Forecasting Carbon Emissions Across Continents

INTRODUCTION

In an era where the imperatives of climate change and environmental sustainability command global attention, the scrutiny of greenhouse gas (GHG) and carbon dioxide (CO_2) emissions assumes unprecedented urgency.

This report unveils a meticulous data analysis endeavor, dissecting the multifaceted dynamics of GHG and CO_2 emissions to glean critical insights and inform strategic decision-making.

The study is structured into three distinct parts, each addressing crucial elements in understanding the global environmental impact.

Part I: Data Exploration

delves into the descriptive statistics of GHG and CO_2 emissions, offering insights into current trends and contributors. This section includes a detailed examination of CO_2 emissions by sector, identifies the top 10 CO_2 contributors in 2022, and provides a continental ranking of these contributors. It also explores the changes in the ranking of countries over the years in terms of CO_2 and GHG emissions, thereby highlighting the evolving landscape of environmental impact.

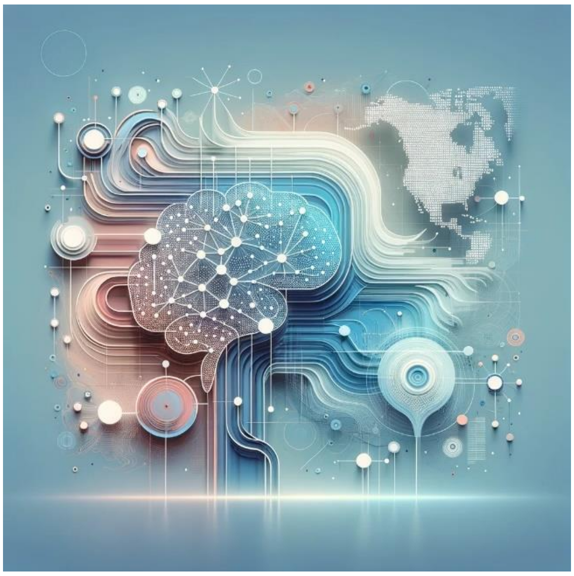
AI generated picture to illustrate - data not representative – free to use

Part II: Correlation Analysis

shifts the focus to understanding the interrelationships between various factors. It examines the correlation between CO_2 and GHG emissions, and between different sectors contributing to these emissions. Furthermore, it investigates the correlation between the emissions of CO_2/GHG and the Gross Domestic Product (GDP), as well as the correlation between continents in terms of their emission profiles. This analysis is pivotal in identifying the economic and geographical patterns associated with environmental impact.

AI generated picture to illustrate – free to use

Part III: Predictive Modelling represents the culmination of the study, where a Long Short-Term Memory (LSTM) model is developed. This model aims to predict the emissions of GHG and CO_2 by continent, offering a valuable tool for forecasting and planning purposes. The predictive model is not only a testament to the power of data science in environmental studies but also serves as a critical instrument for policymakers and environmentalists in strategizing future actions.



AI generated picture to illustrate – free to use

This report, through its rigorous analysis and sophisticated modelling, is not merely an academic exercise but a clarion call to action. It underscores the potential of data science as an ally in the environmental crusade, offering actionable intelligence that can steer policy formulation and environmental advocacy in a direction that safeguards the health of our planet for generations to come.

Enjoy and do not hesitate to contact me for future enquires.

Table of Contents

INTRODUCTION 1

PART I: Explanatory Data Analysis: 4

1) Descriptive statistics: 4
2) Emission trends by sector: 9
3) Top 10 contributors: 11
4) Rank of the contributors by continent 13
5) Insights from the exploratory data analysis: 19

PART II: Correlation Analysis: 20

1) Correlation between CO2 and GHG emissions: 20
2) Correlation between sectors: 21
3) Correlation between emissions and PIB: 24
4) Correlation of emissions between continents: 27

PART III: Predictive Modelling: 30

1) For CO2: 30
2) For GHG: 32

Conclusions and Insights: 34

PART I: Explanatory Data Analysis:

1) Descriptive statistics:

First let us import the libraries needed.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
# from tensorflow.keras.models import Sequential
# from tensorflow.keras.layers import LSTM, Dense
# from tensorflow.keras.optimizers import Adam
from sklearn.metrics import mean_absolute_error, mean_squared_error
```

- **CO2 emissions:**

Importation of the datasets "CO2"

```
co2_sector=pd.read_csv('fossil_CO2_by_sector_country_su.csv',delimiter=";")
co2_capita=pd.read_csv('fossil_CO2_per_capita_by_countr.csv',delimiter=';')
co2_gdp=pd.read_csv('fossil_CO2_per_GDP_by_country.csv',delimiter=';')
co2_total=pd.read_csv('fossil_CO2_totals_by_country.csv',delimiter=';')
```

As an example, here is the co2 by sector dataset:

co2_sector ···

	Substance	Sector	EDGAR Country Code	Country	1970	1971	1972	1973	1974	1975	...	2013	2014	2015	2016	20
0	CO2	Agriculture	AFG	Afghanistan	0,029228567	0,029228567	0,029228567	0,029228567	0,039966661	0,045309517	...	0,055157133	0,084490461	0,116966646	0,162799971	0,310880
1	CO2	Agriculture	ALB	Albania	0,1133	0,1133	0,1133	0,1133	0,113614286	0,112514286	...	0,032738093	0,056623805	0,058719042	0,049604756	0,056676
2	CO2	Agriculture	ARG	Argentina	0,10434285	0,10434285	0,10434285	0,10434285	0,087214278	0,077314278	...	0,999166539	1,145152229	0,892257036	1,359547443	1,278199
3	CO2	Agriculture	ARM	Armenia	0,055288203	0,055288203	0,055288203	0,055288203	0,059966435	0,059966435	...	0,021685714	0,022628571	0,022628571	0,022471428	0,034257
4	CO2	Agriculture	AUS	Australia	0,311142842	0,311142842	0,311142842	0,311142842	0,311142842	0,268190461	...	2,128866419	2,182923567	2,291771194	2,505223526	2,641204
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1462	CO2	Industrial Combustion	GLOBAL TOTAL	GLOBAL TOTAL	3744,304794	3511,720997	3602,041957	3788,025165	3774,962595	3648,618031	...	6358,823576	6425,251996	6286,63559	6130,934194	6067,820
1463	CO2	Power Industry	GLOBAL TOTAL	GLOBAL TOTAL	3823,699383	3910,981426	4189,105946	4524,711259	4603,893544	4695,749014	...	13626,19878	13686,2035	13387,05167	13441,57878	13754,01
1464	CO2	Processes	GLOBAL TOTAL	GLOBAL TOTAL	915,6702646	921,8790237	990,1614658	1030,394452	1007,746632	968,0634448	...	2823,739311	2886,012067	2857,742279	2961,481736	3024,544
1465	CO2	Transport	GLOBAL TOTAL	GLOBAL TOTAL	2796,286627	2876,504749	3045,881595	3221,973378	3191,503489	3275,104998	...	7373,902587	7497,303796	7732,101875	7879,176335	8078,337
1466	CO2	Waste	GLOBAL TOTAL	GLOBAL TOTAL	7,6053133	7,779282139	7,957964985	8,139190481	8,32005196	8,498052386	...	17,58225734	16,00245791	16,964262	17,13900045	16,98883

1467 rows × 57 columns

Now we will deal with missing values and cleaning in these datasets.

```
# Load the datasets
file_paths_CO2 = [
    'fossil_CO2_by_sector_country_su.csv',
    'fossil_CO2_per_capita_by_countr.csv',
    'fossil_CO2_per_GDP_by_country.csv',
    'fossil_CO2_totals_by_country.csv'
]

# Reading the data from each file
datasets_CO2 = [pd.read_csv(path, delimiter=';') for path in file_paths_CO2]
def convert_to_numeric(df):
    start_col = 4 if df.equals(datasets_CO2[0]) else 3
    for col in df.columns[start_col:]:
        df[col] = pd.to_numeric(df[col].str.replace(',', '.'), errors='coerce')
    return df

# Apply the conversion to all datasets
datasets_CO2 = [convert_to_numeric(df) for df in datasets_CO2]
```

We go through each number and change it to a format that allows us to perform mathematical operations, like adding or averaging. Any numbers that don't make sense or can't be converted (perhaps due to being written incorrectly in the original files) are noted as errors for now.

The last step is like a quality check. We apply our conversion to all datasets, ensuring that every number is ready for analysis.

```
# Deleting rows with NaN in the 'Sector' column from the first dataset
datasets_CO2[0].dropna(subset=['Sector'], inplace=True)

# Removing rows where 'Country' is "GLOBAL TOTAL" or "International shipping" from all datasets
countries_to_remove = ["GLOBAL TOTAL", "International Shipping","International Aviation","EU27"]
datasets_CO2 = [df[~df['Country'].isin(countries_to_remove)] for df in datasets_CO2]
# Calculating the number of missing values for each column in each dataset
missing_values_by_column = [{f"Dataset {i+1} - {col}": df[col].isnull().sum() for col in df.columns} for i, df in enumerate(datasets_CO2)]

missing_values_by_column
```

We want to understand the quality of our data before performing any analysis.

To do this, we iterate through each dataset and for each dataset, we count the number of missing values (NaN) in each column. We organize this information in a structured way, with labels indicating which dataset and column each count belongs to.

This helps us identify which columns might have a significant amount of missing data and may require special attention during our analysis.

```
# Replacing missing values with the mean of their respective column for each dataset
datasets_filled = [df.fillna(df.mean()) for df in datasets_CO2]
```

The last but important part is replacing values that are missing. Now we have clean datasets. Now we go to Exploratory Data Analysis.

```
# Descriptive statistics for each dataset
descriptive_stats = [df.describe() for df in datasets_filled]

# Displaying descriptive statistics
descriptive_stats[0]
```

	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	115.148079	115.558602	119.934932	125.615706	125.820040	126.245471	131.724566	134.938780	137.799739	141.088076	139.698347
std	465.431020	459.992077	480.038429	497.415410	489.447778	483.162252	506.143146	520.553233	527.228211	534.523319	521.597724
min	0.003798	0.003816	0.003355	0.003624	0.003801	0.003459	0.002984	0.004530	0.004006	0.004182	0.004416
25%	1.916147	1.901725	1.998970	2.051191	2.111660	2.153369	2.152902	2.236070	2.396665	2.447398	2.418233
50%	13.253005	13.340130	13.872984	14.055133	14.664751	14.625766	15.747926	16.143006	16.675062	16.405768	17.147582
75%	61.158763	62.099446	67.954679	69.630755	62.321724	68.354118	70.889297	71.520979	73.920943	73.692515	76.150963
max	5750.029633	5624.320655	5894.315890	6093.637117	5915.957818	5699.331436	6014.557728	6170.764571	6163.384588	6230.314310	5997.688955

1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
137.401509	136.579957	137.807751	141.617530	143.016805	145.189483	148.741286	153.223416	155.980084	156.786372	157.252465	156.722283
515.253823	503.192726	506.133102	525.679902	527.737970	531.897103	548.164790	568.879002	577.202215	576.632749	578.146064	581.401937
0.003266	0.003066	0.004220	0.003908	0.004221	0.004548	0.005032	0.006483	0.006553	0.007048	0.006707	0.006408
2.303035	2.322327	2.315208	2.466088	2.273849	2.304551	2.519709	2.524752	2.587932	2.581364	2.533689	2.581667
17.461849	17.430439	19.050473	19.960358	19.698381	20.507551	20.632531	22.187709	22.079739	21.568532	21.222688	20.715016
73.449748	74.308978	73.429487	75.819855	78.613547	78.715555	79.104497	79.432687	81.698496	85.341427	80.555274	79.529426
5917.983071	5649.111980	5618.754033	5845.645879	5854.256613	5807.734393	5967.301421	6204.137294	6270.475123	6163.741598	6115.600587	6215.020469

1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
157.248574	158.779846	162.901508	165.830457	167.923524	168.227346	169.174962	173.567758	175.288465	177.456316	184.864462	192.292303
590.003985	598.725286	618.441720	627.031785	641.318934	646.414485	646.355303	664.843964	668.949862	675.090025	711.106230	756.722128
0.006978	0.006825	0.007094	0.004714	0.005083	0.005077	0.005344	0.008666	0.008957	0.008977	0.009439	0.010764
2.666164	2.944619	2.970766	3.014901	3.407918	3.361892	3.352283	3.395249	3.403039	3.387964	3.460821	3.619254
20.215193	19.703701	20.487787	21.215961	21.732724	22.285026	21.475401	22.906581	22.738860	22.132043	22.609002	23.697083
79.499488	80.561973	82.633112	84.672295	85.081760	85.517776	84.830793	86.170994	87.656073	89.537718	96.160970	96.657137
6341.283283	6448.603540	6517.886171	6687.967131	6971.156595	7020.480586	7023.791489	7188.178567	7105.939400	6941.589640	7007.774015	7556.106084

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
10.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
98.499276	204.108745	210.209075	211.649532	209.903507	220.318180	227.054477	230.449682	234.302313	235.889189	235.092319
00.077659	837.875206	875.992457	882.215605	899.886555	960.712328	1015.729110	1032.599722	1069.675013	1082.334630	1069.302790
0.011396	0.010140	0.011661	0.012314	0.012533	0.013450	0.013986	0.014149	0.017046	0.017013	0.017133
3.759061	4.033137	4.345404	4.586828	4.598608	4.574225	4.908997	4.962135	5.013522	5.414774	5.276699
25.328487	24.614281	26.125557	25.217793	26.620850	28.436879	29.373424	30.166350	30.471654	30.884386	31.700618
96.545234	101.576755	102.418576	109.493327	100.442383	106.263625	107.706102	104.919647	106.576356	109.699591	113.008164
31.922006	9232.261373	9845.302236	10069.845690	10696.090420	11565.470010	12572.884150	12928.229140	13485.440500	13650.133430	13479.880370

2016	2017	2018	2019	2020	2021	2022
210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
235.899184	239.694667	245.336324	246.131314	238.754907	250.003703	252.929692
1064.771065	1081.992751	1124.926151	1140.316168	1139.239148	1198.280216	1206.608214
0.017902	0.018552	0.019561	0.019302	0.017970	0.018294	0.020334
5.253936	4.775778	4.998782	5.161539	5.105144	5.317732	5.405390
31.940118	34.278911	34.723999	34.588260	34.557860	35.438866	36.337149
112.916306	116.350988	117.899410	118.360883	113.254149	119.294986	116.488403
13447.136420	13710.100290	14296.571170	14606.127590	14879.556510	15632.894610	15684.626760

First we do the descriptive statistics for each datasets to understand better our data.

The total emissions data shows an increasing trend with high variability. The mean annual emissions are typically in the hundreds to thousands of metric tons range, with maximum values being extremely high, indicating major emitters.

- **GHG emissions:**

As the process is the very same for GHG datasets I will not go into details. Please refer to the code snippets here or the full code that you can find on GitHub.

Importation of the datasets "GHG"

```
ghg_sector=pd.read_csv('GHG_by_sector_and_country.csv',delimiter=";")
ghg_capita=pd.read_csv('GHG_per_capita_by_country.csv',delimiter=';')
ghg_gdp=pd.read_csv('GHG_per_GDP_by_country.csv',delimiter=';')
ghg_total=pd.read_csv('GHG_totals_by_country.csv',delimiter=';')
```
✓ 0.3s

An example, ghg_sectors:

```
ghg_sector
```
✓ 0.0s Python

	Substance	Sector	EDGAR Country Code	Country	1970	1971	1972	1973	1974	1975	...	2013	2014	2015	2016
0	CO2	Agriculture	AFG	Afghanistan	0,029228567	0,029228567	0,029228567	0,029228567	0,039966661	0,045309517	...	0,055157133	0,084490461	0,116966646	0,162799971
1	CO2	Agriculture	ALB	Albania	0,1133	0,1133	0,1133	0,1133	0,113614286	0,112514286	...	0,032738093	0,056623805	0,058719042	0,049604756
2	CO2	Agriculture	ARG	Argentina	0,10434285	0,10434285	0,10434285	0,10434285	0,087214278	0,077314278	...	0,999166539	1,145152229	0,892257036	1,359547443
3	CO2	Agriculture	ARM	Armenia	0,055288203	0,055288203	0,055288203	0,055288203	0,059966435	0,059966435	...	0,021685714	0,022628571	0,022628571	0,022471428
4	CO2	Agriculture	AUS	Australia	0,311142842	0,311142842	0,311142842	0,311142842	0,311142842	0,268190461	...	2,128866419	2,182923567	2,291771194	2,505223526
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4825	GWP_100_AR5_N2O	Industrial Combustion	GLOBAL TOTAL	GLOBAL TOTAL	12,39798771	11,58425896	11,87122101	12,28790018	12,35481257	12,25050671	...	26,16643323	26,218647	25,89538635	25,36552432
4826	GWP_100_AR5_N2O	Power Industry	GLOBAL TOTAL	GLOBAL TOTAL	11,10027029	11,24027815	11,85362348	12,61078462	12,75119864	13,21432186	...	72,9179465	74,19085611	71,79233571	72,93407658
4827	GWP_100_AR5_N2O	Processes	GLOBAL TOTAL	GLOBAL TOTAL	412,8907148	418,0588772	429,1064357	456,8013894	471,3035986	455,4526854	...	376,2542516	377,5672528	376,2034825	376,5507517
4828	GWP_100_AR5_N2O	Transport	GLOBAL TOTAL	GLOBAL TOTAL	34,18007147	34,79176768	37,09558497	39,28263414	39,57646006	40,01327091	...	97,81854719	99,39652114	101,0189339	102,3744487
4829	GWP_100_AR5_N2O	Waste	GLOBAL TOTAL	GLOBAL TOTAL	43,72323944	44,41630687	45,0691399	46,14326618	46,9732261	48,02445676	...	108,4414333	110,2666779	112,4801215	114,776275

4830 rows × 57 columns

Descriptive statistics:

	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	115.148079	115.558602	119.934932	125.615706	125.820040	126.245471	131.724566	134.938780	137.799739	141.088076	139.698347
std	465.431020	459.992077	480.038429	497.415410	489.447778	483.162252	506.143146	520.553233	527.228211	534.523319	521.597724
min	0.003798	0.003816	0.003355	0.003624	0.003801	0.003459	0.002984	0.004530	0.004006	0.004182	0.004416
25%	1.916147	1.901725	1.998970	2.051191	2.111660	2.153369	2.152902	2.236070	2.396665	2.447398	2.418233
50%	13.253005	13.340130	13.872984	14.055133	14.664751	14.625766	15.747926	16.143006	16.675062	16.405768	17.147582
75%	61.158763	62.099446	67.954679	69.630755	62.321724	68.354118	70.889297	71.520979	73.920943	73.692515	76.150963
max	5750.029633	5624.320655	5894.315890	6093.637117	5915.957818	5699.331436	6014.557728	6170.764571	6163.384588	6230.314310	5997.688955

1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
137.401509	136.579957	137.807751	141.617530	143.016805	145.189483	148.741286	153.223416	155.980084	156.786372	157.252465	156.722283
515.253823	503.192726	506.133102	525.679902	527.737970	531.897103	548.164790	568.879002	577.202215	576.632749	578.146064	581.401937
0.003266	0.003066	0.004220	0.003908	0.004221	0.004548	0.005032	0.006483	0.006553	0.007048	0.006707	0.006408
2.303035	2.322327	2.315208	2.466088	2.273849	2.304551	2.519709	2.524752	2.587932	2.581364	2.533689	2.581667
17.461849	17.430439	19.050473	19.960358	19.698381	20.507551	20.632531	22.187709	22.079739	21.568532	21.222688	20.715016
73.449748	74.308978	73.429487	75.819855	78.613547	78.715555	79.104497	79.432687	81.698496	85.341427	80.555274	79.529426
5917.983071	5649.111980	5618.754033	5845.645879	5854.256613	5807.734393	5967.301421	6204.137294	6270.475123	6163.741598	6115.600587	6215.020469

1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
157.248574	158.779846	162.901508	165.830457	167.923524	168.227346	169.174962	173.567758	175.288465	177.456316	184.864462	192.292303
590.003985	598.725286	618.441720	627.031785	641.318934	646.414485	646.355303	664.843964	668.949862	675.090025	711.106230	756.722128
0.006978	0.006825	0.007094	0.004714	0.005083	0.005077	0.005344	0.008666	0.008957	0.008977	0.009439	0.010764
2.666164	2.944619	2.970766	3.014901	3.407918	3.361892	3.352283	3.395249	3.403039	3.387964	3.460821	3.619254
20.215193	19.703701	20.487787	21.215961	21.732724	22.285026	21.475401	22.906581	22.738860	22.132043	22.609002	23.697083
79.499488	80.561973	82.633112	84.672295	85.081760	85.517776	84.830793	86.170994	87.656073	89.537718	96.160970	96.657137
6341.283283	6448.603540	6517.886171	6687.967131	6971.156595	7020.480586	7023.791489	7188.178567	7105.939400	6941.589640	7007.774015	7556.106084

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
198.499276	204.108745	210.209075	211.649532	209.903507	220.318180	227.054477	230.449682	234.302313	235.889189	235.092319
800.077659	837.875206	875.992457	882.215605	899.886555	960.712328	1015.729110	1032.599722	1069.675013	1082.334630	1069.302790
0.011396	0.010140	0.011661	0.012314	0.012533	0.013450	0.013986	0.014149	0.017046	0.017013	0.017133
3.759061	4.033137	4.345404	4.586828	4.598608	4.574225	4.908997	4.962135	5.013522	5.414774	5.276699
25.328487	24.614281	26.125557	25.217793	26.620850	28.436879	29.373424	30.166350	30.471654	30.884386	31.700618
96.545234	101.576755	102.418576	109.493327	100.442383	106.263625	107.706102	104.919647	106.576356	109.699591	113.008164
8431.922006	9232.261373	9845.302236	10069.845690	10696.090420	11565.470010	12572.884150	12928.229140	13485.440500	13650.133430	13479.880370

2016	2017	2018	2019	2020	2021	2022
210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
235.899184	239.694667	245.336324	246.131314	238.754907	250.003703	252.929692
1064.771065	1081.992751	1124.926151	1140.316168	1139.239148	1198.280216	1206.608214
0.017902	0.018552	0.019561	0.019302	0.017970	0.018294	0.020334
5.253936	4.775778	4.998782	5.161539	5.105144	5.317732	5.405390
31.940118	34.278911	34.723999	34.588260	34.557860	35.438866	36.337149
112.916306	116.350988	117.899410	118.360883	113.254149	119.294986	116.488403
13447.136420	13710.100290	14296.571170	14606.127590	14879.556510	15632.894610	15684.626760

The mean emission values exhibit a notable upward trajectory, indicative of a consistent increase in GHG output over the years.

This rise is paralleled by the standard deviation, which suggests a widening dispersion in emission quantities, pointing towards a growing heterogeneity in the data. The minimum and maximum values highlight the extremities of the data, with the maximum showing a significant escalation, reflecting the increased capacity of the highest emitters.

The quartile values, particularly the median (50th percentile), further confirm the central tendency towards higher emissions, while the interquartile range (IQR) between the 25th and 75th percentiles provide insights into the central concentration of the data.

These statistics collectively underscore the critical need for environmental scrutiny and targeted policy interventions to address the escalating challenge of GHG emissions.

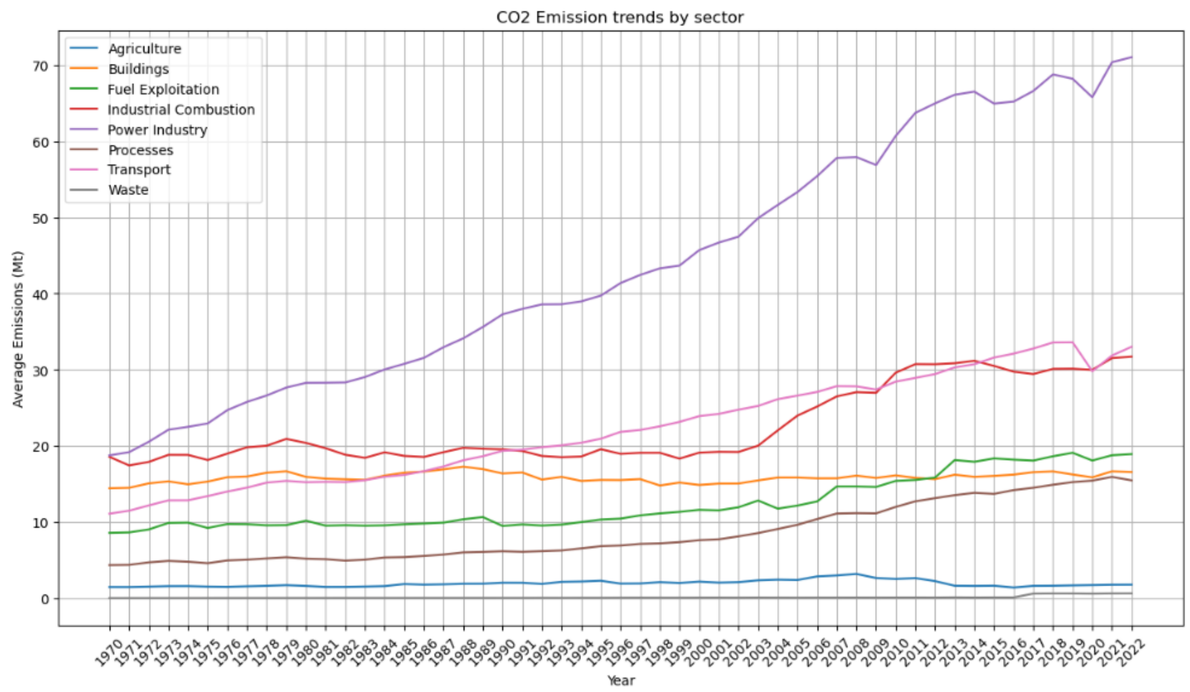
2) Emission trends by sector:

- **CO2:**

CO2 Emission trends by sector

```
# Exploring the dataset for emissions by sector

sectors = datasets_filled[0]['Sector'].unique()

# Function to plot sector-wise trends over time
def plot_sector_trends(df, sectors):
    plt.figure(figsize=(15, 8))

    for sector in sectors:
        sector_df = df[df['Sector'] == sector]
        mean_values = sector_df.iloc[:, 4:].mean()
        plt.plot(mean_values, label=sector)

    plt.title("CO2 Emission trends by sector")
    plt.xlabel("Year")
    plt.ylabel("Average Emissions (Mt)")
    plt.xticks(rotation=45)
    plt.legend()
    plt.grid(True)
    plt.show()

# Plotting sector-wise trends
plot_sector_trends(datasets_filled[0], sectors)
```

We will here display the CO2 emissions trends over the years by sector.



The Power Industry sector is the most significant contributor to CO2 emissions, showing a steep and consistent increase over the decades. This sector far outpaces the others, suggesting that electricity and heat production remain heavily reliant on carbon-intensive sources.

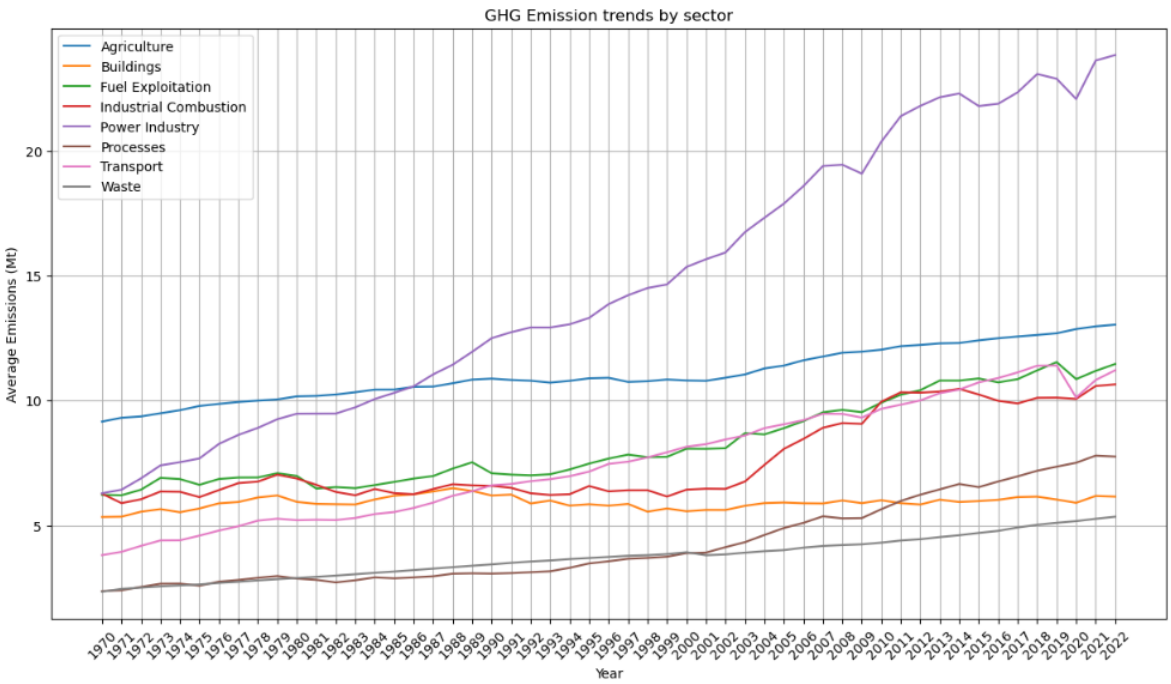
The Transport sector also demonstrates a notable upward trend, reflecting the growing demand for vehicular travel and freight transport.

Agriculture, Industrial Combustion, and Buildings show moderate yet steady increases in emissions over time, indicating a persistent reliance on fossil fuels in these sectors as well. Fuel Exploitation, Processes, and Waste have relatively lower and flatter trends, suggesting a more stable emission profile, but these sectors still contribute to the overall CO2 footprint.

The trajectory of these lines not only underscores the challenges in mitigating CO2 emissions but also highlights the critical sectors where policy and innovation must focus to achieve significant reductions. The data from the graph sends a compelling message for the urgent transformation of energy systems, particularly within the Power Industry and Transport sectors, to address the escalating climate crisis.

- **GHG:**

The code itself is very similar to the CO2 one.



The Power Industry sector emerges as the leading source of GHG emissions, with a pronounced and steady growth, underscoring its significant role in global GHG output. This sector's emissions rise markedly from the 1970s, reflecting the expanding dependence on energy production that is not yet fully sustainable or renewable.

Agriculture maintains the second-highest level of emissions throughout the timeline, indicative of the substantial environmental impact of farming practices and livestock management. The steady growth in this sector suggests an increasing contribution to the global GHG footprint, possibly due to intensified agricultural activities to meet the demands of a growing population.

Other sectors like Transport, Industrial Combustion, and Buildings exhibit progressive increases, although at a more gradual pace compared to the Power Industry. This indicates a steady rise in emissions associated with industrial activities, transportation, and energy use in buildings.

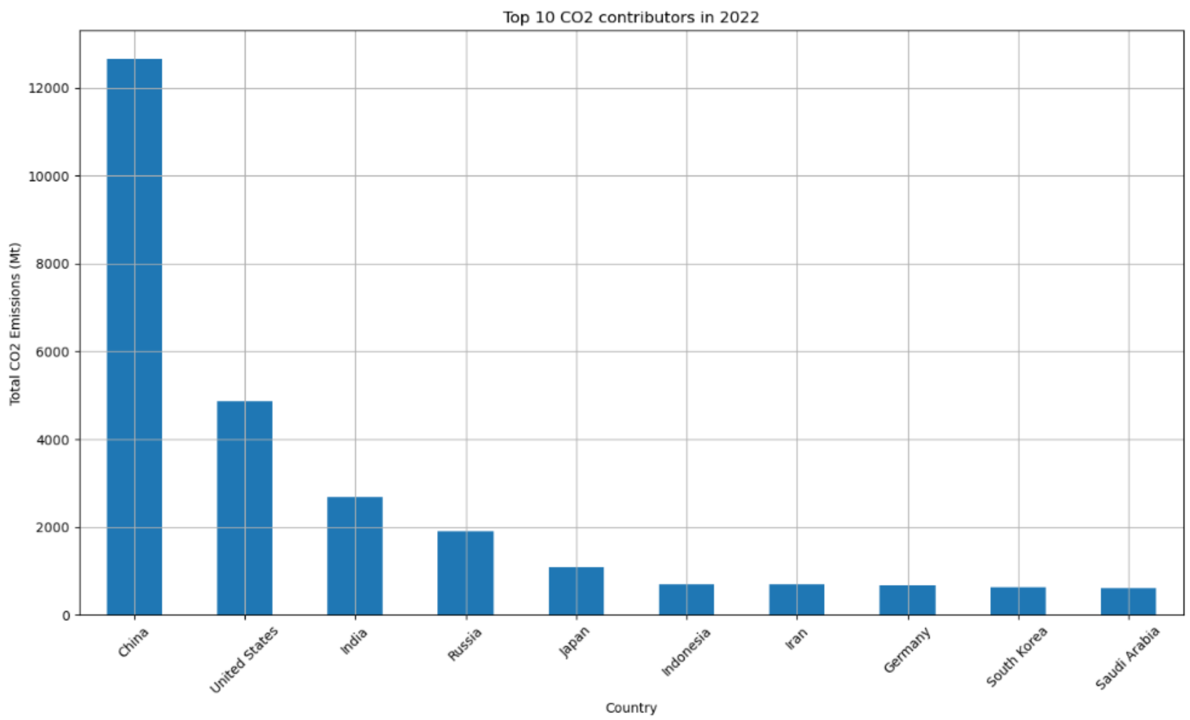
Conversely, Processes, Fuel Exploitation, and Waste demonstrate relatively lower emission levels, with less pronounced increases over time. These sectors appear to have a more stable emission pattern, possibly reflecting the effectiveness of waste management improvements and efficiency gains in fuel processing and industrial operations.

3) Top 10 contributors:

Now we will look at the top 10 countries that contributed to these emissions, for both CO2 and GHG.

- **CO2:**

```
latest_year_column = datasets_filled[3].columns[-1]

# Grouping by country and summing emissions for the latest year
country_emissions_latest = datasets_filled[3].groupby('Country')[latest_year_column].sum()

# Sorting the emissions in descending order
sorted_emissions_latest = country_emissions_latest.sort_values(ascending=False)

# Taking the top 10 countries for clarity in the plot
top_countries_emissions = sorted_emissions_latest.head(10)

# Plotting
plt.figure(figsize=(15, 8))
top_countries_emissions.plot(kind='bar')
plt.title(f"Top 10 CO2 contributors in {latest_year_column}")
plt.xlabel("Country")
plt.ylabel("Total CO2 Emissions (Mt)")
plt.xticks(rotation=45)
plt.grid(True)
plt.show()
```



This bar graph delineates the total CO2 emissions for the year 2022 by the top 10 contributing countries, offering a stark visualization of the unequal distribution of emissions globally.

China leads by a significant margin, with its emissions towering over all other countries, a reflection of its vast industrial base and energy production predominantly reliant on coal.

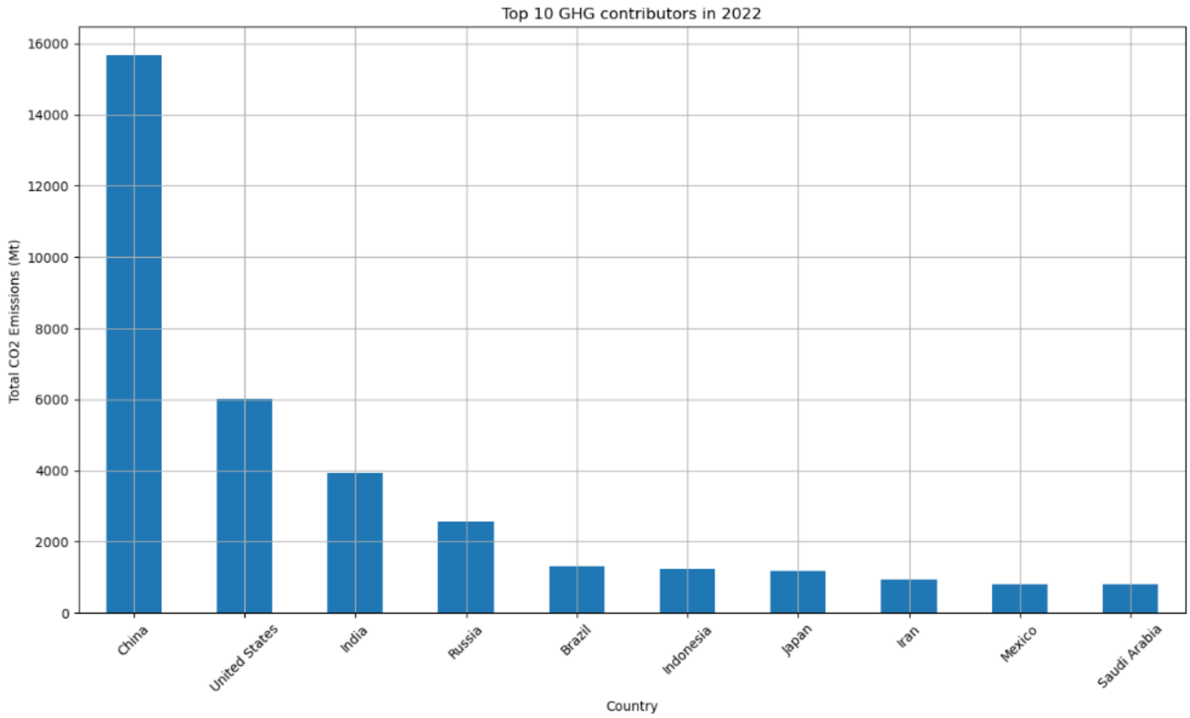
The United States follows as the second-largest emitter, with nearly half the emissions of China, indicating a high per capita emission rate given its size and economy.

India, Russia, and Japan form the middle tier, each contributing substantial but markedly less than the top two emitters.

The lower tier comprises Indonesia, Iran, Germany, South Korea, and Saudi Arabia, each showing considerable emissions but on a comparatively smaller scale.

This distribution underscores the disparate contributions to global CO2 emissions and highlights the necessity for tailored climate policies that consider both the absolute and per capita emissions of each country.

- **GHG:**



The bar chart presents the total greenhouse gas (GHG) emissions in 2022 for the top 10 contributing nations, illustrating a pronounced disparity among countries. China is depicted as the predominant GHG emitter, far exceeding the output of other countries, which reflects its industrial scale and extensive use of fossil fuels. The United States is positioned as the second-largest contributor, with emissions significantly lower than China's but still substantial in the global context. India and Russia follow, with contributions that are notable but not as extensive, highlighting their role in global emissions.

4) Rank of the contributors by continent

I decided to add a file to match all countries with their continent.

The csv file is downloadable here:

https://worldpopulationreview.com/country-rankings/list-of-countries-by-continent

This file is called list-of-countries-by-continent-2023.csv and allows us to map all countries back to their continent/region.

```
# Function to load dataset and get unique countries
def get_unique_countries(file_path):
    dataset = pd.read_csv(file_path, delimiter=';')
    return set(dataset['Country'].unique())

# Getting unique countries from each dataset
unique_countries = {name: get_unique_countries(path) for name, path in {**dataset_file_paths_1970, **dataset_file_paths_1990}.items()}

# Identifying countries not present in all datasets
all_countries = set()
for countries in unique_countries.values():
    all_countries |= countries  # Union of all unique countries

# Load the provided list of countries by continent
file_path_countries_continents = 'list-of-countries-by-continent-2023.csv'
countries_continents = pd.read_csv(file_path_countries_continents)

# Creating a new country-to-continent mapping using the provided dataset
country_continent_map = dict(zip(countries_continents['country'], countries_continents['region']))
```

We had to hand map some of them that were not captured in the file.

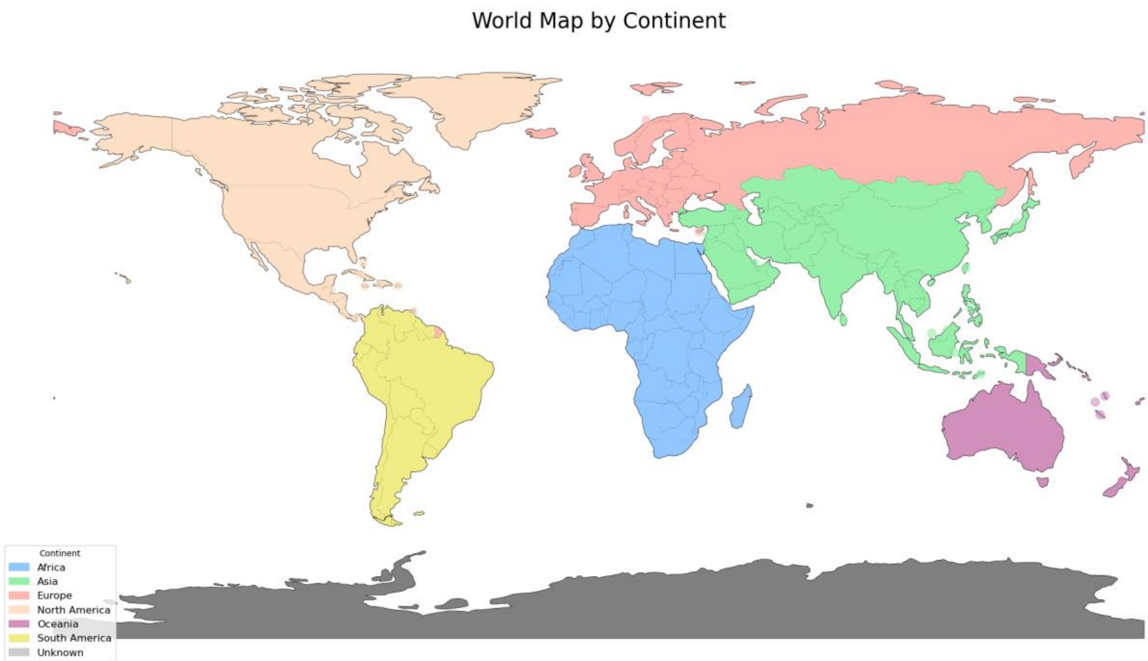
```
# Manual assignments for specific countries
additional_mappings = {
    'Curaçao': 'North America',
    'Switzerland and Liechtenstein': 'Europe',
    'Côte d'Ivoire': 'Africa',
    'Democratic Republic of the Congo': 'Africa',
    'Congo': 'Africa',
    'Cabo Verde': 'Africa',
    'Czechia': 'Europe',
    'Spain and Andorra': 'Europe',
    'France and Monaco': 'Europe',
    'Faroes': 'Europe',
    'The Gambia': 'Africa',
    'Israel and Palestine. State of': 'Asia',
    'Italy. San Marino and the Holy See': 'Europe',
    'Macao': 'Asia',
    'Myanmar/Burma': 'Asia',
    'Réunion': 'Africa',
    'Serbia and Montenegro': 'Europe',
    'Sudan and South Sudan': 'Africa',
    'Saint Helena. Ascension and Tristan da Cunha': 'Africa',
    'São Tomé and Príncipe': 'Africa',
    'Türkiye': 'Asia'  # Turkey is mostly in Asia
}
```

We also use geopandas to verify the world map is fully colored and all countries are in the right region:

```python
# Load the world map
world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))

continent_colors = {
    'Africa': '#92c6ff',  # pastel blue
    'Asia': '#97f0aa',  # pastel green
    'Europe': '#ffb7b2',  # pastel red
    'North America': '#fde0c5',  # pastel orange
    'Oceania': '#d291bc',  # pastel purple
    'South America': '#f0ec86',  # pastel yellow
    'Unknown': '#cccccc',  # light gray for unknown or unattributed
}


# Plotting each country on the map
fig, ax = plt.subplots(1, 1, figsize=(20, 12))
base = world.plot(ax=ax, color='white', edgecolor='black')

# Add a legend for continents
legend_handles = [mpatches.Patch(color=color, label=continent) for continent, color in continent_colors.items()]
ax.legend(handles=legend_handles, title='Continent', loc='lower left', fontsize='large')

# Enhancing the map
ax.set_title('World Map by Continent', fontsize=25, pad=20)
ax.set_axis_off()

plt.tight_layout()
plt.show()
```

Python

Here the only exception is French Guyana that shows in south America, of course, but its emission is in France as it is part of it.



- **CO2:**

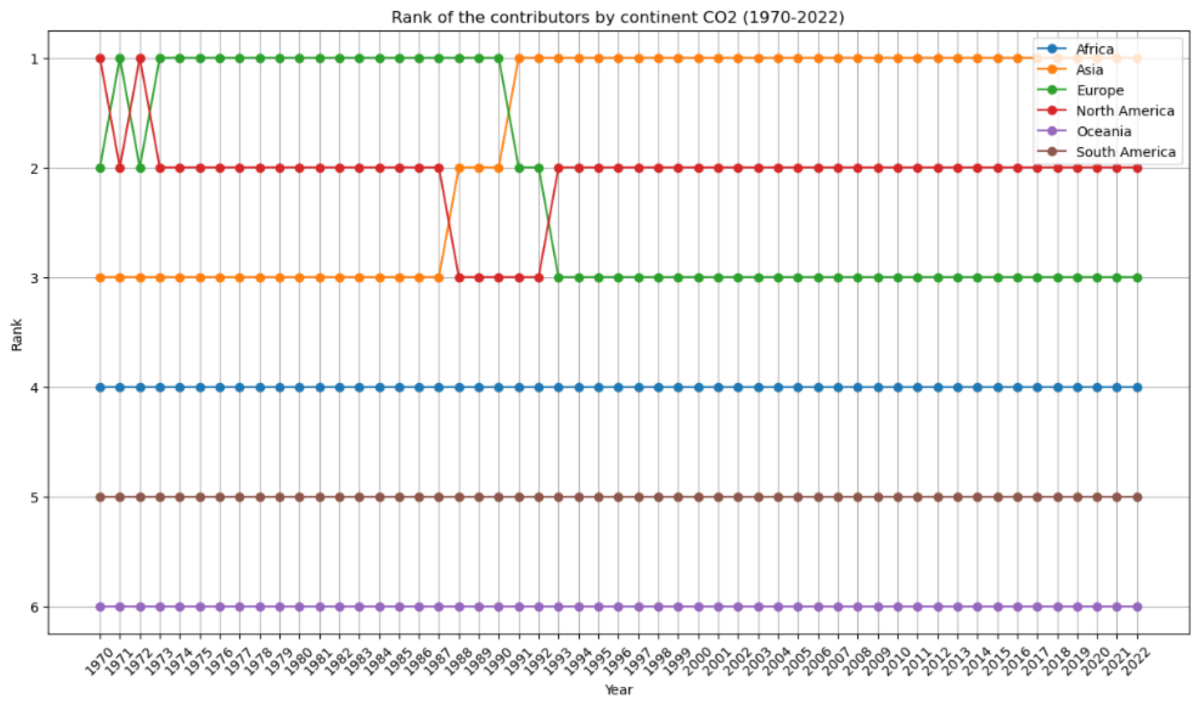
Thanks to this new file and matching, we can study the contributions by continent.

Rank of the contributors by continent

```
country_continent_df = continent[['country', 'region']]

country_continent_df.rename(columns={'country': 'Country', 'region': 'Continent'}, inplace=True)

merged_df = pd.merge(datasets_filled[3], country_continent_df, on='Country', how='left')

continent_emissions = merged_df.groupby('Continent').sum()
continent_emissions.head()
```

Here we aggregate the data by continent, to then be able to study how the ranking evolved over the years:

```
# Function to plot the rankings of continents in CO2 emissions over the years
def plot_continent_emissions_rankings(emissions_df):
    plt.figure(figsize=(15, 8))

    years = emissions_df.columns
    ranks = emissions_df.rank(ascending=False).T  # Transpose to get years as rows for plotting

    for continent in emissions_df.index:
        plt.plot(years, ranks[continent], label=continent, marker='o')

    plt.title("Rank of the contributors by continent CO2 (1970-2022)")
    plt.xlabel("Year")
    plt.ylabel("Rank")
    plt.gca().invert_yaxis()  # Invert y-axis to show the top rank at the top
    plt.legend()
    plt.xticks(rotation=45)
    plt.grid(True)
    plt.show()

# Plotting the rankings
plot_continent_emissions_rankings(continent_emissions)
```



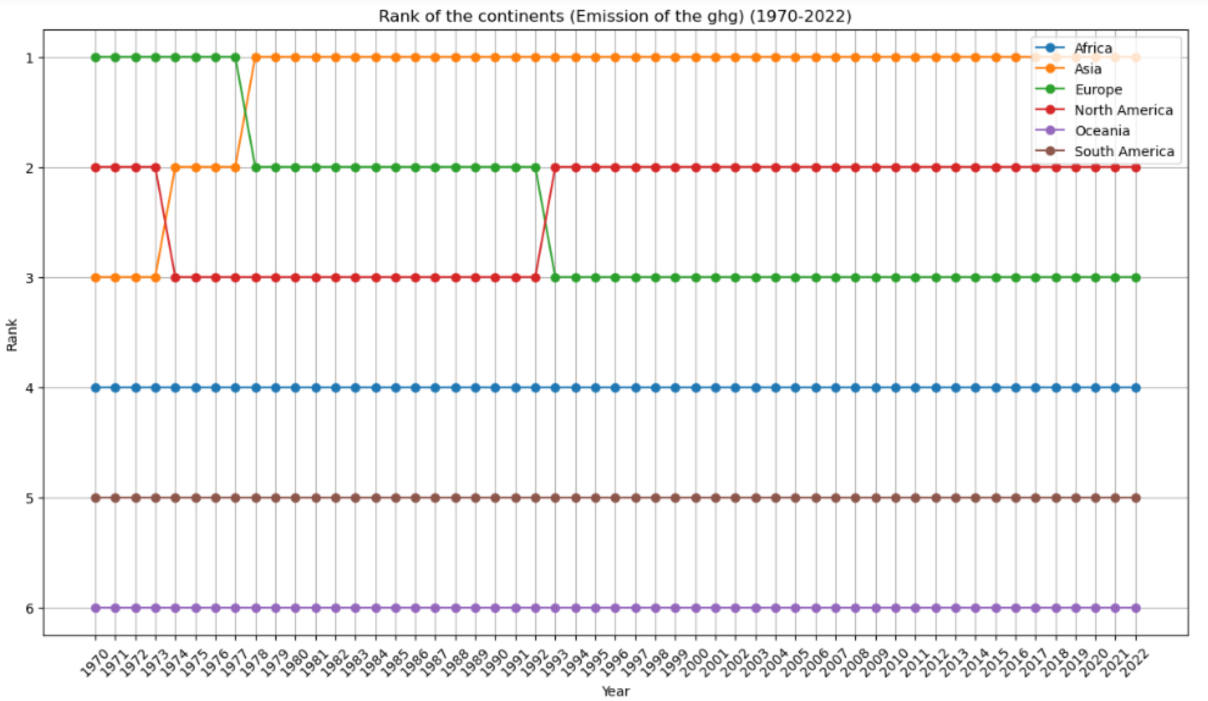
The line graph displays the ranking of continents by their CO2 emissions from 1970 to 2022, illustrating a clear hierarchy in the contributors to global CO2 output.

Asia holds the top rank consistently, which can be attributed largely to the rapid industrialization and economic growth of countries like China and India. North America, primarily driven by the United States, also shows a high rank, reflecting its significant industrial activity and energy consumption. Europe's ranking fluctuates but remains in the top three, indicating its historical industrialization and current energy use patterns.

Africa, South America, and Oceania maintain lower ranks, with Africa and South America showing occasional shifts in their positions but generally remaining at the bottom of the ranking. This suggests a smaller contribution to global CO2 emissions, which correlates with their smaller industrial bases and, for some countries in these continents, their developing economic status. Oceania's position is consistently the lowest, aligning with its smaller economic footprint compared to other continents.

- **GHG:**

Now we do the same graph on overall GHG. The process itself stays the same.



The provided line graph portrays the ranking of continents by greenhouse gas (GHG) emissions from 1970 to 2022, offering a longitudinal perspective on their contributions to global emissions.

Asia maintains the highest rank throughout the timeline, reflecting the substantial emissions from rapid industrialization and population growth. Europe and North America fluctuate between second and third places, indicative of their developed economies and substantial energy consumption.

Notably, there are instances of rank interchanges between Europe and North America, likely reflecting policy changes, economic shifts, or advancements in green technology.

The lower ranks are consistently held by Africa, Oceania, and South America, with Africa occasionally outpacing South America, which aligns with their relatively smaller industrial activities and economic outputs.

The graph highlights the consistent pattern of emissions over the last half-century and emphasizes the disparities in GHG emissions across continents, with the developed world and rapidly developing regions contributing the most.

Something very interesting to see is the rise of Asia: on overall GHG it happens almost 15 years before the CO2 rankings. The earlier rise of Asia's GHG emissions ranking compared to its CO2 emissions ranking can be attributed to the broader scope of GHGs, which include not only CO2 but also other gases like methane (CH4), nitrous oxide (N2O), and fluorinated gases. These gases can come from a variety of sources, not limited to the combustion of fossil fuels which predominantly contributes to CO2 emissions,

Asia has a vast agricultural sector, which can emit significant amounts of methane and nitrous oxide. As countries like China and India intensified their agricultural output to feed large populations, the emissions from this sector would have contributed to the rise in GHG rankings earlier than CO2.

Many GHGs are byproducts of industrial processes other than those that emit CO2. As Asian economies began to grow and industrialize, the expansion of these industries would contribute to a rise in GHGs.

Inadequate waste management, particularly in rapidly urbanizing areas in Asia, can lead to increased methane production, which would affect GHG rankings but not CO2 rankings.

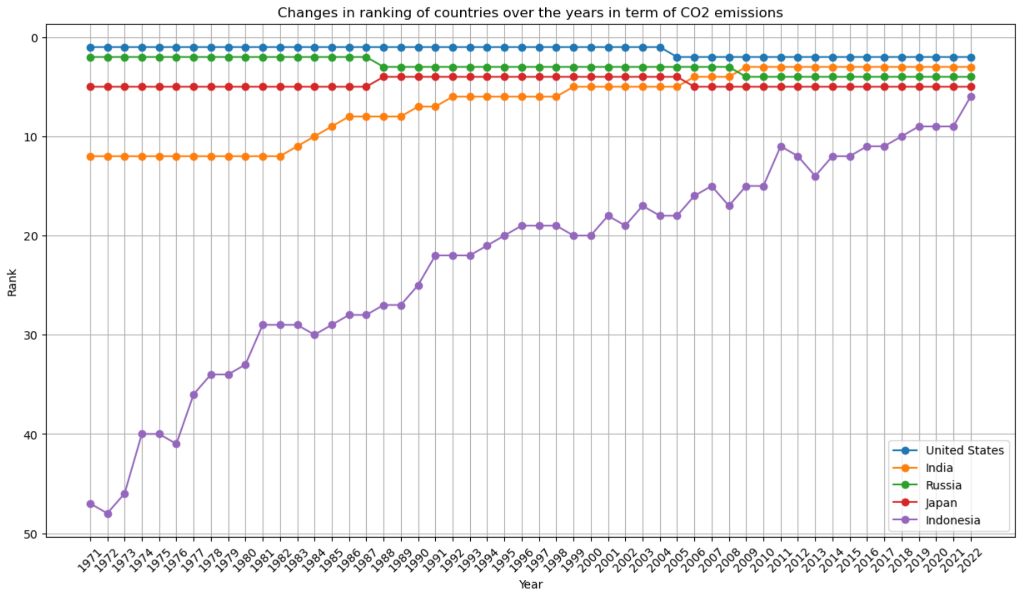
The differential between the rise in rankings of GHG and CO2 emissions for Asia reflects the complexity of emissions sources and the variety of activities contributing to overall greenhouse gas emissions. It underlines the importance of considering all types of emissions when forming environmental policies and not solely focusing on CO2, despite it being the most prominent greenhouse gas.

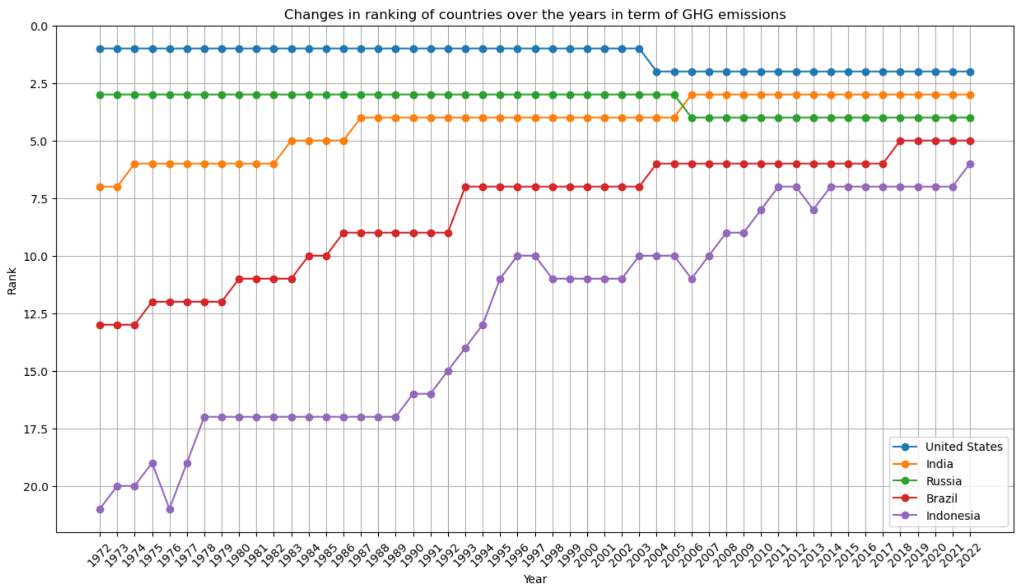
After this very interesting part, we will look at the top 5 countries for both CO2 and GHG over the years:

Changes in ranking of countries over the years in term of CO2 emissions

```python
def calculate_rankings(dataset, year_columns):
    rankings = {}
    for year in year_columns:
        year_data = dataset[['Country', year]].copy()
        year_data.sort_values(by=year, ascending=False, inplace=True)
        year_data['Rank'] = np.arange(1, len(year_data) + 1)
        rankings[year] = year_data[['Country', 'Rank']].set_index('Country')
    return rankings

total_emissions_rankings = calculate_rankings(datasets_filled[3], datasets_filled[3].columns[4:])
```

```python
# Revised function to plot the rankings of countries over years, handling missing data
def plot_country_rankings_revised(rankings, title, top_n=5):
    plt.figure(figsize=(15, 8))

    top_countries = rankings[latest_year].index[1:top_n+1]  # Skipping the first entry which is 'GLOBAL TOTAL'

    for country in top_countries:
        ranks_over_years = [rankings[year].loc[country, 'Rank'] if country in rankings[year].index else np.nan for year in year_columns]
        plt.plot(year_columns, ranks_over_years, label=country, marker='o')

    plt.title(title)
    plt.xlabel("Year")
    plt.ylabel("Rank")
    plt.gca().invert_yaxis()
    plt.legend()
    plt.xticks(rotation=45)
    plt.grid(True)
    plt.show()
latest_year = datasets_filled[3].columns[-1]
year_columns = datasets_filled[3].columns[4:]

plot_country_rankings_revised(total_emissions_rankings, "Changes in ranking of countries over the years in term of CO2 emissions", top_n=5)
```





Here is an overall comparison of those two:

- Both graphs show a general trend of increasing rank (where a lower numerical rank indicates a higher position as an emitter) for developing countries like India and Indonesia, which aligns with their economic growth trajectories.
- The difference in the trends between CO2 and GHG rankings for countries like Brazil and Indonesia suggests significant contributions from non-CO2 GHGs, possibly from land use, agriculture, and deforestation activities.
- The relatively stable or declining CO2 rankings for countries like the United States and Japan could indicate successful mitigation efforts, such as cleaner energy sources and efficiency improvements, which may not be as evident in the overall GHG trends.
- India's earlier and more pronounced rise in GHG rankings compared to CO2 could imply substantial emissions from sectors other than those primarily contributing to CO2, such as agriculture emitting methane.

5) Insights from the exploratory data analysis:

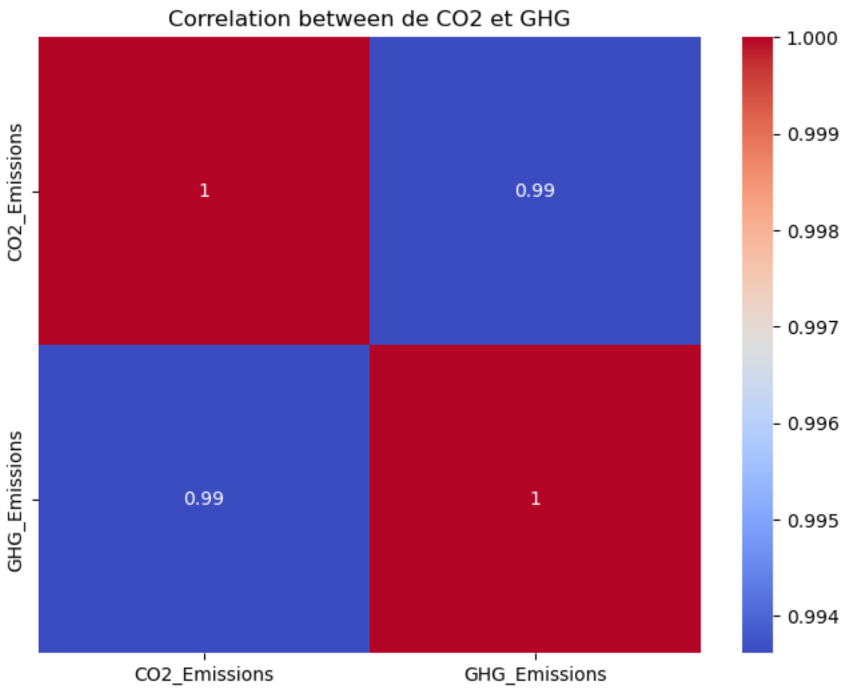
1. **Disparity in Global Emissions**: There is a clear global disparity in CO2 and GHG emissions, with Asia, led by China, and North America, particularly the United States, being the most significant contributors. This highlights the need for targeted emission reduction policies in these regions.
2. **Sector-Specific Emission Trends**: The Power Industry and Transport sectors show the most substantial increase in emissions over time, suggesting a critical focus area for reducing global CO2 and GHG emissions. Agriculture also remains a significant contributor, emphasizing the importance of sustainable practices in this sector.
3. **Increasing Emissions Over Time**: Both CO2 and GHG emissions have shown an overall increase over the observed period. The upward trend in emissions indicates that current efforts to reduce emissions are not yet sufficient to reverse this trend.
4. **Continental Contributions to Emissions**: The rank of continents by emissions reveals a persistent pattern where developed or rapidly industrializing continents contribute more significantly to global emissions compared to less industrialized continents, such as Africa and South America.
5. **Top Emitting Countries**: The top emitting countries list has remained relatively consistent with countries like China, the USA, and India leading. The significant difference between the top emitter and the rest underscores the role that national policies and measures can play in global emission reduction efforts.
6. **Rank Shifts Over Time**: Slight fluctuations in the ranks of continents over the years indicate changes in economic, policy, and technological landscapes, such as the adoption of renewable energy sources and improvements in energy efficiency.

PART II: Correlation Analysis:

1) Correlation between CO_2 and GHG emissions:

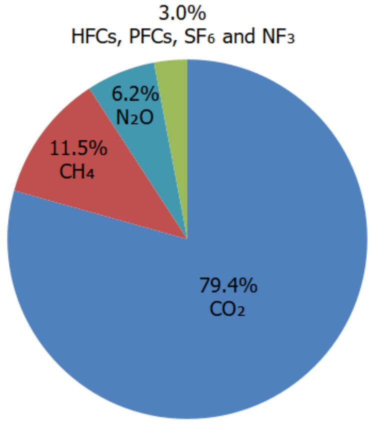
We start the correlation part with a quite simple one: GHG and CO2. As CO2 is a part of , if not the main part of GHG in a lot of contexts, it should be quite obvious they will be correlated.

```
co2_emissions_df = datasets_filled[3]
ghg_emissions_df = datasets_filled_ghg[3]

# Melting the CO2 emissions dataset to long format correctly
co2_long_df = co2_emissions_df.melt(id_vars=['EDGAR Country Code', 'Country'],
                                    var_name='Year',
                                    value_name='CO2_Emissions')

# Melting the GHG emissions dataset to long format correctly
ghg_long_df = ghg_emissions_df.melt(id_vars=['EDGAR Country Code', 'Country'],
                                    var_name='Year',
                                    value_name='GHG_Emissions')
# Merging the datasets
merged_emissions_df = pd.merge(co2_long_df, ghg_long_df, on=['Country', 'Year'])

# Dropping rows with NaN values
merged_emissions_df.dropna(inplace=True)

# Calculating correlation between CO2 and GHG emissions
correlation_matrix = merged_emissions_df[['CO2_Emissions', 'GHG_Emissions']].corr()


# Plot the correlation
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation between de CO2 et GHG")
plt.show()
```



The heatmap indicates a strong positive correlation between CO2 emissions and overall greenhouse gas (GHG) emissions, with a correlation coefficient of 0.99.

This suggests that in the dataset analyzed, as CO2 emissions increase or decrease, GHG emissions tend to follow very closely. Given that CO2 is a major component of GHG emissions, this high correlation is expected.



U.S. Environmental Protection Agency (2023). Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2021

This strong interdependency signals that measures aimed at reducing CO2 emissions are likely to have a directly proportional effect on the reduction of total GHG emissions, thereby reinforcing the importance of CO2 mitigation strategies in the broader context of combating climate change.

2) Correlation between sectors:

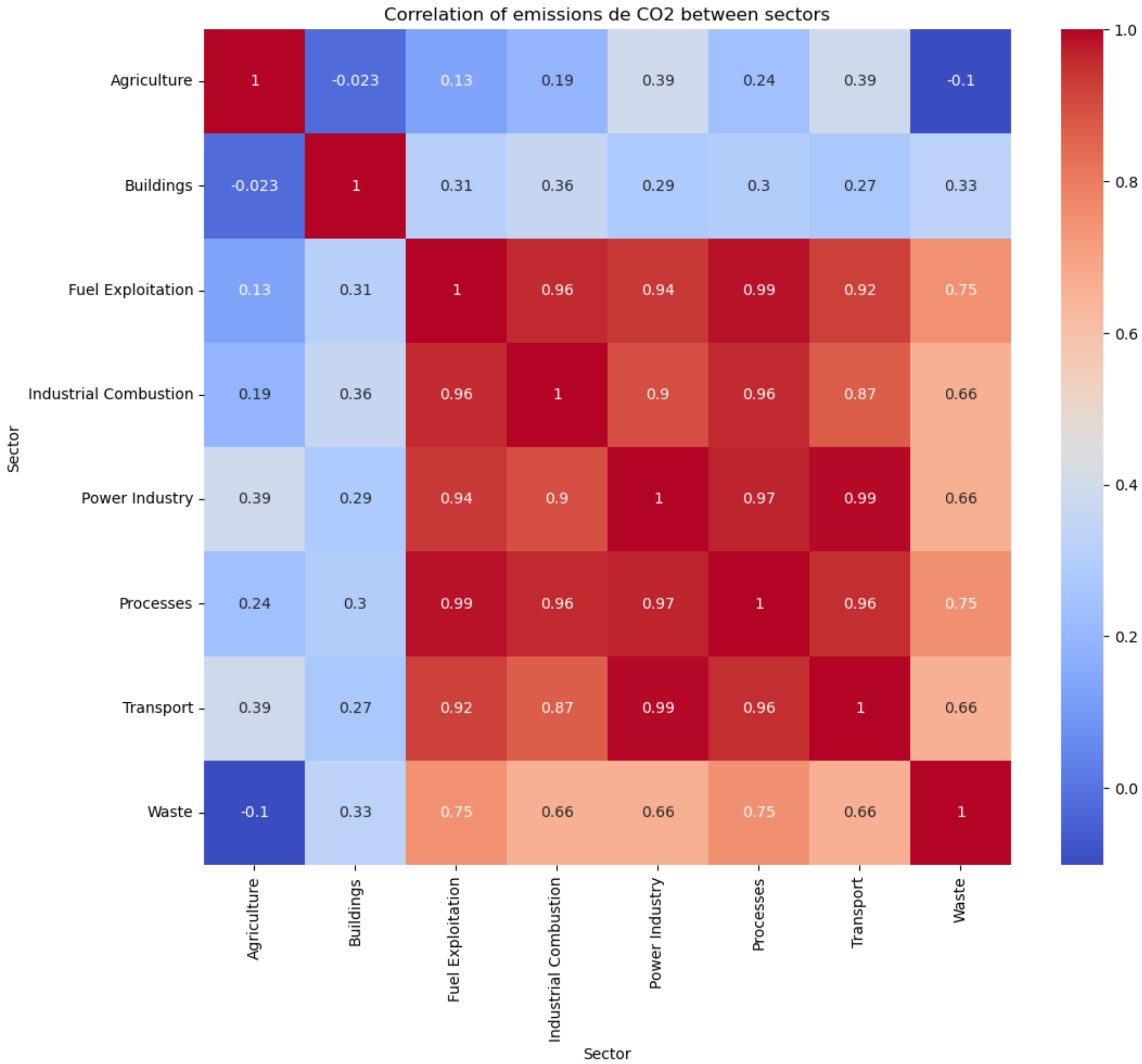
- **CO2:**

We will now do the correlation between sectors.

```
# Grouping by sector and summing emissions for each year
sector_grouped_co2_df = datasets_filled[0].groupby('Sector').sum()
sector_grouped_co2_df=sector_grouped_co2_df.T
# Calculating the correlation matrix across sectors for each year
sector_correlation_matrix = sector_grouped_co2_df.corr()

# Plotting the correlation matrix
plt.figure(figsize=(12, 10))
sns.heatmap(sector_correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation of emissions de CO2 between sectors")
plt.show()

# Displaying the correlation matrix
sector_correlation_matrix
```

Sector	Agriculture	Buildings	Fuel Exploitation	Industrial Combustion	Power Industry	Processes	Transport	Waste
Sector								
Agriculture	1.000000	0.388617	0.955550	0.905505	0.980431	0.959890	0.968542	0.983037
Buildings	0.388617	1.000000	0.241425	0.290622	0.257482	0.238415	0.243654	0.294276
Fuel Exploitation	0.955550	0.241425	1.000000	0.955986	0.969261	0.991226	0.958052	0.945598
Industrial Combustion	0.905505	0.290622	0.955986	1.000000	0.897355	0.953115	0.863743	0.852111
Power Industry	0.980431	0.257482	0.969261	0.897355	1.000000	0.962673	0.994500	0.982383
Processes	0.959890	0.238415	0.991226	0.953115	0.962673	1.000000	0.948127	0.952953
Transport	0.968542	0.243654	0.958052	0.863743	0.994500	0.948127	1.000000	0.981885
Waste	0.983037	0.294276	0.945598	0.852111	0.982383	0.952953	0.981885	1.000000

The correlation heatmap for CO_2 emissions between various sectors reveals a network of interrelated activities impacting carbon output. High positive correlations are observed between sectors such as Fuel Exploitation, Industrial Combustion, Power Industry, and Transport, suggesting that activities in these sectors move in tandem—when one sector's emissions increase, so do the others.



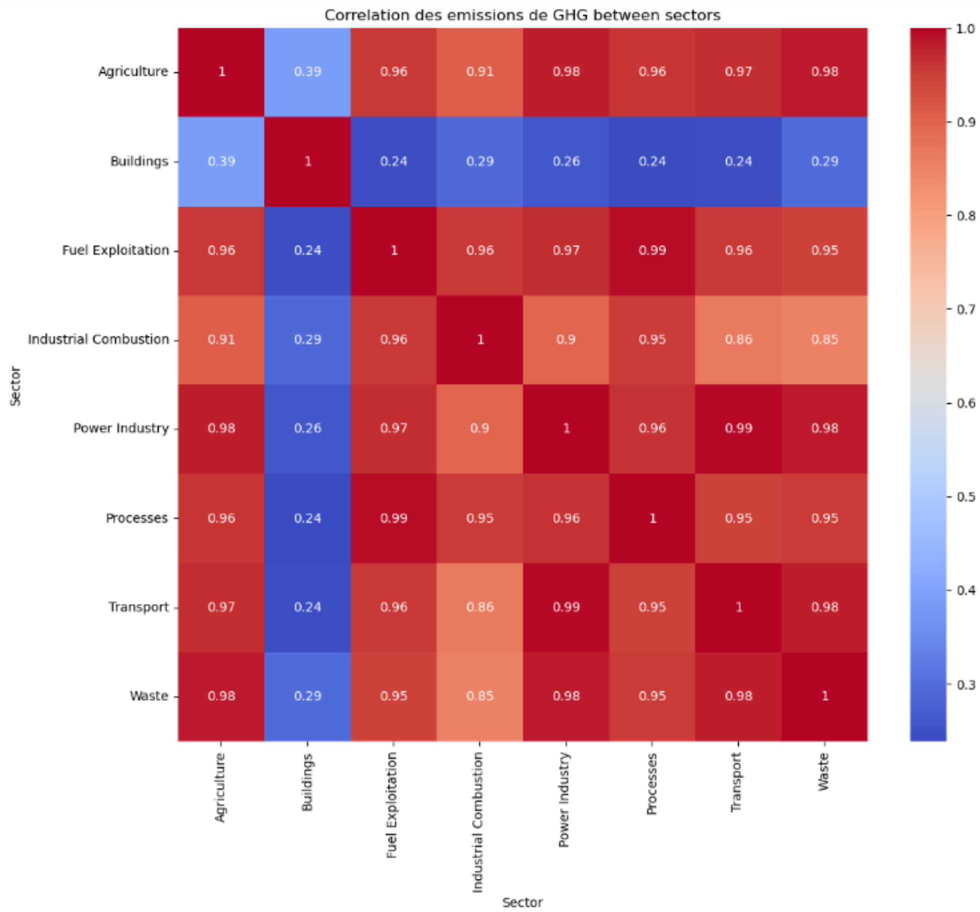
This is particularly true for Fuel Exploitation and the Power Industry, which share a near-perfect correlation, likely due to the direct reliance of power generation on fuel sources.

The relatively lower correlations involving Agriculture and Buildings indicate more independent emission profiles, which could be due to varying operational practices or regulations in these sectors. Interestingly, Waste shows a negative correlation with Agriculture, suggesting that as agricultural emissions increase, waste emissions may decrease, or vice versa, potentially reflecting effective waste management practices in agricultural sectors.

These insights highlight potential areas where policy interventions could be cross-sectional, aiming to address multiple sources of emissions simultaneously.

The same goes for GHG:

- **GHG:**



The correlation matrix for greenhouse gas (GHG) emissions across various sectors reveals a strong interconnectedness, with particularly high positive correlations between Agriculture, Fuel Exploitation, Industrial Combustion, Power Industry, Processes, Transport, and Waste.

These high correlations suggest that as emissions increase in one of these sectors, they tend to increase across the others, indicating a collective contribution to the overall GHG emissions. Notably, the Power Industry shows very high correlations with all sectors except Buildings, reflecting its central role in GHG emissions through energy production and consumption patterns.

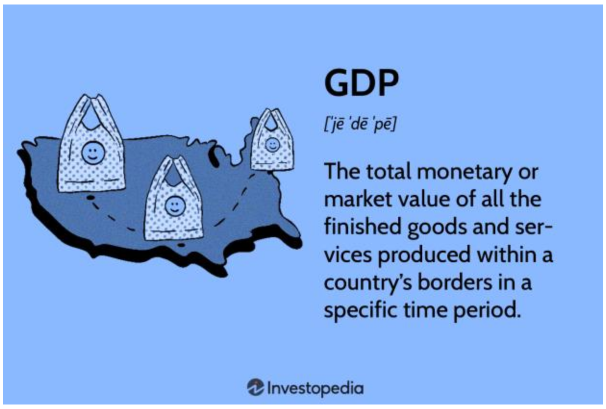
The lower correlation between Buildings and other sectors may indicate unique factors affecting emissions in the building sector, such as localized energy efficiency measures or heating requirements that do not scale with industrial or transport activities.

These insights point to the potential for comprehensive climate action plans that target energy production and consumption across multiple sectors to effectively reduce overall GHG emissions.

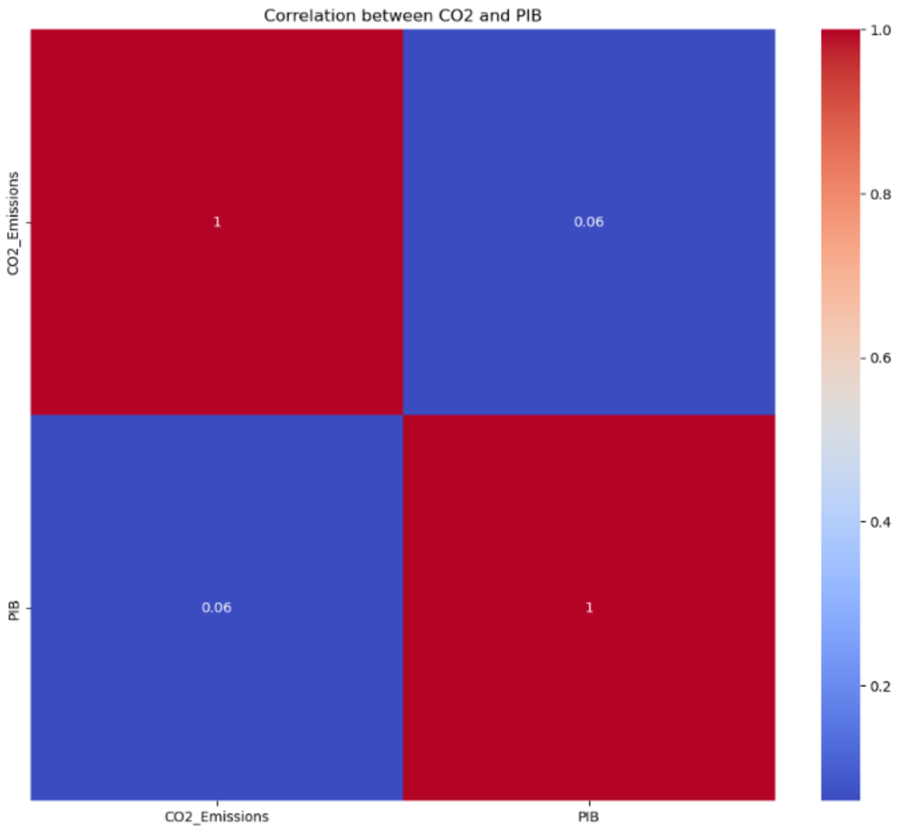
3) Correlation between emissions and PIB (also called GDP):

- **CO2:**

The correlation being studied here is between CO2/GHG and PIB (GDP).




```python
# Melting the CO2 emissions dataset to long format correctly
co2_long_df = datasets_filled_co2.melt(id_vars=['EDGAR Country Code', 'Country'],
                                        var_name='Year',
                                        value_name='CO2_Emissions')

# Melting the GHG emissions dataset to long format correctly
pib_long_df = datasets_filled[2].melt(id_vars=['EDGAR Country Code', 'Country'],
                                        var_name='Year',
                                        value_name='PIB')
# Merging the datasets
merged_emissions_df = pd.merge(co2_long_df, pib_long_df, on=['Country', 'Year'])

# Dropping rows with NaN values
merged_emissions_df.dropna(inplace=True)

# Calculating correlation between CO2 and GHG emissions
correlation_matrix = merged_emissions_df[['CO2_Emissions', 'PIB']].corr()
# Plotting the correlation matrix
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation between CO2 and PIB")
plt.show()
```



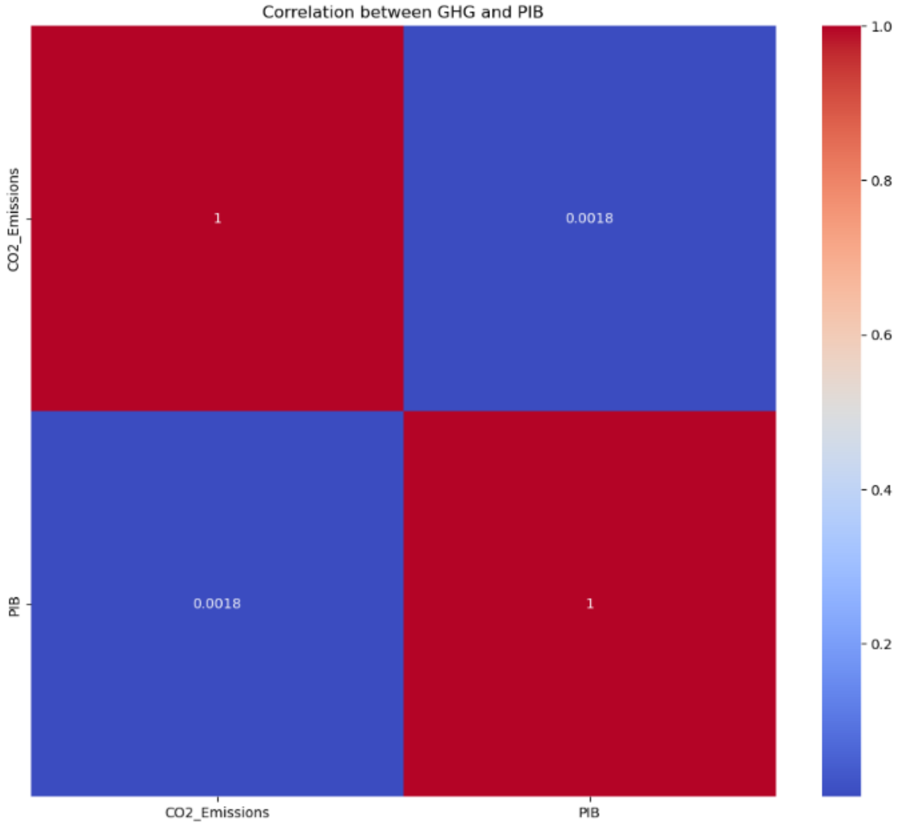
The heatmap illustrates a very weak correlation (0.06) between CO2 emissions and GDP (PIB - Product Internal Bruto), suggesting that there is no significant direct relationship between a country's economic output and its carbon emissions within the dataset analyzed.

This indicates that higher wealth does not necessarily equate to higher CO2 emissions, and vice versa, which could reflect successful decoupling efforts in some economies, where growth has become less carbon-intensive due to advances in technology, shifts towards service-based economies, or more stringent environmental regulations.

It also opens the possibility that other factors, such as energy mix, industrial efficiency, and environmental policies, play a more substantial role in determining a nation's CO2 emissions than GDP alone. This finding is important for policymakers aiming to achieve economic growth without proportionate increases in carbon emissions.

- **GHG:**

```python
# Plotting the correlation matrix
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation between GHG and PIB")
plt.show()
```



The correlation heatmap depicts an extremely weak correlation (0.0018) between greenhouse gas (GHG) emissions and GDP, indicating that within the dataset analyzed, there is no discernible link between the economic activity of a country and its GHG emissions.

This surprising finding suggests that economic growth, as measured by GDP, does not necessarily drive an increase in GHG emissions, which could imply that some economies may be effectively employing green technologies or adopting sustainable practices. This weak correlation may also reflect a global shift towards cleaner energy sources, increased energy efficiency, or the implementation of stringent environmental policies that allow for economic expansion with minimal impact on GHG emissions.

The insight is pivotal for economic and environmental policy, suggesting that economic development can, in some cases, be decoupled from environmental degradation.

4) Correlation of emissions between continents:

For this correlation I was free to choose I decide to go with correlation per continent.

It can be very interested to see how these continents interact between themselves and if for example we see a difference between North and South.

- **CO2:**

```python
merged_df_co2 = pd.merge(datasets_filled[3], country_continent_df, on='Country', how='left')
```

Python

```python
merged_df_co2.head()
```

Python

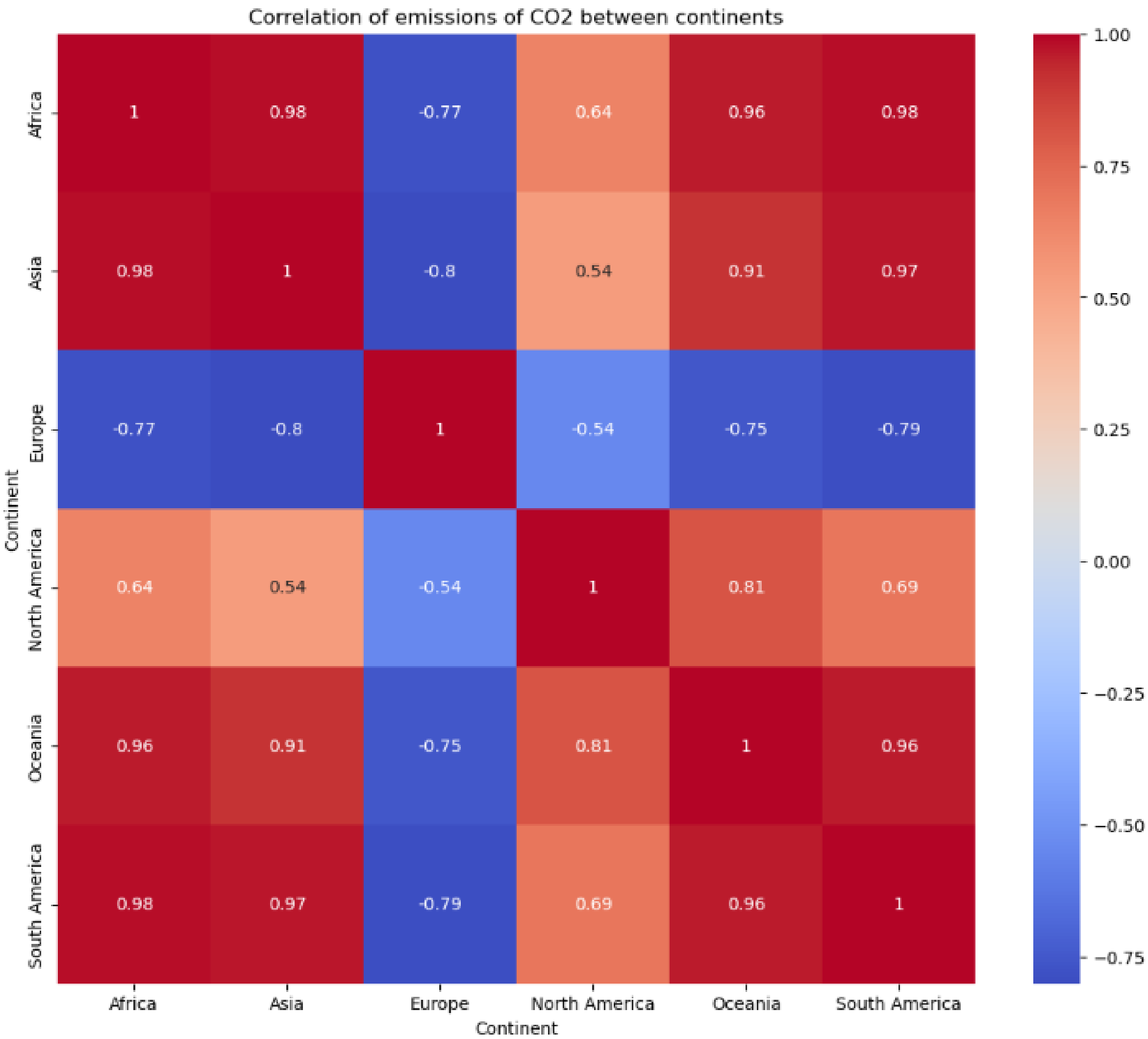
	Substance	EDGAR Country Code	Country	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	198
0	CO2	ABW	Aruba	0.025214	0.028828	0.039472	0.044289	0.043469	0.057396	0.056423	0.067100	0.071937	0.075919	0.079800	0.082922	0.082165	0.084463	0.09862
1	CO2	AFG	Afghanistan	1.734053	1.733842	1.693672	1.733883	2.190254	2.028878	1.892591	2.282429	1.933974	2.059180	2.006283	2.259793	2.309461	3.001547	3.22453
2	CO2	AGO	Angola	8.948153	8.533779	10.383704	11.367327	11.827856	10.924099	7.310833	12.051076	14.232405	14.197356	14.334664	13.248204	12.787518	13.243240	13.38168
3	CO2	AIA	Anguilla	0.002178	0.002178	0.002273	0.002118	0.002360	0.002594	0.002444	0.002547	0.002911	0.003223	0.004422	0.006649	0.007028	0.006098	0.00581
4	CO2	ALB	Albania	4.848780	4.841912	5.523182	4.956735	5.333672	5.428345	5.801022	6.205497	6.826005	7.941258	8.078065	7.075387	7.275600	7.795043	8.43441

After some data manipulation to get what we need we calculate the correlations like before and plot the matrix.

```python
continent_emissions_ghg = merged_df_ghg.groupby('Continent').sum()
continent_emissions_ghg=continent_emissions_ghg.T
correlation_matrix=continent_emissions_ghg.corr()
```

```python
# Plotting the correlation matrix
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation of emissions of GHG between continents")
plt.show()
```

My guess was that these varied correlations would underscore the importance of regional context in understanding global CO2 emissions trends and the need for continent-specific approaches to climate change mitigation.



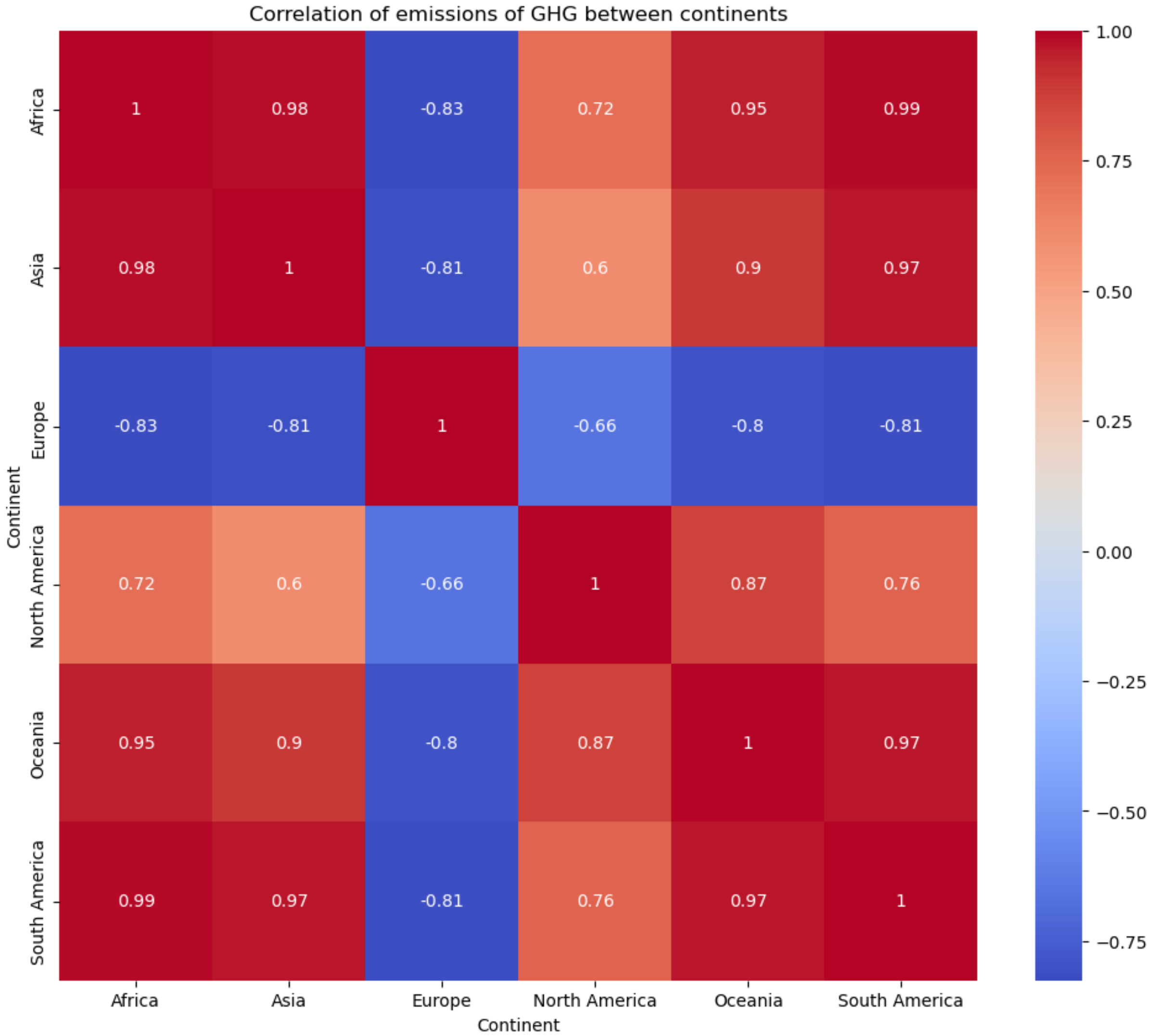
The correlation matrix for CO2 emissions between continents presents a complex interplay of relationships. High positive correlations are seen between Africa, Asia, and South America, suggesting that these continents' emissions patterns move in tandem—likely due to similar developmental stages and industrialization processes.

In stark contrast, there are strong negative correlations between these continents and Europe, indicating inverse relationships; as CO2 emissions increase in Europe, they tend to decrease in Africa, Asia, and South America, or vice versa.

This could reflect differing economic structures, energy dependencies, and environmental policies. North America and Oceania show a mix of positive and negative correlations with other continents, suggesting that factors influencing CO2 emissions in these regions may be more varied or transitional.

- **GHG:**

```
# Plotting the correlation matrix
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation of emissions of GHG between continents")
plt.show()
```



The heatmap for greenhouse gas (GHG) emissions correlations between continents indicates a strong positive relationship between Africa, Asia, and South America, suggesting synchronized trends in GHG emissions, possibly linked to shared developmental patterns or economic growth phases. Conversely, Europe shows a strong negative correlation with Africa and Asia, pointing to divergent GHG emission trends, which could be attributed to Europe's more aggressive climate policies, energy efficiency measures, and shifts toward renewable energy.

North America and Oceania exhibit mixed correlation patterns, indicating variable and possibly transitional influences on GHG emissions. The high positive correlation between Oceania and South America might reflect similar environmental or economic factors affecting GHG emissions.

These correlations underscore the varying impacts of regional developmental policies, energy consumption, and industrialization levels on GHG emissions, highlighting the need for tailored and region-specific strategies in global climate policy and emissions management.

PART III: Predictive Modelling:

1) For CO2:

AFRICA

```
african_countries_co2 = merged_df_co2[merged_df_co2['Continent'] == 'Africa']
african_countries_co2=african_countries_co2.drop('Substance',axis=1)
```

```
african_countries_co2.head()
african_countries_co2_long=african_countries_co2.melt(id_vars=['EDGAR Country Code', 'Country','Continent'],
                                                      var_name='Year',
                                                      value_name='CO2_Emissions')
african_countries_co2_long

african_countries_co2_grouped=african_countries_co2_long.groupby('Year').sum()
african_countries_co2_grouped.head()
```

	EDGAR Country Code	Country	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
2	AGO	Angola	8.948153	8.533779	10.383704	11.367327	11.827856	10.924099	7.310833	12.051076	14.232405	14.197356	14.334664	13.248204	12.787518	13.243240	13.381683	14.132242
13	BDI	Burundi	0.058337	0.058437	0.058246	0.060050	0.064229	0.066496	0.068760	0.071071	0.072277	0.082255	0.081172	0.088375	0.073732	0.089807	0.106418	0.106388
15	BEN	Benin	0.311380	0.361691	0.461619	0.476488	0.489024	0.548639	0.365945	0.427151	0.454084	0.509511	0.535370	0.445586	0.517180	0.509755	0.523587	0.617490
16	BFA	Burkina Faso	0.226220	0.226475	0.225612	0.233199	0.251036	0.260680	0.270121	0.279405	0.284087	0.326271	0.320718	0.307853	0.287174	0.313493	0.303591	0.293698
30	BWA	Botswana	0.153417	0.153517	0.196993	0.262298	0.453163	0.560579	0.817830	0.933457	1.054538	1.200593	1.267802	1.385679	1.445884	1.416804	1.411821	1.511194

```
X = african_countries_co2_grouped.index.values.reshape(-1, 1)  # Year as the feature
y = african_countries_co2_grouped['CO2_Emissions'].values      # CO2 Emissions as the target

# Normalize the feature and target
scaler_X = MinMaxScaler()
X_scaled = scaler_X.fit_transform(X)

scaler_y = MinMaxScaler()
y_scaled = scaler_y.fit_transform(y.reshape(-1, 1))

# Reshape for LSTM input - LSTM expects input to be 3D (num_samples, num_time_steps, num_features)
# Here we have a single time step and a single feature
X_scaled = X_scaled.reshape((X_scaled.shape[0], 1, 1))

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_scaled, test_size=0.2, random_state=42)

# Create the LSTM model
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(1, 1)))  # Adjust the number of neurons and input_shape as needed
model.add(Dense(1))
model.compile(optimizer=Adam(), loss='mean_squared_error')

# Train the model
model.fit(X_train, y_train, epochs=100, batch_size=32, verbose=1)
```

We create here a LSTM (Long Short-Term Memory) model, which is a sophisticated neural network architecture used for emissions forecasting. It operates by analyzing historical emissions data and various environmental factors to make predictions about future emissions levels.

This model employs a recurrent neural network structure with specialized memory cells, allowing it to capture and learn complex temporal patterns and dependencies within the data, which enables it to provide insights into emissions trends and their predictive accuracy across different continents.

```
Epoch 1/100
2/2 [==============================] - 3s 19ms/step - loss: 0.3181
Epoch 2/100
2/2 [==============================] - 0s 6ms/step - loss: 0.3120
Epoch 3/100
2/2 [==============================] - 0s 8ms/step - loss: 0.3060
Epoch 4/100
2/2 [==============================] - 0s 15ms/step - loss: 0.3003
Epoch 5/100
2/2 [==============================] - 0s 8ms/step - loss: 0.2947
Epoch 6/100
2/2 [==============================] - 0s 12ms/step - loss: 0.2889
Epoch 7/100
2/2 [==============================] - 0s 6ms/step - loss: 0.2835
Epoch 8/100
2/2 [==============================] - 0s 9ms/step - loss: 0.2783
Epoch 9/100
2/2 [==============================] - 0s 8ms/step - loss: 0.2727
Epoch 10/100
2/2 [==============================] - 0s 8ms/step - loss: 0.2677
Epoch 11/100
2/2 [==============================] - 0s 8ms/step - loss: 0.2624
Epoch 12/100
2/2 [==============================] - 0s 4ms/step - loss: 0.2575
Epoch 13/100
...
Epoch 99/100
2/2 [==============================] - 0s 11ms/step - loss: 0.0314
Epoch 100/100
2/2 [==============================] - 0s 10ms/step - loss: 0.0311
```
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

The model gets more and more precise with training.

```
# Evaluate the model
loss = model.evaluate(X_test, y_test, verbose=0)
print(f'Test Loss: {loss}')
```
Python

```
Test Loss: 0.035253848880529404
```

We do the same on the different continents to have very precise models.

Continent	Loss
Africa	0.035253849
Asia	0.056542493
Europe	0.151943073
North America	0.154656351
South America	0.156622395
Oceania	0.14317733

The LSTM model's loss metrics across the continents provide insights into the predictive accuracy of the emissions forecasting. Africa and Asia show the lowest loss figures, indicating a higher predictive accuracy for these regions, which could be due to less variability in the factors affecting emissions or a more consistent trend that the model could learn from.

On the other hand, Europe, North America, and South America exhibit higher loss values, suggesting a less accurate forecast that could result from greater variability or complexity in the emissions data or the influence of more erratic factors not captured by the model.

Oceania's loss is moderate, indicating a fair level of prediction accuracy. These metrics not only inform the reliability of the model's forecasts but also potentially reflect the unique emission profiles and trends of each continent, which could be critical for targeted climate strategies.

Continent	Predicted values for next 3 years
Africa	1184.9694/1196.6945/1208.4069
Asia	14773.625/14931.797/15090.275
Europe	5374.2456/5380.6323/5387.0073
North America	5389.055/5398.336/5407.648
South America	5330.9775/5336.9175/5342.8564
Oceania	5357.3423/5362.8804/5368.408

Thanks to this model we get the forecast of emissions for the next 3 years for each continent with a very good accuracy.

2) For GHG:

We do the same kind of model and run it on all continents to have a very precise model.

```python
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_absolute_error, mean_squared_error

X=asia_countries_ghg_grouped.index.values.reshape(-1, 1)  # Year as the feature
y = asia_countries_ghg_grouped['CO2_Emissions'].values      # CO2 Emissions as the target

# Normalize the feature and target
scaler_X = MinMaxScaler()
X_scaled = scaler_X.fit_transform(X)

scaler_y = MinMaxScaler()
y_scaled = scaler_y.fit_transform(y.reshape(-1, 1))

# Reshape for LSTM input - LSTM expects input to be 3D (num_samples, num_time_steps, num_features)
# Here we have a single time step and a single feature
X_scaled = X_scaled.reshape((X_scaled.shape[0], 1, 1))

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y_scaled, test_size=0.2, random_state=42)

# Create the LSTM model
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(1, 1)))  # Adjust the number of neurons and input_shape as needed
model.add(Dense(1))
model.compile(optimizer=Adam(), loss='mean_squared_error')

# Train the model
model.fit(X_train, y_train, epochs=100, batch_size=32, verbose=1)
```

```
Epoch 1/100
2/2 [==============================] - 3s 8ms/step - loss: 0.3050
Epoch 2/100
2/2 [==============================] - 0s 6ms/step - loss: 0.2992
Epoch 3/100
2/2 [==============================] - 0s 3ms/step - loss: 0.2934
Epoch 4/100
2/2 [==============================] - 0s 2ms/step - loss: 0.2877
Epoch 5/100
2/2 [==============================] - 0s 7ms/step - loss: 0.2824
Epoch 6/100
2/2 [==============================] - 0s 3ms/step - loss: 0.2765
Epoch 7/100
2/2 [==============================] - 0s 2ms/step - loss: 0.2712
Epoch 8/100
2/2 [==============================] - 0s 7ms/step - loss: 0.2658
Epoch 9/100
2/2 [==============================] - 0s 8ms/step - loss: 0.2603
Epoch 10/100
2/2 [==============================] - 0s 10ms/step - loss: 0.2548
Epoch 11/100
2/2 [==============================] - 0s 8ms/step - loss: 0.2498
Epoch 12/100
2/2 [==============================] - 0s 7ms/step - loss: 0.2444
Epoch 13/100
...
Epoch 99/100
2/2 [==============================] - 0s 11ms/step - loss: 0.0327
Epoch 100/100
2/2 [==============================] - 0s 11ms/step - loss: 0.0325
```
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Continent	GHG Loss	Predicted values for the next 3 years
Africa	0.038331583	2637.5273/2659.4287/2681.327
Asia	0.041782886	2539.5925/2559.1094/2578.7222
Europe	0.149747267	6889.95/6896.384/6902.8086
North America	0.152614772	6857.1387/6858.5894/6859.995
South America	0.04852245	1936.2035/1947.4519/1958.7338
Oceania	0.143945023	6812.8364/6814.5938/6816.3223

The LSTM model demonstrates variable predictive accuracy across continents for GHG emissions, with Africa and Asia showing the lowest losses, indicative of more reliable forecasts in these regions.

This could suggest that the emission trajectories in Africa and Asia are relatively stable or that they follow a pattern that the model can capture effectively. Conversely, Europe and North America show higher loss values, which may point to more complex emission patterns or greater unpredictability in the factors influencing GHG emissions. South America, while showing a slightly higher loss than Africa and Asia, still indicates a relatively accurate model performance. Oceania's loss is comparable to Europe and North America, suggesting similar challenges in accurately predicting GHG emissions.

These variations in model loss highlight the differences in GHG emission dynamics across continents and may reflect distinct environmental policies, economic development stages, and energy use practices, which are critical considerations for future modeling and policy-making efforts.

We also get the predicted values for the next 3 years for GHG.

Conclusions and Insights:

Conclusions:

1. **Disparity in Emissions**:
   - There is a significant disparity in both CO2 and GHG emissions among continents, with Asia and North America consistently ranking as the highest emitters. This suggests the need for region-specific strategies to address emissions.
2. **Sectoral Impact**:
   - The Power Industry and Transport sectors are the most significant contributors to emissions, highlighting the critical areas where interventions could yield substantial reductions in overall emissions.
3. **Economic Development and Emissions**:
   - The weak correlation between GDP and emissions suggests that economic growth does not always equate to higher emissions, indicating potential decoupling due to energy efficiency, cleaner technologies, or service-oriented economic shifts.
4. **Predictive Model Performance**:
   - The LSTM model showed varying levels of prediction accuracy across different continents for both CO2 and GHG emissions, suggesting the presence of distinct factors influencing emissions in each region.

Insights:

1. **Policy Implications**:
   - The need for differentiated climate policies is clear, as emissions are not uniform across sectors or regions. Tailored approaches are necessary to address the unique characteristics of each sector and region.
2. **Technological Advancements**:
   - The relatively weak correlation between emissions and economic output could imply successful adoption of green technologies in certain regions. There is potential to scale such technologies to reduce emissions further.
3. **International Cooperation**:
   - Given the interdependence of emissions among some continents, international cooperation is crucial. Shared strategies and technologies could benefit multiple regions simultaneously.
4. **Data-Driven Decisions**:

- The insights from the predictive models can be used to inform policy decisions, focusing on areas where the model predicts higher emissions and thus where interventions might be most needed.

5. **Future Research**:
   - Areas with higher model loss indicate a need for further research to understand the underlying complexities that affect emissions, which may include economic policies, energy sources, and land-use practices.
6. **Investment in Sustainability**:
   - To sustain economic growth while minimizing environmental impact, investment in sustainable infrastructure and renewable energy is essential, as shown by the lower correlation of emissions with GDP in some regions.
7. **Monitoring and Reporting**:
   - Continuous monitoring and reporting of emissions are vital for tracking progress and the effectiveness of policies. This project reinforces the importance of accurate data collection and analysis in the fight against climate change.

Overall, the project underscores the complexity of global emissions and the importance of leveraging data analysis and predictive modeling for strategic planning in climate action. It also emphasizes that while the challenge is global, solutions must be adapted to regional and sectoral contexts.

AI generated picture to illustrate – free to use