CO_2 Emissions and Trends;
Continent and Country Level Case Study

Mid-Term Report

STAT 650

Submitted by:

Group 9

826005266, Alexander Peter

326000139, Henry Ruiz

523008789, Sabahat Zahra

433003589, Sai Manisha Duvvada

532008512, Sandeena Shrestha

10/09/2022

Contents

List of Figures

List of Tables

List of Abbreviations

GDP	:	Gross Domestic Product
GHG	:	Green House Gases
VIF	:	Variance Inflation Factor

1. Introduction:

Planet earth is experiencing tremendous global climate change effects in forest fires, droughts, floods, and extreme weather conditions. These impacts are mainly associated with increased carbon dioxide, methane, and other greenhouse gasses in our atmosphere caused by human activities.

According to the United States, Environmental Protection Agency, the largest source of greenhouse gas (GHG) emission in the United States is energy usage, in terms of burning fossil fuels for electricity, heat, and transportation. The rise in energy-related CO_2 emissions has pushed greenhouse gas emissions from energy to their highest level in 2018.

This case study presents global trends in net CO_2 emission across different continents over time.

2. Objectives and Questions

Under the assumption that CO_2 indicators in the atmosphere result from energy-related factors, this study aims to analyze periodic changes in CO_2 emissions by different countries.

Moreover, our objective is to examine the statistical relationship between potential factors responsible for GHG emissions, such as energy consumption and per capita CO_2 production by each country on all the continents.

In our project, we studied the following key points

I. Examination of significant increases and decreases in CO_2 emission by countries from 2000 to 2018,
II. The relationship between population and CO_2 emission,
III. The relationship between CO_2 and Gross Domestic Product (GDP)
IV. Create a dashboard to visualize the results obtained in the study

3. Significance of the Study

Our study expects insight into general trends in energy consumption and associated CO_2 emissions by countries and continents.

These outcomes will be further helpful in understanding the relationship between CO_2 emissions driving factors across countries and continents to present a future framework for controlling environmental pollution in terms of GHG gas and, ultimately, its impact on climate change.

4. Methodology

Dataset

In this study, we used the CO2 and Greenhouse Gas Emissions dataset from GitHub: https://github.com/owid/co2-data, a collection of key metrics maintained by Our World in Data. It is updated regularly and includes data on CO2 emissions (annual, per capita, cumulative, and consumption-based), other greenhouse gases, energy mix, and other relevant metrics.

Pre-processing

The original CO2 and Greenhouse Gas emission dataset has in total of 26000 observations and 60 columns. Approximately 50% of the columns had missing values. In order to reduce the dimensions of the data and the scope of our analysis, we implemented a function that removed the columns in the dataset where the proportion of missing values was greater than 70%. See figure 1. Three data imputation techniques were explored for the remaining columns containing missing values. Column and row mean, and removing the rows with a high proportion of missing values. After analyzing how each method affected the distribution of the values, we decided to drop the rows with a high ratio of missing values, and we moved forward with the analysis.
To have two extra grouping levels, we added additional columns to the dataset, Continent and GPS location (latitude and longitude).

Data visualization and analysis

To address the questions in the objectives section and help us identify patterns and relationships in the data, we explored different visualization and filtering techniques. A dashboard (https://stat650-dashboard-sqdrjafctq-uc.a.run.app/) was created to incorporate all the generated plots.

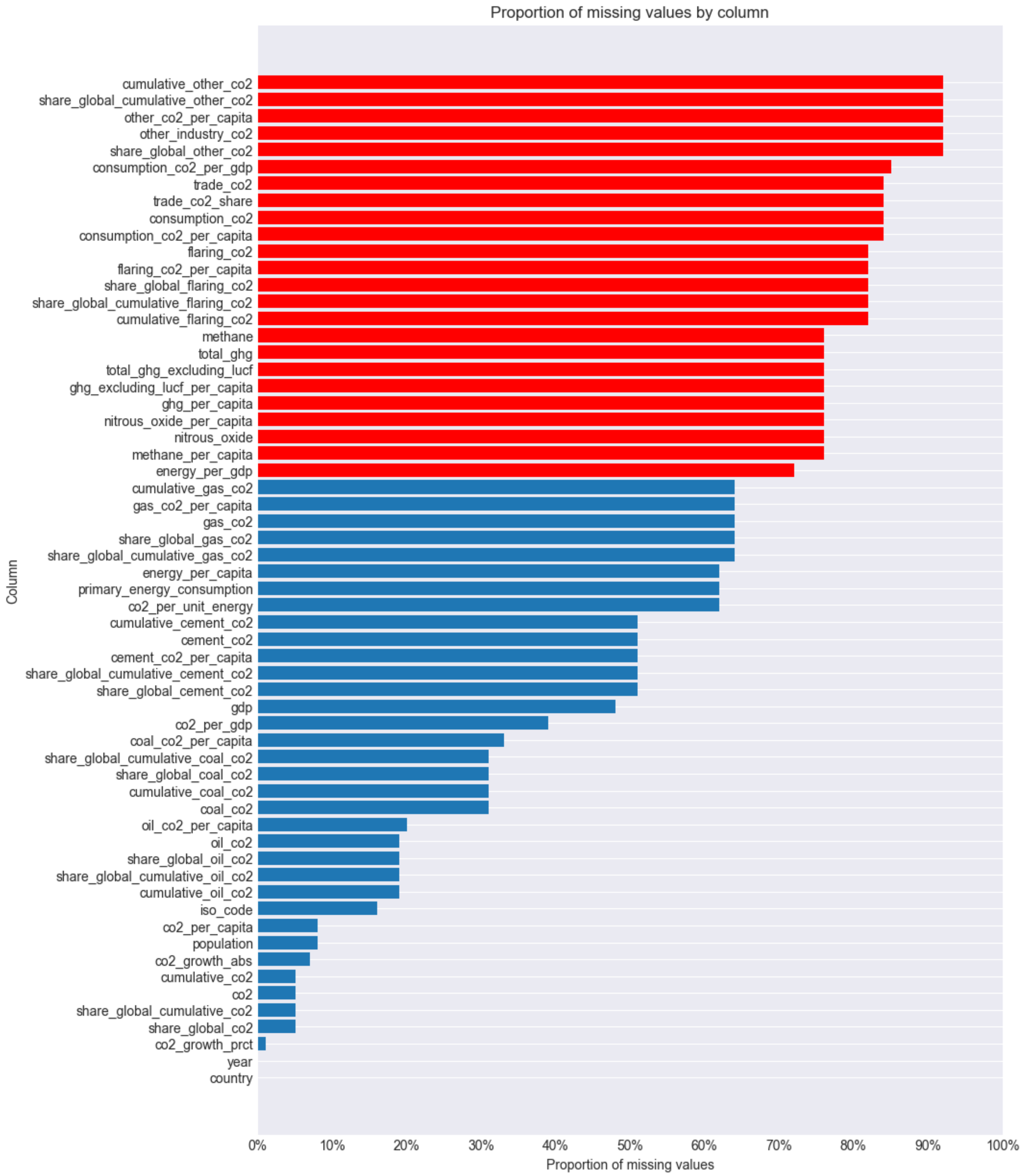

Figure 1: Proportion of missing values by column

Step 1: Data exploration

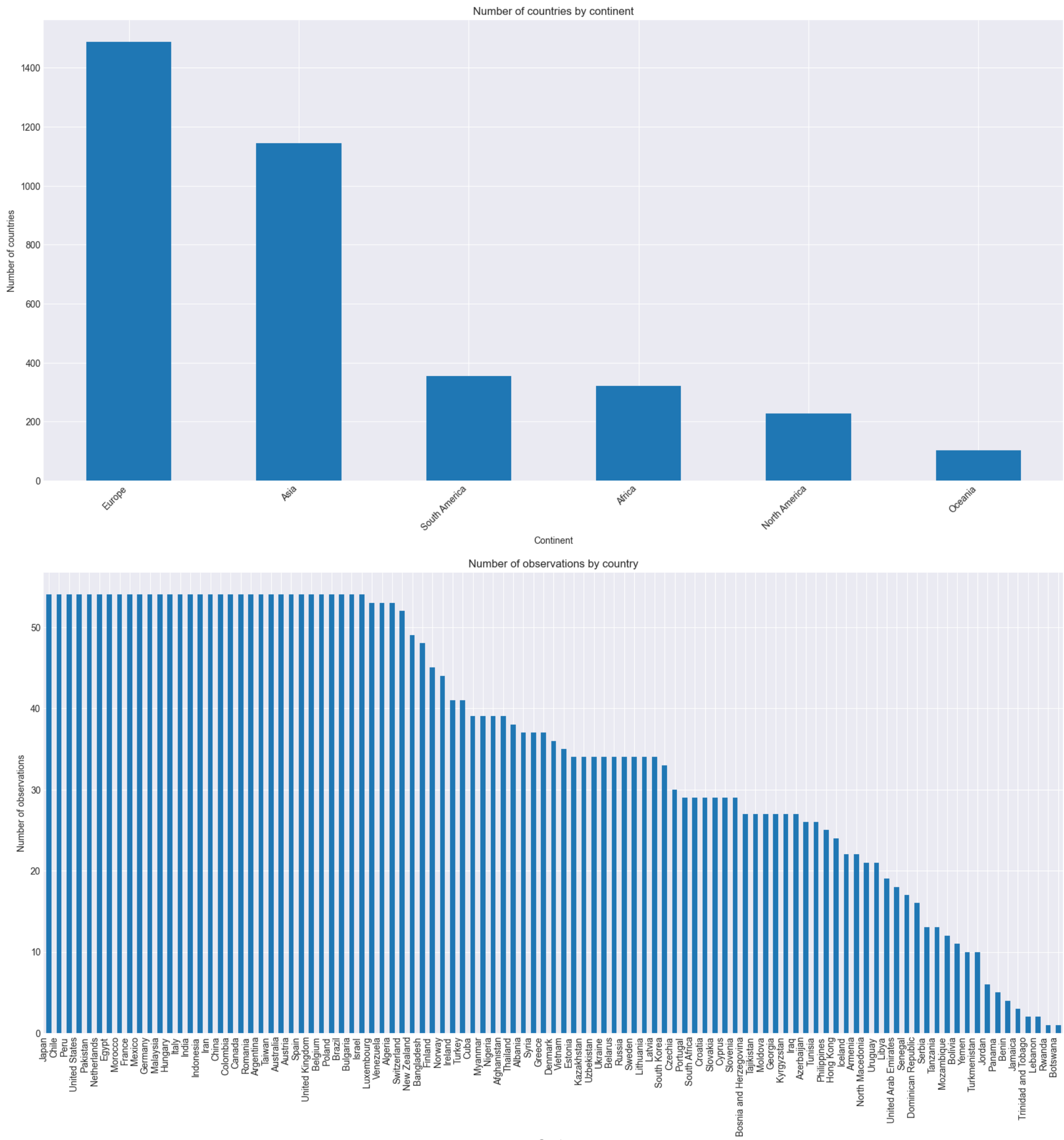

Figure 2: Number of countries by continent and number of observations by country

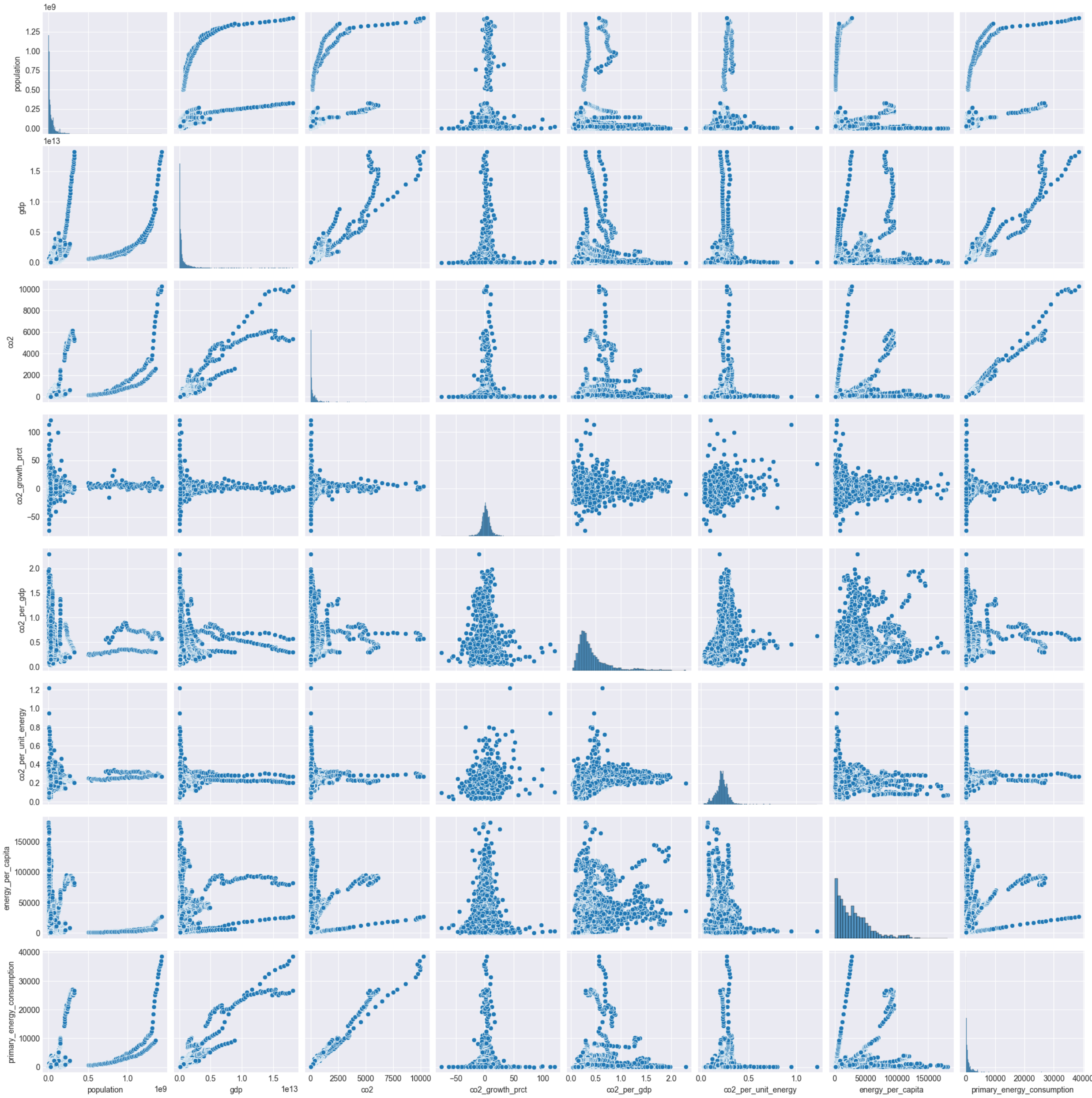

Figure 3: Selected variables Distribution

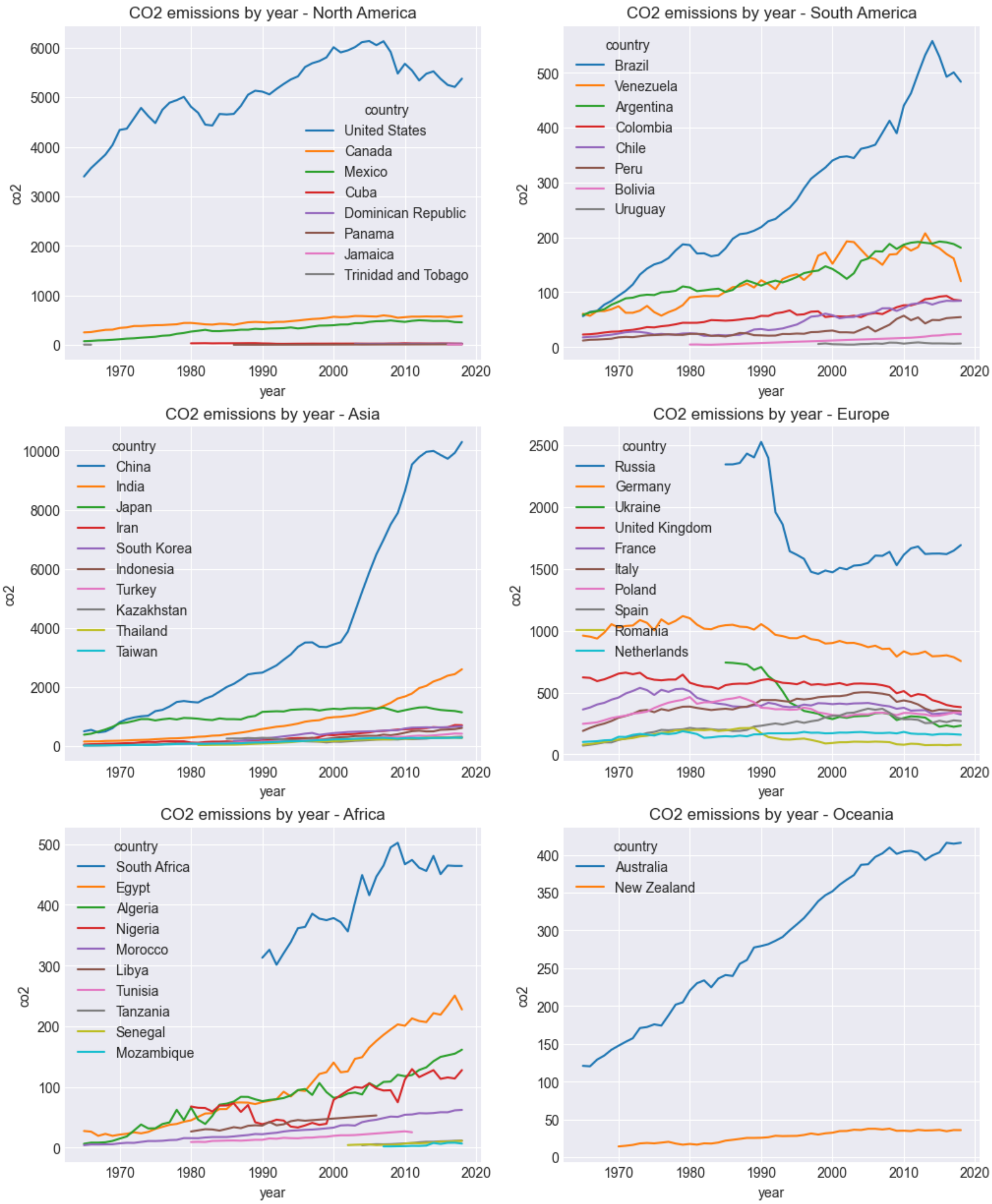

Figure 4: Top 10 countries by continent based on CO2 emissions

Country-level observations

country	year	iso_code	population	gdp	cement_co2	cement_co2_per_capita	co2	co2_growth_abs	co2_growth_prct	..
United States	1965	USA	199733664.0	4.156141e+12	33.090	0.166	3399.342	135.231	4.14	..
United States	1966	USA	201895760.0	4.428300e+12	34.360	0.170	3571.208	171.867	5.06	..
United States	1967	USA	203905072.0	4.538979e+12	33.387	0.164	3705.254	134.046	3.75	..
United States	1968	USA	205805744.0	4.754926e+12	35.102	0.171	3840.702	135.448	3.66	..
United States	1969	USA	207659280.0	4.903770e+12	35.477	0.171	4034.926	194.223	5.06	..
United States	1970	USA	209513344.0	4.912636e+12	34.709	0.166	4339.471	304.545	7.55	..
United States	1971	USA	211384080.0	5.065682e+12	35.291	0.167	4365.247	25.776	0.59	..
United States	1972	USA	213269808.0	5.334297e+12	36.299	0.170	4572.791	207.544	4.75	..
United States	1973	USA	215178800.0	5.637203e+12	36.690	0.170	4784.823	212.032	4.64	..
United States	1974	USA	217114896.0	5.621366e+12	36.580	0.169	4620.820	-164.003	-3.43	..
United States	1975	USA	219081248.0	5.605795e+12	30.276	0.138	4478.039	-142.782	-3.09	..
United States	1976	USA	221086416.0	5.899591e+12	32.171	0.146	4747.563	269.525	6.02	..
United States	1977	USA	223135664.0	6.166912e+12	33.770	0.151	4889.398	141.835	2.99	..
United States	1978	USA	225223312.0	6.518624e+12	35.397	0.157	4941.143	51.744	1.06	..
United States	1979	USA	227339328.0	6.740172e+12	35.719	0.157	5008.358	67.216	1.36	..
United States	1980	USA	229476352.0	6.743208e+12	32.719	0.143	4808.296	-200.063	-3.99	..
United States	1981	USA	231636064.0	6.911865e+12	31.765	0.137	4686.171	-122.124	-2.54	..
United States	1982	USA	233821840.0	6.782207e+12	28.266	0.121	4447.080	-239.092	-5.10	..

Table 1: Observations by country (USA)

Step-2 Data mining

Assessing multicollinearity of parameters

To accomplish this, variance inflation factors (VIFs) were considered, which are only defined for numeric variables. So, the categorical columns were dropped, and a new data frame was created with only numeric data. In the next step, a correlation map of the variables with exceptionally high VIFs was created to assess which needed to be removed in the final analysis.

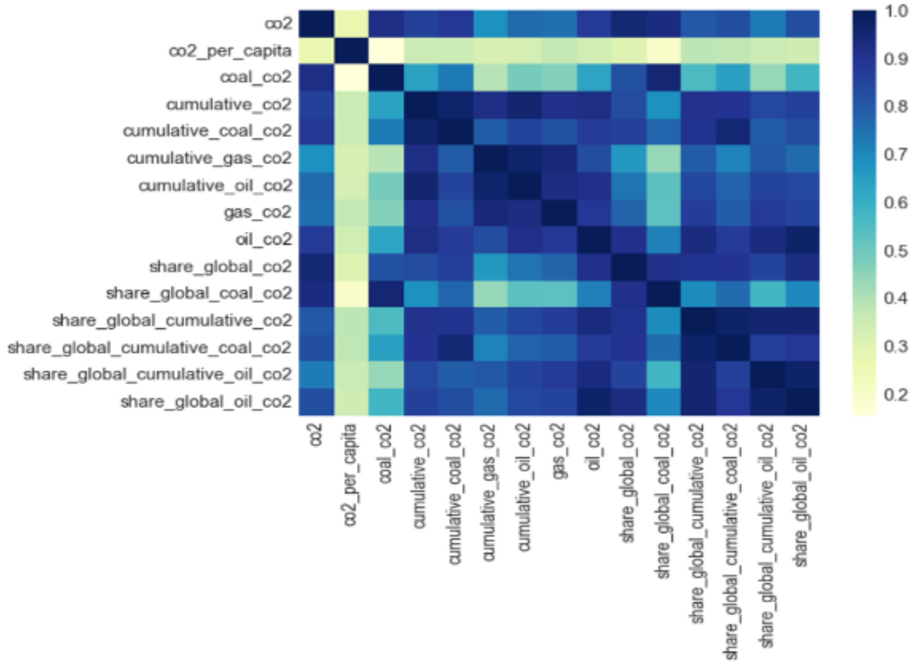

Figure 5: Correlation map for variables with high VIF

According to our findings, CO_2 emissions correlate with cumulative CO_2 and CO_2 emitted from different sources. Similarly, it was expected that the share of overall global CO_2 would be highly associated with cumulative global share, both general and from various sources. Since all these variables could be considered significant responses, we decided to keep overall CO_2 emissions and the percentage of global CO_2 emissions and drop the rest.

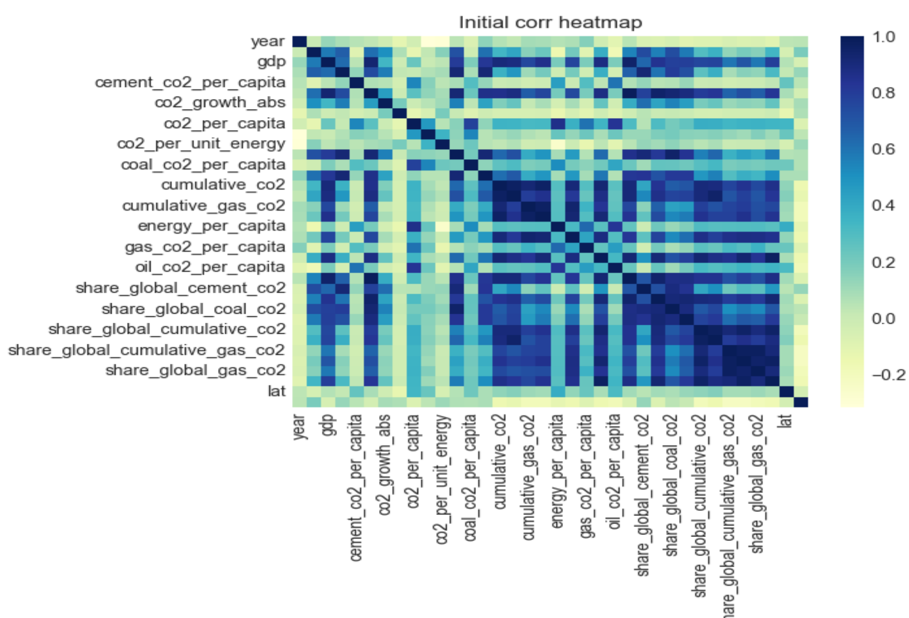

Figure 6: Correlation map for variables with low VIF

So, after this filtration, we have settled on this subset of features to analyze.

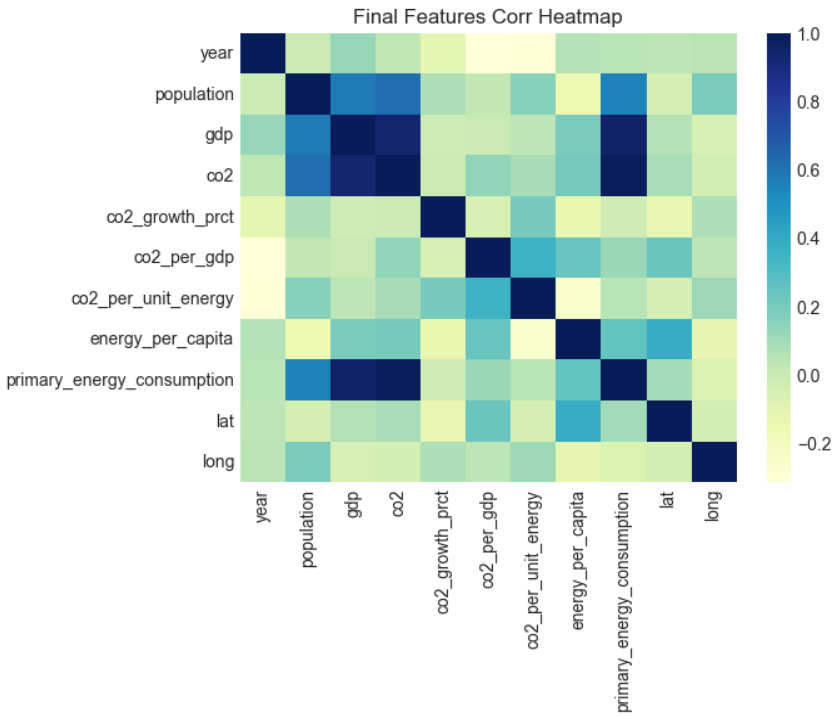

Figure 7: Correlation map for final variables

Step-3 Data visualization and analysis

<u>**Question 1:** Which countries from each continent have seen the greatest increases and decreases in CO_2 efficiency from 2000 to 2018?</u>

Change in CO_2 efficiency was measured by looking at the percent change in co2_per_unit_energy from 2000 to 2018. First, we consider the best-performing countries from each continent, or the countries with the largest drop in CO_2 per unit energy:

Table 2: Countries with the largest increase in CO_2 efficiency (most significant % drop in CO_2 per unit energy)

	continent	country	co2_per_unit_energy
10	Asia	Hong Kong	-0.421569
52	Europe	North Macedonia	-0.381074
3	Africa	Nigeria	-0.247863
75	South America	Peru	-0.184466
67	North America	Mexico	-0.139918
69	Oceania	Australia	-0.025926

Here is the trend in CO_2 per unit energy for each of these well-performing countries:

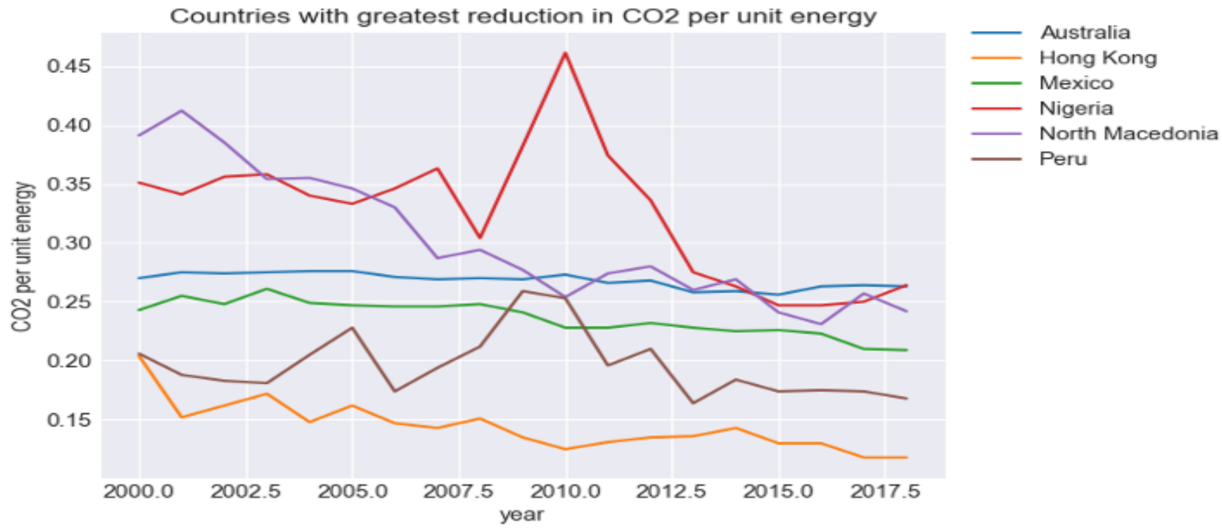

Figure 8: Countries with largest decline in CO2 per unit energy from 2000 to 2018

Since these countries had the largest drop in CO_2 per unit energy consumption, they had the largest increases in CO_2 efficiency by our definition.

Next, we can consider the countries from each continent that performed the worst in terms of CO_2 efficiency.

Here are the top countries that saw little decline or an increase in CO_2 per unit from 2000 to 2018.

	continent	country	co2_per_unit_energy
25	Asia	Tajikistan	2.368421
34	Europe	Bosnia and Herzegovina	0.419811
4	Africa	South Africa	0.028213
70	Oceania	New Zealand	-0.013889
73	South America	Chile	-0.021164
65	North America	Canada	-0.088050

Table 3: Countries with the least increase in CO_2 efficiency (least significant % reduction in CO_2 per unit energy)

Similarly, we can visualize the trend of these country's CO_2 per unit energy consumption as follows:

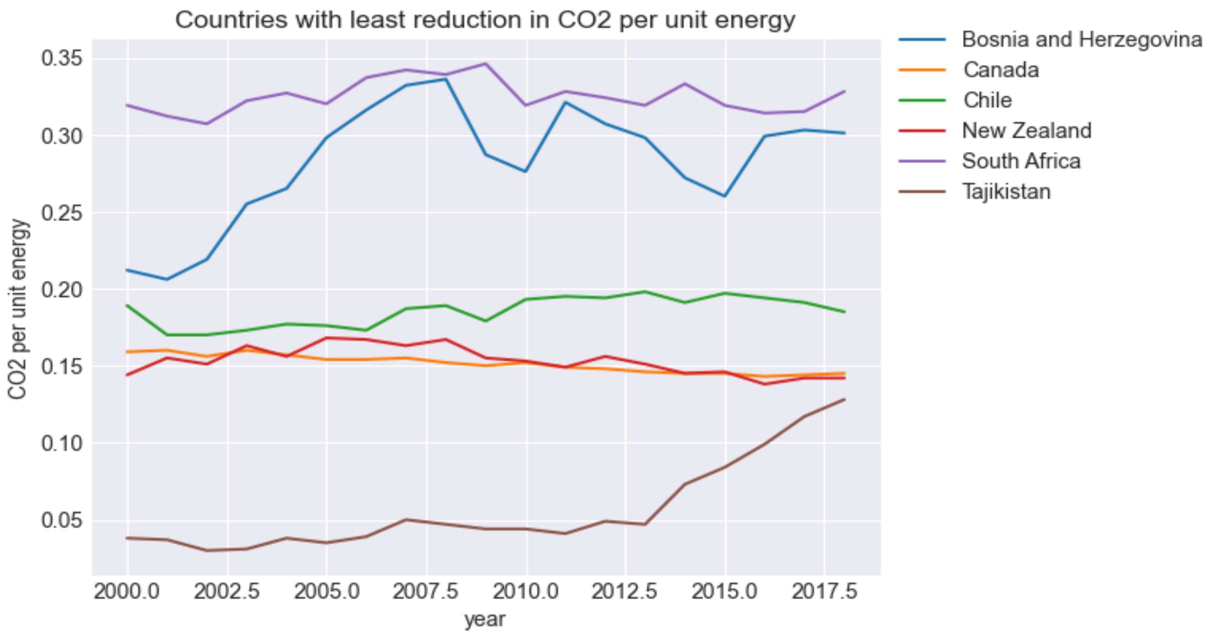

Figure 9: Countries with least decline or increase in CO_2 per unit energy from 2000 to 2018

From these results, we can see that although not every country reduced their CO_2 consumption per unit energy, even the worst performing countries did not increase their usage by large margins. So, it seems that overall, countries are aiming to be more CO_2 efficient and output less CO_2 per unit energy.

<u>Question 2:</u> <u>What is the relationship between population and CO_2 emission?</u>

First, we will look at the relationship between population and CO_2 emissions for all countries that had measurements in 2018.

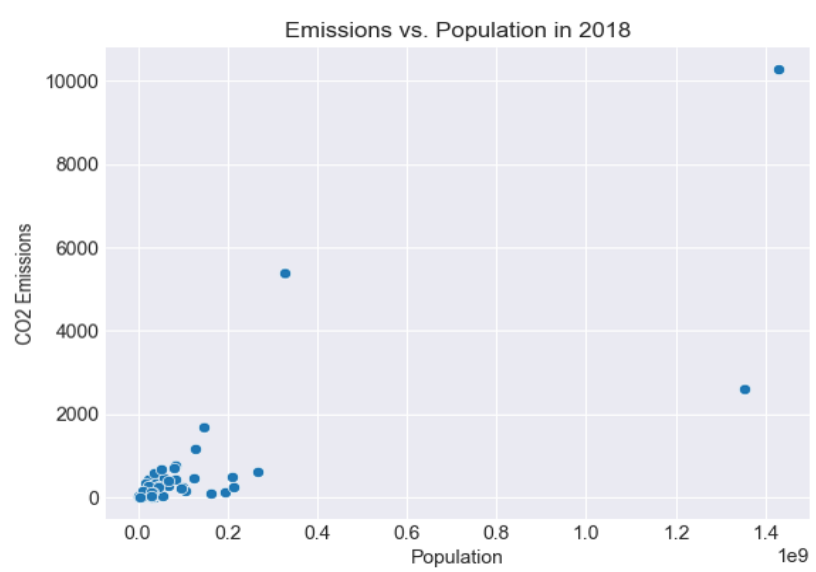

Figure 10: Relationship between CO2 emission and population size

Since most of the data in Figure 9 is concentrated near the minimum and variance in CO_2 emissions looks to increase with population, we try a log-log transformation.

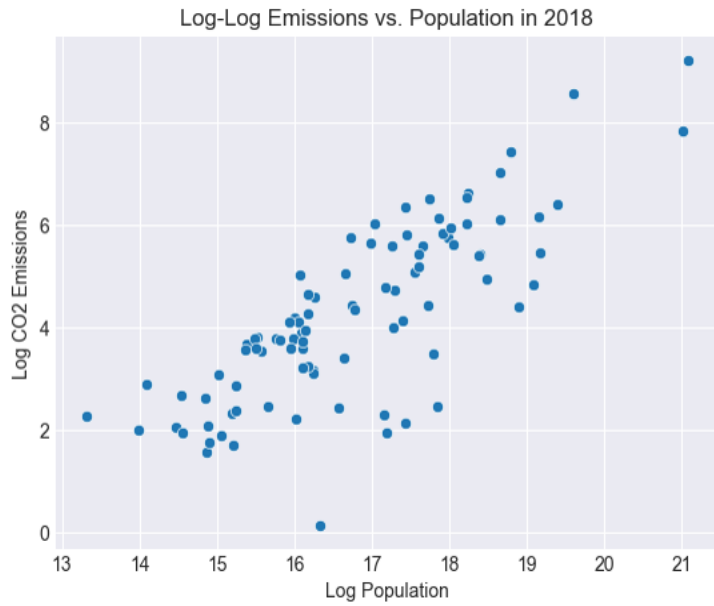

Figure 11: Relationship between CO2 emissions and population with log-log transformation

Now that we have a linear, equal-variance relationship (Fig 10) to learn, we can do linear regression and use the learned parameters to answer questions about the relationship between population and CO_2 emissions.

When we train a linear regression model, we get a slope coefficient of 0.89. Because we used a log-log regression model, the interpretation of the model parameters is different than with an ordinary linear regression model. In this case, we see that if the population increases by 8.9%, then CO_2 emissions increase by 10%. Lastly, we can plot the learned linear model and the slope's 95% confidence interval.

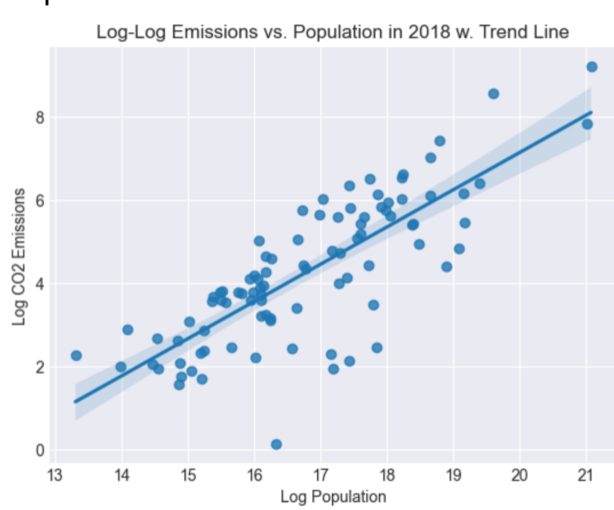

Figure 12: Regression model for log CO_2 and log population

<u>Question 3:</u> <u>What is the relationship between GDP and CO_2 emissions?</u>

The plot (as shown in Figure 12) showed that there are some outliers in the data, so we used the log-log transformation of the GDP and the CO_2 emission and a plot as shown in Figure 13 was observed. A linear regression plot was used to define the relationship between the two values (Figure 14).

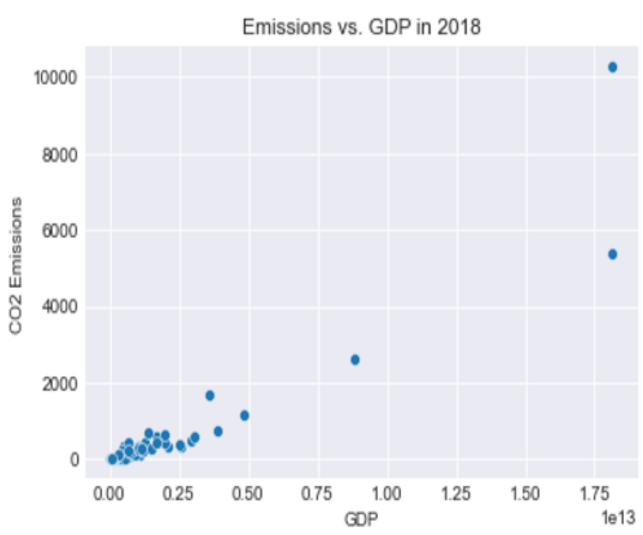

Figure 13: Relationship between CO2 Emission and GDP

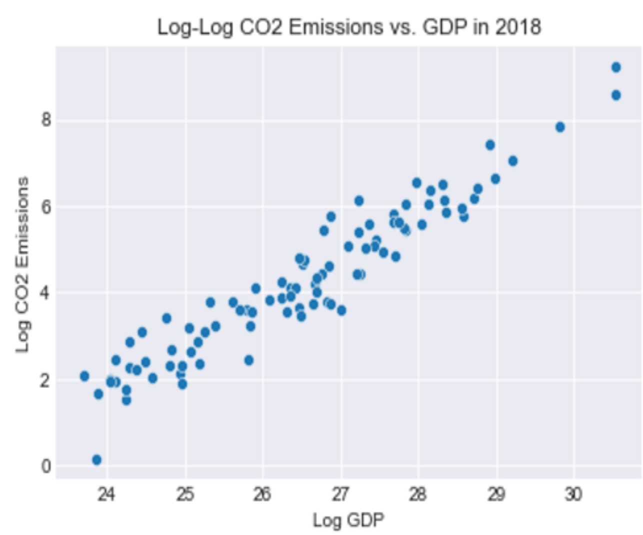

Figure 14: Relationship between CO2 Emission and GDP after log-log transformation

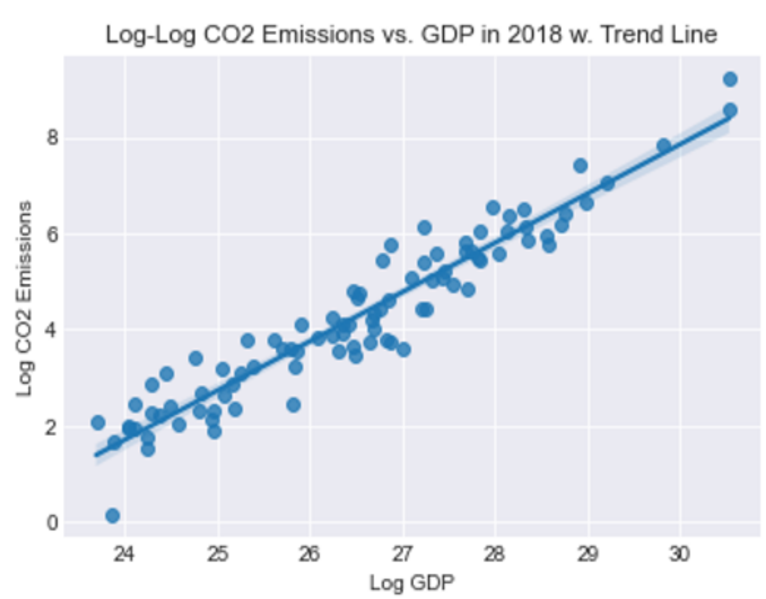

Figure 15: Regression model for log CO_2 and Log GDP

Training a linear regression model gives a slope of 1.02, from which we can determine that a 10.2% increase in GDP will increase the CO_2 emission by 10%.

5. Discussion and Conclusion

Several steps were followed in our analysis, starting from data mining techniques in the preprocessing stage to filtering the provided data. We examined the number of countries, the number of observations, all the parameters available, times series, and missing values and imputation.

Afterward, we analyzed and visualized the data we had, compared different attributes, and attempted to plot the data in the manner we believed was appropriate.

We computed each feature's variance inflation (VI) factors and removed or filtered the elements with high VI factors relative to this dataset.

VIFs measure the degree of multicollinearity between different features, which helps remove all the redundant features. A few more features were removed that did not seem relevant to the questions we wanted to ask. And finally, we were left with a small subset of the original features which we used to analyze the overall trends, rise and fall for various factors like GDP, population, and per unit energy consumption.

After data cleaning and processing, results were analyzed to study our main objectives of the proposed study. According to our results from the GDP versus CO_2 graphs, the advancement in the economy by 10.2% on a per capita basis also increases CO_2 emission by 10%. Similarly, population size was also found to be directly correlated with CO_2 emission. This is because increased demand for energy in more populated countries resulted in more contribution towards global GHG emissions.

In conclusion, our analysis successfully derived a general relationship between GHG indicators across regions.

6. Annex

Link for the original dataset.
https://github.com/owid/co2-data

Link to all the code produced and including the dashboard and the analysis notebook.
https://github.com/haruiz/STAT650-midterm

Link to the dashboard has been provided which visualizes the plots.
https://stat650-dashboard-sqdrjafctq-uc.a.run.app/