Climate Change: Analysis of Emissions and their Relationship to GDP

Zagaria Simone
Matr. 2145389

Statistical Methods for Official Statistics

AA 2024/2025

Introduction

What?

- Our goal is to **Analyze Total Emissions** from various countries around the world between 2000 and 2020
- **Predict Future Emissions** and **Compare Emissions and GDP**

Why?

- Understand global emission trends, forecast future impacts, and assess the link between economic growth and sustainability

How?

- Dataset used (Kaggle):

1. Total Emissions Per Country (2000-2020)[1]
2. GDP Data (1964-2017)[2]

[1] ref: Total Emission Dataset
[2] ref: GDP Dataset

1. Preprocessing

Dataset loading, Data cleaning

Dataset exploration

Emissions Dataset

- Around 50 k rows
- 25 columns (2000-2020)
- “Wide” Format

```
[4]  1  df_emissions.info()
     ✓ 0.0s

...  Data columns (total 25 columns):
      #   Column   Non-Null Count  Dtype
     ---  ------   --------------  -----
      0   Area     49473 non-null  object
      1   Item     49473 non-null  object
      2   Element  49473 non-null  object
      3   Unit     49473 non-null  object
      4   2000     47143 non-null  float64
      5   2001     45614 non-null  float64
      6   2002     45708 non-null  float64
      7   2003     45729 non-null  float64
```

GDP Dataset

- 264 rows
- 63 columns (1960-2017)
- “Wide” Format

```
[35]  1  df_gdp.info()

...  <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 264 entries, 0 to 263
     Data columns (total 63 columns):
      #   Column          Non-Null Count  Dtype
     ---  ------          --------------  -----
      0   Country Name    264 non-null    object
      1   Country Code    264 non-null    object
      2   Indicator Name  264 non-null    object
      3   Indicator Code  264 non-null    object
      4   1960            124 non-null    float64
      5   1961            124 non-null    float64
```

Dataset Cleaning

However, the data presented numerous problems:

Problem 1

Useless columns, such as:

- 'Country Code', 'Indicator Name', 'Indicator Code'
- 'Unit' (all rows were Kilotonnes)

Solution:
The listed columns were **removed** from the dataset

Problem 2

Many negative values in emissions dataset

Solution:
The data was **filtered** out of negative values

Problem 3

Many Missing and NaN values

Solution:
Temporal Interpolation was applied for both datasets

Problem 4

Wide format is not optimal for data analysis

Solution:
The dataset was transformed in **'Long' format**

Final Dataset

Cleaned version: no more NaN or negative values, data was interpolated, transformed into 'Long' format through adding 'Year' column

```
1  df_emissions_long.info()
6]  ✓ 0.1s

<class 'pandas.core.frame.DataFrame'>
Index: 997996 entries, 206 to 1037545
Data columns (total 5 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   Country Name  997996 non-null  object
 1   Item          997996 non-null  object
 2   Element       997996 non-null  object
 3   Year          997996 non-null  datetime64[ns]
 4   Emissions     997996 non-null  float64
dtypes: datetime64[ns](1), float64(1), object(3)
memory usage: 45.7+ MB
```

```
1  df_gdp_long.info()
7]  ✓ 0.0s

<class 'pandas.core.frame.DataFrame'>
Index: 11796 entries, 1 to 15311
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Country Name  11796 non-null  object
 1   Year          11796 non-null  datetime64[ns]
 2   GDP           11796 non-null  float64
dtypes: datetime64[ns](1), float64(1), object(1)
memory usage: 368.6+ KB
```

```
1  print(df_emissions_long.isna().any())

Country Name    False
Item            False
Element         False
Year            False
Emissions       False
dtype: bool
```

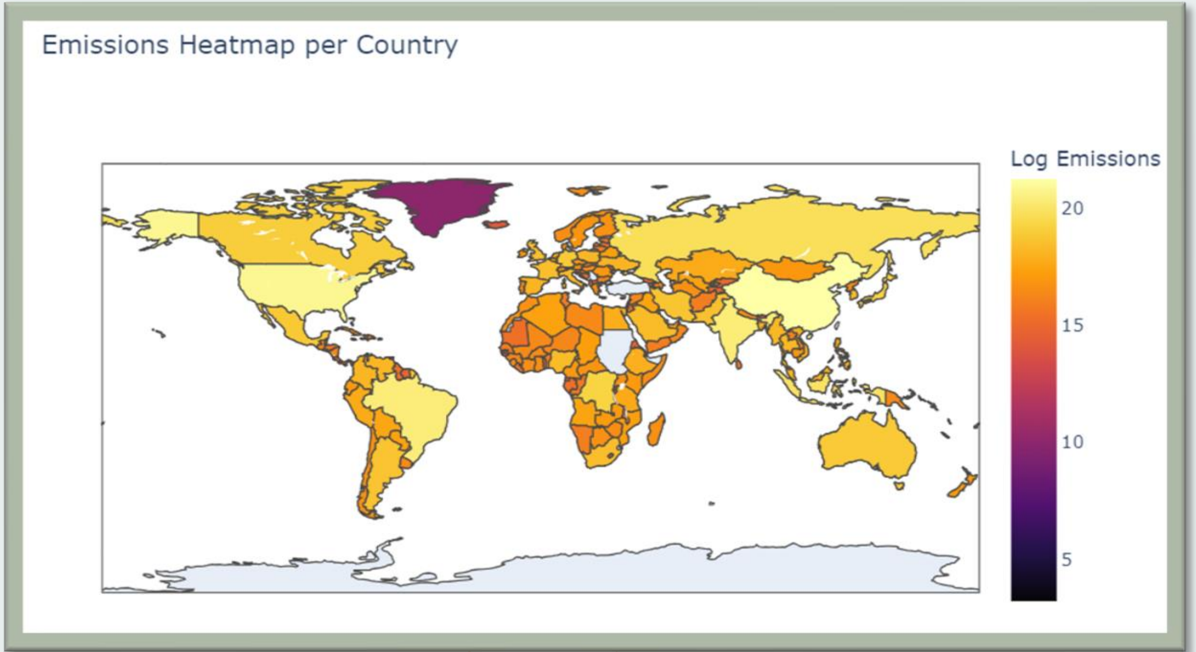
```
1  print(df_gdp_long.isna().any())

Country Name    False
Year            False
GDP             False
dtype: bool
```

2. Exploratory Data Analysis (EDA)

analyzing key distribution patterns,
understanding temporal and regional trends

Emissions Heatmap



We applied a **logarithmic scale** to the emissions data to enhance visibility and emphasize differences across countries

The map highlights that **North America** and **East Asia** are the regions with the highest emissions, while **Africa** has the lowest

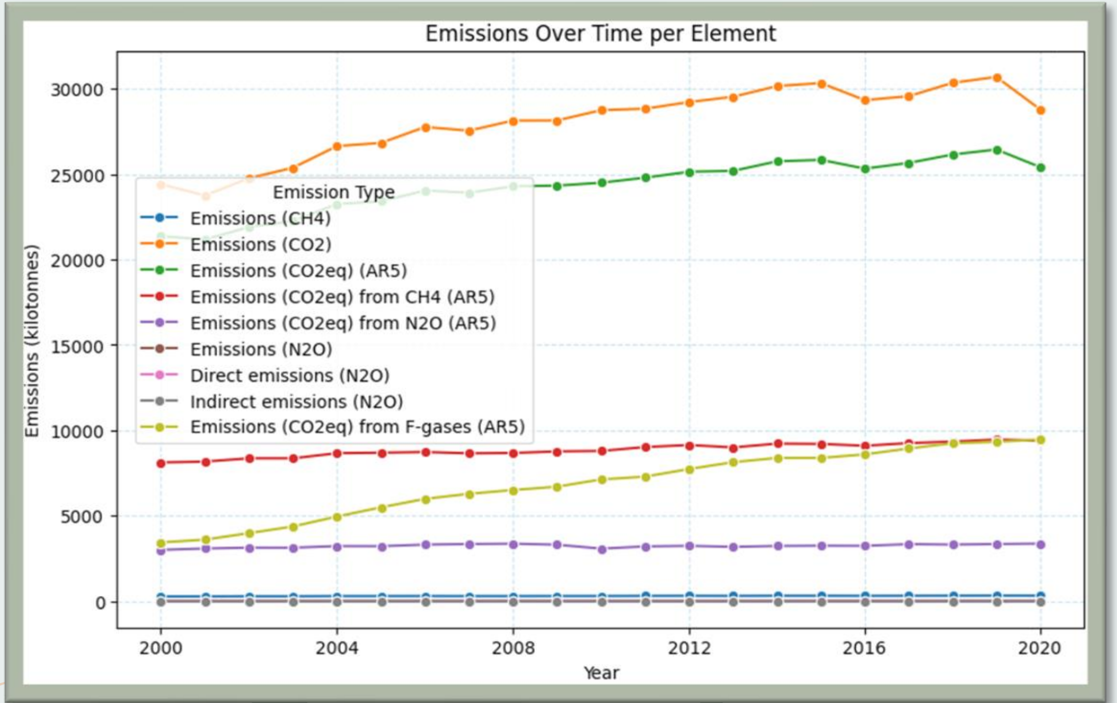
It can indicate how economic development and industrialization strongly influence a country's emission levels

Emissions 2000-2020 by Element

This chart illustrates the **trend of gas emissions by type** from 2000 to 2020

CO_2 emissions dominate, showing a steady increase over time, rising by approximately 20% over the last 2 decades

Minor contributors, such as **methane (CH_4)**, **nitrous oxide (N_2O)** and indirect emissions, tend to remain relatively stable.



Total Emissions by Element

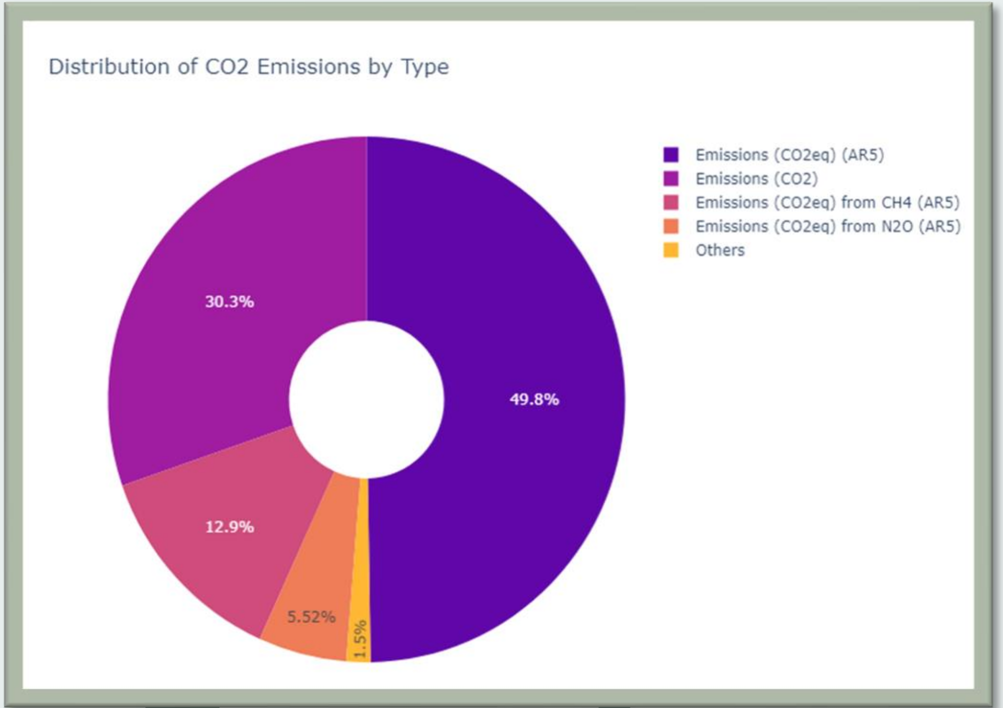
Following up to the previous chart, this pie chart in fact shows that:

The main polluting gas is:

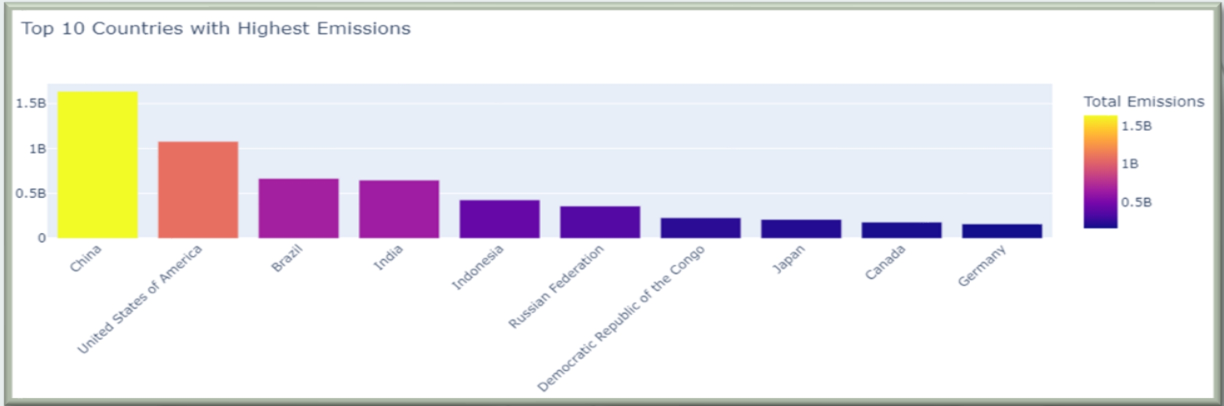
CO_2
(80.1 %)

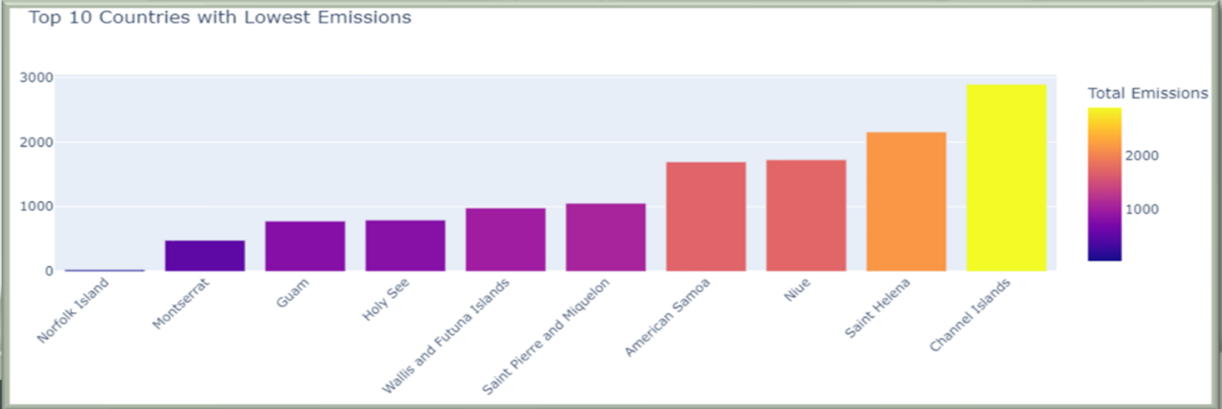
Followed by:

CH_4
(12.9 %)
And
N_2O
(5.52 %)



Top 10, Bottom 10 Emissions Countries





The top 10 countries include major emerging and industrialized countries, with **China** leading as expected, followed by the **United States** and **Brazil**

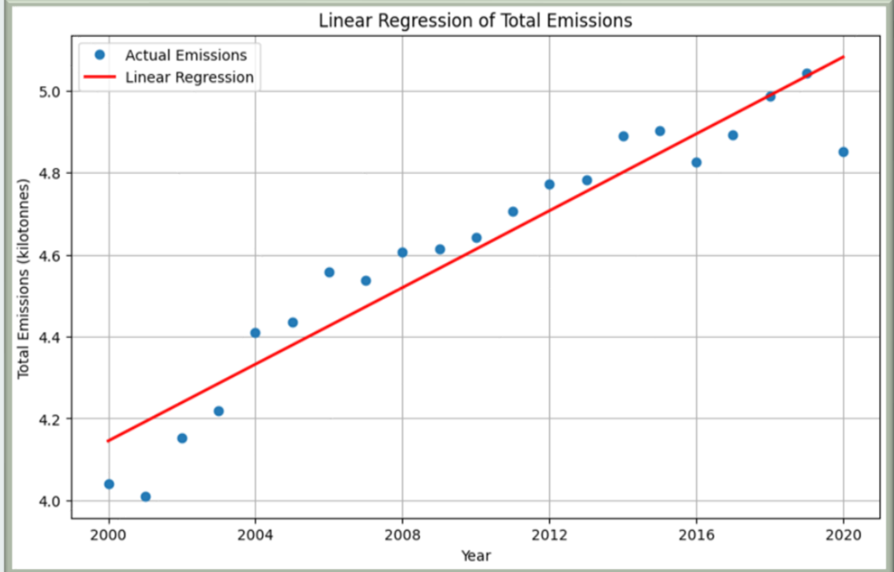
The countries with the lowest emissions are small, independent states or territories, such as the **Vatican**, **Montserrat**, and **Norfolk Island**

We can observe how **global powers contribute disproportionately to global emissions**.

3. Analyses

Linear Regression, Classification, Time Series Predictions, etc.

Linear Regression




```
Coefficient (slope): 4629636.56599357
```

The aim of this Linear Regression is to understand how emissions have evolved over time and identify a general trend.

We have put the **Years** on the X-Axis and the **Total global emissions** on the Y-Axis to observe the relationship between these two variables

The resulting graph shows a clear, steady increase in emissions from 2000 to 2020. The regression suggests that the rise in emissions has been nearly **linear**.

Furthermore, the **slope** value is **positive**, indicating that emissions have risen over the years.

Classification

Next, let's perform a classification of the countries based on their total emission level:

Classification based on **quantiles**:
90-100th = Dark Red (very high)
70-89th = Red (high)
40-69th = Yellow (medium)
10-39th = Green (low)
0-9th = Gray (very low)

This classification allows us to easily **categorize countries** based on their emission levels and assign them to specific groups

The map on the left provides a **zoom on Europe**. As previously mentioned, the most industrialized countries dominate, **implying a potential correlation between emissions and GDP**.

Correlation between Emissions and GDP

Let's then find evidence on the correlation of emissions and GDP:

Correlation between GDP and Emissions: 0.733492688345309, p-value: 0.0

We **merged the two dataset** of Total Emissions per Country and the GDP per country by the 'Year column

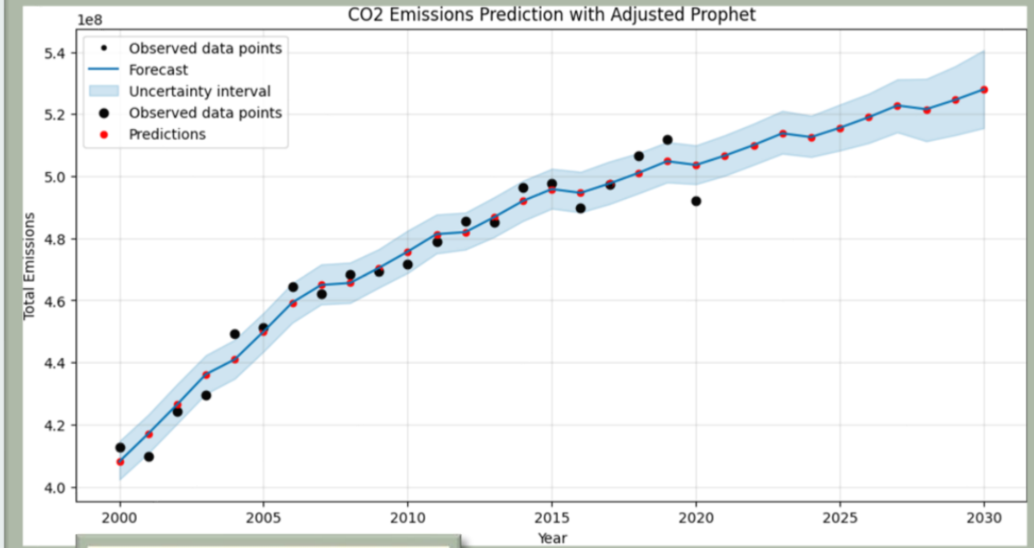
We then calculated the Pearson Correlation. The results presented a **correlation value of over 0.7**, indicating an **overall strong correlation** between the two variables.

However, while this was true for most countries, but some had minor or even negative correlation, indicating that even other factors may influence this relationship.

The **p-value** equals to zero indicates that the observed correlation is **statistically significant**, and it is highly unlikely to be random.

# index	Country Name	# Correlation
	Mancante: 0 (0%) Distinto: 174 (100%) 174 Valori distinti	Mancante: 0 (0%) Distinto: 174 (100%) Min -0.972623708... Max 0.989816691...
0	Afghanistan	0.9529203419003796
1	Albania	0.08736417516229479
2	Algeria	0.7534579183494262
3	American Samoa	0.706160520080175
4	Andorra	-0.01634357224231186
5	Angola	0.7208898558948503
6	Antigua and Barbuda	0.942384238929272
7	Argentina	-0.42739710561042427
8	Armenia	0.9486125288343719
9	Aruba	-0.39008452988684117
10	Australia	-0.645474130912841
11	Austria	0.680605097622988
12	Azerbaijan	0.691863623550387
13	Bahrain	0.9631452064763626
14	Bangladesh	0.9505426160076648
15	Barbados	0.531387544048864

Time Series Prediction



Year	Predicted Emissions
2021-12-31	5.065949e+08
2022-12-31	5.099935e+08
2023-12-31	5.137884e+08
2024-12-31	5.125815e+08
2025-12-31	5.155765e+08
2026-12-31	5.189751e+08
2027-12-31	5.227700e+08
2028-12-31	5.215631e+08
2029-12-31	5.245581e+08
2030-12-31	5.279567e+08

Here we can observe the exact numbers of the predictions made by Prophet.

For this analysis we will use **Prophet**, a forecasting tool specialized in time series data, to analyze the trend in total annual emissions predict the **future trend**.

The model predicts a **continuation of the linear increase** in total emissions over the next decade.
This forecast highlights the possible **worsening of climate catastrophes** unless significant interventions occur.

The rise in emission in 2030, according to the forecast, indicates a **8.83% rise** compared to **2020**, and a **30.70% rise** compared to **2000**.

4. Conclusions

Final Conclusions drawn from the Data Analyses

In Conclusion:

1.

Economic development and industrialization strongly influence a country's emission levels. In fact, the top polluting countries, like China and USA, are all major emerging or highly industrialized nations.

2.

Over the last two decades, emissions have shown an **almost-linear increase**, rising by approximately **20% since 2000**. The forecasts indicate this upward trend is likely to continue, with further **increases expected by 2030**.

3.

Carbon dioxide (CO_2) is the dominant contributor to pollution, accounting for over **80%** of total emissions. It is followed by **methane (CH_4)** and **nitrous oxide (N_2O).**

4.

Analyzing the relationship between emissions and GDP**, we found a significant correlation**. The data indicates a **strong positive relationship**, confirming that **emission levels tend to be closely tied to economic development.**

Thank you for the attention!