Global CO2 Emissions Analysis

Data Source

This data is publicly disclosed open data (external source).
I downloaded it on GitHub: https://github.com/owid/co2-data

Collected by: Our World in Data

The CO2 and Greenhouse Gas Emissions dataset is updated regularly and is built upon a number of datasets and processing steps:

- Statistical review of world energy (BP)
- International energy data (EIA)
- Primary energy consumption (Our World in Data based on BP's Statistical review of world energy & EIA's International energy data)
- Global carbon budget - Fossil CO2 emissions (Global Carbon Project)
- Global carbon budget - Global carbon emissions (Global Carbon Project)
- Global carbon budget - National fossil carbon emissions (Global Carbon Project)
- Global carbon budget - National land-use change carbon emissions (Global Carbon Project)
- Global carbon budget (Our World in Data based on the Global Carbon Project's Fossil CO2 emissions, Global carbon emissions, National fossil carbon emissions, and National land-use change emissions)
- Greenhouse gas emissions (including methane and nitrous oxide) by sector (CAIT)
- CO2 dataset (Our World in Data based on all sources above)

Additionally, to construct variables per capita and per GDP, the following datasets and processing steps were used:

- Population (Our World in Data based on a number of different sources).
- GDP (University of Groningen GGDC's Maddison Project Database, Bolt and van Zanden, 2020).

I chose this dataset because I'm very interested in climate change, pollution and sustainability. These are relevant and important topics that effect our everyday life. Therefore, it is crucial to understand this kind of data to encourage data-driven decisions to plan ahead for the future of our planet.

Data Profile

Data Contents: This dataset includes data on CO2 emissions (annual, per capita, cumulative and consumption-based), other greenhouse gases, energy mix, and other relevant metrics for the years 1750 – 2021.

Data Shape (before cleaning):

The dataset contains 46523 rows and 74 columns.

Data Cleaning:

Wrangling Steps:

- Dropped 33 columns not needed for the analysis:
  'co2_including_luc', 'co2_including_luc_growth_abs', 'co2_including_luc_growth_prct', 'co2_including_luc_per_capita', 'co2_including_luc_per_gdp', 'co2_including_luc_per_unit_energy', 'co2_per_unit_energy', 'consumption_co2', 'consumption_co2_per_capita', 'consumption_co2_per_gdp', 'cumulative_co2_including_luc', 'cumulative_luc_co2', 'ghg_excluding_lucf_per_capita', 'ghg_per_capita', 'land_use_change_co2', 'land_use_change_co2_per_capita', 'methane', 'methane_per_capita', 'nitrous_oxide', 'nitrous_oxide_per_capita', 'share_global_co2_including_luc', 'share_global_cumulative_co2_including_luc','share_global_luc_co2', 'share_global_cumulative_luc_co2', 'total_ghg', 'total_ghg_excluding_lucf', 'trade_co2', 'trade_co2_share', 'cumulative_other_co2', 'other_co2_per_capita', 'other_industry_co2', 'share_global_cumulative_other_co2', 'share_global_other_co2'

- Rounded the “population” column to 0 decimal places and rounded all other columns containing floats to 2 decimal places.
- After dropping the columns not needed for the analysis, a subset only containing data for the years 2012-2021 was created to work with more recent data: containing 2600 rows and 41 columns

Consistency Checks:

- 1 mixed-type data column was detected:” iso_code” and converted to a string since it contains abbreviations for countries/regions
- 0 duplicates were found
- 893787 values are missing (47% of total)
- 13194 values are missing for the subset dataframe (years 2012-2021) making up 13% of the total values

For the time being I’ m not deleting any missing value/observation.

Data Shape (after cleaning):

The dataset contains 46523 rows and 41 columns.
The subset dataframe (years 2012-2021) contains 2600 rows and 41 columns.

Data Dictionary:

Column	Description	Time Variant/ Invariant	Type
country	Geographic location.	Invariant	Qualitative, nominal
year	Year of observation.	Variant	Qualitative, ordinal

iso_code	ISO 3166-1 alpha-3, three-letter country codes	Invariant	Qualitative, nominal
population	Population of each country or region.	Variant	Quantitative, discrete
gdp	Gross domestic product measured in international-$ using 2011 prices to adjust for price changes over time (inflation) and price differences between countries. Calculated by multiplying GDP per capita with population.	Variant	Quantitative, continuous
cement_co2	Annual production-based emissions of carbon dioxide (CO_2) from cement, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
cement_co2_per_capita	Annual production-based emissions of carbon dioxide (CO_2) from cement, measured in tonnes per person. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
co2	Annual total production-based emissions of carbon dioxide (CO_2), excluding land-use change, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
co2_growth_abs	Annual growth in total production-based emissions of carbon dioxide (CO_2), excluding land-use change, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
co2_growth_prct	Annual percentage growth in total production-based emissions of carbon dioxide	Variant	Quantitative, continuous

	(CO_2), excluding land-use change. This is based on territorial emissions, which do not account for emissions embedded in traded goods.		
co2_per_capita	Annual total production-based emissions of carbon dioxide (CO_2), excluding land-use change, measured in tonnes per person. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
co2_per_gdp	Annual total production-based emissions of carbon dioxide (CO_2), excluding land-use change, measured in kilograms per dollar of GDP (2011 international-$). Production-based emissions are based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
coal_co2	Annual production-based emissions of carbon dioxide (CO_2) from coal, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
coal_co2_per_capita	Annual production-based emissions of carbon dioxide (CO_2) from coal, measured in tonnes per person. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
cumulative_cement_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from cement since the first year of data availability, measured in million tonnes. This is based on territorial emissions, which do not account for	Variant	Quantitative, continuous

	emissions embedded in traded goods.		
cumulative_co2	Total cumulative production-based emissions of carbon dioxide (CO_2), excluding land-use change, since the first year of data availability, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
cumulative_coal_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from coal since the first year of data availability, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
cumulative_flaring_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from flaring since the first year of data availability, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
cumulative_gas_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from gas since the first year of data availability, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
cumulative_oil_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from oil since the first year of data availability, measured in million tonnes. This is based on territorial emissions, which do not account for		Quantitative, continuous

	emissions embedded in traded goods.		
energy_per_capita	Primary energy consumption per capita, measured in kilowatt-hours per person per year.	Variant	Quantitative, continuous
energy_per_gdp	Primary energy consumption per unit of gross domestic product, measured in kilowatt-hours per international-$.	Variant	Quantitative, continuous
flaring_co2	Annual production-based emissions of carbon dioxide (CO_2) from flaring, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
flaring_co2_per_capita	Annual production-based emissions of carbon dioxide (CO_2) from flaring, measured in tonnes per person. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
gas_co2	Annual production-based emissions of carbon dioxide (CO_2) from gas, measured in million tonnes. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
gas_co2_per_capita	Annual production-based emissions of carbon dioxide (CO_2) from gas, measured in tonnes per person. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
oil_co2	Annual production-based emissions of carbon dioxide (CO_2) from oil, measured in million tonnes. This is based on territorial emissions, which do not account for	Variant	Quantitative, continuous

	emissions embedded in traded goods.		
oil_co2_per_capita	Annual production-based emissions of carbon dioxide (CO_2) from oil, measured in tonnes per person. This is based on territorial emissions, which do not account for emissions embedded in traded goods.	Variant	Quantitative, continuous
primary_energy_consumption	Primary energy consumption, measured in terawatt-hours per year.	Variant	Quantitative, continuous
share_global_cement_co2	Annual production-based emissions of carbon dioxide (CO_2) from cement, measured as a percentage of global production-based emissions of CO_2 from cement in the same year. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from cement has been calculated by Our World in Data using global CO_2 emissions from cement provided in the Global Carbon Budget dataset.	Variant	Quantitative, continuous
share_global_co2	Annual total production-based emissions of carbon dioxide (CO_2), excluding land-use change, measured as a percentage of global production-based emissions of CO_2 in the same year. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions has been calculated by Our World in Data using global CO_2 emissions provided in the Global Carbon Budget dataset. Global emissions include all country	Variant	Quantitative, continuous

	emissions as well as emissions from international aviation and shipping.		
share_global_coal_co2	Annual production-based emissions of carbon dioxide (CO_2) from coal, measured as a percentage of global production-based emissions of CO_2 from coal in the same year. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from coal has been calculated by Our World in Data using global CO_2 emissions from coal provided in the Global Carbon Budget dataset.	Variant	Quantitative, continuous
share_global_cumulative_cement_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from cement since the first year of data availability, measured as a percentage of global cumulative production-based emissions of CO_2 from cement since the first year of data availability. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from cement has been calculated by Our World in Data using global CO_2 emissions from cement provided in the Global Carbon Budget dataset.	Variant	Quantitative, continuous
share_global_cumulative_co2	Total cumulative production-based emissions of carbon dioxide (CO_2), excluding land-use change, since the first year of data availability, measured as a percentage of global total	Variant	Quantitative, continuous

	cumulative production-based emissions of CO_2 since the first year of data availability. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions has been calculated by Our World in Data using global CO_2 emissions provided in the Global Carbon Budget dataset. Global emissions include all country emissions as well as emissions from international aviation and shipping.		
share_global_cumulative_coal_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from coal since the first year of data availability, measured as a percentage of global cumulative production-based emissions of CO_2 from coal since the first year of data availability. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from coal has been calculated by Our World in Data using global CO_2 emissions from coal provided in the Global Carbon Budget dataset.	Variant	Quantitative, continuous
share_global_cumulative_flaring_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from flaring since the first year of data availability, measured as a percentage of global cumulative production-based emissions of CO_2 from flaring since the first year of data availability. This	Variant	Quantitative, continuous

	is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from flaring has been calculated by Our World in Data using global CO_2 emissions from flaring provided in the Global Carbon Budget dataset.		
share_global_cumulative_gas_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from gas since the first year of data availability, measured as a percentage of global cumulative production-based emissions of CO_2 from gas since the first year of data availability. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from gas has been calculated by Our World in Data using global CO_2 emissions from gas provided in the Global Carbon Budget dataset.	Variant	Quantitative, continuous
share_global_cumulative_oil_co2	Cumulative production-based emissions of carbon dioxide (CO_2) from oil since the first year of data availability, measured as a percentage of global cumulative production-based emissions of CO_2 from oil since the first year of data availability. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from oil has been calculated by Our World in Data using	Variant	Quantitative, continuous

	global CO_2 emissions from oil provided in the Global Carbon Budget dataset. Global oil emissions include all country emissions as well as emissions from international aviation and shipping.		
share_global_flaring_co2	Annual production-based emissions of carbon dioxide (CO_2) from flaring, measured as a percentage of global production-based emissions of CO_2 from flaring in the same year. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from flaring has been calculated by Our World in Data using global CO_2 emissions from flaring provided in the Global Carbon Budget dataset.	Variant	Quantitative, continuous
share_global_gas_co2	Annual production-based emissions of carbon dioxide (CO_2) from gas, measured as a percentage of global production-based emissions of CO_2 from gas in the same year. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from gas has been calculated by Our World in Data using global CO_2 emissions from gas provided in the Global Carbon Budget dataset. Global gas emissions include all country emissions as well as emissions from international aviation and shipping.	Variant	Quantitative, continuous

share_global_oil_co2	Annual production-based emissions of carbon dioxide (CO_2) from oil, measured as a percentage of global production-based emissions of CO_2 from oil in the same year. This is based on territorial emissions, which do not account for emissions embedded in traded goods. Each country's share of global CO_2 emissions from oil has been calculated by Our World in Data using global CO_2 emissions from oil provided in the Global Carbon Budget dataset. Global oil emissions include all country emissions as well as emissions from international aviation and shipping.	Variant	Quantitative, continuous

Noteworthy Information about Values:

“country”: contains 269 unique values for the full dataframe.
There are 9 countries/regions less for the subset dataframe (2012-2021) due to missing values for the past 10 years or obsolete countries.
“year”: contains 272 unique values

There are also a lot of “0” values in the dataset. This doesn’t necessarily mean that all these values are missing. It could also be the case that some numbers didn’t reach the minimum threshold to be recorded (CO2 emissions are measured in millions of tons, there could be countries or source of energy that produced less than 0,1 million of tons CO2).

Missing Values Distribution:

Missing Values full Dataframe:

country	0
year	0
iso_code	6661
population	7949
gdp	31972
cement_co2	21549
cement_co2_per_capita	23809
co2	15174
co2_growth_abs	17579
co2_growth_prct	21491
co2_per_capita	19608
co2_per_gdp	30232
coal_co2	21454
coal_co2_per_capita	22102
cumulative_cement_co2	21650
cumulative_co2	17167
cumulative_coal_co2	21555
cumulative_flaring_co2	21749
cumulative_gas_co2	21681
cumulative_oil_co2	21582
energy_per_capita	36536
energy_per_gdp	39364
flaring_co2	21648
flaring_co2_per_capita	22251
gas_co2	21580
gas_co2_per_capita	22183
oil_co2	21481
oil_co2_per_capita	22156
primary_energy_consumption	36438
share_global_cement_co2	25693
share_global_co2	17167
share_global_coal_co2	21555
share_global_cumulative_cement_co2	25693
share_global_cumulative_co2	17167
share_global_cumulative_coal_co2	21555
share_global_cumulative_flaring_co2	30543
share_global_cumulative_gas_co2	24533
share_global_cumulative_oil_co2	23102
share_global_flaring_co2	30543
share_global_gas_co2	24533
share_global_oil_co2	23102

Missing Values Subset (2012-2021):

country	0
year	0
iso_code	0
population	198
gdp	1442
cement_co2	257
cement_co2_per_capita	257
co2	120
co2_growth_abs	240
co2_growth_prct	220
co2_per_capita	240
co2_per_gdp	1337
coal_co2	240
coal_co2_per_capita	240
cumulative_cement_co2	257
cumulative_co2	230
cumulative_coal_co2	240
cumulative_flaring_co2	240
cumulative_gas_co2	240
cumulative_oil_co2	230
energy_per_capita	616
energy_per_gdp	1442
flaring_co2	240
flaring_co2_per_capita	240
gas_co2	240
gas_co2_per_capita	240
oil_co2	230
oil_co2_per_capita	240
primary_energy_consumption	604
share_global_cement_co2	257
share_global_co2	230
share_global_coal_co2	240
share_global_cumulative_cement_co2	257
share_global_cumulative_co2	230
share_global_cumulative_coal_co2	240
share_global_cumulative_flaring_co2	240
share_global_cumulative_gas_co2	240
share_global_cumulative_oil_co2	230
share_global_flaring_co2	240
share_global_gas_co2	240
share_global_oil_co2	230

Descriptive Analysis:

	year	population	gdp	cement_co2	cement_co2_per_capita	co2	co2_growth_abs	co2_growth_prct	co2_per_capita	co2_
count	46523.000000	3.857400e+04	1.455100e+04	24974.000000	22714.000000	31349.000000	28944.000000	25032.000000	26915.000000	16291
mean	1925.686478	6.005374e+07	2.679977e+11	8.386319	0.063867	379.988026	5.713787	20.490187	3.668287	0
std	61.042693	3.280828e+08	2.104075e+12	63.013589	0.123098	1799.875838	58.766155	699.580069	14.947738	0
min	1750.000000	2.100000e+01	4.998000e+07	0.000000	0.000000	0.000000	-1818.470000	-100.000000	0.000000	0
25%	1882.000000	3.808912e+05	7.530493e+09	0.000000	0.000000	0.120000	0.000000	-0.540000	0.120000	0
50%	1930.000000	2.509282e+06	2.605900e+10	0.010000	0.000000	3.110000	0.020000	3.790000	0.890000	0
75%	1977.000000	9.996447e+06	1.134711e+11	0.690000	0.090000	43.660000	0.770000	10.580000	4.070000	0
max	2021.000000	7.909295e+09	1.136302e+14	1672.590000	2.570000	37123.850000	1859.760000	102318.510000	824.460000	37

Descriptive statistics look okay. It's noticeable that a lot of the variables have zero as a minimum value which could mean that values are missing or CO2 emissions maybe didn't reach the threshold to be recorded. Also, there is a huge gap between the minimum value for GDP and the maximum value, indicating huge differences between poor countries and rich countries.

Data Limitations and Ethics:

- Data is from 1750-2021
- Missing data makes up 52% of the grand total for the full dataframe and 13% of the subset dataframe (2012-2021) which is going to be quite challenging
- There is no PII data

Questions to explore:

- What were the top 5 or 10 countries that produced the most CO2 total and per capita over the years?
- What countries are the top 5 polluters in 2021?
- How does the ranking of the top 5 or 10 countries change in terms of annual CO2 emissions?
- What is the overall emissions trend of all the countries?
- Which factors/predictors contribute the most to the top three CO2 emitter countries?
- Which country has emitted the most CO2 in total since 2012?
- Which country has made the best progress in reducing CO2 emissions during the last decade?
- Which sectors contributed most heavily to emissions?
- Is there a correlation between CO2 emission and GDP or CO2 emissions and population size?