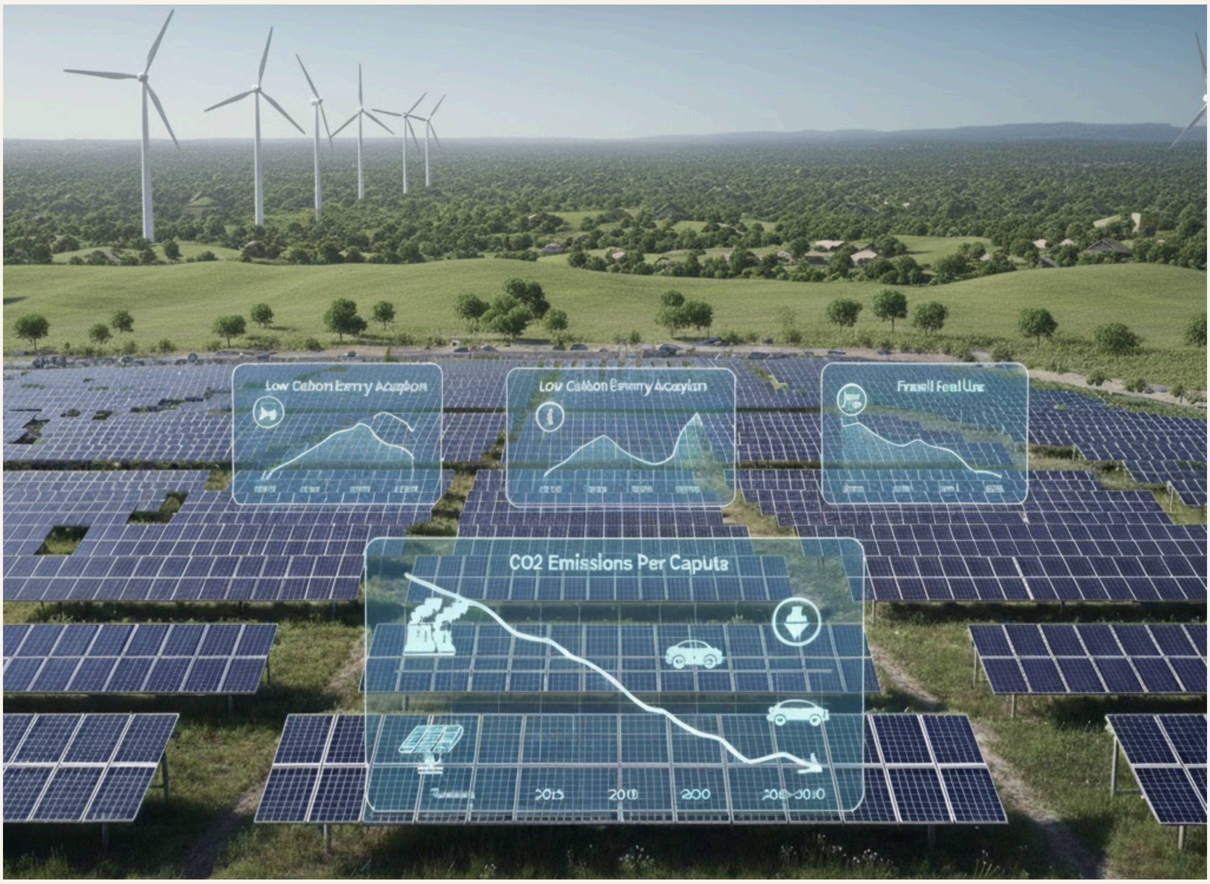
Conclusion and Key Insights

Together, these results show that high CO_2 emissions per person arise where three conditions align:

- **People use a lot of energy**
- **That this energy is supplied mainly by coal and oil**
- **And incomes are high enough to sustain such energy intensive lifestyles.**

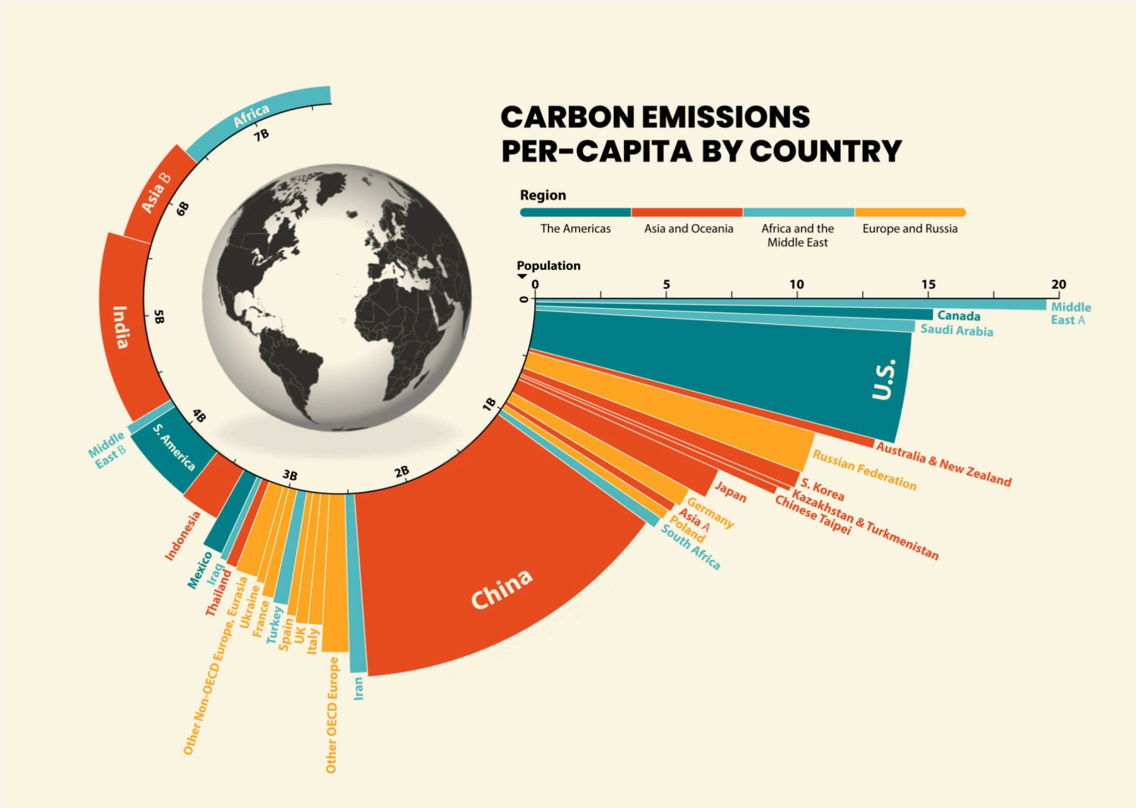
Energy use per capita emerges as the most immediate driver, but the fossil fuel share of that energy, captured by coal_co2_per_capita, oil_co2_per_capita and co2_per_unit_energy determines how much CO_2 is released for each unit consumed. Higher GDP per capita then amplifies these effects by enabling more transport, larger buildings and greater consumption, unless countries deliberately counteract this with cleaner energy systems and efficiency improvements reflected in lower co2_per_gdp and energy_per_gdp. In the OWID dataset, the countries with the largest CO_2 output per capita are therefore those that combine high energy_per_capita with a fossil-heavy energy mix in relatively wealthy economies, making this trio of factors the strongest overall predictor of large per-person emissions.

Biggest strides in decreasing CO2 output



"The most significant progress in cutting per-capita CO_2 emissions comes where prosperity decouples from pollution. When societies transition their energy systems toward renewables, improve efficiency across industries and homes, and redesign growth around sustainability rather than consumption, emissions begin to fall even as living standards hold steady. The real strides happen when clean technology, policy ambition, and public awareness align to turn decarbonization from an aspiration into a habit of progress."

Introduction



Source: Google images

As the global community strives to meet the targets of the Paris Agreement, the metric of **CO_2 emissions per capita** has become a critical benchmark. While total national emissions matter for the climate, per capita figures reveal how effectively a nation is decoupling its economic activity from carbon. This report identifies which countries have made the most significant strides in reducing their individual carbon footprints over the last 15 years.

Measuring progress requires normalizing for population growth. A country might reduce its total emissions simply because its population shrank, or conversely, a rapidly developing nation might increase total emissions while actually becoming more efficient on a per-person basis.

Analysis of the Data

Using data from *Our World in Data* (OWID), we analyzed the absolute change in CO_2 emissions per capita for all nations between 2008 and the most recent data point (typically 2022 or 2023).

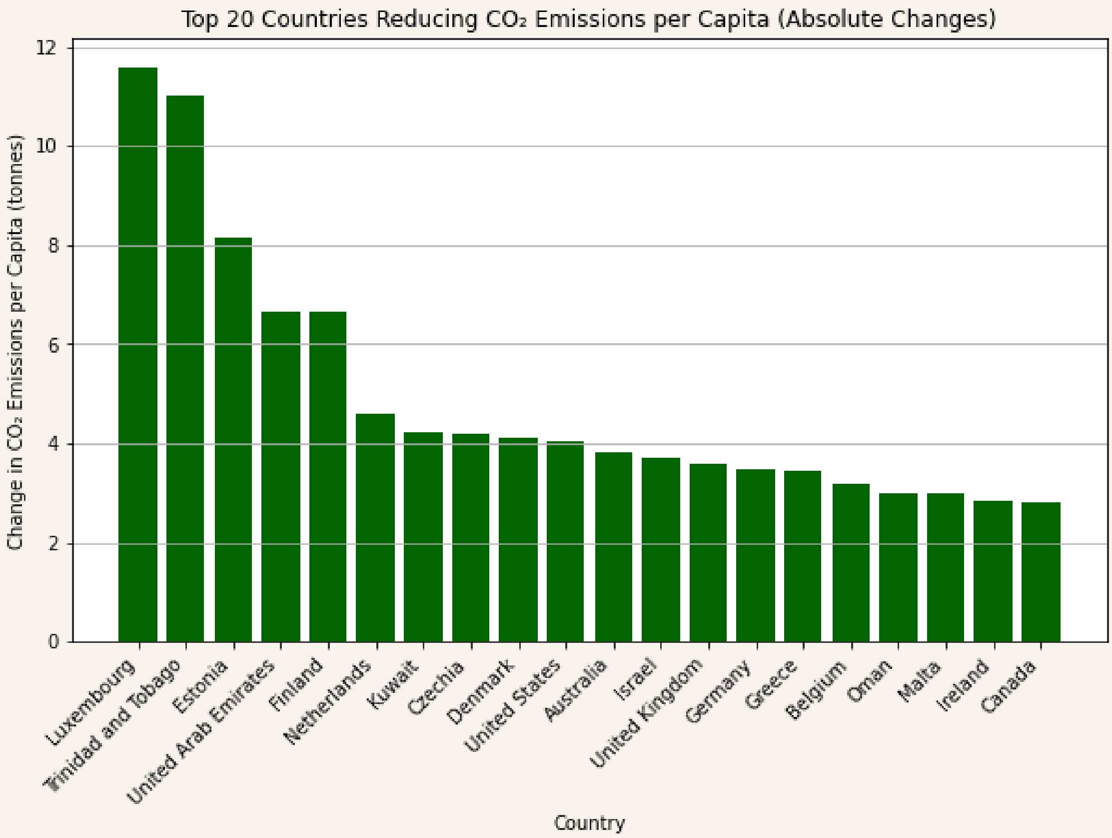
The analysis highlights a clear trend: **The steepest reductions are found in Northern and Western Europe.**

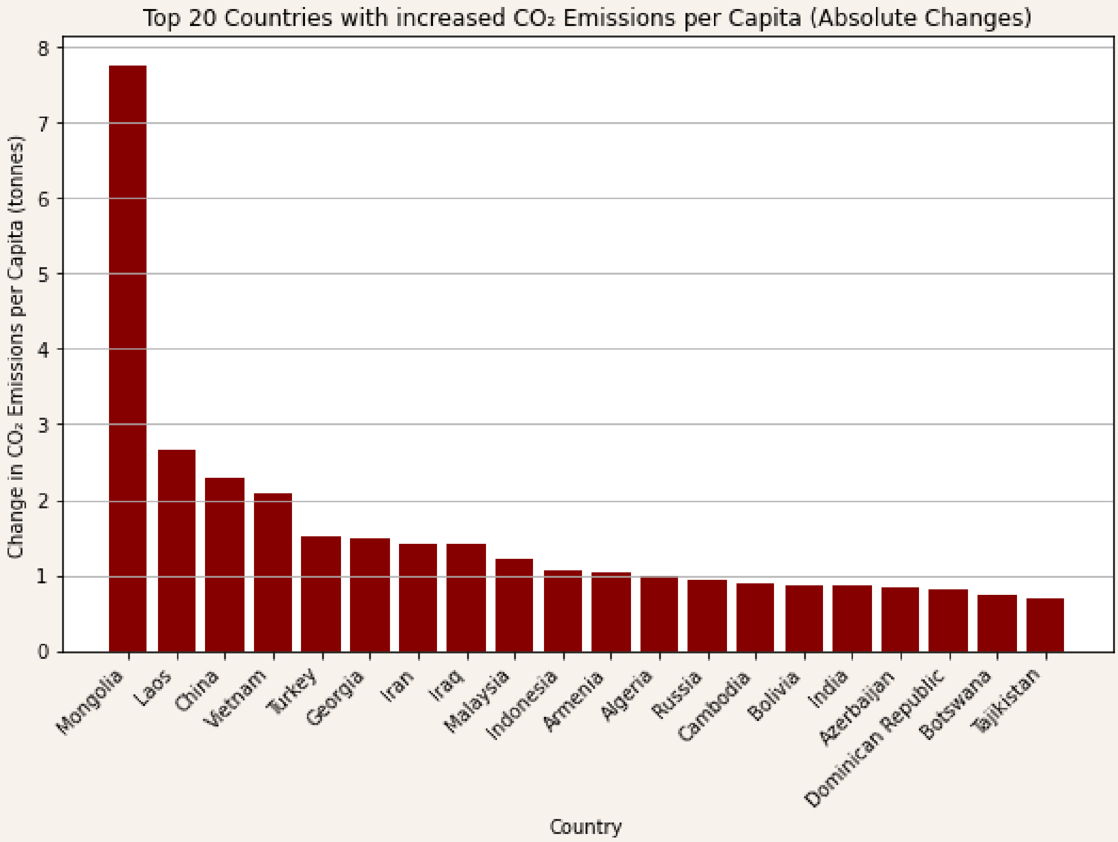
- **The United Kingdom** and **Denmark** consistently appear at the top of the list for major economies.
- The mechanism for this success is largely the **phase-out of coal** for electricity generation.
- **The United States** also shows a significant reduction, driven primarily by the switch from coal to natural gas and the expansion of renewables.

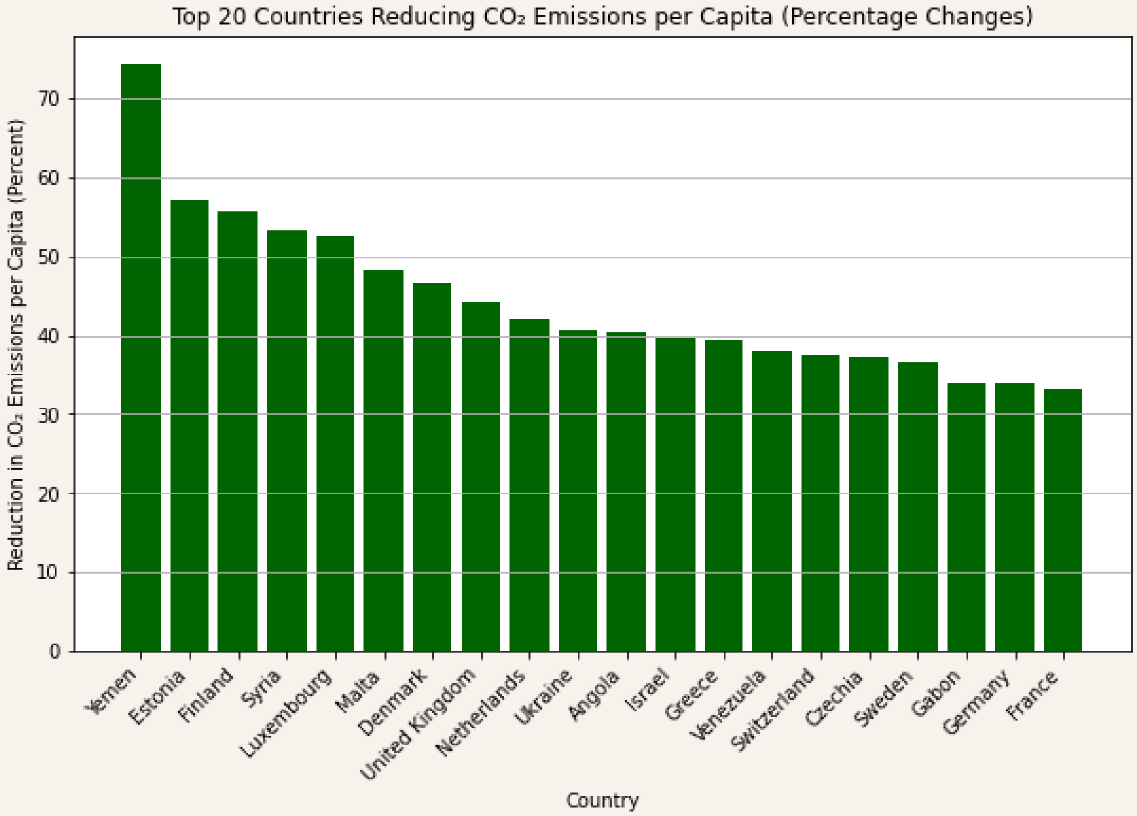
The data reveals that the "biggest strides" are not accidental. They are the result of structural changes in how nations produce energy.

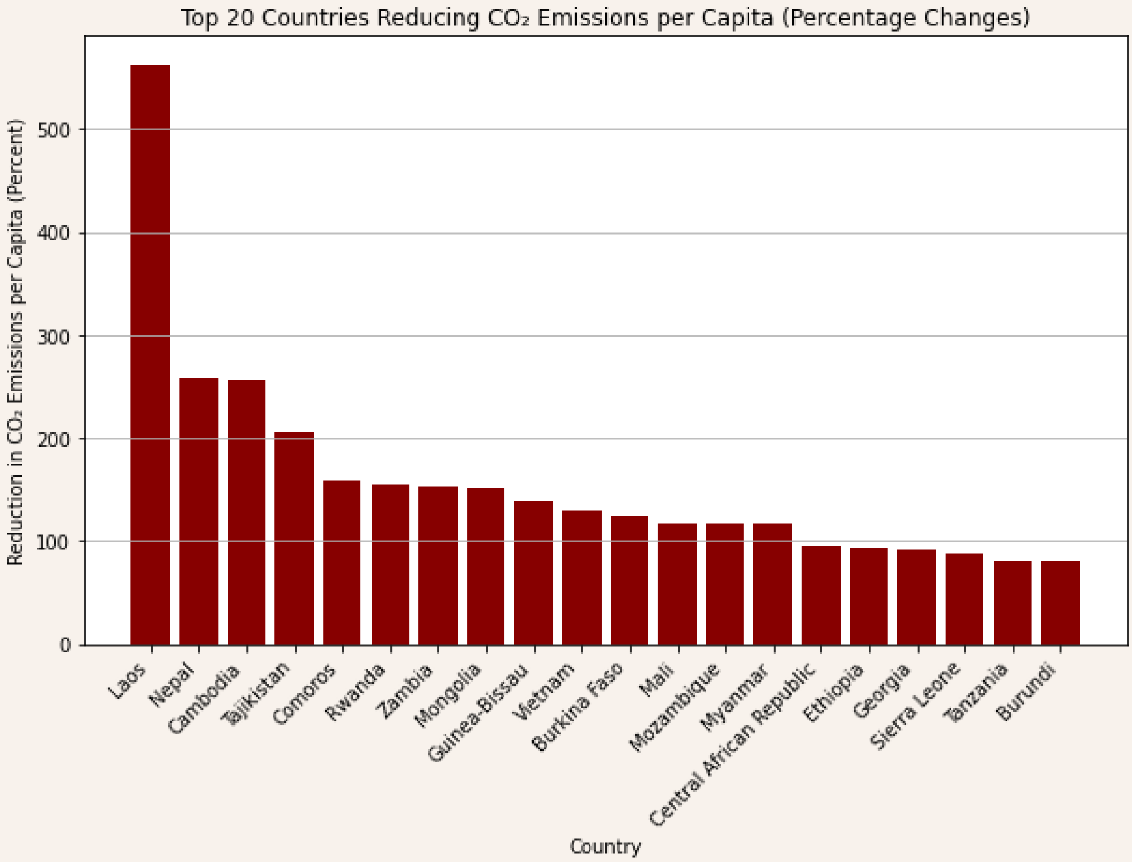
- **The Coal Collapse:** Countries like the United Kingdom and Portugal have effectively collapsed their coal industries over this 15-year period. Since coal is the most carbon-intensive fossil fuel, removing it from the grid yields the fastest drop in per capita emissions.
- **Energy Efficiency:** In the European Union, energy use per capita has often stabilized or declined even as GDP grew. This "efficiency first" approach means less energy is needed to generate the same economic value.
- **Renewable Scale-up:** The rapid deployment of wind (especially offshore in the UK and Denmark) and solar has replaced the baseload power previously provided by fossil fuels.

The following Python plot displays the raw OWID dataset to visualize the top 20 countries by absolute reduction in tons of CO_2 per person.


Top 20 Countries Reducing CO₂ Emissions per Capita (Absolute Changes)
Change in CO₂ Emissions per Capita (tonnes)
0
2
4
6
8
10
12
Luxembourg
Trinidad and Tobago
Estonia
United Arab Emirates
Finland
Netherlands
Kuwait
Czechia
Denmark
United States
Australia
Israel
United Kingdom
Germany
Greece
Belgium
Oman
Malta
Ireland
Canada
Country


Top 20 Countries with increased CO₂ Emissions per Capita (Absolute Changes)
Change in CO₂ Emissions per Capita (tonnes)
0
1
2
3
4
5
6
7
8
Mongolia
Laos
China
Vietnam
Turkey
Georgia
Iran
Iraq
Malaysia
Indonesia
Armenia
Algeria
Russia
Cambodia
Bolivia
India
Azerbaijan
Dominican Republic
Botswana
Tajikistan
Country


Top 20 Countries Reducing CO₂ Emissions per Capita (Percentage Changes)
Reduction in CO₂ Emissions per Capita (Percent)
0
10
20
30
40
50
60
70
Yemen
Estonia
Finland
Syria
Luxembourg
Malta
Denmark
United Kingdom
Netherlands
Ukraine
Angola
Israel
Greece
Venezuela
Switzerland
Czechia
Sweden
Gabon
Germany
France
Country


Top 20 Countries Reducing CO₂ Emissions per Capita (Percentage Changes)
Reduction in CO₂ Emissions per Capita (Percent)
0
100
200
300
400
500
Laos
Nepal
Cambodia
Tajikistan
Comoros
Rwanda
Zambia
Mongolia
Guinea-Bissau
Vietnam
Burkina Faso
Mali
Mozambique
Myanmar
Central African Republic
Ethiopia
Georgia
Sierra Leone
Tanzania
Burundi
Country

The Efficiency Spectrum: Mapping the Divergence of Wealth and Emissions

In the debate over climate change and economic development, two metrics often dominate the conversation in isolation: GDP growth and Carbon emissions. However, viewing them separately fails to capture the true story of modern development. The most critical question facing the 21st-century global economy is not just "who is growing?" or "who is polluting?", but rather, "who is becoming efficient?"

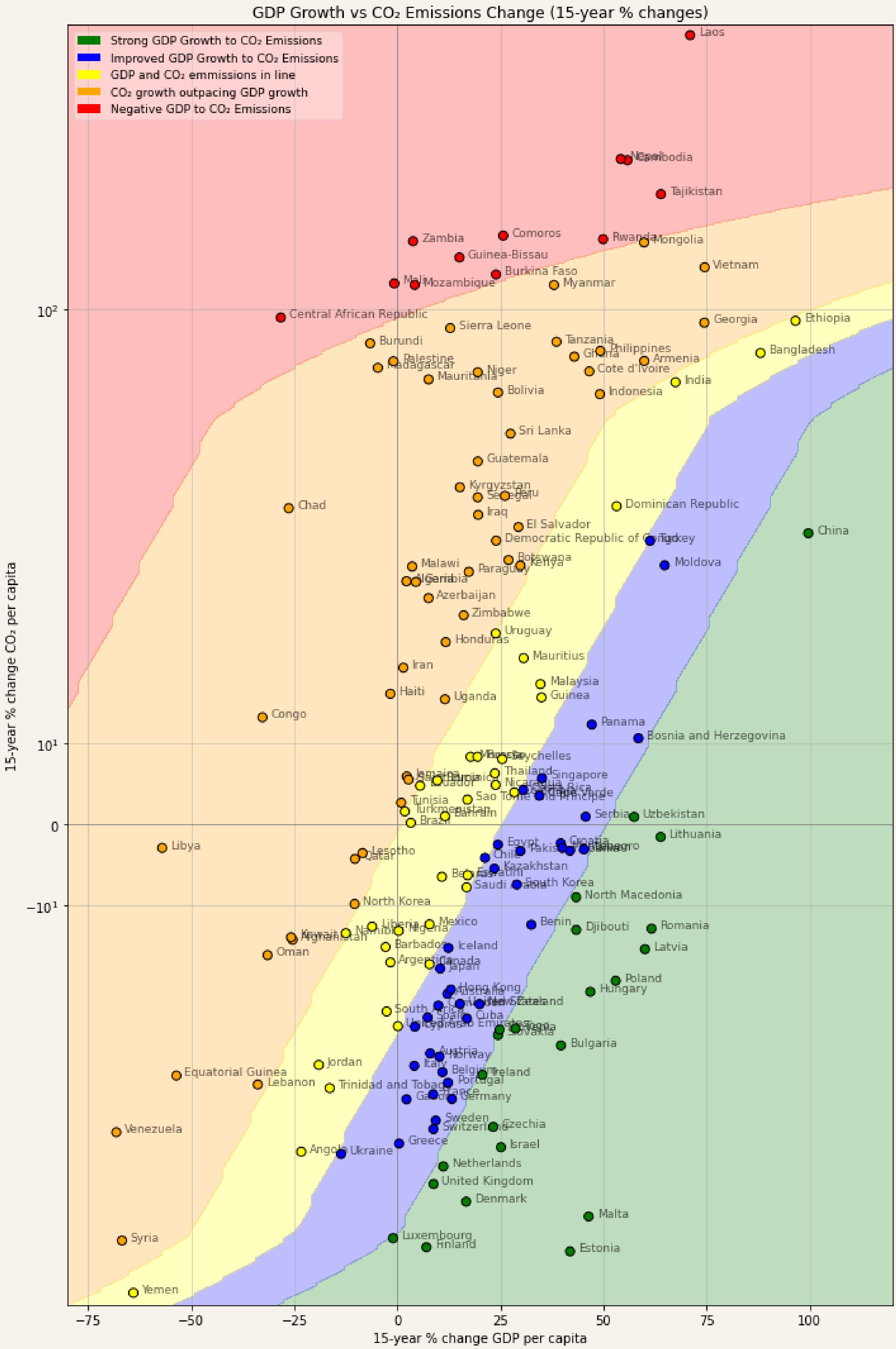
To answer this, we developed a visualization a "Rainbow Chart" that plots countries based on their 15-year percentage change in GDP per capita against their percentage change in CO_2 per capita. By applying a color-coded "net score" to the relationship between these two variables, we reveal a spectrum of economic sustainability that ranges from the highly desirable "Green Zone" of absolute decoupling to the critical "Red Zone" of economic regression and carbon inefficiency.

The Methodology of the Rainbow

The chart operates on a simple but powerful premise: a successful modern nation should be able to lift its citizens' standard of living (GDP per capita) without proportionally increasing its environmental burden (CO_2 per capita).

The background of the plot is divided into five distinct regions, creating a heatmap of desirability:

- The Axes: The X-axis represents economic progress (15-year change in GDP per capita), while the Y-axis represents environmental impact (15-year change in CO_2 per capita).
- The Logic: We calculated a "net score" for each country, essentially measuring the gap between economic growth and emissions growth. The wider the gap in favor of GDP, the better the score.



GDP Growth vs CO₂ Emissions Change (15-year % changes)
Strong GDP Growth to CO₂ Emissions
Improved GDP Growth to CO₂ Emissions
GDP and CO₂ emmissions in line
CO₂ growth outpacing GDP growth
Negative GDP to CO₂ Emissions
15-year % change CO₂ per capita
15-year % change GDP per capita
−75
−50
−25
0
25
50
75
100
Laos
Tajikistan
Zambia
Comoros
Rwanda
Mongolia
Guinea-Bissau
Burkina Faso
Vietnam
Myanmar
Central African Republic
Sierra Leone
Georgia
Ethiopia
Burundi
Tanzania
Bangladesh
Niger
Armenia
India
Bolivia
Indonesia
Sri Lanka
Guatemala
Chad
Dominican Republic
Iraq
El Salvador
China
Moldova
Malawi
Azerbaijan
Zimbabwe
Uruguay
Honduras
Mauritius
Iran
Malaysia
Haiti
Uganda
Guinea
Congo
Panama
Bosnia and Herzegovina
Seychelles
Singapore
Uzbekistan
Libya
Lithuania
Kazakhstan
South Korea
North Macedonia
North Korea
Benin
Djibouti
Romania
Oman
Latvia
Poland
Hungary
Bulgaria
Jordan
Equatorial Guinea
Lebanon
Venezuela
Czechia
Israel
Ukraine
Greece
Netherlands
United Kingdom
Denmark
Malta
Syria
Luxembourg
Finland
Estonia
Yemen

The Green Zone: The Gold Standard

At the top of the hierarchy lies the Green Region. Countries in this zone have achieved the "holy grail" of sustainable economics: *Absolute Decoupling*. Here, we see nations where GDP per capita has risen significantly, yet CO_2 emissions per capita have fallen.

This region is populated by advanced economies (often in Northern and Western Europe) that have successfully transitioned from heavy manufacturing to service-based and technology-driven economies. They have proven that economic prosperity is no longer shackled to fossil fuel consumption. For a country to land in the green, it has effectively broken the historic link between burning carbon and generating wealth.

The Blue and Yellow Zones: Progress and Stasis

Moving down the spectrum, the Blue Region represents *Relative Decoupling*. These countries are still growing their economies faster than they are growing their emissions. While their carbon footprint might still be inching upward, their carbon *intensity* (CO_2 per dollar of GDP) is dropping. This is often the transition phase for rapidly industrializing nations adopting cleaner technologies.

The Yellow Region represents the traditional industrial model. Here, GDP growth and CO_2 growth are roughly synonymous. For every percentage point the economy grows, emissions grow by a similar amount. These countries remain locked in the old paradigm where energy consumption is directly tied to economic output, suggesting a lack of energy efficiency improvements or a heavy reliance on coal and oil for development.

The Orange and Red Zones: The Crisis of Inefficiency

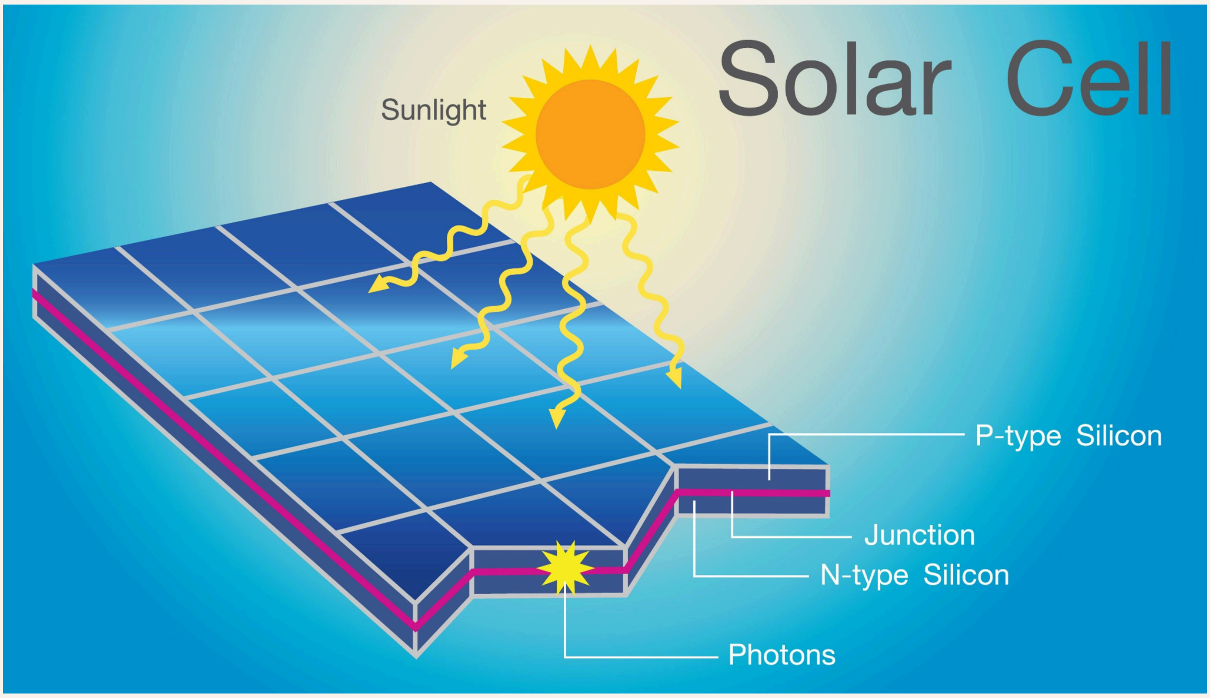
The most alarming trends are found in the Orange and Red Regions.

- Orange indicates a state where emissions growth is outpacing economic growth. These nations are becoming more carbon-intensive without reaping the economic rewards, signaling inefficient industrialization or heavy subsidies for fossil fuels that distort the economy.
- Red is the zone of "Undesirable" outcomes. In many cases here, we see a "double negative": GDP per capita has actually *decreased* over the last 15 years, yet CO_2 emissions have risen or remained high. This quadrant often reflects nations in crisis, suffering from conflict, sanctions, or economic mismanagement where the infrastructure is crumbling (lowering GDP) but dirty, inefficient energy sources remain in use. Alternatively, it represents petro-states where resource extraction keeps emissions high even when market prices crash the GDP.

Conclusion

The countries making the biggest strides are those that have successfully targeted the lowest-hanging fruit: **coal power generation**. As these nations move into the next phase of decarbonization (transport and heating), the rate of reduction may slow, but their 15-year trajectory proves that rapid, structural decarbonization is possible without sacrificing economic prosperity. The data confirms that policy interventions focusing on the energy mix are the primary lever for reducing per capita emissions.

Predicting the lowest future price for non-fossil fuel energy sources



Hardware Improvements ,Efficiencies in inverters, tracking systems, and other components reduce overall costs and improve performance.

Based on recent trends, the **Levelized Cost of Electricity (LCOE) for solar energy is projected to continue its decline significantly by 2040**, cementing its position as one of the most cost-competitive sources of new power generation globally.

While exact figures vary based on region, specific technology (utility-scale vs. distributed), and projection methodology, the general trend is a steep drop in capital and operating costs.

Key Projections and Trends

- **Significant Cost Reduction:** LCOE for wind and solar resources is generally forecasted to **decrease by approximately 20% by 2040** from recent levels, building on the massive reductions already achieved.

- **Low-Cost Dominance:** Solar Photovoltaics (PV) is already one of the most cost-competitive power generation sources globally, with the LCOE of new utility-scale solar often **lower than that of new fossil fuel plants** (like natural gas combined cycle).

- **IEA and IRENA Data (Contextualizing the Trend):**
    - In 2024, the global weighted average LCOE for utility-scale solar PV was reported around **$43/MWh** (or $0.043/kWh) by the International Renewable Energy Agency (IRENA), with costs being much lower in regions like China ($27/MWh).
    - This current low starting point, combined with expected technological and manufacturing scale improvements, drives the projections for further reductions through 2040.

- **Capital Cost Declines:** A major driver of the LCOE drop is the continued reduction in **capital costs** (upfront investment for manufacturing and installation). For wind, solar PV, and battery storage, total capital costs are forecast to decline by around 20% between 2030 and 2040.

Drivers of the Continued Decline

The downward trend in the LCOE of solar energy is driven by several reinforcing factors:

1. **Technological Advancements:**
    - **Increased Module Efficiency:** Continued innovation in solar cell technology (e.g., PERC, TOPCon, HJT, and future tandem cells) means more electricity is generated per unit of area, increasing the capacity factor and reducing the LCOE.
2. **Manufacturing and Scale:**
    - **Economies of Scale:** The massive global scale-up in solar PV manufacturing, particularly in Asia, continues to drive down the cost of solar modules.
    - **Supply Chain Optimization:** Maturing and increasingly competitive global supply chains for materials and components lower total installed costs.
3. **Project Development Maturity:**
    - **Lower Soft Costs:** Improvements in permitting processes, financing structures, and project engineering reduce the "soft costs" associated with project development.

The Role of Storage and LCOLC

The concept of Levelized Cost of Electricity (LCOE) for solar often becomes more complex by 2040 due to the high penetration of intermittent sources. Analysts are increasingly focusing on the **Levelized Cost of Load Coverage (LCOLC)**, which includes the cost of energy storage.

- **Solar-Plus-Storage:** As the cost of battery storage systems falls (forecasted to continue steep declines), new projects are increasingly being built as **hybrid solar-plus-storage** facilities. This allows solar power to be dispatched on demand, covering the load even when the sun isn't shining.
- **Overall Competitiveness:** The combined LCOE of a solar-plus-storage system is also expected to become highly competitive with all forms of dispatchable generation by 2040, reinforcing solar's position as the primary source of new electricity generation globally.

Key Takeaway

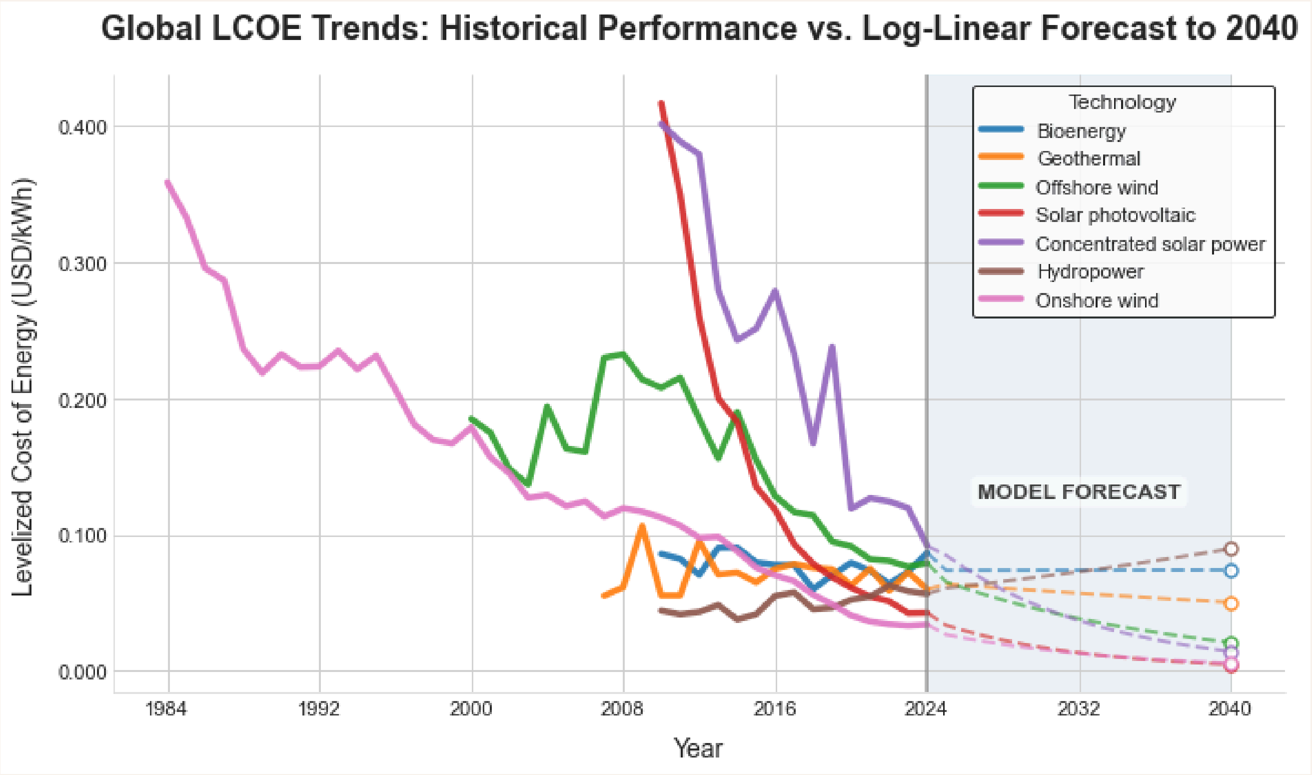
By 2040, utility-scale solar PV is predicted to be **substantially cheaper** than almost all forms of new fossil fuel generation, including new Combined Cycle Gas Turbine (CCGT) plants, whose LCOE is expected to **increase** due to rising fuel costs and carbon pricing.

The continued cost drop solidifies solar's role as the foundation of future electricity grids.

Would you like to see the comparison of the projected 2040 LCOE for solar against onshore or offshore wind?

Forecasting the Future of Energy: A Log-Linear Approach to LCOE

The challenge of accurately forecasting the **Levelized Cost of Electricity (LCOE)** for rapidly advancing renewable technologies, like solar and wind, requires moving beyond simple linear extrapolation. Our approach utilizes historical data from the Our World in Data (OWID) LCOE figures and applies a **log-linear regression** model to project prices for years up to 2040. This methodology is strategically chosen because it accurately captures the fundamental economic drivers governing technology costs.



The Superiority of the Log-Linear Model

Simple **linear regression** assumes a constant absolute dollar reduction in LCOE each year. As demonstrated, this flawed assumption inevitably leads to physically impossible negative costs when extrapolated, because the absolute dollar reduction must decelerate as the cost approaches zero.

Conversely, the **log-linear regression** (or learning curve model) assumes a constant **percentage reduction** for every unit increase in the independent variable (in our case, time, or more commonly, cumulative installed capacity). This translates the exponential, decelerating cost-decline curve into a straight line when plotted on a semi-log graph.

This model respects the fundamental economic principle known as the **"Learning Curve"** or **"Experience Curve"**: as manufacturing scale doubles, production costs drop by a constant percentage. By modeling the logarithm of LCOE against time, we achieve a forecast that:

1. **Is Always Positive:** The model asymptotically approaches, but never crosses, zero, thus maintaining physical realism.

2. **Reflects Market Maturity:** The model naturally slows the rate of cost decline as technologies mature, correctly anticipating that it is harder to shave one cent off a price of three cents than it was to shave ten cents off a price of forty cents.

Implementation and Constraints

Our implementation uses the sklearn.linear_model.LinearRegression trained on the **natural logarithm** of the LCOE values from a specific recent window (e.g., the last 10 years). This focuses the forecast on the most recent, relevant market dynamics, rather than diluting the trend with outdated data from the early years of deployment. The final predicted values are then obtained by taking the **exponent** of the log-linear results.

To maintain robust forecasting, two important constraints are imposed:

1. **Clamping:** We enforce a minimum predicted value (e.g., 10% of the last actual price) to prevent excessively aggressive and unrealistic drops, providing a necessary guardrail against model over-optimism.
2. **Data Filtering:** Only technologies with sufficient historical data are modeled, ensuring the regression is statistically meaningful.

In conclusion, our method of employing log-linear regression on OWID LCOE data offers a robust, realistic framework for forecasting the future cost of renewable energy. It correctly anticipates the exponential nature of technological advancement and provides credible price targets for years up to 2040.