CO2 EMISSIONS ANALYSIS

DATA SOURCE

The dataset is publicly disclosed open data (external source).

I downloaded it on Kaggle (here).

Collected by: US Energy Administration (administrative data).

Collection method: All datasets were pulled from the US Energy Information Administration, and joined them all together into one dataset.

I choose this dataset because pollution, climate change and sustainability are topics in which I am very interested. Understanding this kind of data and taking data-driven decision is our only way to build a bright future for the entire world.

DATA PROFILE

DATA CONTENTS

Dataset contains information about CO2 emissions, energy production and consumption, GDP and number of population, divided by Countries and years.

DATA SHAPE (before cleaning)

The dataset contains 55440 rows and 11 columns.

DATA CLEANING

WRANGLING STEPS

- Dropped an unnecessary column (Unnamed: 0)
- Rounded the following columns (containing floats) to 2 decimal places:
  - Energy_consumption
  - Energy_production
  - GDP
  - Energy_intensity_per_capita
  - Energy_intensity_by_GDP
  - CO2_emission
- Rounded the following column to 0 decimal places:
  - Population

CONSISTENCY CHECKS

- 0 mixed-type data columns detected.
- 0 duplicates found.
- 61.134 values are missing (11% of total).
  For the time being, I am not deleting any missing value / observation.

DATA SHAPE (after cleaning)

The dataset contains 55440 rows and 10 columns.

DATA DICTIONARY

Column	Description	Time Variant / Invariant	Type
Country	Name of Country	Invariant	Qualitative, nominal
Energy_type	Type of energy source	Invariant	Qualitative, nominal
Year	Year in which data was recorded	Variant	Qualitative, ordinal
Energy_consumption	Amount of consumption for the specific energy source (measured in Quads)	Variant	Quantitative, continuos
Energy_production	Amount of production for the specific energy source (measured in Quads)	Variant	Quantitative, continuos
GDP	Countries GDP at purchasing power parities (measured in billions $)	Variant	Quantitative, continuos
Population	Population of Country (measured in thousands)	Variant	Quantitative, discrete
Energy_intensity_per_capita	calculated as units of energy per unit of capita (measured in millions BTU / person)	Variant	Quantitative, continuos
Energy_intensity_by_GDP	calculated as units of energy per unit of GDP measured in 1000 BTU / billions $ GDP PPP)	Variant	Quantitative, continuos
CO2_emission	amount of C02 emitted (measured in milliontonnes)	Variant	Quantitative, continuos

VALUES NOTEWORTHY INFORMATIONS

Country: contains 231 unique values, each of them recorded 240 times)

Energy_type: contains 6 unique values, each of them recorded 9240 times)

Year: contains 40 unique values (1980-2019), each of them recorded 1386 times)

Energy_consumption: contains 17722 "0" values.

Energy_production: contains 25497 "0" values.

Energy_intensity_per_capita: contains 5784 "0" values.

Energy_intensity_by_GDP: contains 11442 "0" values.

CO2_emission: contains 27359 "0" values.

There are a lot of “0” values. However, this doesn't mean they're actually missing values.

Probably numbers just didn't reach the minimum threshold to be recorded (e.g. CO2 emissions are measured in millions of tonnes, so there could be countries or source of energy that produced less than 0,1 million tonnes of CO2).

MISSING VALUES DISTRIBUTION

Country	0
Energy_type	0
Year	0
Energy_consumption	11153
Energy_production	11151
GDP	15414
Population	9426
Energy_intensity_per_capita	5082
Energy_intensity_by_GDP	5082
CO2_emission	3826

DESCRIPTIVE ANALYSIS:

	Year	Energy_consumption	Energy_production	GDP	Population	Energy_intensity_per_capita	Energy_intensity_by_GDP	CO2_emission
count	55440.0000	44287.000000	44289.000000	40026.000000	4.601400e+04	50358.000000	50358.000000	51614.000000
mean	1999.5000	1.537627	1.532586	827.144181	6.263020e+04	71.898907	3.695095	78.800072
std	11.5435	15.456603	15.303575	5981.703197	4.562088e+05	113.728744	4.590756	902.221446
min	1980.0000	-0.160000	0.000000	0.120000	1.100000e+01	0.000000	0.000000	-0.010000
25%	1989.7500	0.000000	0.000000	9.740000	1.142000e+03	3.800000	0.900000	0.000000
50%	1999.5000	0.020000	0.000000	47.760000	6.158000e+03	29.780000	2.990000	0.000000
75%	2009.2500	0.210000	0.110000	263.690000	2.004300e+04	95.520000	4.970000	4.320000
max	2019.0000	601.040000	611.510000	127690.250000	7.714631e+06	1139.320000	166.910000	35584.930000

The minimum value in “Energy_consumption” and “CO2_emission” is negative.

There is a huge gap between the minimum value for GDP and the maximum value, hinting at huge differences between poor countries and rich countries.

The same applies to energy and emissions: the data is very spread and there is a great difference in countries energy consumption and emissions.

75% of total values in “Energy_consumption”, “Energy_production” is less than 1.

This will be sooner investigated through charts and deeper analysis.

NEW VARIABLE:

Considering how much the GDP seems it will be important in future analysis, I decided to create a flag to divide GDP into four categories, using the quartiles range.

Those were the conditions:

```
# Creating first condition
df_co2.loc[df_co2['GDP'] <= 9.74, 'GDP_category'] = 'Low'
```

```
# Creating second condition
df_co2.loc[(df_co2['GDP'] > 9.74) & (df_co2['GDP'] <= 47.76), 'GDP_category'] = 'Medium'
```

```
# Creating third condition
df_co2.loc[(df_co2['GDP'] > 47.76) & (df_co2['GDP'] <= 263.69), 'GDP_category'] = 'Medium-high'
```

```
# Creating last condition
df_co2.loc[df_co2['GDP'] > 263.69, 'GDP_category'] = 'High'
```

If GDP is less or equal to 9.74 then: ‘Low’.

If GDP is major than 9.74 and less or equal to 47.76 then: ‘Medium’.

If GDP is major than 47.76 and less or equal to 263.69 then: ‘Medium-high’.

If GDP is major than 263.69 then: ‘High’.

The dataframe now has 11 columns.

DATA LIMITATIONS:

Dealing with all these missing values will be challenging, considering they are the 11% of the grand total.

Additionally, the time span of data is too wide. It even contains obsolete countries, such as “West Germany” and “East Germany” (reunited in 1989) or the “Former U.S.S.R.” or “Former Yugoslavia”.

For this reason, I decided to export two versions of the dataframe.

- One containing the cleaned version (after all the steps explained above).
- The second one contains all the data cleaned, but only the records from 2015 to 2019, to keep only more timely data.

ETHICAL CONSIDERATIONS:

Data does not contains sensitive informations or bias of every kind.

However, the collection method used by US Energy Administration is not clear and is not provided any explanation for the huge quantity of missing values.

KEY QUESTIONS

Those are the questions that I would like to explore.

More questions will be added as soon as I investigate deeper the data.

- Is there a correlation between energy consumption and CO2 emissions? If yes, what kind of relationship is?
- Is there a correlation between GDP and CO2 emissions? If yes, what kind of relationship is?
- Is there a correlation between GDP and energy consumption? If yes, what kind of relationship is?
- How spread is the data regarding GDP?
- Which are the countries with highest CO2 emissions?
- Which are the continents with highest CO2 emissions? (To answer this question a new variable should be derived).
- What's the trend in CO2 emissions in the last 5 years? Is the average increasing or decreasing over the years?