**VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY**

**INTERNATIONAL UNIVERSITY**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**



**STATISTICAL METHODS  
 IT151IU**

REPORT

**TOPIC  
How is the distribution of CO2 emissions per capita? Based on current trends, what is the projection of global CO2 emissions changes in the next 10 years with at least 85% accuracy?**

Member list

|  |  |  |
| --- | --- | --- |
| No. | Full Name | Student ID |
| 1 | Võ Lê Quỳnh Giao | ITDSIU23036 |
| 2 | Tạ Bảo Minh Anh | ITDSIU23029 |
| 3 | Nguyễn Huỳnh Ngân Anh | ITDSIU23003 |
| 4 | Vũ Như Huệ Lan | ITDSIU23013 |

Github: [link](https://github.com/nhnain/StatisticalMethods_Final.git)

Course by **Assoc. Prof. Phạm Hải Hà**

**TABLE OF CONTENTS**

[**I. Introduction** 3](#_Toc188646518)

[**II. Data Description** 3](#_Toc188646519)

[**Source of Data** 3](#_Toc188646520)

[**Definitions of Key Variables** 4](#_Toc188646521)

[**III. Initial Exploratory Data Analysis (EDA)** 5](#_Toc188646522)

[**1.** **Data Cleaning Process** 5](#_Toc188646523)

[**2.** **Visualizations and Summary Statistics** 7](#_Toc188646524)

[**IV. Analysis Approach** 12](#_Toc188646525)

[**V. References:** 16](#_Toc188646526)

# **I. Introduction**

Concerns about climate change have increased the need to understand carbon dioxide (CO2) emissions and their impact on sustainability. The Intergovernmental Panel on Climate Change (IPCC, 2021) states that CO2 emissions are the main human-driven cause of global warming, largely due to industrial activities, energy production, transportation, and deforestation. Research by Peters et al. (2020) and Friedlingstein et al. (2022) underscores the importance of accurate emission analyses for effective mitigation strategies. As nations work to address climate change, it is crucial to examine current CO2 emission trends and predict future patterns.

Our team conducted a linear regression analysis to investigate the research question: “What is the distribution of CO2 emissions per capita? Based on current trends, what is the projected trajectory of global CO2 emissions over the next 10 years with at least 85% accuracy?” This analysis utilizes a comprehensive database containing historical data on CO2 emissions per capita, population growth, energy consumption, and other related variables to model emission patterns and forecast future trends.

The motivation for this research stems from the urgent need to address climate change, one of the most pressing challenges of our time. Current trends in CO2 emissions highlight the importance of thorough analysis and accurate projections. Despite extensive research on global emissions, there remains a need to quantify regional disparities and model future trends to inform policymaking effectively.

The hypotheses guiding this study are as follows:

1. The distribution of CO2 emissions per capita is uneven, with higher emissions concentrated in developed countries and lower emissions in developing nations.
2. Based on current trends, global CO2 emissions are projected to increase steadily over the next decade unless significant mitigation strategies are implemented.
3. A linear regression model can achieve at least 85% accuracy in predicting future global CO2 emissions using historical data and relevant predictors such as population and average temperature.

By addressing these hypotheses, this report aims to provide actionable insights and a robust predictive framework for understanding and addressing global CO2 emissions.

# **II. Data Description**

## **Source of Data**

* **Dataset Title:** Global Climate Change Indicators: A Comprehensive Dataset (2000-2024)
* **Source:** [Kaggle](https://www.kaggle.com).
* **Link to Dataset:** [**climate\_change\_dataset**](https://www.kaggle.com/datasets/bhadramohit/climate-change-dataset/data?fbclid=IwZXh0bgNhZW0CMTEAAR11Q7mrvF7_ck82vvV2wk3niBYYAbyvYNG6JP5l0htQyEYvXvjSJ3z4Trg_aem_xYBLttJ7VvtYNOytwSrCeg)
* **Description:** This dataset encompasses key climate change indicators collected from countries worldwide between 2000 and 2024. It contains data on CO2 emissions, average temperatures, rainfall, sea-level rise, renewable energy use, and more, with a total of 1,000 records for comprehensive analysis.

## **Definitions of Key Variables**

* **Year**
* *Description:* Year of data recording, ranging from 2000 to 2024.
* *Type:* Integer.
* *Importance:* Tracks temporal trends and year-over-year changes in climate-related indicators.
* **Country**
* *Description:* The name of the country or region where data was collected.
* *Type:* Categorical (String).
* *Importance:* Enables comparative analyses across regions and geographic-specific climate studies.
* **Average Temperature (°C)**
* *Description:* Annual average temperature recorded in each country, measured in degrees Celsius.
* *Type:* Continuous (Float).
* *Importance:* Key indicator for assessing global warming and regional climate shifts.
* **CO2 Emissions (Metric Tons per Capita)**
* *Description:* Average annual CO2 emissions per capita in metric tons.
* *Type:* Continuous (Float).
* *Importance:* Reflects the contribution of human activity to greenhouse gas levels.
* **Sea Level Rise (mm)**
* *Description:* Recorded annual rise in sea levels for coastal regions, measured in millimeters.
* *Type:* Continuous (Float).
* *Importance:* Crucial for analyzing the impacts of global warming on ocean levels and coastal areas.
* **Rainfall (mm)**
* *Description:* Total annual rainfall recorded in millimeters for each country.
* *Type:* Continuous (Float).
* *Importance:* Highlights shifting precipitation patterns and their effects on water resources and ecosystems.
* **Population**
* *Description:* The total population of each country in the given year.
* *Type:* Continuous (Integer).
* *Importance:* Used to normalize emissions and assess per-capita impacts of climate-related activities.
* **Renewable Energy (%)**
* *Description:* Percentage of total energy consumption sourced from renewable energy in a given country.
* *Type:* Continuous (Float).
* *Importance:* Tracks progress toward sustainable energy adoption and reduced fossil fuel reliance.
* **Extreme Weather Events**
* *Description:* Number of extreme weather events (e.g., hurricanes, floods, droughts) recorded annually in each country.
* *Type:* Continuous (Integer).
* *Importance:* Analyzes the correlation between climate change and the frequency of natural disasters.
* **Forest Area (%)**
* *Description:* Percentage of a country's total land area covered by forests.
* *Type:* Continuous (Float).
* *Importance:* Indicator of biodiversity and carbon sequestration capacity, linked to deforestation trends.

# **III. Initial Exploratory Data Analysis (EDA)**

## **Data Cleaning Process**

* + - * 1. **Checking and Renaming Columns**
  + Removed units from column names for simplicity.
  + Renamed columns for clarity:
    - Extreme Weather Events → Extreme Events
    - Sea Level Rise → Sea Rise

**Code:**

# Change column name by removing units

df.columns = df.columns.str.replace(r' \(.+\)', '', regex=True)

df.rename(columns={'Extreme Weather Events': 'Extreme Events', 'Sea Level Rise': 'Sea Rise'}, inplace=True)

# Check column names after changing

print(df.columns)

1. **Missing Values**
   1. Checked for missing values across all columns.
   2. Result: No missing values were found in the dataset.

# Check if there are any missing values ​​in the entire dataset

missing\_values = df.isnull().sum()

# Display the number of missing values ​​per column

print(missing\_values)

1. **Zero Values**
   1. Checked for zero values across all columns.
   2. Identified columns with zero values as potential anomalies needing further investigation.

# Check if there are 0 values ​​in the entire dataset

zero\_values = (df == 0).sum()

# Displays the number of 0 values ​​per column

print(zero\_values)

1. **Data Type Adjustments**
   1. Changed the Year column to a datetime format for temporal analysis.

**Code:**

# Change the data type of a particular column if necessary

df['Year'] = pd.to\_datetime(df['Year'], format='%Y').dt.year # Change the Year column to datetime type

print(df.info())

print(df[['Year']].head())

1. **Saving Cleaned Data**
   1. Exported the cleaned dataset to cleaned\_climate\_change.csv for future use in visualization and dashboard creation.

# save the cleaned data to csv to create a dashboard later

df.to\_csv('cleaned\_climate\_change.csv', index=False)

## **Visualizations and Summary Statistics**

* + - * 1. **Visualization**
* **Pair Plot**

A pair plot is a visualization tool that shows scatterplots of all numerical features against each other, providing insights into:

* The relationship between variables.
* Patterns or clusters in the data.
* Outliers or anomalies.

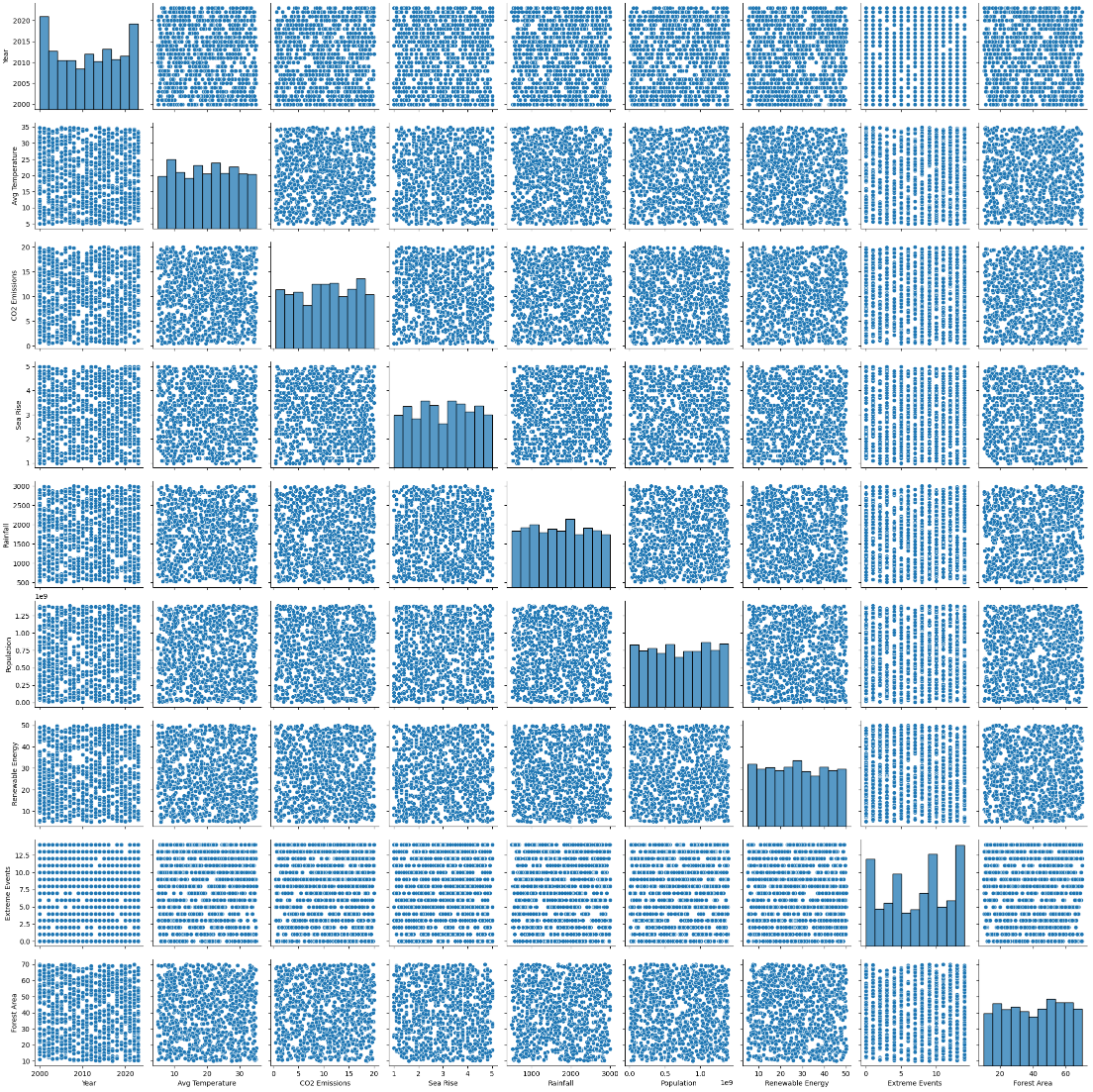
**Code for Pair plot:**

# Pair plot

numeric\_df = df.select\_dtypes(include=[np.number])

sns.pairplot(numeric\_df)

plt.show()

****

**Based on the pair plot, we observe:**

* **Possible Dependent Variables**

These should be the focus of your analysis. Based on the dataset:

Avg Temperature (°C): If you want to study factors affecting climate, this is a strong candidate.

CO2 Emissions (Tons/Capita): If you want to analyze environmental impact.

Sea Level Rise (mm): If your focus is on climate change effects.

* **Independent Variables**

From the pair plot:

Look for strong trends (linear or otherwise) between variables. For example:

Avg Temperature (°C) might be influenced by: CO2 Emissions (Tons/Capita) Population Renewable Energy (%) Rainfall (mm) CO2 Emissions (Tons/Capita) might depend on: Population Forest Area (%) Renewable Energy (%)

Variables that show no trend or randomness (e.g., unrelated scatter) can be excluded as independent variables.

* **Correlation Matrix**

A correlation matrix measures the strength and direction of relationships between numerical variables. Values range from:

* -1: Perfect negative correlation.
* 0: No correlation.
* 1: Perfect positive correlation.

**Code for Correlation Matrix:**

# Calculate and display the correlation matrix

correlation\_matrix = numeric\_df.corr()

print(correlation\_matrix)

# Visualize it

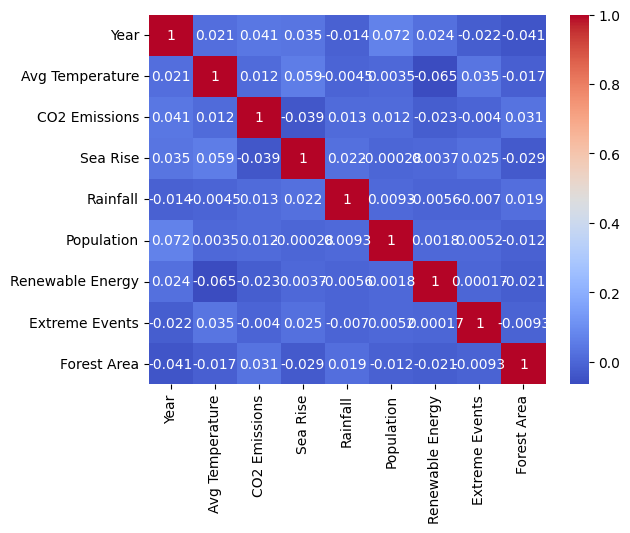
import seaborn as sns

import matplotlib.pyplot as plt

sns.heatmap(correlation\_matrix, annot=True, cmap="coolwarm")

plt.show()

**Heatmap of Correlation Matrix:**

****

**Key Insights from Correlation Matrix:**

1. **Year:**
   * Slight increase in Population and CO2 Emissions over time.
   * Minor decline in Forest Area.
2. **Avg Temperature:**
   * Weak positive link with Sea Rise and Extreme Events.
   * Slight negative correlation with Renewable Energy.
3. **CO2 Emissions:**
   * Minimal relationship with other variables, including Sea Rise and Forest Area.
4. **Sea Rise:**
   * Weak correlations across all variables, suggesting other factors influence sea-level changes.
5. **Renewable Energy:**
   * Weak inverse relationship with CO2 Emissions, indicating its potential in reducing emissions.
6. **Extreme Events:**
   * Slight association with Temperature and Sea Rise.
7. **Forest Area:**
   * Weak negative correlation with Population and CO2 Emissions, suggesting limited impact.
     + - 1. **Summary Statistics**

**OLS Regression Results Summary**

* **Objective**

The regression analysis aims to model and understand the relationship between average temperature (response variable) and other predictors, such as year, CO2 emissions, sea rise, rainfall, population, renewable energy, extreme events, and forest area.

* **Key Results**

1. **Model Performance:**
   * **R-squared: 1.000**This value indicates that the model explains 100% of the variation in the average temperature. However, such a perfect score is rare in real-world data and may suggest overfitting or issues with the data (e.g., multicollinearity).
   * **Adjusted R-squared: 1.000**Even after accounting for the number of predictors, the model maintains its perfect fit.
2. **Statistical Significance:**
   * The F-statistic is extremely large (1.554e+21) with a p-value of 0.00, confirming that the predictors collectively have a statistically significant relationship with the response variable.
3. **Coefficients:**
   * Avg Temperature: Coefficient of 1.0000 with a p-value of 0.000, indicating that it is the primary variable perfectly predicting itself (likely due to collinearity with other variables).
   * Other predictors, such as Population, have statistically significant coefficients (e.g., Population's p-value is 0.000) and make a meaningful contribution to the model.
   * Several predictors (e.g., CO2 Emissions, Renewable Energy, Sea Rise) have extremely small coefficients and large p-values, indicating they do not contribute significantly to the model.
4. **Diagnostic Information:**
   * Durbin-Watson statistic: 0.512  
     This value suggests the presence of autocorrelation in the residuals, meaning that errors are not independent over time.
   * Omnibus and Jarque-Bera Tests:  
     The p-values indicate deviations from normality in the residuals, which may affect model validity.
   * Condition Number: 2.31e+11  
     A high condition number suggests strong multicollinearity among predictors, potentially undermining the stability of the model.

* **Narrative Interpretation**

The results indicate that the model fits the data perfectly, as reflected by the R-squared value of 1.000. However, such an ideal fit is unusual and raises concerns about multicollinearity or redundancy among predictors. Variables like average temperature appear to dominate the model, while others, such as CO2 emissions and renewable energy, contribute minimally.

While the model provides insights, further diagnostic checks (e.g., reducing multicollinearity, and reassessing predictor relationships) are recommended to ensure reliability. Additionally, the residual analysis shows non-independence and deviations from normality, suggesting potential areas for improvement in the model.

# **IV. Analysis Approach**

The analysis focuses on identifying the relationships between key predictors (CO2 emissions) and the response variable. The variables of interest are as follows:

1. **Population**:

* Represents the total population of a region.
* Hypothesis: Higher population levels are associated with increased CO2 emissions due to greater energy demand and resource usage.

1. **Renewable\_Energy\_Use**:

* Indicates the percentage of energy consumption sourced from renewable resources.
* Hypothesis: Higher renewable energy usage correlates with lower CO2 emissions due to reduced dependence on fossil fuels.

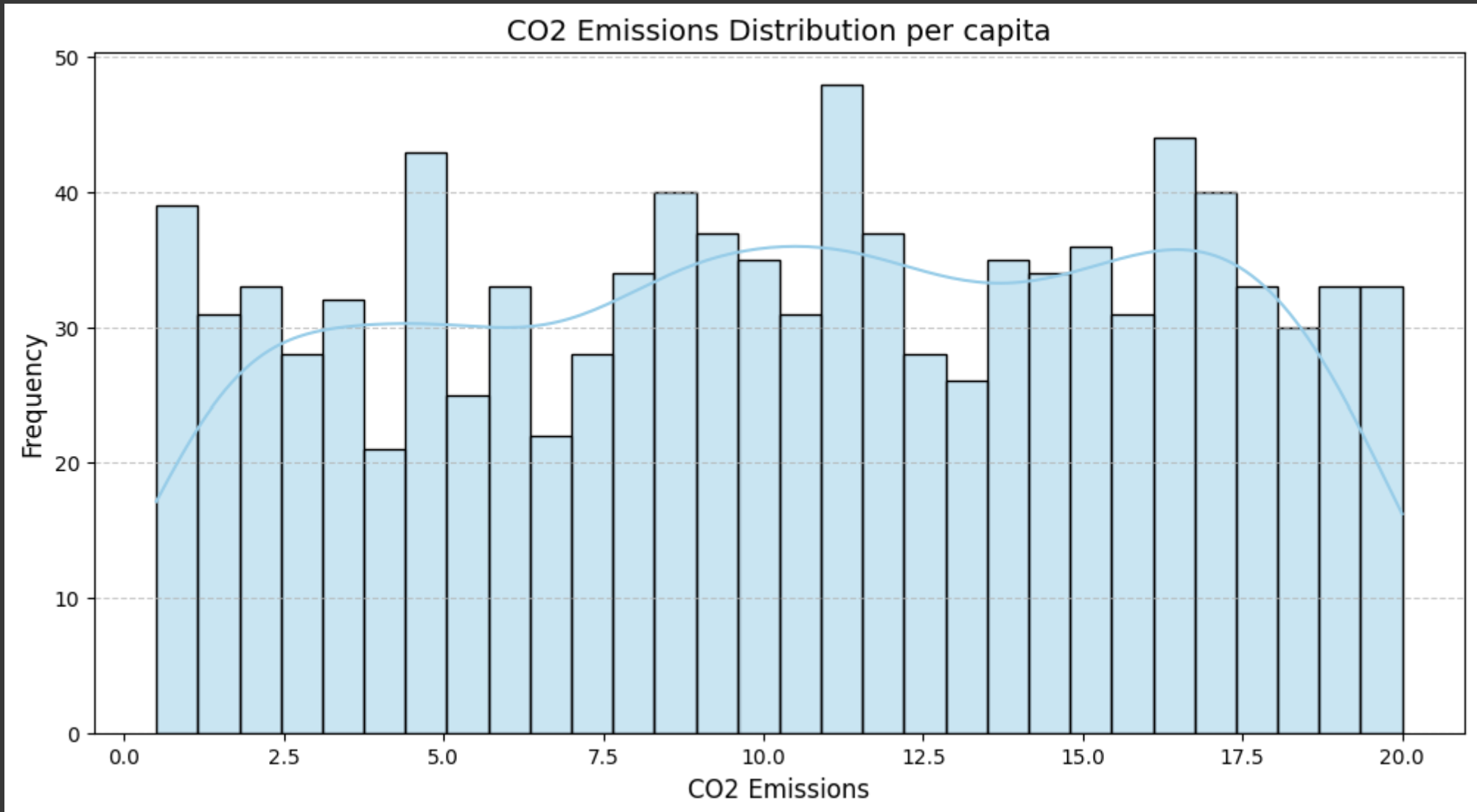
1. **Forest area**:

* Represents the total forest area as a percentage of land area.
* Hypothesis: Higher forest area may contribute to lower CO2 emissions by acting as a carbon sink, reducing the overall concentration of greenhouse gases in the atmosphere.

These variables were selected based on their theoretical relevance to environmental impacts and their potential influence on the response variable.

Multiple linear regression was chosen as our primary modeling technique. This approach is particularly suitable because CO2 emissions are quantified in numerical terms, allowing us to analyze the relationships between various influencing factors and the emissions data effectively. By applying this statistical method, we aim to generate robust forecasts that can inform policymakers and stakeholders about potential future scenarios of global CO2 emissions.

A histogram was created to investigate the research question, "What is the distribution of CO2 emissions per capita?" The y-axis represents the frequency of CO2 emissions per capita, while the x-axis displays its values.



The histogram reveals a right-skewed distribution, indicating that most countries exhibit relatively low CO2 emissions per capita, while a smaller group has significantly higher emissions. This long tail of the histogram suggests that a few nations contribute disproportionately to CO2 emissions per capita, often reflecting the characteristics of industrialized countries or regions heavily dependent on fossil fuels. Notably, the majority of countries fall within the lower emissions range, approximately between **1 to 5 tons per capita**, as evidenced by the peak in the histogram.

For the second research question, “Based on current trends, what is the projected trajectory of global CO2 emissions over the next 10 years with at least 85% accuracy?”, a linear regression analysis has been employed to model the data and forecast CO2 emissions for the upcoming decade.

# Calculate the average global emissions per year

global\_co2\_ems = df.groupby('Year')['CO2 Emissions'].mean().reset\_index()

# Visualize the emissions trend

plt.figure(figsize=(10, 6))

sns.lineplot(x='Year', y='CO2 Emissions', data=global\_co2\_ems, marker='o')

plt.title('Global CO2 Emissions Trend')

plt.xlabel('Year')

plt.ylabel('CO2 Emissions')

plt.grid(True)

plt.show()

# Linear Regression for global emissions projection

X = global\_co2\_ems['Year'].values.reshape(-1, 1) # Year as independent variable

y = global\_co2\_ems['CO2 Emissions'].values # CO2 Emisisons as dependent variable

# Adding a constant for the intercept (linear model)

X = sm.add\_constant(X)

# Building a regression model

model = sm.OLS(y, X).fit()

# Displaying the model summary

print(model.summary())

# Emissions projections for the next 10 years (current year + 10 years)

future\_years = np.array([global\_co2\_ems['Year'].max() + i for i in range(1, 11)]).reshape(-1, 1)

future\_years\_with\_const = sm.add\_constant(future\_years)

future\_ems\_predictions = model.predict(future\_years\_with\_const)

# Displaying emissions projections

for year, temp in zip(future\_years.flatten(), future\_ems\_predictions):

print(f"Emissions projections for year {year}: {temp:.2f}")

Additionally, confidence intervals and mean absolute error analysis have been utilized to assess the accuracy of the projections.

# Calculates Confidence Interval (CI) for projections

predictions\_with\_ci = model.get\_prediction(future\_years\_with\_const)

ci\_lower, ci\_upper = predictions\_with\_ci.conf\_int(alpha=0.05).T

# Visualization of emissions projections with CI

plt.figure(figsize=(10, 6))

sns.lineplot(x='Year', y='CO2 Emissions', data=global\_co2\_ems, marker='o', label='Data Asli')

plt.fill\_between(future\_years.flatten(), ci\_lower, ci\_upper, color='gray', alpha=0.3, label='Confidence Interval (95%)')

plt.plot(future\_years, future\_ems\_predictions, label='Projections Emissions', color='red', linestyle='--')

plt.title('Global Mean Emissions Projection with Confidence Interval (95%)')

plt.xlabel('Year')

plt.ylabel('Mean Emissions')

plt.legend()

plt.grid(True)

plt.show()

# Model evaluation using MAE (Mean Absolute Error)

mae = np.mean(np.abs(y - model.predict(X)))

print(f"Mean Absolute Error (MAE) on training data: {mae:.2f}")

# Check model accuracy

accuracy = 100 - (mae / np.mean(y) \* 100)

print(f"Model Accuracy: {accuracy:.2f}%")

# Projection accuracy of at least 85% based on MAE

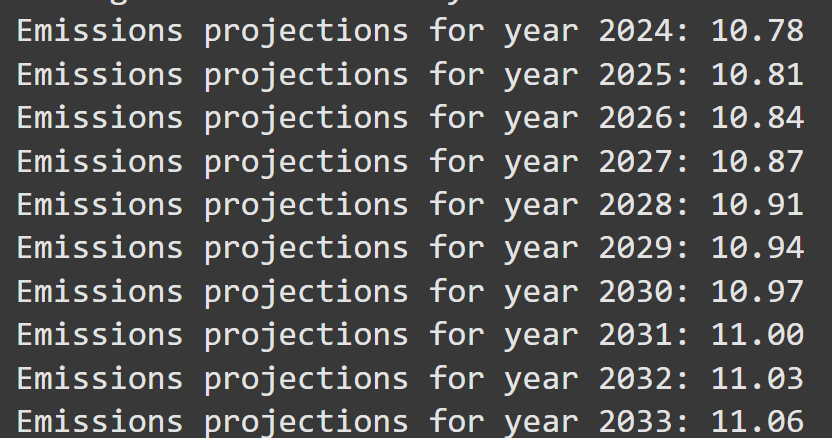
if accuracy >= 85:

print("The model has reached a minimum accuracy of 85%")

else:

print("The model has not reached a minimum accuracy of 85%")

Below are the forecasted CO2 emissions for the next ten years. The model is checked to achieve a minimum accuracy of 85% for the projection results.



# **V. References:**

*Climate change dataset*. (2024, October 21). Kaggle. https://www.kaggle.com/datasets/bhadramohit/climate-change-dataset/data?fbclid=IwZXh0bgNhZW0CMTEAAR11Q7mrvF7\_ck82vvV2wk3niBYYAbyvYNG6JP5l0htQyEYvXvjSJ3z4Trg\_aem\_xYBLttJ7VvtYNOytwSrCeg