**CO2 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 iii](#_Toc116211141)

[List of Tables iii](#_Toc116211142)

[List of Abbreviations iv](#_Toc116211143)

[1. Introduction: 1](#_Toc116211144)

[2. Objectives and Questions 1](#_Toc116211145)

[3.Significance of the Study 1](#_Toc116211146)

[4. Methodology 2](#_Toc116211147)

[Step 1: Data filtration 2](#_Toc116211148)

[Step-2 Data mining 3](#_Toc116211149)

[Step-3 Data visualization and analysis 5](#_Toc116211150)

[5. Discussion and Conclusion 10](#_Toc116211151)

[6. Annex 10](#_Toc116211152)

# List of Figures

[Figure 1: Proportion of missing values by column 3](#_Toc116223210)

[Figure 2: Number of countries by continent and number of observations by country 4](#_Toc116223211)

[Figure 3: Selected variables Distribution 5](#_Toc116223212)

[Figure 4: Top 10 countries by continent based on CO2 emissions 6](#_Toc116223213)

[Figure 5: Correlation map for variables with high VIF 8](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223214)

[Figure 6: Correlation map for variables with low VIF 8](#_Toc116223215)

[Figure 7: Correlation map for final variables 9](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223216)

[Figure 8: Countries with largest decline in CO2 per unit energy from 2000 to 2018 10](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223217)

[Figure 9: Countries with least decline or increase in CO2 per unit energy from 2000 to 2018 11](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223218)

[Figure 10: Relationship between CO2 emission and population size 11](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223219)

[Figure 11: Relationship between CO2 emissions and population with log-log transformation 12](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223220)

[Figure 12: Regression model for log CO2 and log population 12](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223221)

[Figure 13: Relationship between CO2 Emission and GDP 13](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223222)

[Figure 14: Relationship between CO2 Emission and GDP after log-log transformation 13](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223223)

[Figure 15: Regression model for log CO2 and Log GDP 13](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223224)

# List of Tables

[Table 1: Observations by country (USA) 7](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223237)

[Table 2: Countries with the largest increase in CO2 efficiency (most significant % drop in CO2 per unit energy) 9](#_Toc116223238)

[Table 3: Countries with the least increase in CO2 efficiency (least significant % reduction in CO2 per unit energy) 10](file:////Users/haruiz/Downloads/MidTerm_Report_Group9.docx#_Toc116223239)

# 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 CO2 emissions has pushed greenhouse gas emissions from energy to their highest level in 2018.

This case study presents global trends in net CO2 emission across different continents over time.

# 2. Objectives and Questions

Under the assumption that CO2 indicators in the atmosphere result from energy-related factors, this study aims to analyze periodic changes in CO2 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 CO2 production by each country on all the continents.

In our project, we studied the following key points

1. Examination of significant increases and decreases in CO2 emission by countries from 2000 to 2018,
2. The relationship between population and CO2 emission,
3. The relationship between CO2 and Gross Domestic Product (GDP)
4. 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 CO2 emissions by countries and continents.

These outcomes will be further helpful in understanding the relationship between CO2 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.

## Chart, bar chart Description automatically generated

Figure 1: Proportion of missing values by column

## Step 1: Data exploration

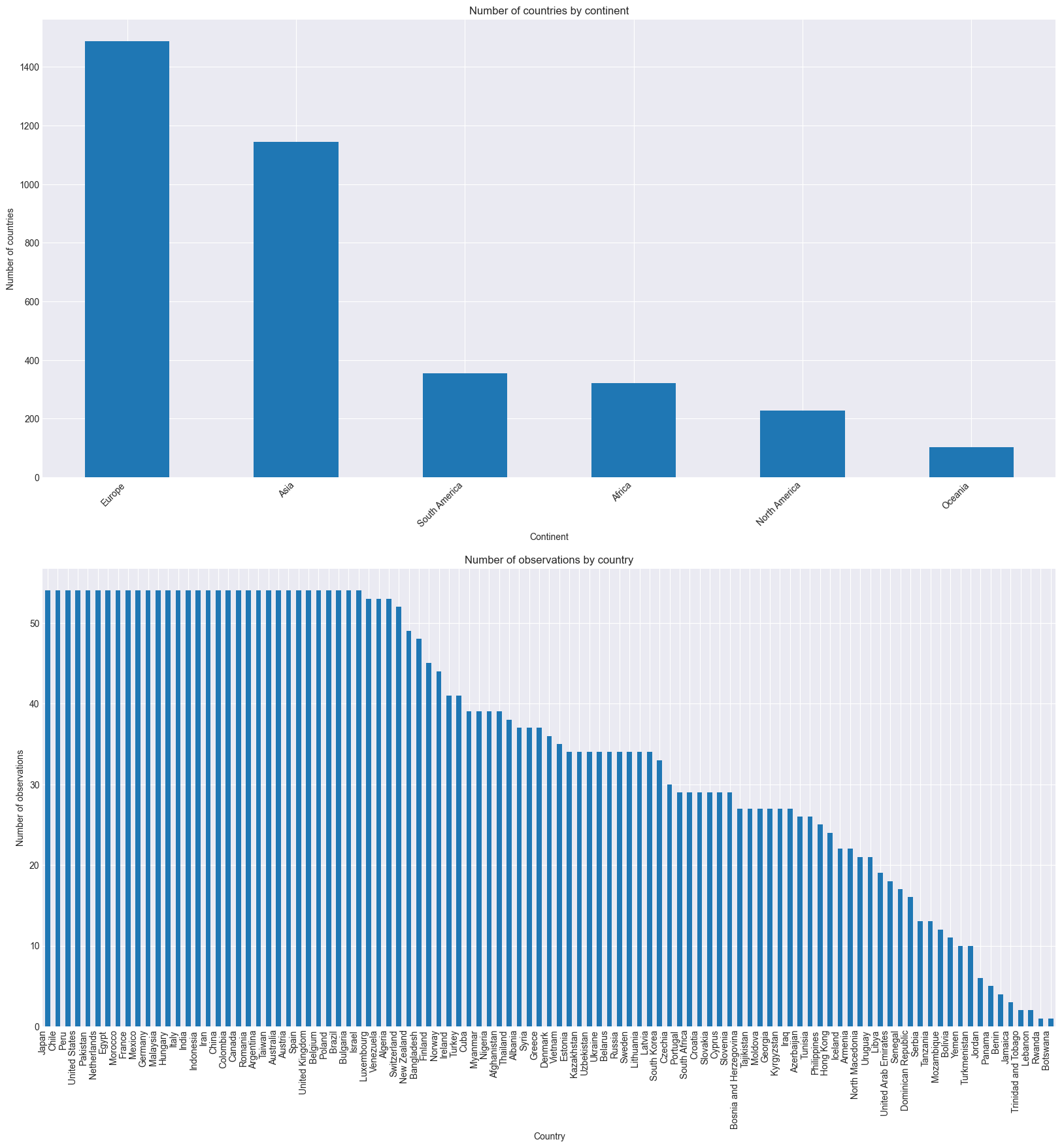


Figure 2: Number of countries by continent and number of observations by country

Diagram

Description automatically generated with medium confidence

Figure 3: Selected variables Distribution

Chart, line chart

Description automatically generated

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

**Country-level observations**

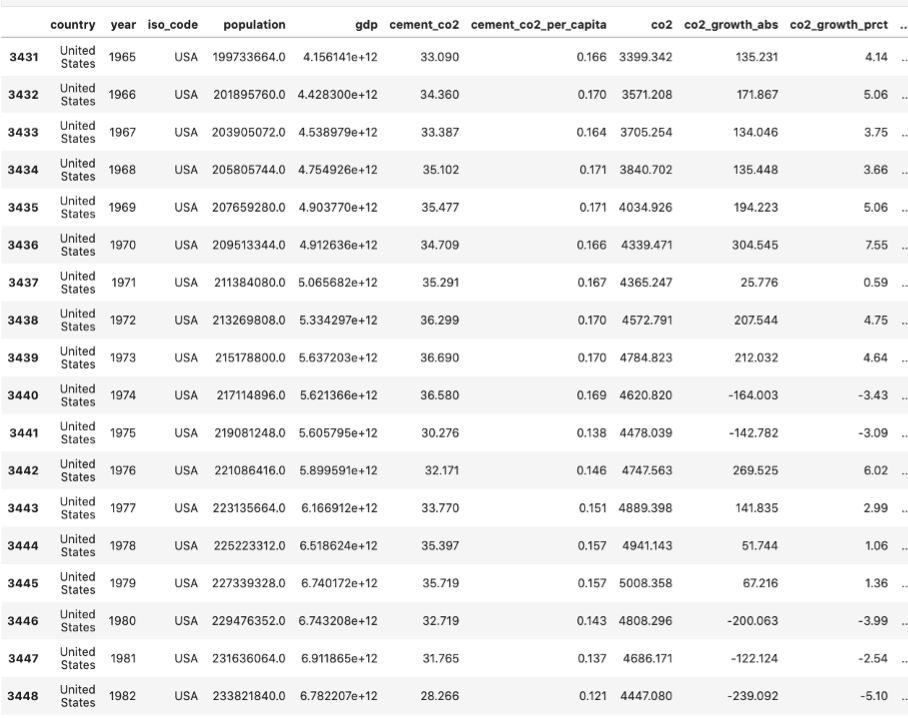


Table 1: Observations by country (USA)

## 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.

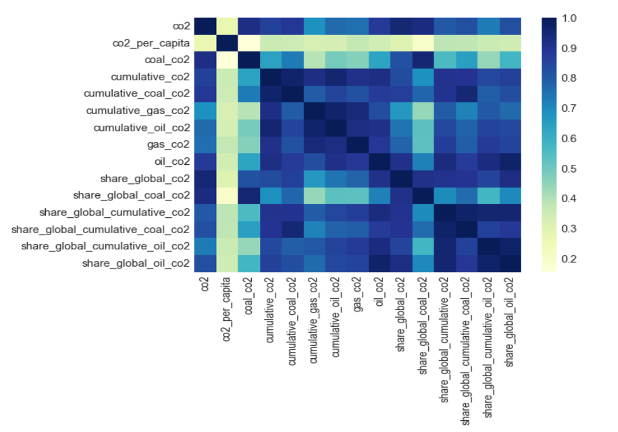
According to our findings, CO2 emissions correlate with cumulative CO2 and CO2 emitted from different sources. Similarly, it was expected that the share of overall global CO2 would be highly associated with cumulative global share, both general and from various sources. Since all these variables could be considered significant responses, we decided to keep overall CO2 emissions and the percentage of global CO2 emissions and drop the rest.

Figure 5: Correlation map for variables with high VIF

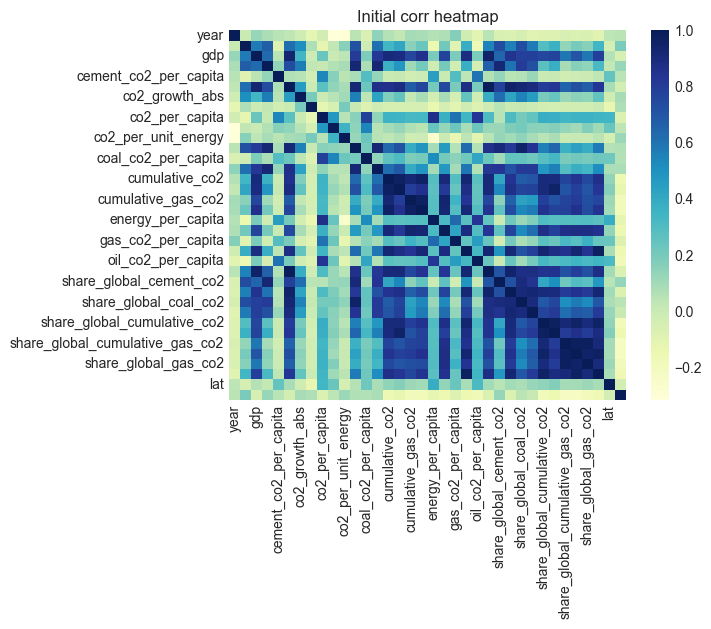


Figure 6: Correlation map for variables with low VIF

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

Chart

Description automatically generated

Figure 7: Correlation map for final variables

## Step-3 Data visualization and analysis

**Question 1:** Which countries from each continent have seen the greatest increases and decreases in CO2 efficiency from 2000 to 2018?

Change in CO2 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 CO2 per unit energy:

Table 2: Countries with the largest increase in CO2 efficiency (most significant % drop in CO2 per unit energy)

Table

Description automatically generated

Here is the trend in CO2 per unit energy for each of these well-performing countries:

Chart, line chart

Description automatically generatedSince these countries had the largest drop in CO2 per unit energy consumption, they had the largest increases in CO2 efficiency by our definition.

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

Next, we can consider the countries from each continent that performed the worst in terms of CO2 efficiency.

Here are the top countries that saw little decline or an increase in CO2 per unit from 2000 to 2018.

Table

Description automatically generated

Table 3: Countries with the least increase in CO2 efficiency (least significant % reduction in CO2 per unit energy)

Similarly, we can visualize the trend of these country’s CO2 per unit energy consumption as follows:

Chart, line chart

Description automatically generatedFrom these results, we can see that although not every country reduced their CO2 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 CO2 efficient and output less CO2 per unit energy.

Figure 9: Countries with least decline or increase in CO2 per unit energy from 2000 to 2018

**Question 2:** What is the relationship between population and CO2 emission?

Chart, scatter chart

Description automatically generatedFirst, we will look at the relationship between population and CO2 emissions for all countries that had measurements in 2018.

Figure 10: Relationship between CO2 emission and population size

Chart, scatter chart

Description automatically generatedSince most of the data in Figure 9 is concentrated near the minimum and variance in CO2 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 CO2 emissions.

Chart, scatter chart

Description automatically generatedWhen 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 CO2 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 CO2 and log population

**Question 3:** What is the relationship between GDP and CO2 emissions?

Chart, scatter chart

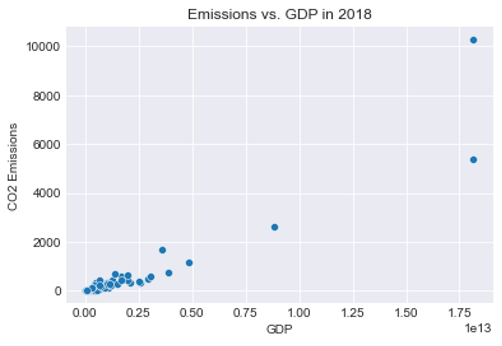
Description automatically generatedThe plot (as shown in Figure 12) showed that there are some outliers in the data, so we used the log-log transformation of the GDP and the CO2 emission and a plot as shown in Figure 13 was observed. A linear regression plot was used to define the relationship between 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

Chart, scatter chart

Description automatically generated

Figure 15: Regression model for log CO2 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 CO2 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 CO2 graphs, the advancement in the economy by 10.2% on a per capita basis also increases CO2 emission by 10%. Similarly, population size was also found to be directly correlated with CO2 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/>