CO_2 Emissions Exploratory Data Analysis - Research Note

Tim Cross

September 12, 2025

This research note summarises the initial steps taken to understand the factors that affect global and national CO_2 emissions. This is part of a larger project that aims to analyse data from the World Bank's World Development Indicators which covers 1496 variables from diverse areas such as finance, health, education, energy and environmental factors. Ultimately our goal is to use machine learning to model this large dataset and look for possible complex relations between CO_2 emission and other variables. In this research note we perform an exploratory data analysis with a small subset of the data: CO_2 emissions, GDP (Gross Domestic Product) and population. Here we look at CO_2 emissions per county and its possible correlation with GDP and population over a period of 15 years. We also look at conditional correlations by grouping countries into two GDP per capita bands. All three variables show a high and positive correlation (0.8-1) for most countries. Nevertheless we find that when looking at countries with GDP per capita $>$ \$ 30,000 this trend is reversed. Most of these countries show a decrease in their CO_2 emission over the 15 year period.

Data

We source our data from World Bank's World Development Indicators database[1]. We chose data collected over 15 years from 2004 to 2019, since the CO_2 data is only available up to 2020, at the moment. However, due to the COVID pandemic and lock-down we observed that there was a drop in the CO_2 emission for this year, which would skew the analysis. So for this initial exploratory analysis we remove this year. The data covers all countries as recorded on the World Bank database.

Data Cleaning

We use the PANDAS PYTHON library for data cleaning and to prepare the data for further processing and analysis. First, we remove all countries which did not have any CO_2 data which removes 28 countries (see App A), and leaves 189 countries. We then changed the layout of the data using MELT and PIVOT_TABLE commands to have countries and years as rows and the main variables as columns. Next we changed all empty cells to 'Nan' to allow PANDAS to interact with the data without error, and the years were changed to integers to allow them to be manipulated correctly. Finally, we reordered the data by country name to allow easy manual access to the data.

Methodology

For this exploratory analysis we took the cleaned data into EXCEL, to quickly analyse and visualise the data[2]. The first step was to look at the correlations between the three variables: CO_2 Emissions, GDP and Population for each country.

We used the CORREL function within EXCEL which generates a number between -1, fully anti-correlated and 1, fully correlated. A value of zero corresponds to no correlation between the variables. The correlation is defined as follows,

$$Correl(X, Y) = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sqrt{\sum(x - \overline{x})^2}\sqrt{\sum(y - \overline{y})^2}}\ , \tag{1}$$

[1] https://databank.worldbank.org/source/world-development-indicators

[2] This can also be done within PYTHON, but we took this path to gain more experience with EXCEL

where X and Y are the two arrays of variables and the sum runs over all years for each country. For example, X can be the UK's CO_2 Emission in 2004-2019 and Y can be UK's GDP for the same years. In total, we measure 3 correlations for each country: CO2 and GDP, CO2 and Population, GDP and Population.

We also group countries based on their GDP per capita, to find conditional correlations. To do so we calculate the average GDP per capita over the 15 years per country. We then use this averaged value and divide the countries into a high and low GDP per capita bands.

To aid in the correlation analysis, we additionally calculate the change in each variable, X as $\Delta X = X_{2019} - X_{2004}$. Using ΔX in combination with the correlation value, we can better understand the relations between the different variables and find out if the positive/negative correlations are due to an increase or decrease in a variable.

Results

Figure 1 shows CO_2 , GDP and population changes over the study period, with all variables being scaled to 1 at the start of the study period for ease of comparison. It shows the variables change over the 15 year period, from 2004 to 2019. GDP generally rapidly increases over the period with two decreasing periods 2008-2009 and 2014-2015, almost doubling at the end of this period. CO_2 also increases but at a slower rate with a small decrease in 2008-2009 and a flattening off from 2013-2016, ending with an increase of 30%. Population increase over the period in a linear manner, by 2019 having increased 20%. WE find that the correlations between the three variables are all close to 1, CO_2 and GDP 0.99, CO_2 and Population 0.96 and GDP and Population 0.96.

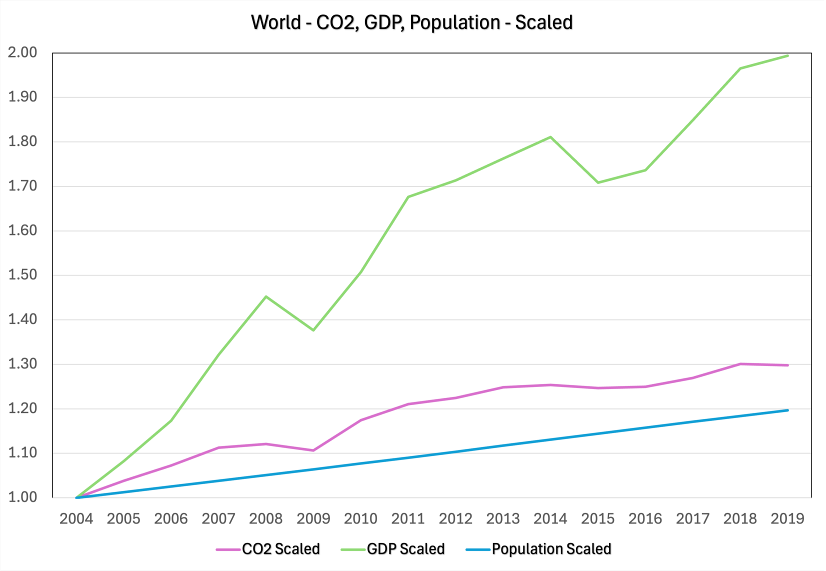

Figure 1: CO_2 emission (magenta), GDP (green) and population (blue) variation over the study period of 15 years between 2004 and 2019. All quantities have been scaled to 1 at year 2004. While GDP and CO_2 emission fluctuate they show an overall increasing trend. Population on the other hand grows linearly.

Figure 2 shows the distribution of correlations between pairs of variables for each country. Here the correlations are divided into 10 linearly spaced bins between their respective minimum and maximum values. CO_2 and GDP show a large concentration of countries with a strong correlation, with 82 countries having correlations above 0.8. The next largest groupings are in the two bins $0.42 - 0.61$ and $0.61 - 0.80$ which contain 45 countries in total. The further 7 bins are fairly evenly spread with countries, ranging from 6 to 11 per bin. CO_2 and population also show a large grouping of countries with a strong correlation, 99 countries with correlations larger than 0.8. This is followed by 18 countries in $0.6 - 0.8$ bin. There is a further peak of 21 countries with a strong anti-correlation of between -0.78 and -0.98. The further 7 bins contain from 3 to 10 countries. GDP and Population shows a large

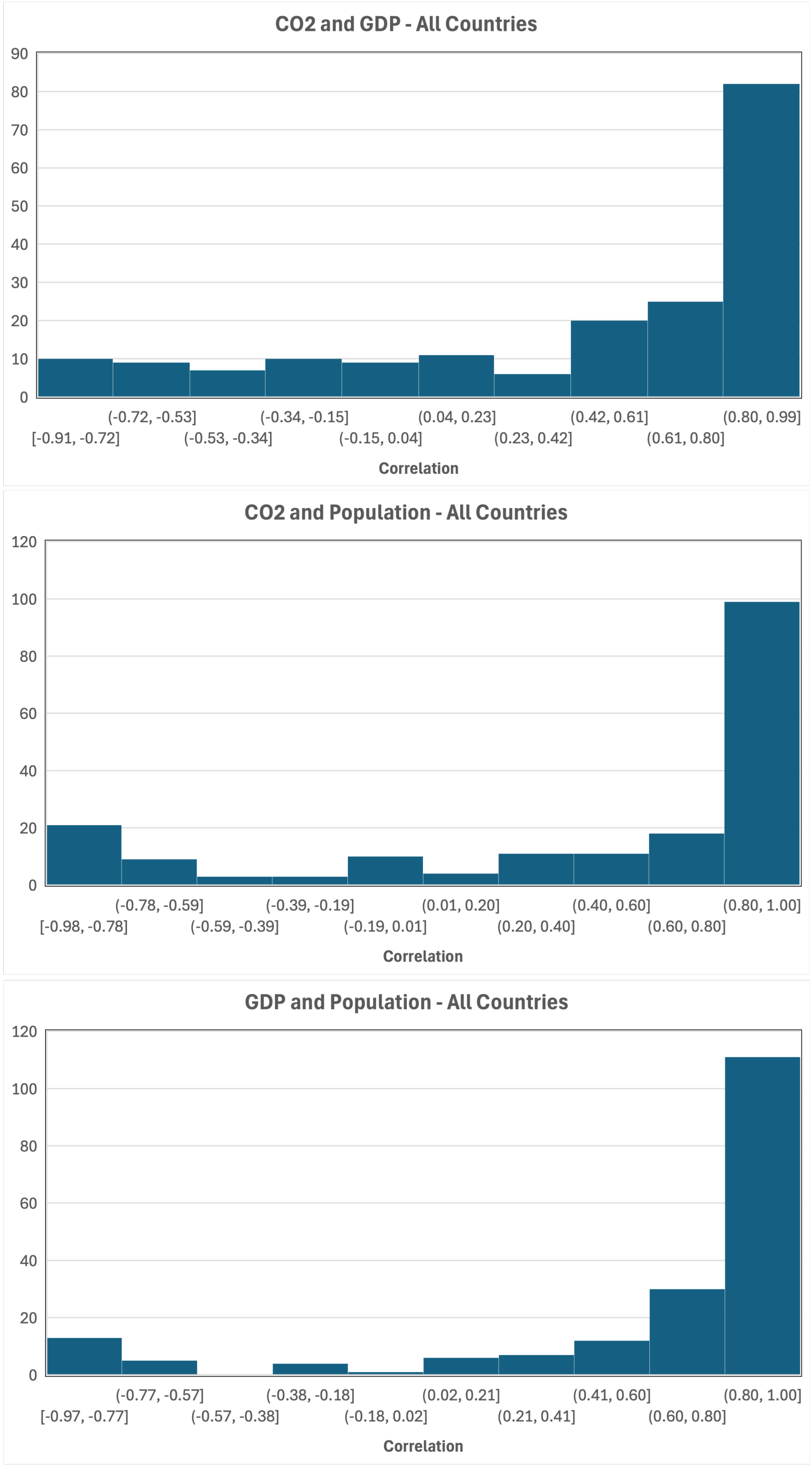

Figure 2: Correlations between CO_2 Emissions, GDP and population for all countries over the study period, 2004 - 2019. The results are sorted into 10 bins, starting from the minimum and ending at the maximum correlation.

number of countries in the highest bin, 111 countries between a correlation of 0.80 and 1.00. The second highest bin contains the second highest grouping of countries with 30 between a correlation of 0.60 and 0.80. There is a further small peak of countries with a high anti-correlation, 13 between -0.77 and -0.97.

Although most countries show highly positive correlations between CO_2 and GDP, there is a significant number of countries with negative correlations. To see if this correlation depends on GDP, we divide the data into a high and a low GDP per capita band and show their respective distributions in Figure 3. The high band includes 28 countries with GDP per capita $> \$30,000$ while the lower band includes the rest of the countries (161). Note that we use GDP per capita instead of GDP (Total) to avoid skewing the results towards more populous countries. In other words, we wanted to avoid dominating the high bin by more populous countries. As can be seen in the figure the high GDP band has a preference for negative correlations in contrast to the low GDP band. Of the 28 countries in the high band 19 have a negative ΔCO_2 , showing a decrease in their CO_2 emission, while their GDP per capita increases.

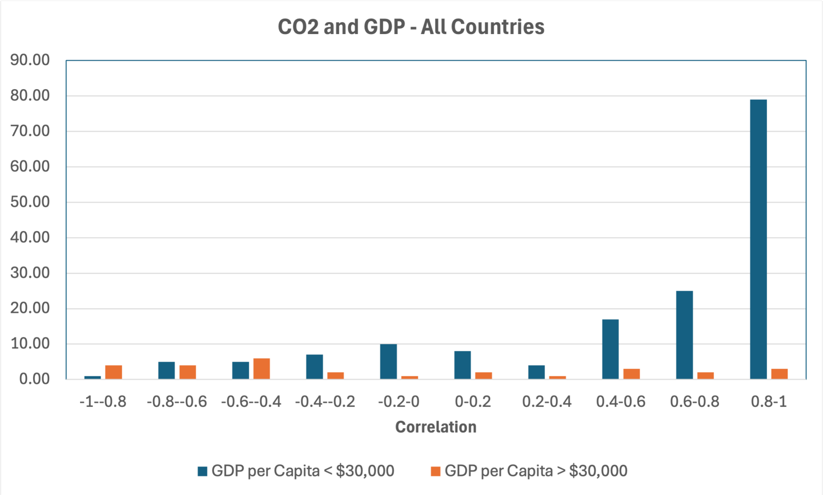

Figure 3: Distribution of correlations between CO_2 and GDP for high ($> \$30k$, orange) and low ($< \$30k$, blue) GDP per capita bands.

Conclusion

In this research note we performed an exploratory data analysis for a subset of our CO_2 data. We limited the analysis to exploring correlations between CO_2 , GDP and population over a 15 year period between 2004 and 2019. We cleaned and prepared our data using Python's Pandas library and then exported it into Excel for further analysis. During the data cleaning we removed 28 countries with missing CO_2 emission data. As these are relatively small countries, we do not expect them to have a large impact on our analysis.

Overall, we find all pairs of variable to be positively correlated for the majority of countries, showing that as population and GDP grow so does the CO_2 emission. Focusing on the correlation between GDP and CO_2 we see that out of the 189 countries studies here 76% show positive correlations with 43% having correlations > 0.8. When measuring this correlation for countries with GDP per capita $> \$30,000$ we find that the majority show negative correlations with only 39% being positively correlated. This might be because as a country's wealth grows it moves to using cleaner energy which require investment into its infrastructure. However, other factors may also be at play here. For example, due to the higher cost of labour in richer countries most products are likely to be imported in which case the CO_2 emission cost of their production is not attributed to these countries. To disentangle these factors in future studies we will include data on energy sources, manufacturing, import and exports.

We also see interesting trends when considering the global emission and GDP, which shows that the variations in global GDP is approximately mirrored in the CO_2 emission. We make a note of this to follow it up in future work.

This analysis sets the tone for our future research where we will explore a wider range of variables to study other possible factors that contribute to a country's CO_2 emission. In particular we note that a correlation between two variables does not necessarily imply a causal connection between them. Our ultimate goal is to learn from countries that have successfully reduced their CO_2 emission or have kept their emissions low. Therefore, in addition to looking at overall trends for all countries, in future studies we will perform case studies of selected countries.

A Removed countries

Removed Country
American Samoa
Aruba
Bermuda
British Virgin Islands
Cayman Islands
Channel Islands
Curacao
Faroe Islands
French Polynesia
Gibraltar
Greenland
Guam
Hong Kong SAR, China
Isle of Man
Korea, Peoples Republic
Kosovo
Macao SAR, China
Monaco
New Caledonia
Northern Mariana Islands
Puerto Rico
San Marino
Saint Maarten (Dutch part)
South Sudan
St. Martin (French part)
Turks and Caicos Islands
Virgin Islands (U.S.)
West Bank and Gaza

Table 1: List of removed countries